論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Development of a protein sequence alignment method for accurate homology modeling
著者(和文)	牧垣秀一朗
Author(English)	Shuichiro Makigaki
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第11681号, 授与年月日:2020年12月31日, 学位の種別:課程博士, 審査員:石田 貴士,秋山 泰,岡﨑 直観,村田 剛志,関嶋 政和
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第11681号, Conferred date:2020/12/31, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Doctoral thesis

Development of a protein sequence alignment method for accurate homology modeling

Author
Shuichiro Makigaki
Supervisor
Takashi Ishida

26 November 2020

Department of Computer Science
School of Computing
Tokyo Institute of Technology

Abstract

A protein structure provides important information that can be used for various practical applications in the biological sciences. Nevertheless the number of entries of the PDB is growing rapidly, the tertiary structure of only 0.1 % of all known proteins has as yet been experimentally characterized. Homology modeling, tertiary structure prediction of a protein using homologous protein structures, is useful if good templates are available. Modern homology detection methods can find remote homologs with high sensitivity. However, the accuracy of homology models generated from homology-detection-based alignments is often lower than that from the ideal alignments. Thus, manual modifications by experts have been applied in practical cases. In this study, we proposed a new pairwise generation method for homology modeling.

Firstly, we proposed a pairwise sequence alignment generation method based on a machine learning model that learns the structural alignments of known homologs. Machine learning has already applied to homology detection. However, they did not generate the sequence alignments and cannot be directly used for homology modeling. Hence, we used dynamic programming during sequence alignment to dynamically predict a substitution score from the learned model instead of a fixed substitution matrix or profile comparison. Then, machine learning was used in this substitution score prediction process. We evaluated the first method by carefully splitting the training and test datasets and comparing the predicted structure's accuracy with that of state-of-the-art methods. It generated more accurate tertiary structure models than those produced from alignments

obtained by the other methods.

Secondly, we proposed a sequence alignment generation method for remote homologs detected by an intermediate sequence search (ISS). ISS is a homology detection method that performs homology searches with different databases iteratively and can detect more remote homologs compared with single sequence searches. However, the method does not generate a sequence alignment between query and template sequences and this is a critical problem in homology modeling. Thus, this is the first study demonstrating the generation and evaluation of the sequence alignment using the ISS results in the context of homology modeling. As the results of the evaluation, model accuracies were improved compared with models using naïve dynamic programming-based alignment. The proposed method generated more accurate homology models, especially for remote homologs.

Finally, we discussed the advantages and disadvantages of each proposed method and integrated these two methods. We found that ISS-based method was useful for a difficult pair of query and template proteins that our machine learning-based alignment method cannot generate accurate sequence alignment for. Thus we developed a pipeline that selected the better alignment based on the alignment length. As a result, the integrated method showed small improvements.

Acknowledgements

Many people who support me are the key to complete this thesis, but I apologies for listing here just those who have been involved deeply in this thesis. I appreciate Takashi Ishida as a supervisor for keeping an excellent research environment and continuous support. I also thank Masakazu Sekijima, Naoaki Okazaki, Tsuyoshi Murata, and Yutaka Akiyama as dissertation committee members for valuable comments and discussions. All reviewers have been crucial, especially Hasić Haris, Jason Kurniawan, Keisuke Yanagisawa, and Masahito Ohue, and I appreciate them for their feedback on this thesis. Continuous support and a good understanding of CyberAgent, Inc have also been crucial. I would like to appreciate family, colleagues, laboratory members, and friends, especially Tomohiro Makigaki, Chinami Makigaki, Koujiro Makigaki, Masato Sato, Rei Yamaguchi, Tatsuya Akutsu, Mayumi Kamada, and Asuka Kato.

Contents

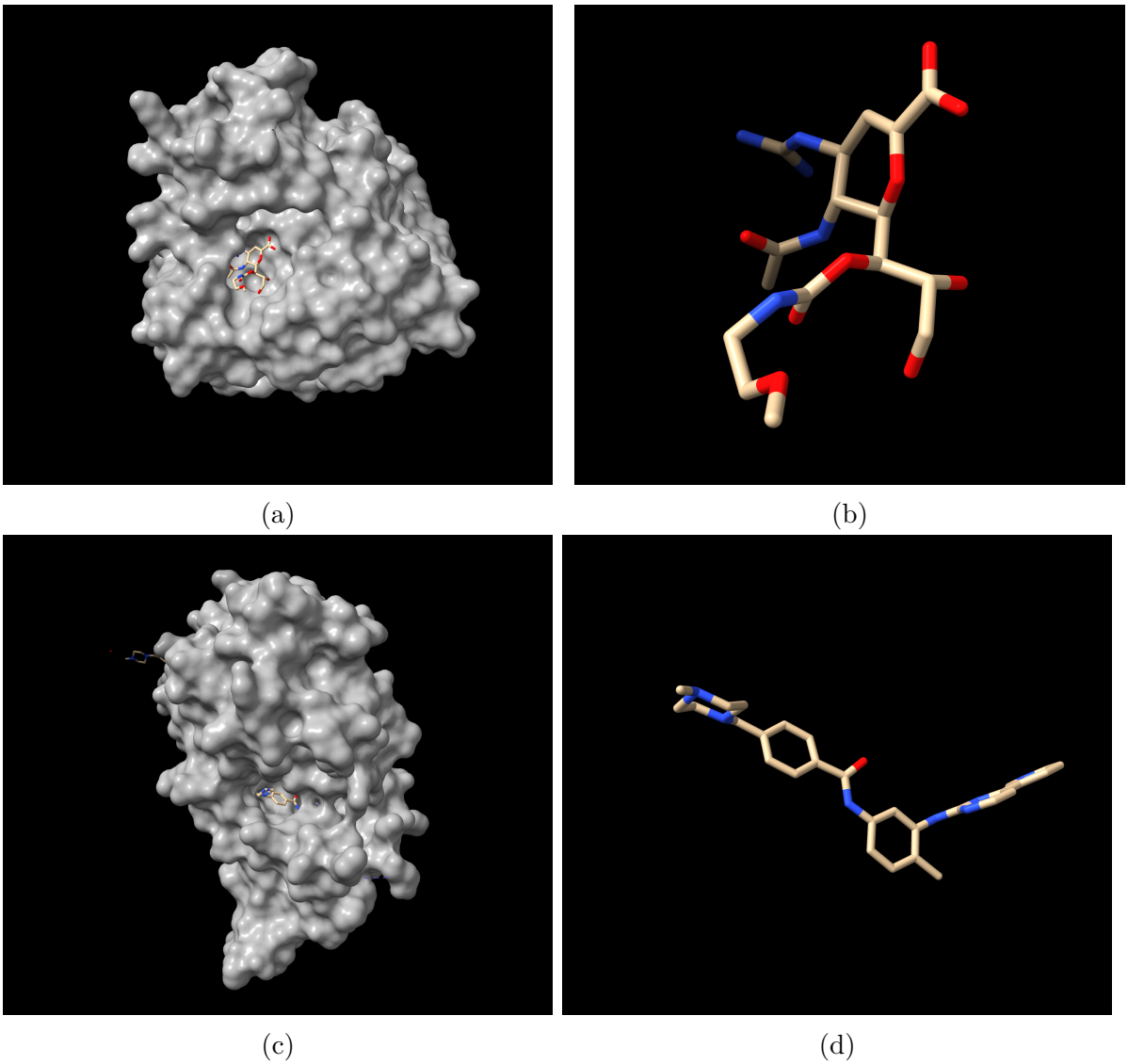
Chapter 1.

Introduction

1.1. Protein structure

Proteins are key molecules in life sciences such as biology, biochemistry, and pharmaceutical sciences. Proteins are chain-like molecules composed of 20 standard amino acids. There are approximately 100,000 different proteins in human cells, and each protein has different function. Different proteins have different amino acids compositions, and the sequence is called amino acid sequence of a protein. Proteins often achieve their functions by forming specific three-dimensional structures according to their amino acid sequences. Thus, many researchers tried to determine and analyze the structures. Such researches about protein structures are often called "structural biology." The structure information significantly contributes to revealing protein functions, insight into protein interactions, and drug development. Therefore, the importance of the three-dimensional structure is increasing.

Protein three-dimensional structures can be determined by experimental methods. Nowday, many protein structures have been determined and often registered to and accessible in the online Protein Databank (PDB; `http://www.rcsb.org/pdb`) [1]. The



(a) (b)

(c) (d)

Figure 1.1.: Structure-based drug design. (1.1a, 1.1c) Protein structure representation by drawing molecule surface area and bound ligands (1.1a, 1.1b) Neuraminidase (PDB: `5JYY`) that is accelarates the lysis of newly generated viral particles in infected cells and zanamivir small molecule shown at the center by ball-and-stick (1.1c, 1.1d) BCR-ABL1 protein, which is chimeric proteins that contribute to the pathophysiology of chronic myeloid leukemia (CML), in complex with imatinib

PDB contains 168,599 protein structures in September 2020. These structures are now used for various application researches and play an important role in the growing world-wide effort in structural genomics [2], [3].

For example, structure-based drug design (SBDD) and designing or improving ligands are well-known examples of application based on protein structure information (figure 1.1). The development of oseltamivir and zanamivir, which are Influenza treatments, is a notable example. After Colman *et al.* succeeded to determine structure of the catalytic and antigenic sites in influenza virus neuraminidase in 1983 [4], which accelarates the lysis of newly generatedviral particles in infected cells, the structure information helps discovery and development of zanamivir and oseltamavir [5]. The development of therapeutics of chronic myelogenous leukemia (CML) is also helped by protein structure. Imatinib, the first therapeutics for CML, had beed approved, but resistance was reported. In 2000, Schindler *et al.* reported the structure of target protein that leads to the uncontrolled growth and survival of the leukaemic cells [6], and structural analysis revealed that the lack of hydrogen bonds due to amino acid substitutions led to the development of nilotinib and dasatinib, which are less resistant to drug resistance. The crystallographic study greatly helps drug-discovery efforts to find new compounds that might inhibit the protein with higher affinity while retaining the excellent kinase selectivity profile [7]. In the SBDD, virtual screening based on docking of protein and small ligands is the representative method [8], [9]. Small molecule drugs usually work by binding to an active pocket of a protein. The relationship between the drug and the pocket is known to have a key-keyhole relationship. Since whether the key fits depends on the shape and chemical complementarity of the keyhole, the three-dimensional structure of the protein is a very useful piece of information.

Protein structure information can be also used for analyzing different type of interaction. Many proteins work in the body by interacting with other proteins: enzyme of

metabolism, signal transduction, and transcription factor. The knowledge of the PPI can provide insight into the mechanisms and is valuable for drug design [10]. Structure information helps the prediction of the protein-protein interaction (PPI) [11]. The proteins bind in a broader are than small ligands, but shape complementarity is an important factor as well. The PPI is classically revealed by high-throughput experimental methods, and computational protein-protein docking methods [12]–[14] improve the prediction accuracy and reduces false positives with increasing the number of resolved protein structure [15].

In addition to the interaction estimation, the function itself can be annotated and assigned by the structure information [16], [17]. This is because three-dimensional structures are often more evolutionarily conserved than amino acid sequences, and functional inferences based on three-dimensional structural comparisons are also being studied. A protein's function is usually experimentally determined. If the experimental determination has failed, the protein's function can often be inferred from sequence similarity. However, when the sequence similarity-based methods fail, the protein structure information can provide hints and clues of the function [18]. The function is predicted by globally or locally comparing structures of function-known and -unknown proteins [19].

1.1.1. Protein structure hierarchy

Protein structures are important information in biology, but there are four levels of protein structure hierarchy: primary, secondary, tertiary, and quaternary (figure 1.2). Although all hierarchical structures are important information, three-dimensional information is most commonly used in structural information applications, and the term "protein structure" refers to a tertiary structure. The target of this research is this three-dimensional structure.

Protein primary structure is often called as a polypeptide chain. These polypeptide

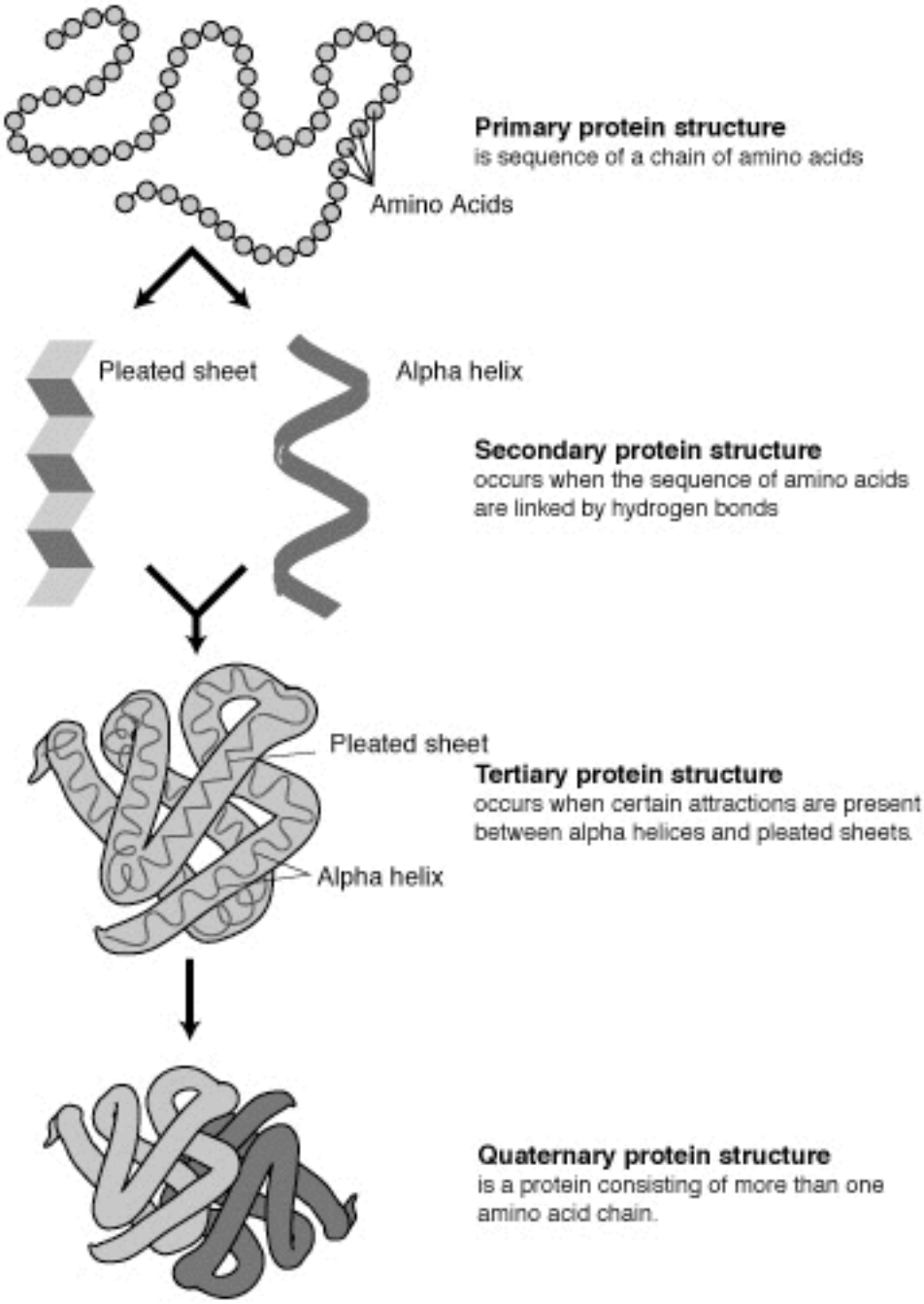


Figure 1.2.: Protein structure hierarchy (image from Wikipedia: `https://commons.wikimedia.org/wiki/File:Protein-structure.png`)

chains are elaborated on ribosomes from a codon sequence template on messenger RNA, and forming linear amino acid sequences.

The secondary structure is built in a protein structure by hydrogen-bonding between amino acids in the polypeptide chain. Helix is the most popular secondary structure in proteins. Hydrogen-bonding is built between residues that are four residues far from each other. Sheet structure is based on five or ten residues in subsequences and another five or ten residues downstream of the subsequence. The secondary structure is useful for local structural information and structural class classification, but it is difficult to use for SBDD. However, complex structures of proteins are characterized by secondary structures, which are often used to visualize the patterns of three-dimensional structures (figure 1.3). The secondary structure is always folded across loops or turns.

Tertiary structure is built when hydrophobic residues slide into a protein. Then, the structure is stabilized by hydrogen-bonding, ion-bonding, or disulfide bond. During the formation of the tertiary structure, there are lots of other reasonable combinations of interactions. Nevertheless, we do not fully understand why only one combination occurs. With a tertiary structure, various applications described above are possible

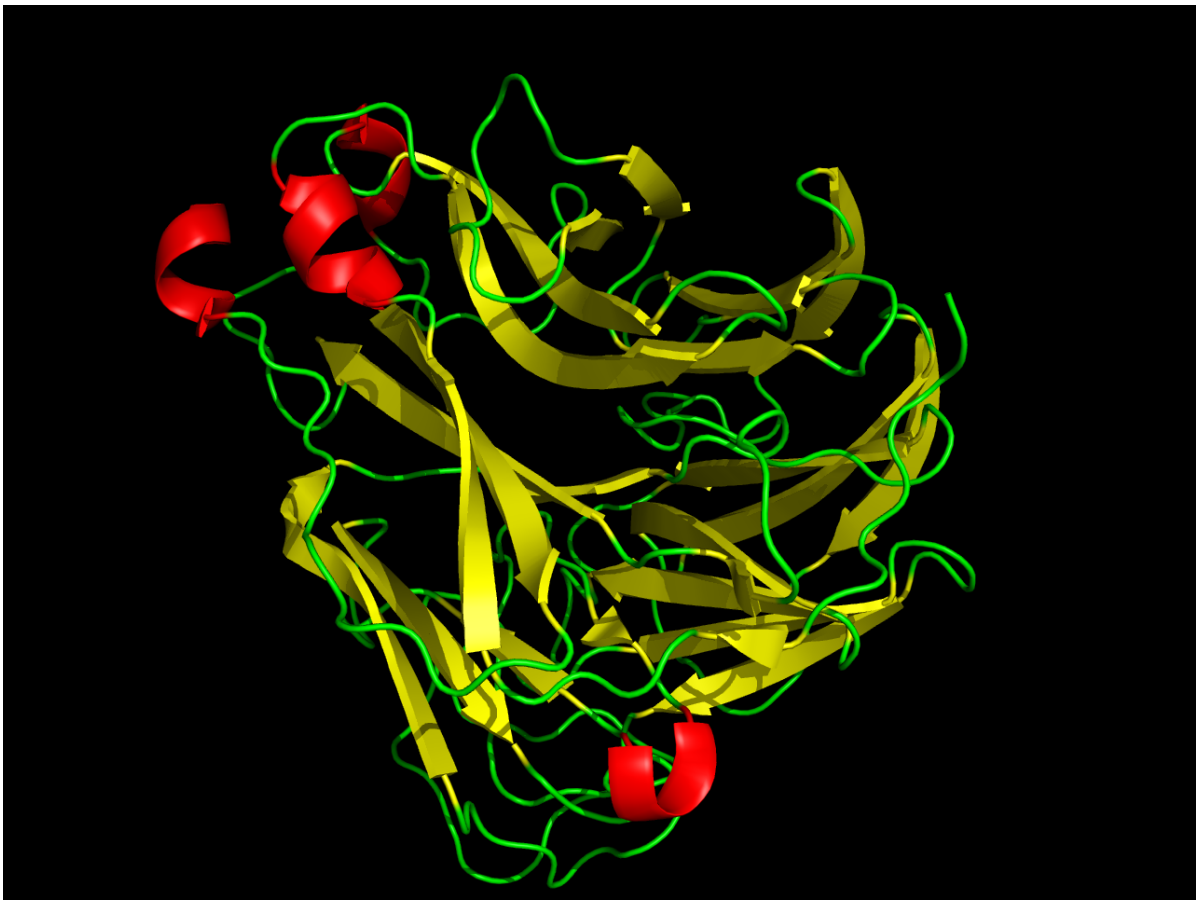
Some proteins function isolated from others. However, many other proteins form a complex and obtain new or advanced functions. These complex proteins are called as quaternary structure. Hydrogen-bonding, ion-bonding, and disulfide bonding, which stabilize the tertiary structure, are also used to stabilize the quarternary structure. The fourth-order structure is also important because proteins often form complexes to function, but docking can also predict from the tertiary structure.

1.1.2. Experimental determination of protein structure

The structure of proteins at all levels can be determined experimentally. In particular, the primary structure and amino acid sequence can be easily determined by the Edman

(a)

```
5JYY  RNFNNLTKGLCTINSWHIYGKDNAVRIGESSDVLVTREPYVSCDPDECRFYALSQGTTIR   60
5JYY  GKHSNGTIHDRSQYRALISWPLSSPPTVYNSRVECIGWSSTSCHDGKSRMSICISGPNNN  120
5JYY  ASAVVWYNRRPVAEINTWARNILRTQESECVCHNGVCPVVFTDGSATGPADTRIYYFKEG  180
5JYY  KILKWESLTGTAKHIEECSCYGERTGITCTCKDNWQGSNRPVIQIDPVAMTHTSQYICSP  240
5JYY  VLTDNPRPNDPNIGKCNDPYPGNNNNGVKGFSYLDGANTWLGRTISTASRSGYEMLKVPN  300
5JYY  ALTDDRSKPIQGQTIVLNADWSGYSGSFMDYWAEGDCYRACFYVELIRGRPKEDKVWWTS  360
5JYY  NSIVSMCSSTEFLGQWNWPDGAKIEYFL   388
```



(b)

Figure 1.3.: Amino acid sequence and the protein secondary structure. The protein is Neuraminidase (PDB: 5JYY) that is accelarates the lysis of newly generated viral particles in infected cells. (1.3a) Amino acid sequence of the protein. Each alphabet is a amino acid; for example, A, C and D are Alanine, Cysteine and Aspartic Acid, respectively. (1.3b) Secondary and tertiary structure representation. Red, yellow and green show helix, sheet and loop region, respectively.

method, and nowadays, automated peptide sequencers are also available, and the secondary structure can be easily determined by spectroscopic methods such as circular dichrois spectrum analysis. On the other hand, three-dimensional structures are still difficult to be determined.

Protein tertiary structures can be determined by experimental methods such as X-ray crystallography or nucleic magnetic resonance. In X-ray crystallography, crystals of proteins are needed in which protein molecules are aligned regularly. The crystal is irradiated with X-ray, and diffusion is observed. Because the protein molecules in a crystal are aligned, the diffusion is observed stronger and weaker because of X-ray's wave nature, which shows discretized spots. After retrieving the phase of X-ray, electronic density will be available. X-ray crystallography enables structure determination high resolution. However, the diffraction intensity of the protein molecule is weak because the number of electrons in the atoms that make up the protein is small. Strong X-ray and highly sensitive sensing devices are needed to get enough electron density for protein structure determination. Recently, the ability of the sensor devices and computing performance have improved. However, the biggest bottleneck is crystallizing the protein molecules. The majority of the structures in the PDB, 89 % (141,568) were determined by X-ray crystallography (as of December 2019).

The X-ray crystallography needs to crystallize proteins and can be used for proteins that cannot be crystallized. In the case, Multi-dimensional Nuclear Magnetic Resonance (NMR) is the method to get the distance of atoms in a molecule by observing nuclear magnetic resonance. By the multi-dimensional peak of the resonance of 1H, 15N, and 13C, the NMR method can determine the protein structure of about 100 residues. The advantage of this method is that there is no need to crystalize phase because resonance can be observed in the solvent. However, it is difficult to determine the whole structure of a protein that is consists of over 100 residues because too many peaks are overlapped.

Hence, there are only 8 % (12,843) of the structures being characterized by NMR spectroscopy.

In the 1970s, electron microscopy (EM) has begin to be used for structure determination. The EM is used for structure determination of proteins that cannot be crystallized and are consists of over 100 residues. After developing electron crystallography and applying phase contrast analysis, single particle analysis with cryo-transmission electron microscopy has shown good results. The cryo-transmission EM observes target materials that are fixed in amorphous ice by a weak electron beam. It avoids the corruption or denaturalization of protein molecules caused by freezing and vacuuming that are widely used by former EM methods. Nevertheless, it requires high sensitive electron beam sensors and a high-performance computing environment for processing a huge number of low S/N images. These requirements are becoming more feasible with the advancement of computing power, and cryo-transmission EM shows promise for the determination of protein structure. 2 % (4284) of PDB are characterized by this method. However, it still has a difficult problem to avoid corruption and denaturalization. This is because frozen molecules, as well as the electron beam, are often corrupted by nearby air in the amorphous ice. How to freeze and fix target materials without corruption is still a trial-and-error procedure. Furthermore, if the protein molecule has many disorder regions, single particle analysis with cryo-transmission EM cannot get high-resolution images.

These experimental methods have advantages and disadvantages, but there is always a financial cost. The cost of novel drug target structure determination, such as human membrane proteins, by X-ray crystallography is about 2.5 millon dollars on average. In the case of soluble human proteins such as kinases and proteases, it still requires 450,000 dollars on average [20]. On the other hand, the complete cryo-transmission EM system needs several million dollars to buy it. Also, several hundred thousand annually to maintain and operate the system [21]. Therefore, determining structure of the representative

protein of new fold or family is prioritised, and the structure determination of homologs is often taken over by structure prediction methods such as homology modeling.

1.2. Protein structure prediction

As we denoted, the experimental determination of protein tertiary structure is still difficult. Thus, despite improvements in experimental methods for determining protein structures, the speed at which amino acid sequences can be revealed has overtaken our ability to ascertain the corresponding proteins' structures. The number of entries of the PDB is growing rapidly by about 30 new entries daily as average. By 2019, the PDB contained 158,958 experimental protein structures. A recent analysis of all protein chains in the PDB shows that these proteins can be grouped into 5553 families and 1486 folds. [22], [23] However, the tertiary structure of only 0.1 % of all known proteins has as yet been experimentally characterized. [3] There is still a huge gap between the number of known sequences (235,561,514 in UniRef as of 15$^{\text{th}}$ September 2020). Therefore, protein structure prediction, that is, the use of computational techniques to generate a tertiary structural model of a given amino acid sequence, has been required.

After the Anfinsen's experiment, it is known that protein structure in nature is stabilized in thermodynamics stable. Theoretically, the whole folding behavior can be computationally simulated and explained. The biophysical field is interested in the simulation of protein folding, and the computer science field is interested in the structure prediction because of the applications of optimization algorithms and machine learning. However, the number of combination of 100 amino acid sequence consists of 20 kinds of amino acids are 20^{100}, which means sequence space is near-infinite. It indicates that the difficulty of tertirary structure prediction based on simple structure optimization.

Before tertiary structure prediction, one-dimensional structure prediction methods were developed widely, which includes secondary structure prediction and solvent acces-

sibility prediction. In the early days of computational structure prediction, the secondary structure prediction is the main topic, but the secondary structure is not enough for applied researches, such as SBDD. Therefore, prediction methods that directly returns the coordinates of atoms belong to a protein have been researched. Tertiary structure prediction includes various methods of different fields that are developed separately. However, the input is basically amino acid sequence of a protein and the output is three-dimensional coordinates of protein atoms. Current prediction methods can be roughly categorized into two methods; *de novo* methods (template-free methods) and homology modeling (template-based methods.)

1.2.1. Protein tertiary structure prediction methods

De novo method

The coordinate of each atom in the protein can be determined computationally, but actually it is hard to search all spaces of structures. Therefore, various methods for predicting a protein structure have been proposed and can be briefly classified as either physicochemical (*de novo*) simulations or template-free modeling methods. Other methods, called template-based or homology modeling, predict structures based on templates and their sequence alignment to a target protein.

The advancement of computing performance and computational optimization algorithms in the computer science field has led to *de novo* method improvement. This is a prediction method that does not use known homologs or fold recognition. ROSETTA [24] and QUARK [25], [26] have existed at the top of this field. However, AlphaFold achieves the state of the art accuracy in 2019 [27], [28]. AlphaFold uses highly accurate contact map predictions powered by deep neural network and main- and side-chain optimization using many high-performance computers.

Structure prediction based on *de novo* methods is still a problem nevertheless the

recent improvement. The search space is too large; for example, suppose that rotation is degree of rotational freedom, the candidate number of protein structures that consist of 100 amino acids becomes 3^{100}. Also, it is difficult to execute perfect molecular dynamics simulation, which contains accurate force field, physiological condition, and enough simulation time for completing the folding process. The accuracy and reliability of models by *de novo* methods are much lower than that of models by homology modeling based on alignments with more than 30 % sequence identity. Indeed, *de novo* method can predict the basic topology of a protein or domain in some cases. For about 40 % of proteins of shorter than 150 amino acids, Rosetta often generates models that have sufficient global similarity to the true structure to recognize it in a search of the protein structure database [2], [3]. However, the accuracy of *de novo* models is often low for problems requiring high-resolution structure information.

Homology modeling

Homology modeling predicts structures based on templates and their sequence alignment to a target protein. Template structures are the structures of homologous proteins, often found with homology detection methods. After experimental structure determination success to determine the structure of the representative protein of a family, homology modeling is often used for proteins of the same family.

Reviewing the results of the critical assessment of structure prediction (CASP) held in 2018, physics-only-based methods were less successful in contributing the best models for specific targets. Instead, machine learning-based methods for predicting residue-residue contacts, and now distances, currently outperform the physics-only methods by far [29]. As one example, among non-template methods, some models from the AlphaFold achieve about 0.8, which are near template-based models. When the templates are available, the template-based structure prediction, such as homology modeling, generates accurate models [30]. In these targets, more than 50 % of the prediction results achieve a high

similarity score ($0 <$ GDTTS < 1; near 1 means near-native) greater than 0.8, and 98 % of the results shows a similarity score greater than 0.5. On the other hand, other prediction methods without explicit templates achieve a similarity score of 0.6 on average for targets that the templates cannot be found.

Currently, homology modeling methods are the most practical because the predicted models are much more accurate if we can find good templates and protein sequence alignments. The resulting models of homology modeling can be used for a wide range of applications. In this field, MODELLER [31] and SWISS-MODEL [32] are the state of the art methods. These homology modeling methods commonly consists of the following procedure:

1. Obtain template protein and the structure by homology detection method.
2. Make alignment of the prediction target sequence and template protein sequence.
3. Predict the structure and optimize the results.

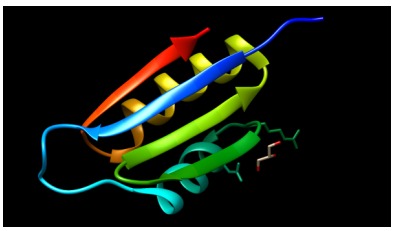
Improvements in experimental methods and discoveries of protocols will continue to determine protein structures one by one. This means that the number of objects for homology modeling will increase, and homology modeling will continue to be used for structure prediction.

1.3. Problem of current homology modeling methods

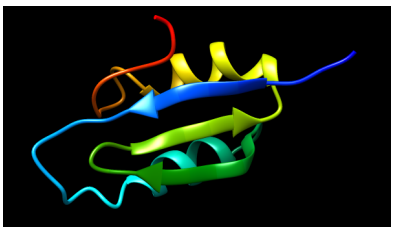
Homology modeling is often accurate and is now widely used. However, there is some room for improving prediction accuracy. The first issue is homology detection. Homology modeling requires homologous proteins as a template for which the structure is already known. If homologous proteins as templates are not found, homology modeling can not be used. In long-term homology detection studies, the first generation was sequence-sequence comparison methods based on a fixed substitution matrix, such as FASTA

[33] and BLAST [34]. To improve the search sensitivity, methods to compare sequences and sequence profiles based on multiple sequence alignments were introduced. Those methods, PSI-BLAST [35] and DELTA-BLAST [36] detected remote homology with high accuracy. As profile-profile comparison methods, FORTE [37], FFAS [38] and SPARKS-X [39] are well known. Then, the hidden Markov model (HMM)-based methods, which are a subset of the sequence profile-based methods, were developed. Now, HMM-HMM comparison methods, such as HHpred [40], SAM [41] and HMMER [42] have performed excellently in structure prediction benchmarks [43], [44] and are considered as the state-of-the-art methods in the field. Homology detection has been the main research target of this field because better templates drastically improve prediction accuracy. However, there is another issue that affects accuracy.

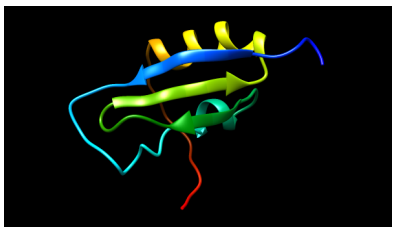
The second issue is quality of sequence alignment. Homology modeling generates structure models based on template structures and a sequence alignments. However, the sequence alignments are often not optimal for accurate homology modeling. In many cases, a sequence alignment obtained from homology detection is used in homology modeling because it generally generates a sequence alignment for calculating the sequence similarity score to be used for judging the homology. As an example of the influence of differences among sequence alignments in homology modeling, figure 1.4 shows pairwise alignments of a target protein and a homologous protein sequence generated by various homology detection tools (FFAS [38], HHpred [46], and SPARKS-X [39].) Figure 1.4a–1.4e and table 1.4f show the predicted models generated based on the alignments and their prediction accuracy. TM-score [47] is a structure similarity measure which extends the approaches used in the Global Distance Test (GDT) [48] and MaxSub [49]; more closer to 1, more accurate prediction. The DALI [45] algorithm is a structural alignment and was used for showing the ideal alignment. All alignments by homology detection tools differed from the structural alignment, and a model based on



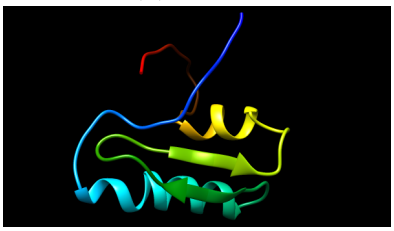
(a) Native



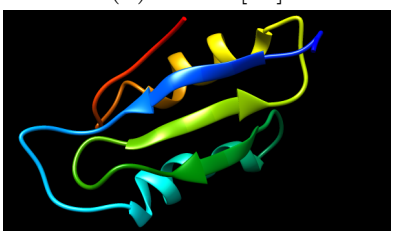
(b) DALI [45]



(c) FFAS [38]



(d) HHpred [46]



(e) SPARKS-X [39]

(1.4a) Native	TM-score
(1.4b) DALI	0.710
(1.4c) FFAS	0.476
(1.4d) HHPRED	0.474
(1.4e) SPARKS-X	0.524

(f) Model accuracy

(g) Ideal sequence alignment (Structural alignment by DALI)

```
4PWU   QQQEATLAIRPVGQ--GIGMPD-GFSVWHHLDANGIRFKSITPQ-KDGLLIKFDSTAQGA    56
3OFE   --RTLMTFVSVTGNPTREESDTITKLWQTSLWNNHIQA-ERYMVDDNRAIFLFKDGTQAW    57

4PWU   AAKEVLGRALPHGYIIALLE    76
3OFE   DAKDFLI-EQERCKGVTIEN    76
```

(h) Sequence alignment 1 (by FFAS)

```
4PWU   QQQEATLAIRPVGQGIGMPDGFSV----WHHLDANGIRFKSITPQKDGLLIKFDSTAQGA    56
3OFE   ---RTLMTFVSVTGNPTREESDTITKLWQTSLWNNHIQAERYMVDDNRAIFLFKDGTQAW    57

4PWU   AAKEVLGRALPHGYIIALLE    76
3OFE   DAKDFLIE------------    65
```

(i) Sequence alignment 2 (by HHpred)

```
4PWU   QQQEATLAIRPVGQ-GIGMPDGFSVWHHLDANGIRFKSITPQKDGLLIKFDSTAQGAAAK    59
3OFE   ------VSVTGNPTREESDTITKLWQTSLWNNHIQAERYMVDDNRAIFLFKDGTQAWDAK    54

4PWU   EVLGRALPHGYIIALLE    76
3OFE   DFL--------------    57
```

(j) Sequence alignment 3 (by SPARKS-X)

```
4PWU   QQQEATLAIRPVGQGIG--MPDGFSVWH-HLDANGIRFKSITPQKDGLLIKFDSTAQGAA    57
3OFE   --RTLMTFVSVTGNPTREESDTITKLWQTSLWNNHIQAERYMVDDNRAIFLFKD---GTQ    55

4PWU   AKEVLGRALPH-GYIIALLE    76
3OFE   AWDAKDFLIEQERCKGVTIE    75
```

Figure 1.4.: Example of alignments that are not suitable for template-based modeling. Existing methods cannot generate whole ideal alignments. The target protein is PDBID: `4PWU`, and the template protein is PDBID: `3OFE`.

the structural alignment was the nearest match to the a native structure.

Even if the alignment by homology detection is not appropriate and better alignment may be generated by other methods, these are often used. The problem occurs in the case of homology modeling by remote homologs. These remote homologous templates could not be used because a static substitution matrix could not generate alignment. If a more accurate model is required, researchers must often edit alignments manually before modeling to improve their quality. For the case, an automatic alignment generation method is required for accurate homology modeling that uses low sequence identity alignments. The alignment results of remote homology detection often do not generate accurate homology modeling results. Recent homology search methods have been able to detect remote homologs, although sometimes sufficiently accurate structure models cannot be obtained because the quality of the sequence alignment generated by the homology detection program is poor. This problem has been mentioned in several studies [50] in which researchers have tried to improve alignments manually based on their knowledge of biology; fully automated methods are still required. If there is a method to make low sequence identity homologs available for homology modeling, more template candidates whose protein structure is already known become available.

In essence, alignment quality is crucial to homology modeling. Thus far, a method's ability to detect remote homologs has been prioritized because models cannot be generated without a template. However, to achieve higher-accuracy homology modeling, the improvement of sequence alignment generation is a critical open problem. Thus, the alignment generation methods for remote homologs are needed.

1.4. Related researches

Historically, the studies of sequence alignment generation in structure prediction have been mostly consistent with studies of sequence homology search. Most sequence ho-

mology search methods perform a sequence alignment between a query sequence and a sequence in a database for scoring the smilarity. Sequence homology search was intended initially to investigate evolutionary relationships and did not evaluate whether sequence alignment is optimized for structural prediction at the time of the search.

The most basic method of sequence alignment generation is dynamic programming. Dynamic programming allows us to generate the optimal alignment for a given score model, such as the Smith-Waterman algorithm [51]. There are two major streams of research, one of which is to increase the speed. Methods such as FASTA [33] and BLAST [34] are known to speed up the process by generating a semi-optimal alignment using heuristics. Another stream is research to improve the score model to increase the sensitivity of the search. There are studies of static score models that derive relative substitution ease from existing sequence alignments during the evolutionary process, such as BLOSUM [52]. This is called an amino acid substitution matrix. It allows us to generate alignments that do not rely solely on simple amino acid matches, which increases the sensitivity of the search. Naïve dynamic programming is computationally expensive ($\mathcal{O}(mn)$ for a length m of the query sequence and n of template sequence) and difficult to use when the sequence database is large. Therefore, semi-optimal but fast sequence alignment generation methods have been studied. Although these methods are useful for searching, they have unique problems, such as short sequence alignments, and their quality is still poor in terms of sequence alignment for structure prediction.

On the other hand, improvements in the score model have led to research in more sensitive detection. The use of sequence profiles incorporating evolutionary information and their effective use has become the main topic. Sequence alignment has also been improved; as a result, to discover distantly related homologous proteins with low sequence similarity. Compared to earlier studies, the quality of sequence alignment for structural predictions is expected to improve, but this has not been validated. If the

goal is to perform sensitive sequence homology searches, using sequence profiles is a good idea. However, the quality of structural prediction using its sequence alignment is a case by case basis. Here, we introduce three highly sensitive detection methods as related researches.

PSI-BLAST [35] is a classical but sensitive homology detection method. The feature is still that it uses the position-specific score matrix, the acronym PSSM, for search. PSSM is a statistically processed matrix for multiple homologous protein sequence alignments created by some method, with the frequency of amino acid occurrences and snapping sequence profiles created for each position. This allows for the inclusion of evolutionary information at each position in the amino acid sequence. PSI-BLAST is a mechanism that searches against a sequence database based on this input. Because searches are performed using the relational information of homologous proteins, distantly related proteins are easily detected. Because sequence alignments are updated with each search, it is possible to create further PSSMs from the updated alignments, repeat the search, and discover more distantly related proteins. The quality of sequence alignment for structural predictions is also improved if closely enough related protein sequences are available.

Then, another method using this PSSM, called FFAS [38], was proposed. FFAS has been improved from PSI-BLAST, which compares sequences and PSSM, to a comparison between sequence profiles. In other words, the database side is changed to a database of array profiles. Because the input and the databases have relational information to each other, the comparison is more sensitive. As a side effect, it is expected that the quality of alignment for structure prediction has been improved. As shown in figure 1.4, this is not sufficient because optimization of alignment for structure prediction is not their goal.

Recently, Soding *et al.* proposed more sensitive method, HHsearch [53]. This is

similar to FFAS, in the lineage of comparisons between sequence profiles but is a non-starter. This method's gist is that sequence alignment between homologous proteins is modeled by hidden Markov models (HMMs). PSSMs become HMMs, and search databases become HMM databases. A method to compare HMMs with each other was proposed, and a search is now possible. This method is clever in many respects, such as the fact that gaps are represented in the model, the method to find maximum likelihood pathways and generate alignments, etc. As a result, it has an excellent ability to find distantly related homologous proteins and is said to be state of the art in search. The tools are integrated with structure prediction by homology modeling and are said to be highly accurate. However, when template proteins are distantly related, there is still room for improvement in structure prediction accuracy.

Current homology search methods are so sensitive but there are still problems in the view print of structure prediction as described in the previous section, . First, the alignment generated is often inappropriate for structure generation. The reason is that template protein discovery is also still important for homology modeling. Sequence alignment is a secondary output to improve search accuracy and is primarily used as a basis for search results. If the search is correct, it does not give any superiority to the appropriateness of the alignment for structural predictions. The fact that some methods do not generate alignments was a problem for developing the method in this study. In this study, the quality of sequence alignment, which has not been directly evaluated in previous studies, is assessed by structure prediction.

1.5. Research purpose

The purpose of this research is to automatically generate computationally optimal alignment for accurate homology modeling, even for remote homologs, which are often detected by modern homology search tools. It aims to make more templates available that

previous methods could not use and increase the number of template candidates. In this research, by using structural alignments of known homologs and other related homologs, we optimized alignments automatically by a data-driven method. We focused on structural alignment-based as supervised machine learning and intermediate sequence-based method and develop alignment generation algorithms.

1.6. Contributions

The main contribution of this thesis is to develop new sequence alignment method for accurate homology modeling especially with low sequence identity pair of query and template proteins. As we denoted, recent homology detection methods have been able to find remote homologs, although sometimes sufficiently accurate structure models cannot be obtained because the quality of the sequence alignment generated by homology detection program is poor. This is becasue the main purpose of homology detection methods is to find homologous proteins from a database and quality of sequence alignment has not been evaluated. If a more accurate model is required, researchers must often edit alignments manually before modeling to improve their quality. Thus, to achieve higher-accuracy homology modeling, the improvement of sequence alignment generation is a critical open problem. This is the first study dedicated to better sequence alignment generation and the proposed automatic and accurate sequence alignment generation method would be helpful for biological researchers utilizing homology modeling.

The novelty of the method described in chapter 3 is that proposing a new pairwise sequence alignment generation method based on a machine learning model that learns the structural alignments of known homologs. In structural alignment, the structural difference between a target protein structure and a template protein structure is minimized. Thus, sequence alignments generated by structural alignment are ideal for homology modeling. Since it is difficult to use machine learning to directly predict sequence align-

ment, we instead use dynamic programming during sequence alignment to dynamically predict the substitution score from the trained model instead of a fixed substitution matrix or profile comparison. Machine learning is used for this substitution score prediction process. Recently, machine learning methods have demonstrated power in homology detection, fold recognition, residue contact map prediction, dihedral prediction, model quality assessment and secondary structure prediction [54]–[59]. Machine learning also seems effective for tackling the problem of alignment generation for homology modeling. However, this topic has not been studied because it is difficult to treat alignment generation as a classification or regression problem.

The novelty of a method proposed in chapter 4 is developing a novel pairwise sequence alignment method for remote homologs from the intermediate sequence search (ISS). The basic idea of ISS is the following: two sequences of remote homologous proteins, which do not have enough sequence identity or a close relationship evolutionally, can be related via another sequence whose characteristics and features are intermediate between the two remotely homologous proteins. To our knowledge, this is the first study demonstrating the generation and evaluation of alignments using ISS results in the context of template-based modeling. In the ISS method, after searching for homologs of the query protein in the database, the results are used as new queries to detect more distantly related homologs by re-running the homology search. By identifying a connection via these intermediate sequences, the ISS method can detect relationships between the original query protein and remote homologs. Unlike other profile-based methods such as PSI-BLAST, the ISS detects remote homologs by these intermediate homologous proteins. This alignment method is effective when generating alignment of remote homologs that are not found by existing alignment-based sequence search methods. We found that this method can be used for these difficult pair of query and template proteins that our machine learning-based alignment method cannot generate alignment.

1.7. Contents of this thesis

The chapter 2 describes background and overview of homology modeling as well as problems of homology modeling. Homology modeling is the most practical structure prediction method, and there is some room for improvement. Specifically, in this chapter, we will focus on alignment quality for accurate homology modeling.

In chapter 3, we propose a new pairwise sequence alignment generation method based on a machine learning model that learns the structural alignment of known homologs. Since it is difficult to directly predict sequence alignment using machine learning, we instead use dynamic programming during sequence alignment to dynamically predict a substitution score from the learned model instead of a fixed substitution matrix or profile comparison. Machine learning is then used in this substitution score prediction process.

In chapter 4, we propose a new sequence alignment generation method for remote homologs detected by an intermediate sequence search for homology modeling. Based on our study, this is the first study that demonstrates the generation and evaluation of alignment using ISS results in the context of homology modeling.

Chapter 5 describes impact of model accuracy improvement for protein interaction estimation methods. We show how much structure prediction accuracy affects accuracies of subsequent applications. Furthermore, we integrate our methods by merging the methods and show how much structure prediction accuracy improves.

Chapter 2.

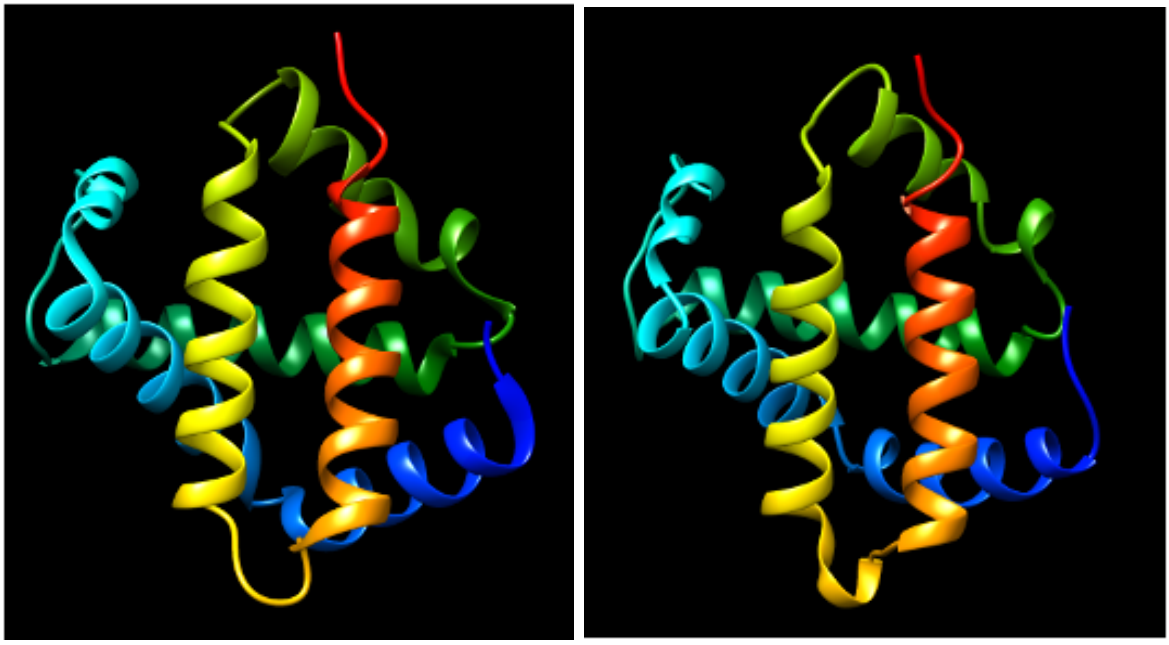
Homology modeling

Homology modeling predicts structures based on templates and their sequence alignment to a target protein. Template structures are the structures of homologous proteins, often found with homology detection methods. Currently, homology modeling methods are the most practical because the predicted models are often accurate if we can find good templates and obtain good sequence alignments between query and templates.

2.1. Background of homology modeling

2.1.1. Homology

The similarity of sequences from the gene of a common ancestor is called homology (example in figure 2.1). Homologous proteins often have a similar structure as well as function. By gene specification and duplication, amino acid sequence coded in the gene is altered, and the protein is also changed, which is called as molecular evolution. Usually, the structure of a protein tends to keep more similarities than the sequence. Therefore, there are homologous proteins that sequence identity is low, but the structure and the function are similar.



(a) Protein structure of human hemoglobin (b) Protein structure of chicken hemoglobin

(c) Sequence alignment between hemoglobin amino acid sequences of human and chicken. "-" in the alignment means a gap.

```
Human     VLSPADKTNV---------HAGEYGAEALERMFLSFPTTKTYFPHFDLSH     41
Chicken   MLTAEDKKLIQQAWEKAASHQEEFGAEALTRMFTTYPQTKTYFPHFDLSP     50


Human     GSAQVKGHGKKV     53
Chicken   GSDQVRGHGKKV     62
```

Figure 2.1.: Homologous proteins and the sequence alignment. Hemoglobins of human and chicken are homologs, and the amino acid sequences are similar each other.

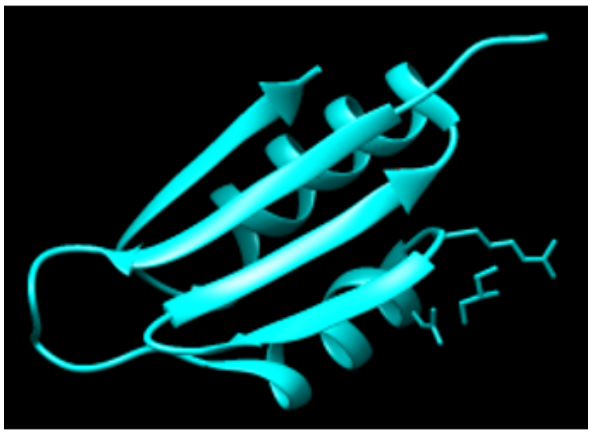
2.1.2. Sequence alignment

Even if two homologous proteins share similar structure, we cannot execute homology modeling by this fact only. Homology modeling requires a mapping of amino acids (residues) between a query sequence and a template sequence; this mapping is called a sequence alignment. The sequence alignment means evolutionary relationship of amino acids of two proteins and is used in homology detection for calculating sequence similarity between two amino acid sequences. Aligned residues indicates one residue was replaced by the other through the evolutionary process. Unfortunately, there is no correct answer of sequence alignment because we cannot directly observe the process of protein evolution. Thus, sequence alignment is estimated by maximizing the similarity score of a given score model or stochastic model.

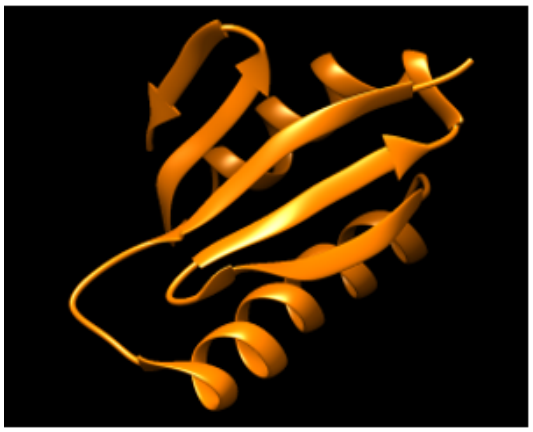
By using dynamic programming that adds high score at the position where similar residues exist and some penalty for gaps, an optimal sequence alignment can be calculated. Smith-Waterman algorithm [51] is classical one of the naïve dynamic programming-based alignment methods. As we denoted, to generate a sequence alignment, we need a score model to quantify amino acid substitutions. The basic score model is amino acid substitution matrix, which provides substitution score between two amino acid types. The matrix is calculated by applying stochastic procedure to known homologous proteins' sequences. PAM and BLOSUM are well known for this purpose. Especially, BLOSUM is the default matrix of BLAST [34]. To improve the alignment quality and protein similarity estimation, various models have been developed

2.1.3. Structural alignment

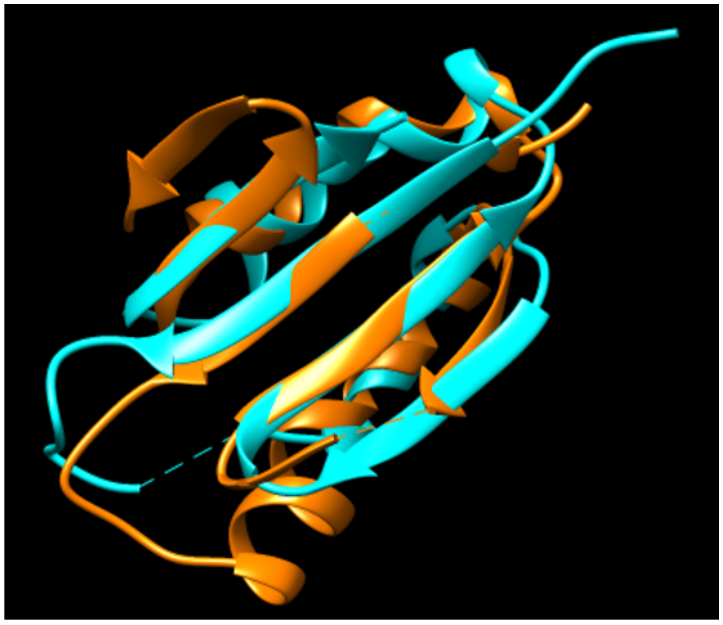
Another protein alignment is a structural alignment (example in figure 2.2). Structural alignment is based on superposition of 3D structure, and this means that structural alignment does not use residue similarity for aligning sequence. In structural align-



(a) 4PWU



(b) 3OFE



(c) Superimposed

(d) Structural alignment

```
4PWU  QQQEATLAIRPVGQ--GIGMPD-GFSVWHHLDANGIRFKSITPQ-KDGLL    46
3OFE  --RTLMTFVSVTGNPTREESDTITKLWQTSLWNNHIQA-ERYMVDDNRAI    47

4PWU  IKFDSTAQGAAAKEVLGRALPHGYIIALLE    76
3OFE  FLFKDGTQAWDAKDFLI-EQERCKGVTIEN    76
```

Figure 2.2.: Structural alignment. Proteins 4PWU and 3OFE are similar each other. The structural alignment is generated by superimposing the two proteins so that maximising locally structural identity. Dashed lines in 2.2c show gaps by structural alignment.

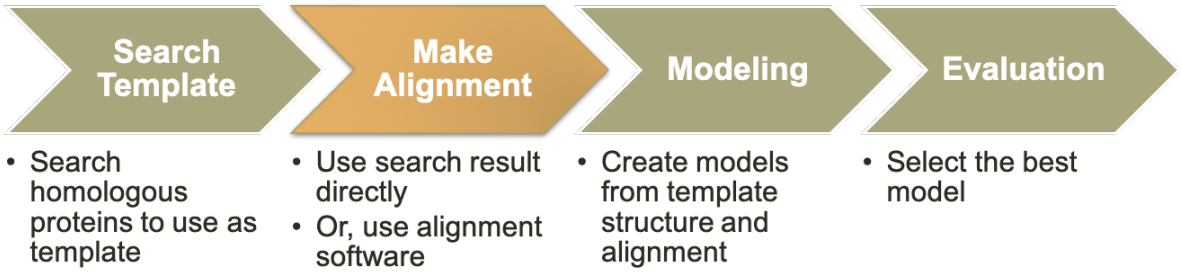


Figure 2.3.: Overview of homology modeling. In this thesis, we focus on alignment generation phase.

ment, the aligned region emphasized that structures are similar in the region. This shows valuable information for relationship of sequence and the structure of homologs. MANMOTH [60], DALI [45] and TMalign [61] are well known methods. By applying structural alignment to big known homolog database, some of these developers create homologous protein database with structural alignment information.

2.2. Protocol of homology modeling

Homology modeling consists of the following procedure:

1. Obtain template protein and the structure by homology detection method.
2. Make alignment of a prediction target sequence and template protein sequence.
3. Predict the structure and optimize the results.

Figure 2.3 is the overview diagram of homology modeling procedure.

Homology modeling uses the homologous proteins as a template for structure prediction. Actually, homologous protein sequences are not identical because of insertion, deletion, and substitution. These regions are modeled by other methods, which often includes knowledge of experienced biologists. Homology modeling predicts structures based on templates and their sequence alignment to a target protein. Template structures are the structures of homologous proteins, often found with homology detection methods. Currently, homology modeling methods are the most practical because the predicted models are much more accurate if we can find good templates and protein sequence alignments. If sequence identity of the alignment is high enough, homology modeling can achieve root mean square distance under 1 Å. After experimental structure determination success to the determinate structure of the representative protein of a family, homology modeling is often used for proteins of the same family.

There is no perfect solution for modeling of non-identical regions of alignment, and homology modeling also has its advantages and disadvantages. Empirically, homologous proteins often conserve sub-sequences of important sites for the function. Because homology modeling successfully can predict the structure of these conserved regions, it becomes one of the widely used methods of protein structure prediction. However, as well as modeling of the functional site, the whole modeling of protein is still needed for protein interaction prediction and functional or docking site prediction.

2.3. Current homology modeling methods

The protocol of homology modeling methods are similar to each other. However, homology modeling basically consists of some procedures which have been researched and developed independently: homology detection, sequence alignment and model generation.

2.3.1. Model generation

In the 1970s, manual homology modeling is shown by plastic ball and stick. In the 1980s, rigid fragment assembly is developed. This method finds conserved regions of protein structure by superposing homologous proteins. The region where the sequence identity of the prediction target sequence and the conserved region is high is the target of structure prediction. As for loops that are non-conserved regions, the most reasonable models are found by searching a structure database. If many templates are available, rigid body assembly achieves near X-ray structure. Even though it is very old, rigid body assembly is still used.

Next, the segment matching method is developed. The idea is that about five residues in a protein form 100 class variety. In the 1990s, spacial restraint satisfaction methods appeared. One is the distance geometry method, which uses upper and lower bound of distances and dihedral angles and predicts all-atom coordinates. Another method is the spacial restraint satisfaction method. Firstly, it generates models that satisfy the restrains of distance and dihedral angles. Secondly, spacial restraint and force field to enforce proper stereochemistry are combined into an optimization function. Lastly, it generates models that minimize the optimization function. The spacial restraint satisfaction method can combine various restrains to the target function, and we can say that it contains other methods of homology modeling.

As the next step, the loop region is predicted. Loop structure means three kinds of regions:

1. Regions between-helix and-sheet
2. Regions of various length configurations
3. Regions of the surface of the structure

Generally, substitution, insertion, and deletion occur in the exposed loop region, which

does not have a specific secondary structure. These exposed regions are important for docking and ligand binding because these regions are exposed to the solvent. Therefore, loop prediction is important for the usage of prediction models. Actually, from the lessons of rigid body assembly, it is difficult to predict the structure of the loop region that contains over five residues. There are various methods, such as combined with de novo method, loop database search, or combination of them.

In the early days, coordinate prediction of the whole atom was tried and developed, but it was unsuccessful. Recently, the structure of only the backbone chain is predicted firstly, the loop region is modeled secondly, and the side chain coordinate is predicted and optimized. On the accurate prediction of the backbone chain, efficient algorithms of side-chain position prediction are known. Side-chain packing and optimization are also important because the side chains have an important role in molecule recognition. Side-chain conformation is also predicted by template structure. Disulfide bridges are treated as a special case because the bonding is strong and important for protein's folding. Basically, during side-chain position prediction, the backbone chain is fixed. Some methods allow shifting backbone chain during side-chain prediction, but they cannot achieve high accuracy. Many methods treat the prediction as an optimization problem.

At the final phase, predicted models are evaluated and selected. If some candidates are available, some of them are selected. In a basic method, the Ramachandran plot is used for checking collision and structural inconsistency. Also, scores from homology modeling software are used, which are pseudo-potential energy function. These evaluation and selection approaches are called as model quality assessment or model accuracy estimation. Recently, machine learning shows its efficiency in this field.

MODELLER

Within existing homology modeling methods, MODELLER is firstly developed by Sali as a spacial restraint satisfaction method. The idea of MODELLER is to generate the maximum likelihood model that satisfies various spacial restraints. MODELLER works as follows:

1. Generate alignments of a target sequence and template sequence.
2. Mapping structural and spacial restraints to a target sequence in the alignment.
3. Define probabilistic density on each restriction and generate the most reasonable model where the maximum likelihood.

Various spacial restraints are defined by researching groups of known homologs, for example, distance distribution of Cα atoms and dihedral angles.

2.3.2. Template search and selection

Homology modeling requires templates. Template structures are the structures of homologous proteins, often found with homology detection methods. The idea of homology modeling is old, but it was unsuccessful and did not see the light of day. However, in the 1990s, remote homology detection such as PSI-BLAST show the light on homology modeling because these methods could detect remote homologs with high sensitivity.

In long-term homology detection studies from FASTA [33] and BLAST [34] as sequence-sequence comparison method, profile-sequence comparison methods based on multiple sequence alignments, such as PSI-BLAST [35] and DELTA-BLAST [36], have detected homology with high accuracy. As sequence profile-profile comparison methods, FORTE [37], FFAS [38] and SPARKS-X [39] are well known. They often achieved higher sensitivity than older methods of sequence-sequence comparison. As state-of-the-art methods of homology detection, hidden Markov model (HMM)-based methods, a subset of sequence

profile-based methods, also detect remote homologs; HMM comparison methods, such as HHpred [40], SAM [41] and HMMER [42] have performed excellently in structure prediction benchmarks [43], [44].

For a long time, many homology detection methods are developed, and these roughly categorized to alignment-based and non-alignment-based. Using sequence alignment based on substitution matrix is reasonable because, when two amino acid sequences share common sub-sequences in each other, the two protein can be homologous. Generally, structural features of protein are usually more pronounced than sequence similarity [62]. This indicates that structural alignment can be more accurate for homology detection, but usually structural information cannot be used because 3D structure is not resolved in many cases of homology detection. Homology databases with structural alignment are also useful for getting homology information. These databases are often integrated with sequence alignment-based homology detection. By showing known structurally similar proteins after alignment-based homology detection, the database can show various information of homologous proteins. Machine leaning-based fold recognition is one of homology detection without sequence alignment. Also, FORTE [37] uses correlation of two position specific score matrices.

The intermediate sequence search (ISS) method has been proposed to provide more distantly remote homology detection [63]. The basic idea of ISS is the following: two sequences of remote homologous proteins, which do not have enough sequence identity or a close relationship evolutionally, can be related via another sequence whose characteristics and features are intermediate between the two remotely homologous proteins. If the match score between both of the first and third sequences and the second and the third sequences is high, it can be concluded that the first and second sequences are related, even though their sequence similarity is low. In the ISS method, after searching for homologs of the query protein in the database, the results are used as new

queries to detect more distantly related homologs by re-running the homology search. By identifying a connection via these intermediate sequences, the ISS method can detect relationships between the original query protein and remote homologs. The idea of the intermediate sequence search itself is not novel [63]. Decades ago, Entrez [64] provided intermediate sequence information. However, the naïve ISS procedure often provides many false positives [65] and requires significant computing resources to evaluate many homology searches. Recently, to overcome the computational demand and occurrence of false positives, approaches that utilize network or graph theory were proposed [66], [67]. In addition, machine learning-based intermediate sequence search methods have demonstrated good results [68], [69].

2.3.3. Alignment generation and correction

Homologous protein sequences are not identical because of insertion, deletion, or mutation of residues. To model these regions, homology modeling requires the alignment of the target sequence and template sequence. Because many homology detection methods generate alignments for scoring, the alignments are also used for homology modeling directly. However, if sequence identity is low, especially for remote homologs, the tools cannot generate alignment, or the length of alignment is too short for the target sequence. In the case, other information such as secondary structure prediction is used for the additional data source, and the alignment is manually edited.

Alignment quality is crucial to accuracy of homology modeling. Thus far, a method's ability to detect remote homologs has been prioritized because models cannot be generated without a template. However, to achieve higher-accuracy homology modeling, the improvement of sequence alignment generation is a critical open problem.

2.4. Problem of homology modeling

The model generation itself has a long research history, and there are no significant improvements after MODELLER [31] and SWISS-MODEL [32]. However, there are some possibilities for improving prediction accuracy in other procedures of homology modeling.

Homology detection is one of them. Homology modeling requires homologous proteins that structure is already known as a template structure. If homologous proteins as templates could not be found, homology modeling could not be used. Therefore, for a long time, the homology detection method for remote homologs is developed. Recent homology detection methods have been able to detect remote homologs. Machine learning methods are recently developed, and some of them achieve state-of-the-art accuracy for homology detection or fold recognition However, they do not generate alignment which is required for homology modeling. In this case, alignment generation phase is separated and executed.

Also, in some cases, sufficiently accurate structure models cannot be generated because the quality of the sequence alignment generated by the homology detection program is poor. Homology detection affects the accuracy of homology modeling. However, the alignment from the tools is often not appropriate for homology modeling because the tools are developed for homology detection itself and put importance on the sensitivity, selectivity, and execution speed. In many cases, alignment from homology detection is used because it often generates alignment for calculating a score. Even if the alignment is not appropriate, these are often used.

The problem occurs in the case of homology modeling by remote homologs. Statistics of structural database shows the average sequence identity of a pair of homologs is 20–30 % on average [71], [72]. These remote homologous templates could not be used because a static substitution matrix could not generate alignment. It is known that

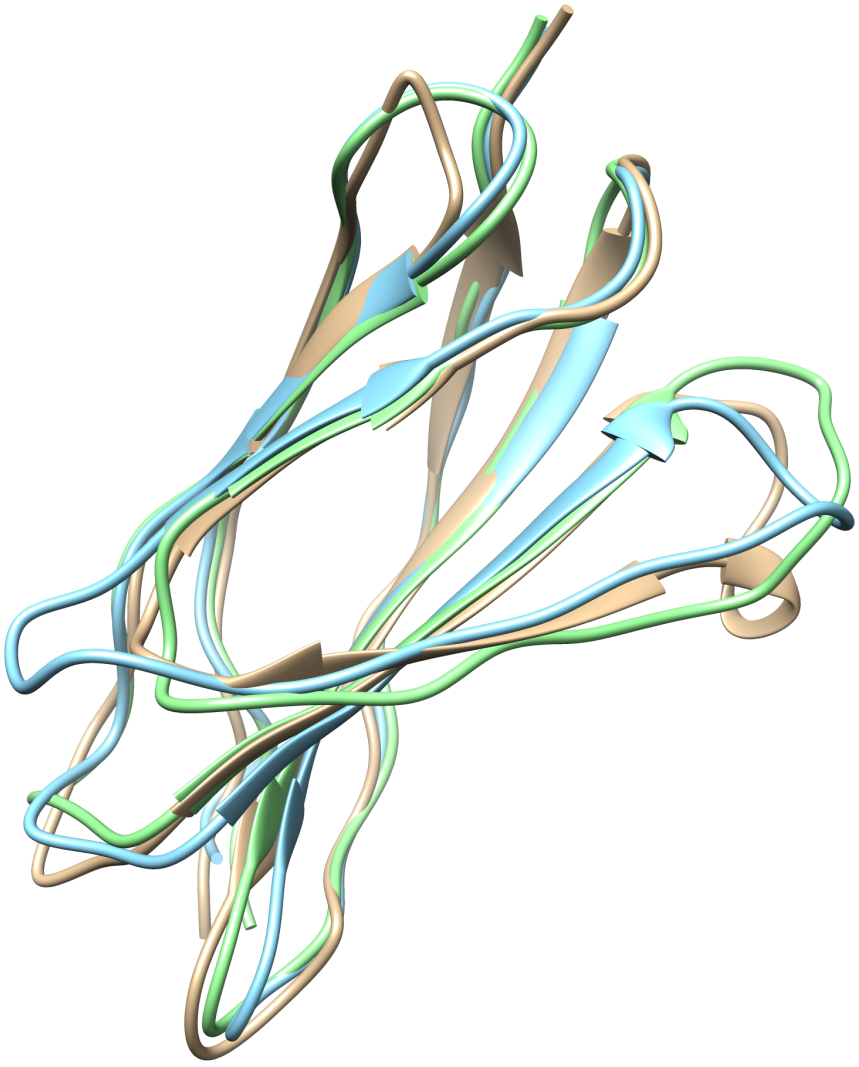


Figure 2.4.: Model differences. Query (yellow) and template proteins are `1QG3A` and `1VA9A`, respectively. The green model is generated from a structural alignment (TM-align), and the blue model is from HHsearch. The TM-scores of HHsearch and structural alignment are 0.801 and 0.881, respectively. [Molecular graphics were performed with the UCSF Chimera [70] package.]

particular amino acids are able to substitute each other at each position in the core of a particular protein. Here, we have two problems. Firstly, we may not be able to get structurally reasonable alignments by standard amino acid substitution matrix, like the PAM matrix or BLOSUM matrix. This is because valid substitutions of amino acids are different among each position in the core of various proteins. The substitution rates of these matrices are obtained by averaging among a lot of sequence alignments, and these substitutions are observed in both core and loop regions of proteins. We can make matrices that indicate conservative regions in multiple alignments of similar proteins' sequences, for example, position-specfic scoring matrix. These matrices can contain information of amino acids that occurred in each line of multiple sequence alignments. This is a powerful method to search sequence patterns in a new protein sequence that are similar to known protein sets. Secondly, gaps should not be inserted into core regions of a protein. The number of gaps that are inserted into core must be small, if any.

Nevertheless, the criteria of whether homology modeling can be used or not is mainly sequence identity of the target sequence and template sequence. This is sometimes reasonable because homologous protein sequences are often similar to each other. Generally speaking, sequence identity of 50 % is the threshold for homology modeling. If sequence identity is high, generating alignment for accurate homology modeling is an easy task because the static substitution matrix works well. In the opposite case, if sequence identity is low, manual adjustment to the alignment is required, or may have to give up homology modeling even if homologs are available. Model accuracy often becomes low if the sequence identity of the target sequence and template sequence is low. In this case, an automatic alignment generation method is required for accurate homology modeling that uses low sequence identity alignments. Currently, if a more accurate model is required, experienced researchers must often edit alignments manually before modeling to improve their quality. This often becomes a kind of craftwork, and it is not

guaranteed that the alignment is optimal or suitable for accurate homology modeling. The craftspersonship may result in low reproducibility of the results.

This problem has been mentioned in several studies [50] in which researchers have tried to improve alignments manually based on their knowledge of biology; fully automated methods are still required. [73] proposed an automated method to improve alignments by optimizing gap penalties. They evaluated the premise of gap location in protein 3D structures by examining large protein structure datasets and found that the distribution of gaps in protein 3D structures differed from previous studies. However, they used the technique mainly for homology detection, and the quality of its alignments for prediction model accuracy was still unclear. The problem of alignment quality for homology modeling is mentioned many times. However, it remains an open problem because:

1. Template search is also important and prioritized.

2. Homology modeling is not used when reasonable alignment cannot be made.

3. Optimal alignment is not clear.

In this research, we offer some contributions to solve these problems.

Chapter 3.

Sequence alignment generation by substitution score prediction using machine learning

3.1. Introduction

When two homologous proteins can be detected and homology modeling can be used, sequence alignment quality is important for accurate homology modeling. The alignment should reflect the structure feature. The sequence alignment described in this chapter can generate alignments that are similar to structural alignment without structural information and appropriate for homology detection.

Recent homology search methods have been able to detect remote homologs, although sometimes sufficiently accurate structure models cannot be obtained because the quality of the sequence alignment generated by homology detection program is poor. If a more accurate model is required, researchers must often edit alignments manually before modeling to improve their quality. In structural alignment, the structural difference between a target protein structure and a template protein structure is minimized; thus, sequence alignments generated by structural alignment are ideal for homology modeling (Figure 2.4). Often, the sequence alignments generated by the homology detection methods are dissimilar to those generated by structural alignment, especially for remote homologs. In essence, alignment quality is crucial to homology modeling. Thus far, a method's ability to detect remote homologs has been prioritized because models cannot be gener-

ated without a template. However, to achieve higher-accuracy homology modeling, the improvement of sequence alignment generation is a critical open problem.

This problem has been mentioned in several studies [50] in which researchers have tried to improve alignments manually based on their knowledge of biology; fully automated methods are still required. [73] proposed an automated method to improve alignments by optimizing gap penalties. They evaluated the premise of gap location in protein 3D structures by examining large protein structure datasets, and found that the distribution of gaps in protein 3D structures differed from previous studies. However, they used the technique mainly for homology detection and the quality of its alignments for prediction model accuracy was still unclear.

Recently, machine learning methods have demonstrated power in homology detection, fold recognition, residue contact map prediction, dihedral prediction, model quality assessment and secondary structure prediction [54]–[59]. Machine learning also seems effective for tackling the problem of alignment generation for homology modeling. However, this topic has not been studied because it is difficult to treat alignment generation as a classification or regression problem.

In this chapter, we propose a new pairwise sequence alignment generation method based on a machine learning model that learns the structural alignments of known homologs. Because it is difficult to directly predict sequence alignment using machine learning, we instead use dynamic programming during sequence alignment to dynamically predict a substitution score from the learned model instead of a fixed substitution matrix or profile comparison. Machine learning is used in this substitution score prediction process. We evaluate the proposed method using a carefully split training and test dataset and compare the accuracy of predicted structure models with those of state-of-the-art methods as a measure of sequence alignment quality.

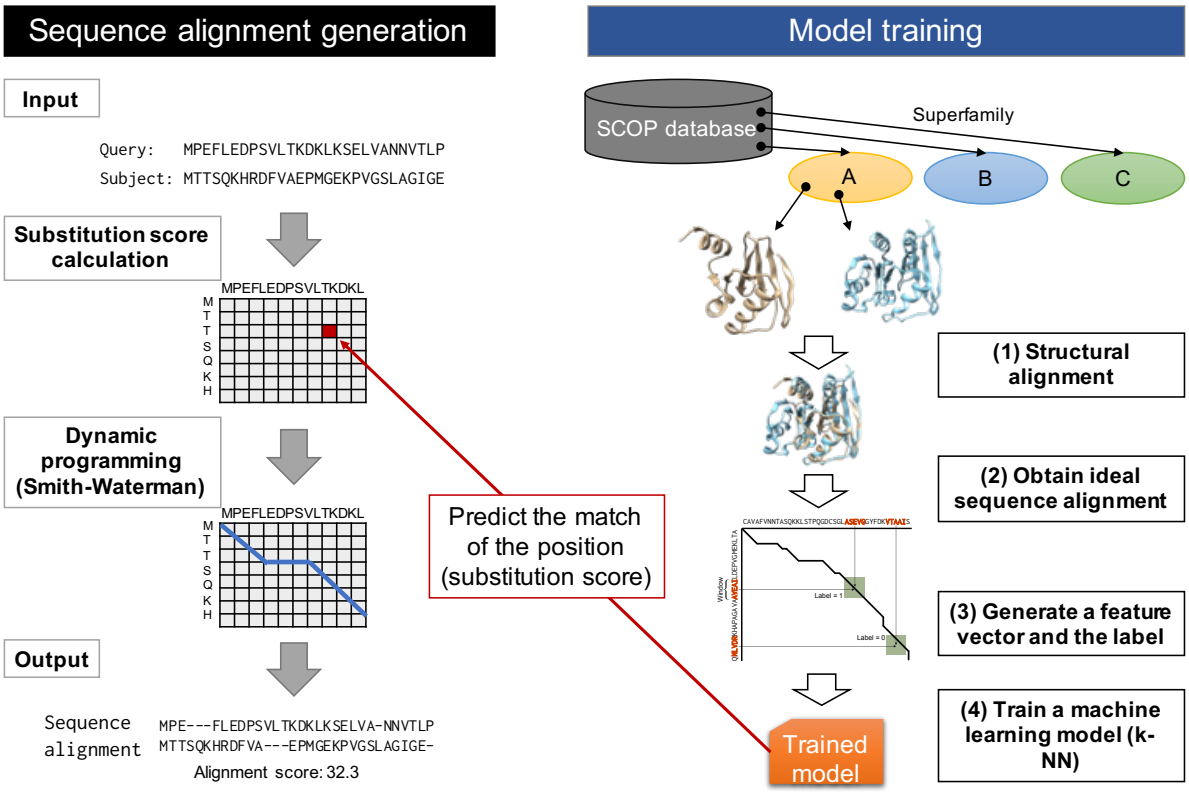


Figure 3.1.: Overview of the proposed method. Two sequences are aligned using the Smith–Waterman algorithm and substitution scores used in the process are estimated by a prediction model. The prediction model is trained to output an alignment similar to the structural alignment.

3.2. Materials and methods

Generally, sequence alignment generation is integrated with the homology detection process and the detection tools output sequence alignments with homology search results from the database. In this study, we focus only on alignment generation. Thus, the inputs are a target's amino acid sequence (query) and another amino acid sequence that was detected as a template by any homology detection method (subject), and the output is an alignment that is more suitable for homology modeling. This process is often called re-alignment. Figure 3.1 shows an overview of our method. The proposed method accepts query and subject amino acid sequences as input, then aligns their sequences using the Smith–Waterman algorithm [51]. In classical dynamic programming,

a substitution matrix such as BLOSUM62 or PAM250, is used to evaluate the match between residue pairs. To improve alignment accuracy, profile comparison methods, including FORTE [37] and FFAS [74], use the similarity between two position-specific score matrices (PSSMs) of a target residue pair. In contrast, we evaluate residue matches based on a supervised machine learning technique. We train a prediction model using pairwise structural alignments of structurally similar protein pairs as the labels of a training dataset. Thus, the method is expected to output similar sequence alignments by structural alignment. The PSSMs of two input sequences are used as input to the prediction model; to predict the match of a residue pair, PSSMs around target residues within a fixed size window are used. Finally, the method returns a sequence alignment and an alignment score as output.

3.2.1. Datasets

Our method needs information about known structurally similar proteins to create structural alignments, for which we used the Structural Classification of Proteins (SCOP) [71], [72] database. The SCOP database classifies proteins by class, folds, superfamily (SF), family and domain based on manually curated function/structure classifications and contains redundant sequences. Thus, we used the SCOP40 database instead, which contains only domains whose sequence identity is $<40\,\%$ to avoid overfitting and reduce execution time. In this study, we define domains that are in the same SF as structurally similar.

For accurate evaluation and parameter optimization, we split training, test and validation datasets from the full dataset. We selected five domains each from seven SCOP classes to cover various protein structure types, selecting test domains only from SFs containing greater than ten domains. We ignored any small SFs and sorted the remaining domains by their PDB revision date, ultimately selecting 35 domains as test data.

For our validation dataset, we split two groups from the remaining dataset. For one group, we selected one domain each from seven SCOP classes for parameter search; the other contains one domain each from the classes for a gap penalty search. Finally, we split 49 domains ($= 35 + 7 + 7$) from all the datasets for test and validation, and the remaining domains were used for training (see appendix B for details).

In the training dataset, we generated structural alignments of every domain pair in the same SF using TM-align [61]. We treated domain pairs whose TM-align score [TM-score [47]] was <0.5 as having low structural similarity and filtered them out [75]. If the SF had only one domain, it was ignored because we could not define a pairwise alignment for it. Finally, 140889 pairwise structural alignments were generated. For PSSM generation, we used three-iteration PSI-BLAST with the UniRef90 [76] database. When the training dataset became too large to process within a reasonable computation time, the training dataset was reduced to 1/10 of its initial size by random selection.

Figure 3.2 shows the change in the number of domains per superfamily of SCOP according to those sequence redundancy-level. In the case of the original dataset (without deduplication) and a dataset that exactly matched sequences were removed (sequence identity $< 100\,\%$), several superfamilies have over 1,000 domain in itself. The number of superfamilies that include many domains decreases in datasets whose sequence identities below 95 % and 40 %. Without deduplication, there are too many domains in a particular superfamily for that superfamily's information to be biased. To reduce the computation time of proposed method by decreasing the amount of data handled by kNN, we used a non-redundant database with sequence identity less than 40 %.

3.2.2. Input vector and label definition

In order to use machine learning methods to predict matching scores, we had to encode information about residue pairs in a numerical vector representation. In addition, we

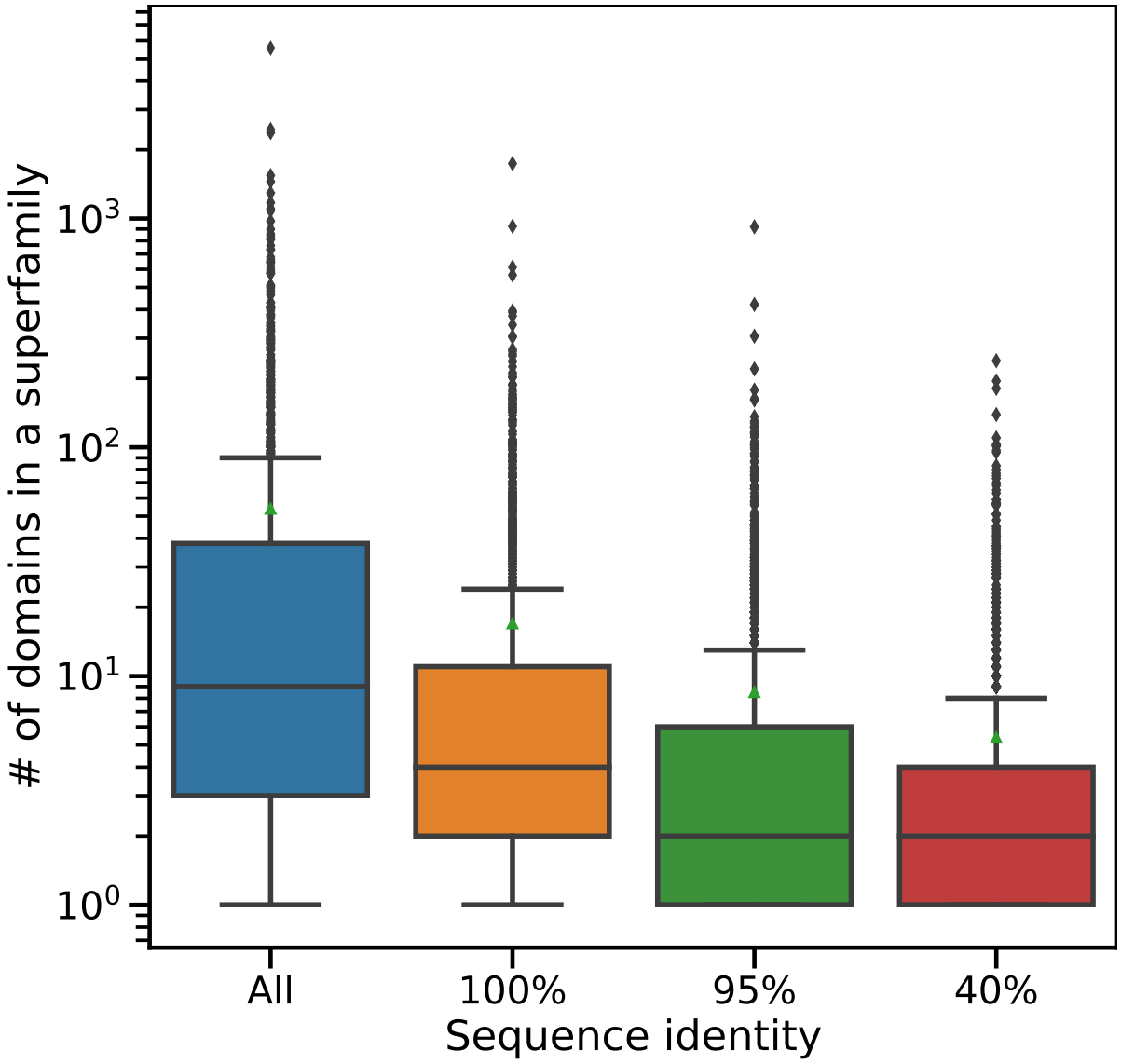


Figure 3.2.: Number of domains per SuperFamily. X axis shows sequence identity for clustering to reduce duplication. Green triangle and line in a box show mean and median, respectively.

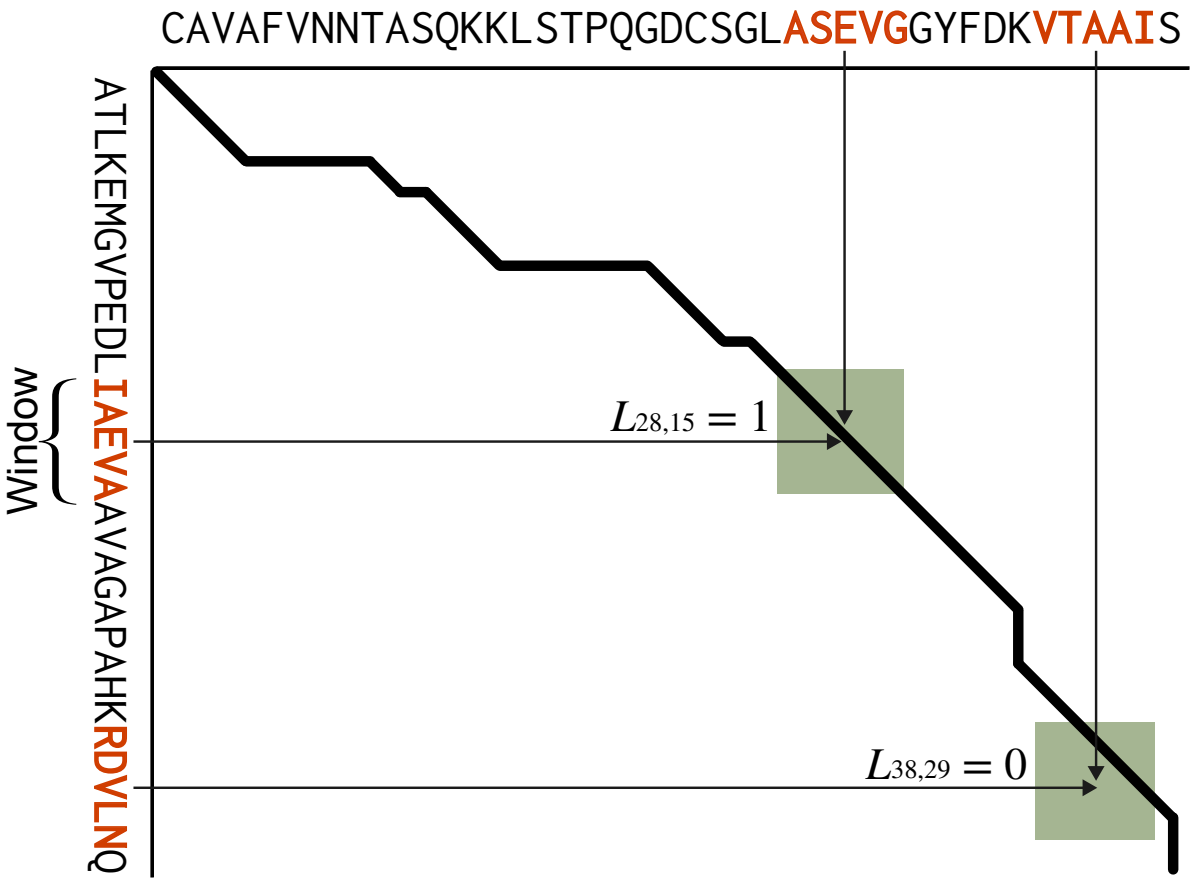


Figure 3.3.: Overview of a feature vector encoding scheme. The X and Y axes show an amino acid sequence. The bold black line shows the structural alignment path between the sequences on the X and Y axes, with the green rectangle indicating the window. The feature vector set is calculated only within this window. The feature vector is the concatenation of the PSSM columns of the window subsequence. If the current column is on the line, the label is 1; otherwise, it is 0.

dealt with the problem as a binary classification problem and used the reliability score of a prediction as a matching score, because structural alignment can only tell us whether a position in a dynamic programming matrix is a match; defining a correct matching score is difficult. Figure 3.3 shows an overview of this design.

Let (Q, T) be the query and target sequences, respectively. Let Q_i be the ith residue of sequence Q and T_i be the ith residue of sequence T. To encode amino acid sequences in a numerical vector, we make PSSMs of the sequences in advance. The column length of a PSSM is 20, which is the number of amino acid types, and the row length is the length of the sequence. Feature vector $\mathbf{V}_{x,y}$ at Q_x and T_y is the concatenation of the query and target residues' feature vectors:

$$\mathbf{V}_{x,y} = (\mathbf{P}_x^{query}, \mathbf{P}_y^{target}). \tag{3.1}$$

$\mathbf{P}$ is the concatenation of PSSM rows around the residue, defined as

$$\mathbf{P}_i = (\mathbf{p}_{i-\frac{w}{2}}, \ldots, \mathbf{p}_i, \ldots, \mathbf{p}_{i+\frac{w}{2}}), \tag{3.2}$$

where w is the window size and $\mathbf{p}_i$ is the ith row of the PSSM. Regarding 'padding' regions defined in $i \leq 0$, $|Q| > i$ and $|T| > i$, we assign $\mathbf{p}_i$ to be $\mathbf{0}$. For example, in the case of $w = 5$, the feature vector dimension is $200 = 20 \times 5 \times 2$.

We can define this feature vector at every residue pair of the query and target sequences. However, we calculate them only within areas where the window moves along with the alignment path because information from residue pairs that are far from the alignment path is not informative.

We assign label $L_{x,y}$ at Q_x and T_y to be 0 or 1:

$$L_{x,y} = \begin{cases} 1, & \text{if } Q_x \text{ matches } T_y \\ 0, & \text{otherwise.} \end{cases} \tag{3.3}$$

The inputs are a pair of query and template PSSMs and a residue position. The outputs are a predicted label and the normalized confidence score ($0 \leq \text{score} \leq 1$).

3.2.3. Alignment calculation

The pairwise sequence alignment of input sequences is calculated using the Smith–Waterman algorithm [51], which requires a substitution score for each residue pair. We predict this score using supervised machine learning and the feature vector defined above. Specifically, we used the k-nearest neighbor (kNN) classification model because it is simple and powerful, especially for large training datasets [77]. kNN calculates the distance between an input feature vector $\mathbf{V}_{x,y}$ and feature vectors in training dataset $\mathbf{V}_{x,y}^{training}$ and checks the labels of the k nearest feature vectors. Generally, the most major label is output as the predicted label from kNN. In this case, 0 or 1 is output because this is a binary classification problem. However, predicting binary labels is too coarse-grained for alignment generation. Thus, the classification confidence score of the kNN algorithm, which is the ratio of a predicted positive label, was used as the substitution score of Q_x and T_y, instead.

3.2.4. Parameter optimization

Our method requires some hyperparameters, which we optimized using the validation dataset. We set the number of nearest neighbors to (10, 100, 1000), the gap open penalty to (-0.0001, -0.001, -0.01, -0.1, -1), and the gap extend penalty to (-0.00001, -0.0001, -0.001, -0.01, -0.1, -1). Using a grid search, we selected 1000 as the number of kNN neighbors. The affine gap penalty optimizations were -0.1 for gap-open and -0.0001 for

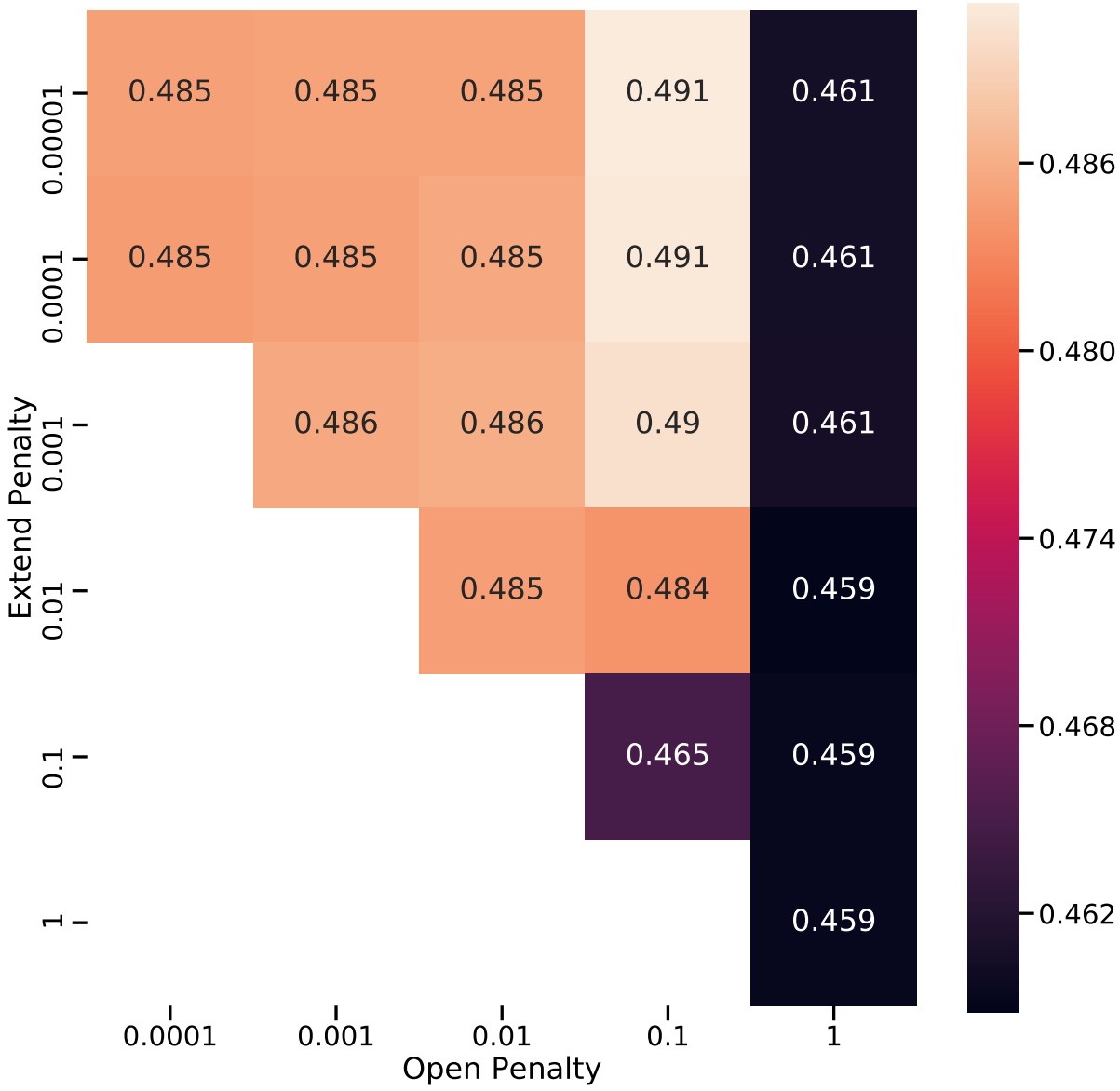


Figure 3.4.: Result of gap penalty grid search. Values in the heatmap show the average TM-score of the validation dataset. Penalties were optimized to -0.1 for gap-open and -0.0001 for gap-extend.

gap-extend (Figure 3.4). These gap penalties were much smaller than the general gap penalties used in other studies because our method's predicted substitution scores were too small for general gap penalties.

3.3. Results

In our method, residue matches at two sequence positions is estimated by kNN. It can be considered as an independent binary classification problem. Thus, we first checked the performance of the label prediction process using the receiver operating curve (ROC) and the area under the ROC (AUC). Figure 3.5 shows the results. The proposed method predicted labels accurately, except for d2axto1, which showed an almost random prediction.

Next, we compared the accuracy of three-dimensional predicted protein models generated from these alignments to evaluate the quality of generated sequence alignments. This step is required because there may not be strong correlation between match prediction and model accuracy and we cannot compare our method with other methods directly. We used MODELLER as a modeling tool [31]. We treated the entire SCOP40 domains, in which sequence similarities in an SF are $<40\,\%$, in the SF where the query is as a structurally similar protein and applied the proposed method to them. Model accuracy can be evaluated by calculating the similarity between an experimentally resolved structure and a predicted structure. For this purpose, we used TM-score [47], which evaluates model accuracy by scoring from 0.0 (least accurate) to 1.0 (most accurate).

We used all domains in the SF of the query as template proteins and generated pairwise alignments. We then compared the accuracy of the proposed method with those of PSI-BLAST, DELTA-BLAST, HHsearch [53], the Smith–Waterman algorithm with a BLOSUM62 substitution matrix, and structural alignment. For PSI-BLAST, which accepts a profile as a query, we made profiles by running three iterative PSI-

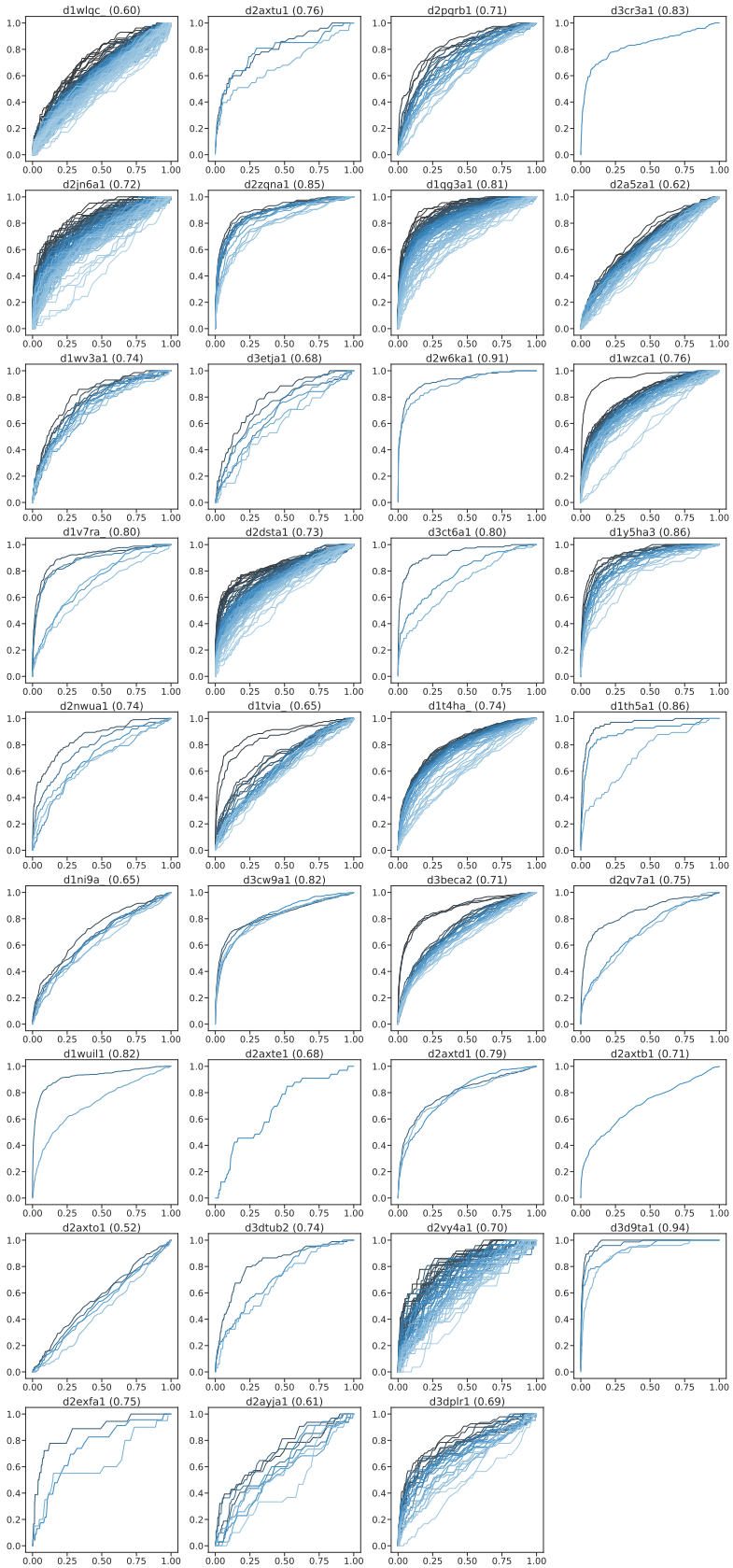


Figure 3.5.: ROC of label prediction. The title is the target name shown in Supplementary Table S1; the average AUC is shown next to the name.

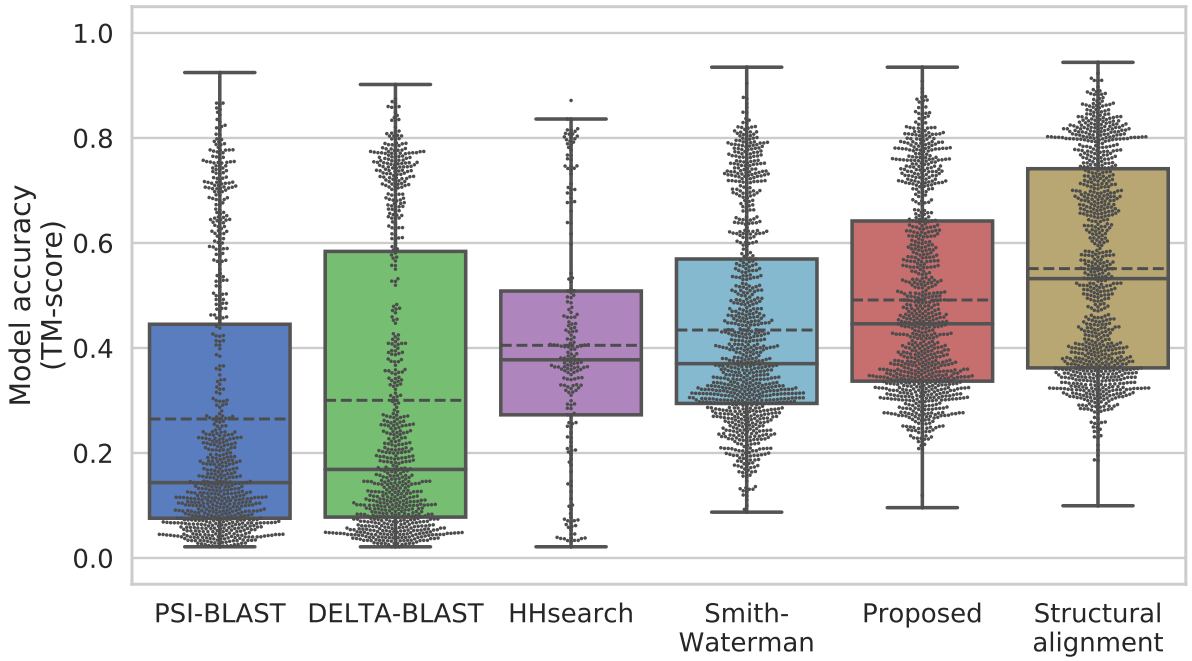


Figure 3.6.: Sequence-independent TM-score of the proposed and competitor methods. The solid line shows the medians and the dashed line shows the means. Dots indicate the data density at each TM-score.

BLAST searches in the UniRef90 database. DELTA-BLAST allows us to use a sequence as a query because it finds profiles from the Conserved Domain Database [78] before searching. For HHsearch, we used Uniclust20 [79] to generate query profiles. We used TM-align [61] for structural alignment.

Figure 3.6 shows the accuracy of protein structure prediction. As expected, structural alignments generated the most accurate models (0.551 on average), although the proposed method achieved results that were nearly as accurate (0.499). The naïve Smith–Waterman algorithm and HHsearch performed the next best; their average scores were 0.432 and 0.472, respectively. From the results of data density, in all methods—including proposed method—these results had two peaks. The top-ranking models of all methods showed similar accuracy, but the worst models' accuracies improved when using the proposed method.

In Figure 3.7, we show as an example one of the generated models and an actual

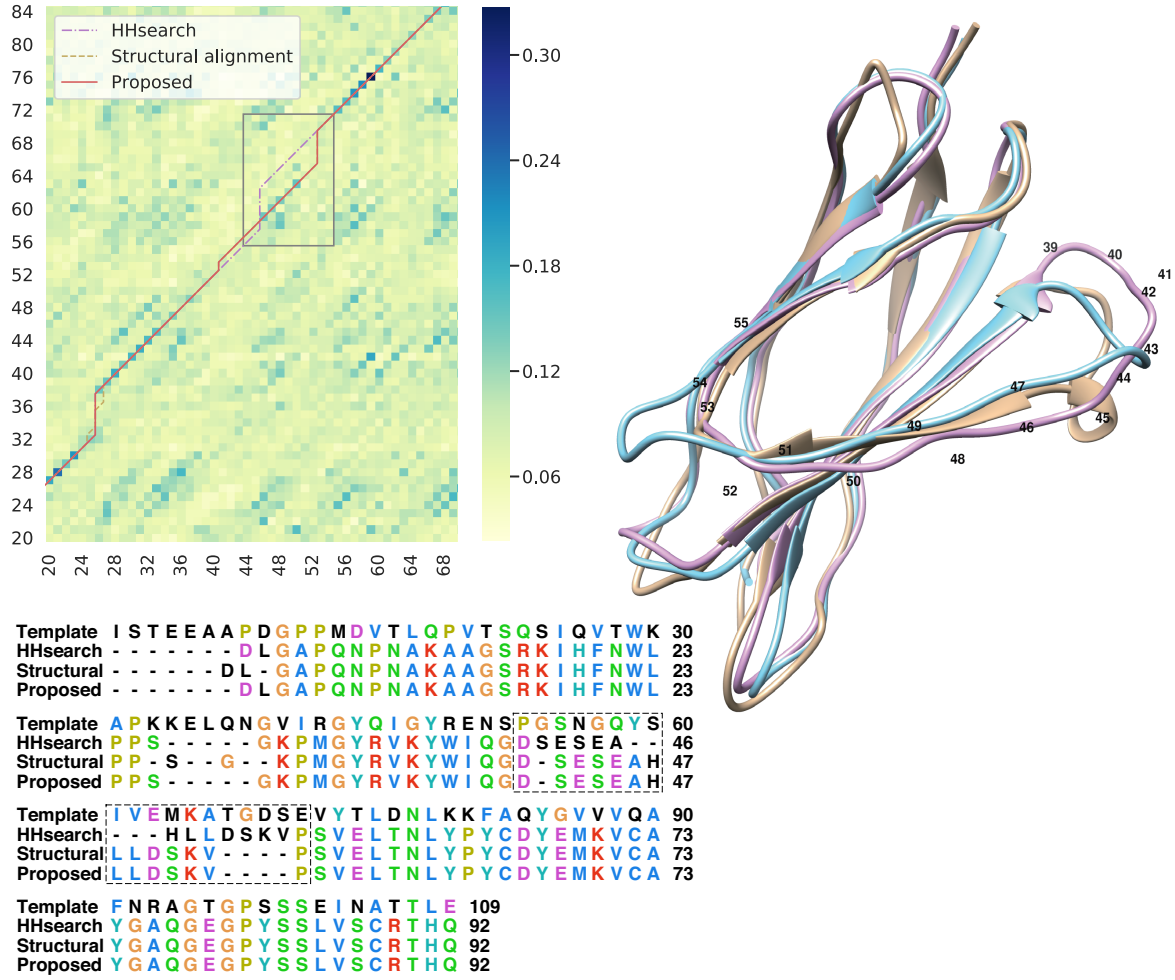


Figure 3.7.: The yellow model in right figure represents the native structure, the red model is generated by the proposed method, and the blue model is from HHsearch. The TM-scores of HHsearch and our method are 0.815 and 0.871, respectively. The left figure is an excerpt of score heatmap and alignment paths. X and Y axes show the query (1QG3A) and template (1VA9A) residue numbers, respectively. HHsearch (dotted dash) generated different alignments between #46 and #55 from the structural alignment (dash), whereas the proposed method (solid) could generate similar alignments to the structural alignment.

alignment, indicating that the proposed method could improve model accuracy. The proposed method succeeded in aligning almost a whole protein and generated very similar alignment results to the structural alignment method. In contrast, HHsearch failed to correctly align a region around the 4th beta strand (residue numbers 40–55) and caused structural differences from the native structure in the loop regions on both sides of the sheet. Figure 3.7 shows how the proposed method correctly aligned the region that HHsearch failed to align. The kNN predicted match score between a query residue with residue number 49 and a template residue with residue number 63 is much higher than the scores around the position. Thus, the proposed method generated an alignment passing through the position.

3.4. Discussion

3.4.1. Application for homology detection

Our method can be used for homology detection by sorting the alignment scores it includes in its result. We investigated the method's homology detection and the top model accuracies of a search result ranking. The proposed method's homology detection performance was compared with those of PSI/DELTA-BLAST and HHsearch, as shown in Table 3.1. To ensure a reasonable computation time, the training dataset was reduced to 1/100 instead of 1/10. ROC_n considered results only up to the nth false positive and AUC_n was regularized by the number of false positives and cutoff n. In this evaluation, we defined true positives as those having the same detected SF as the query and false positives as those having different SFs. Compared with PSI/DELTA-BLAST and HHsearch, the detection sensitivity of the proposed method was lower. The highest average AUC_{50} of HHsearch was 0.706. By contrast, the proposed method had the lowest score, 0.205. We think this is because the proposed method shows many false positive results.

Table 3.1.: Average AUC_{50} and model accuracy (TM-score) of the proposed and competitor methods.

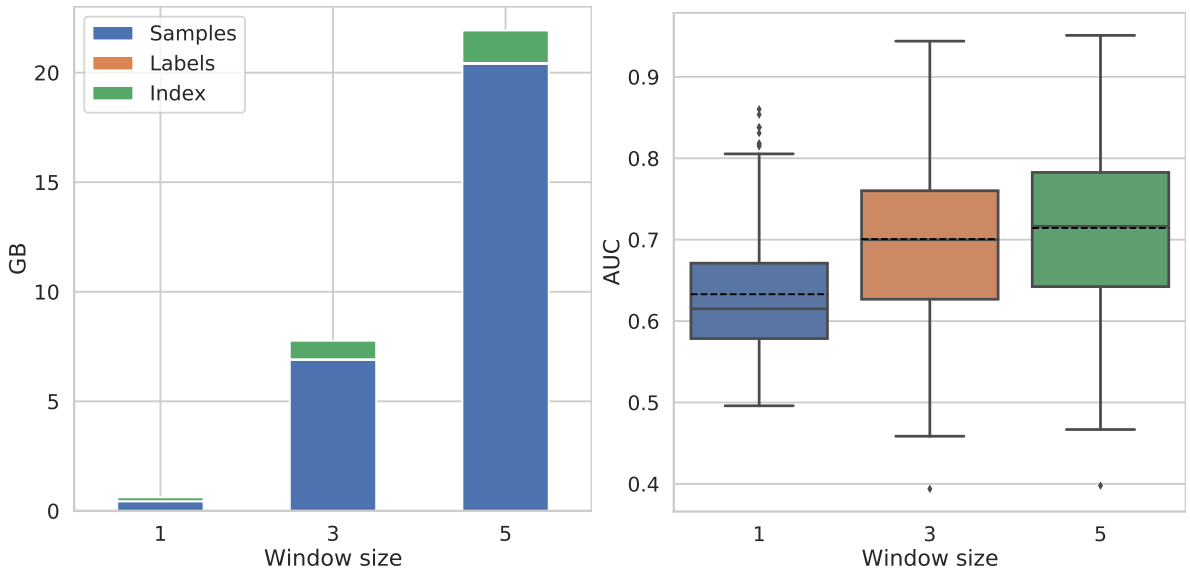
	PSI-BLAST	DELTA-BLAST	HHsearch	Proposed
AUC_{50}	0.323	0.340	0.706	0.205
TM-score	0.278	0.324	0.205	0.298

Using the search results, we applied homology modeling to the top 10 search result and made 3D models; the models' accuracy is mentioned in the second row of table 3.1. The proposed method achieved the second-highest average TM-score, 0.298. From these results, it is difficult to use our method for homology search. Therefore, we consider that our proposed method is currently useful for the alignment generation phase of homology modeling, after template detection.

3.4.2. Optimization of window size and the influence of training data reduction

We tested our method using (1, 3, 5) as window size candidates and compared the label prediction accuracy using AUC. The results of window sizes (1, 3, 5) were 0.640, 0.689 and 0.701, respectively. We also tested data reduction ratios of (0.001, 0.01, 0.1) and compared the label prediction accuracy using AUC. The results for ratios of (0.001, 0.01, 0.1) were 0.635, 0.676 and 0.701, respectively. Although increasing the window size and reduction ratio may increase accuracy, we could not evaluate them because the size of the required training dataset would be bigger than our computing resources can manage.

Figure 3.8a shows the relationship between window size and memory usage. The main reason of the increased memory usage was the increase of negative samples as the window size became larger. Figure 3.8b shows the AUC of MATCH/UNMATCH label prediction as a function of window size. The label prediction accuracy became better



(a) (b)

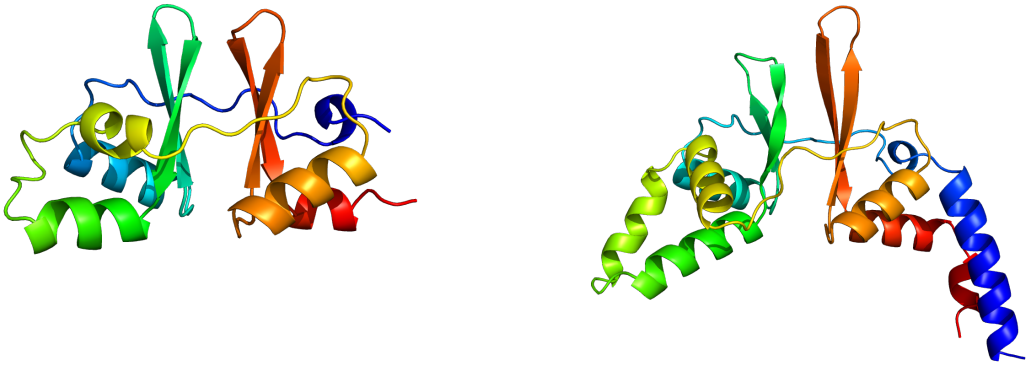
Figure 3.8.: (3.8a) The relationship of window size and memory usage. Samples, labels and index mean training data used in kNN, labels for training and index for fast sample data access, respectively. (3.8b) the AUC of MATCH/UNMATCH label prediction as a function of window size.

by increasing the window size and the best accuracy was obtained with window size equal to 5. However, the improvement of the accuracy from window size equal 3 to 5 was relatively small. Thus, larger window size may improve the prediction accuracy but the improvement would be small. From the analysis by Figure 3.8a, larger window size will require more computational memory. Therefore, we decided to use the window size equal to 5 in proposed method.

3.4.3. Analysis of the proposed machine learning model and feature vectors

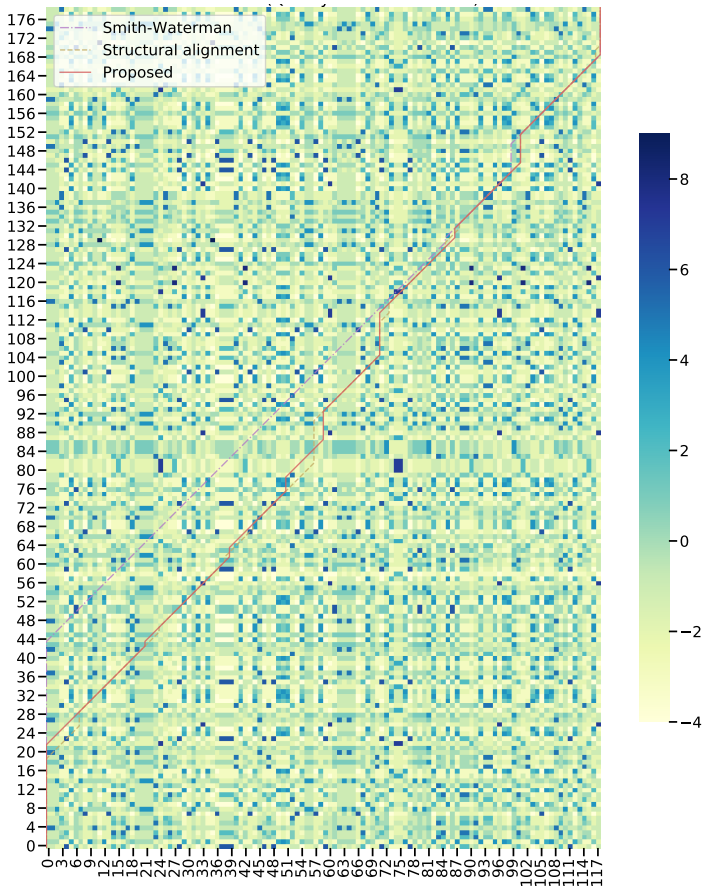
We have shown the better prediction accuracy of proposed method compared with previous methods. However, the reason why our approach significantly improved the accuracy of sequence alignments for homology modeling. Thus, we analyzed which aspect contributed the improvement in the proposed method.

In this analysis, we used SCOP domain `d1y5ha3` (figure 3.9a) as the target protein,



(a) Target protein (`d1y5ha3`)

(b) Template protein (`d2ooxe1`)



(c)

Figure 3.9.: (3.9a, 3.9b): Structure of query and template protein as an example. (3.9c): Heatmap of the BLOSUM62 substitution matrix scores. X and Y axis show positions of query sequence and template sequence, respectively.

and `d2ooxe1` (figure 3.9b) as the template protein. Figure 3.9c shows a heat map of BLOSUM62 substitution scores used in dynamic programming process. Those scores were used in Smith-Waterman sequence alignment. Although the two proteins are classified into the same SCOP SuperFamily, the sequence similarity was low. There were no regions where high score positions were aligned diagonally, which will be aligned without gaps. As the result, the sequence alignment by BLOSUM62 produced a different alignment from structural alignment. Therefore, the homology modeling accuracy by using this alignment was low; TM-scores of models from the Smith-Waterman alignment and the proposed methods were 0.623 and 0.751, respectively.

Next, we focused on the proposed method's feature vector and analyzed how the feature vector design worked. The proposed method has two features: (1) to look at the peripheral information, and (2) supervised machine learning. Compared with the proposed method using kNN, we compared it with the PSSM distance used as a score; let the PSSM distance here be the L^2-norm between the two vectors (query sequence and template sequence) at each position (column) of the PSSM.

First, we considered the significance of incorporating peripheral information. Figures 3.10a, 3.10b and 3.10c show heatmaps of the PSSM distance at each sequence position without considering the peripheral information, the PSSM distance of the concatenated vectors of the feature vectors of the peripheral residues (window size is 5), and the score of the proposed method, respectively. The min-max normalization was used to create the heatmaps of the PSSM distance. The proposed method incorporates the single residue pairs and the peripheral information by concatenating the feature vectors of surrounding amino acids. When peripheral information was not considered, the score heat map was similar to the score heat map of BLOSUM62 (figure 3.9c and 3.10a). Although there were amino acid pairs similar at each position, there were not many places where they were aligned, i.e., high scores are aligned diagonally. On the other hand, when peripheral

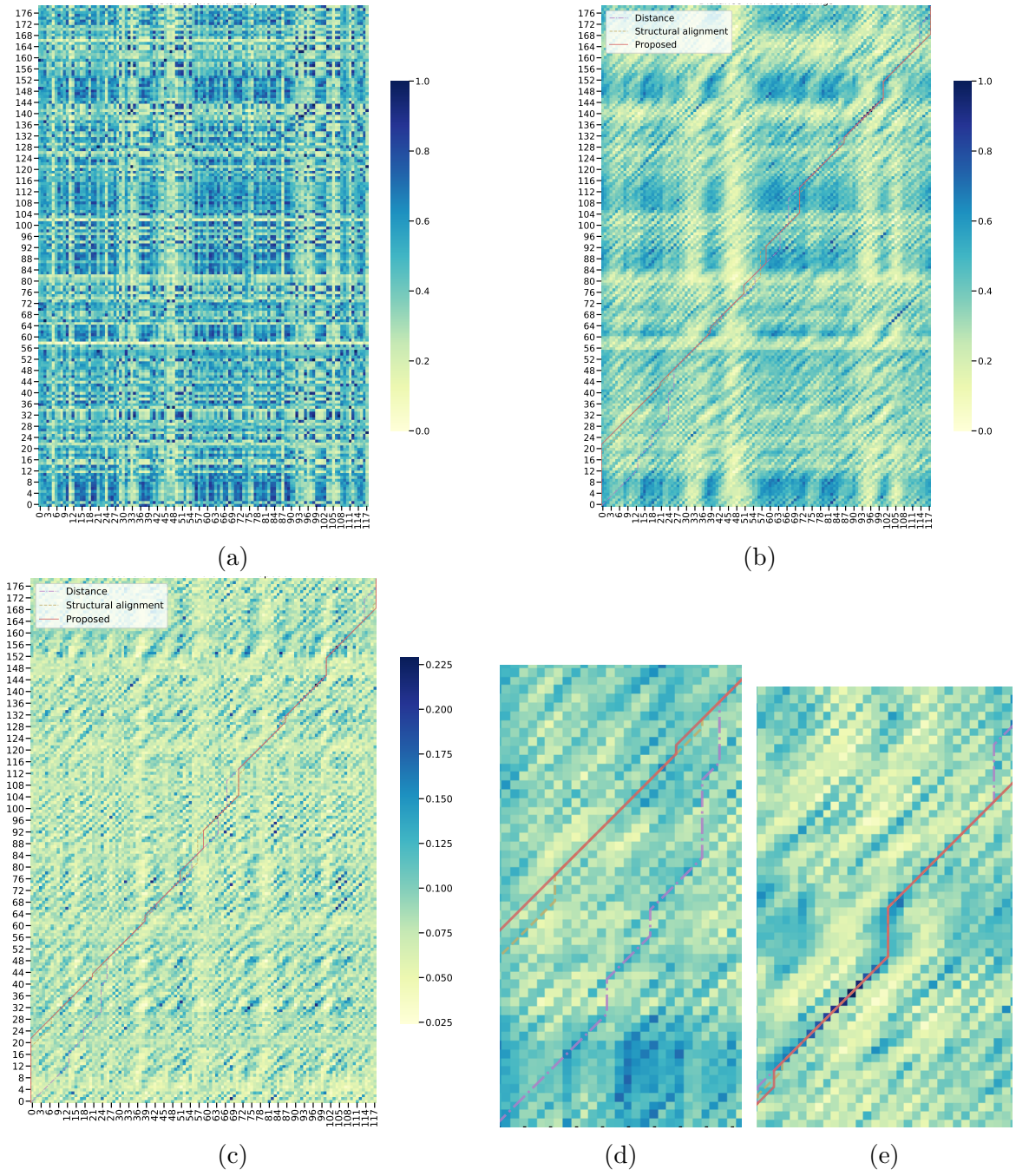


Figure 3.10.: Heatmaps of various scoring. (3.10a): Score map of PSSM distance. (3.10b): Score map of PSSM distance with peripheral information. (3.10c): Score map of the proposed scoring. (3.10d, 3.10e): Excerpts from score map of PSSM distance with peripheral information. Query ranges are #1–#30 for 3.10d and #87–#117 for 3.10e.

information was taken into account, areas with near-neighborhood high scores emerged, and regions with a high degree of profile similarity appeared (figure 3.10b). We could observe the same aspect of the proposed method (figure 3.10c). This means that the high-scoring areas become linear due to the peripheral information, and the aligned area covered by the alignment is likely to increase. This is the effect of the incorporation of peripheral information.

In figures 3.10d and 3.10e, we looked at the differences between the alignment based on the PSSM distance score with peripheral information and the alignment based on the proposed method (the gap penalty was the same as the proposed method). Around the C-terminus (figure 3.10e), the alignment based on the PSSM distance score and the structural alignment was generally consistent. However, on the N-terminus side (figure 3.10d), the relationship between the structural alignment and the PSSM distance was weak, and the alignment was significantly out of the structural alignment.

Next, we examined the effect of incorporating supervised machine learning. In figure 3.11c, the heatmap shows the difference between the min-max normalized proposed method score and the min-max normalized PSSM distance with peripheral information, respectively. A positive difference position in figure 3.11c means that the PSSM distance is far, but the proposed method's score is high. From figure 3.11d, we could see the improvement of the proposed method's alignment in the excerpted area where the PSSM distance was far, but the score of the proposed method was high. We think the alignment error caused by PSSM distance or sequence similarity was corrected in this area. From figure 3.11e, the areas with close PSSM distance and high scores from the proposed method were generally consistent; we believe that the existing methods can be used for this part of the alignment.

Figure 3.12 shows the frequency profiles of the amino acids at each position. Overall, based on amino acid diversity and frequency of occurrence at each position, we believe

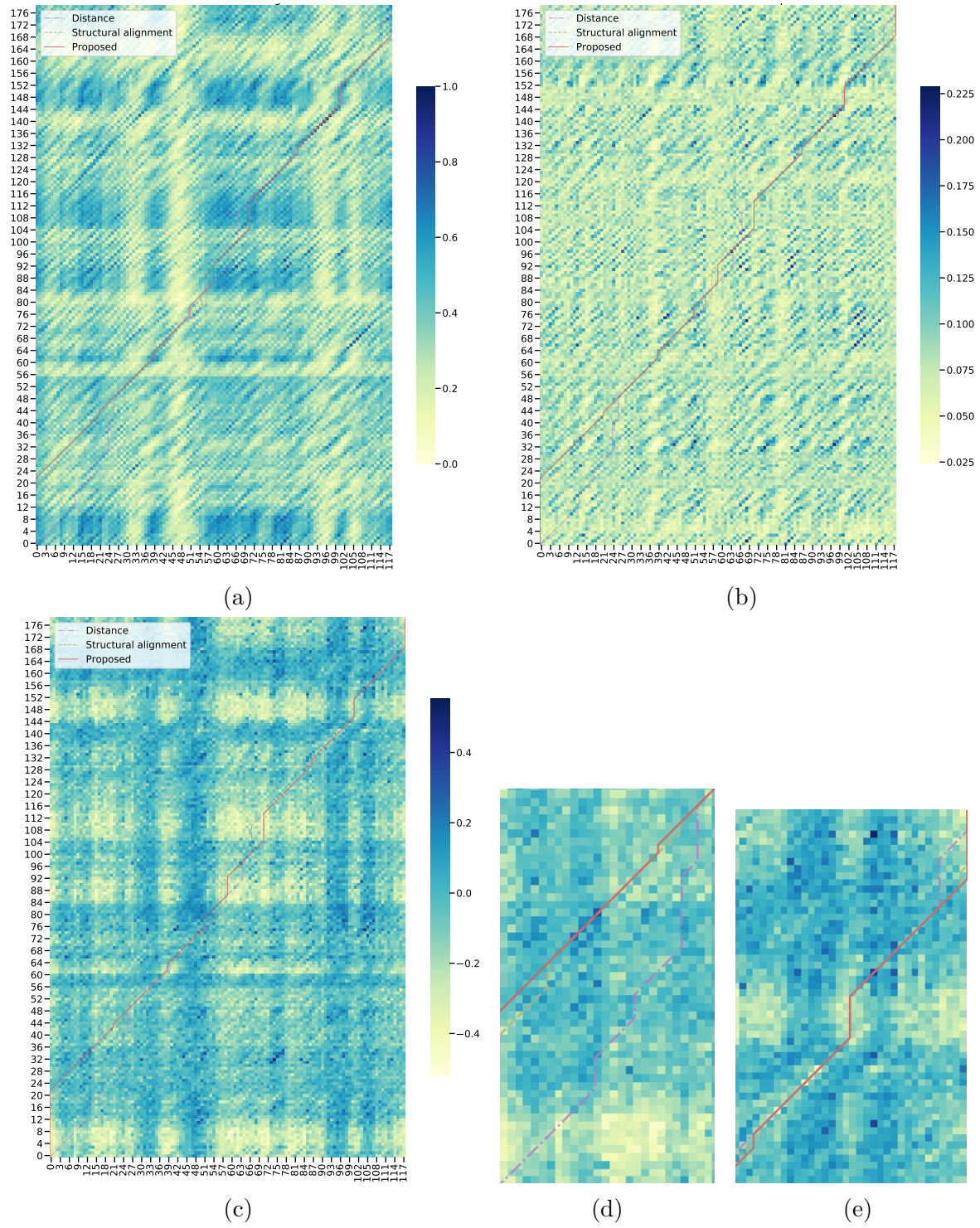
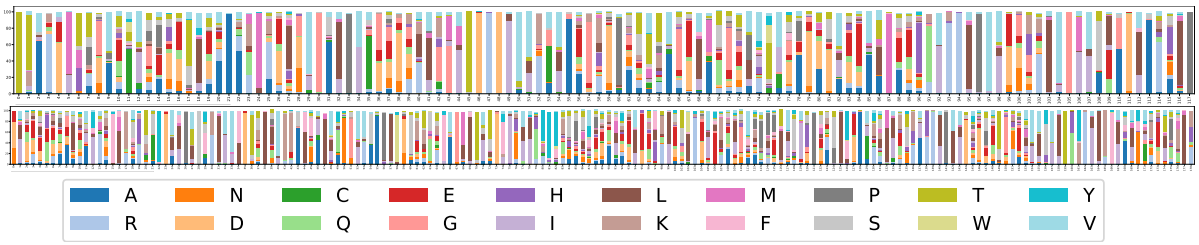
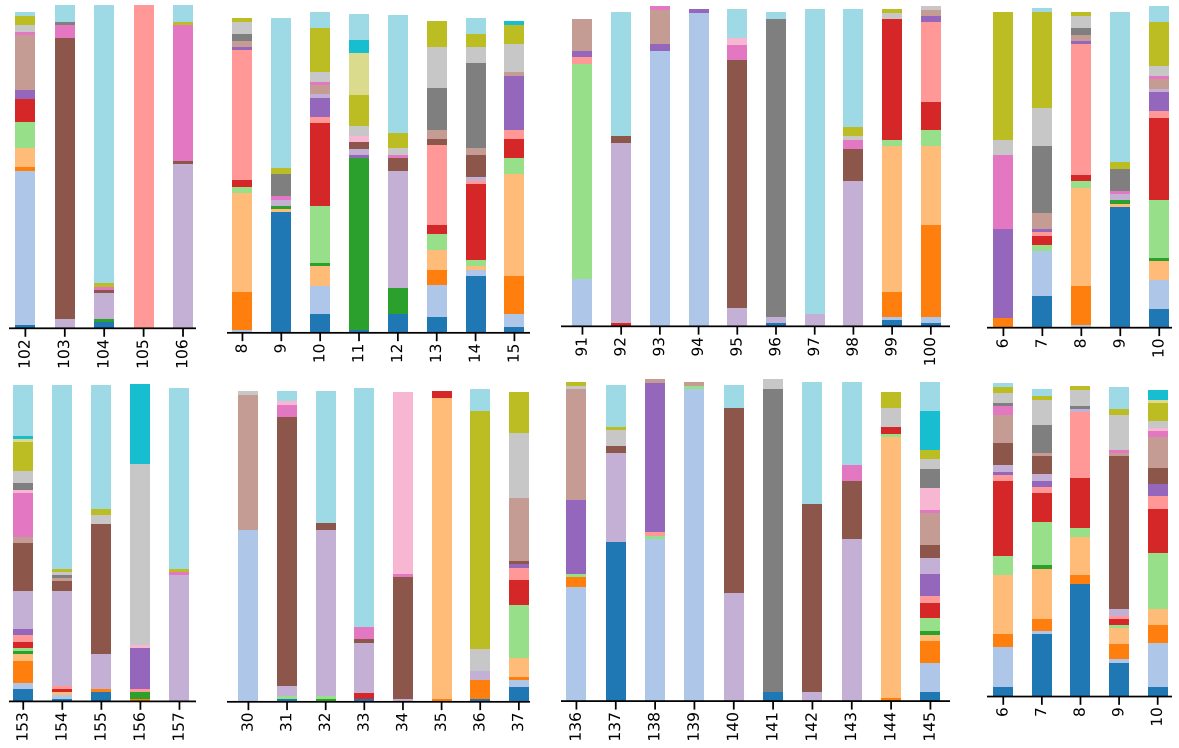


Figure 3.11.: Effect of supervised learning Heatmaps of various scoring. (3.11a): Score map of PSSM distance with peripheral information. (3.11b): Score map of the proposed scoring. (3.11c): the difference between the proposed method score and the PSSM distance. (3.11d, 3.11e): Excerpts from score map of difference between the proposed method score and the PSSM distance. Query ranges are #1–#30 for 3.11d and #87–#117 for 3.11e.



(a) Overview of amino acid frequency profile



(b) (c) (d) (e)

Figure 3.12.: Amino acid frequency profiles of query and template proteins. The upper is the amino acid frequency profile of the query sequence, and the bottom is the profile of template sequence. X and Y axis show sequence position and amino acid frequency % at each position (3.12a) Overview of amino acid frequency profiles. Amino acids are colored by bottom image's rules. (3.12b–3.12e) excerpts that have characteristics of learning. Characteristics of each example are discussed in the main text.

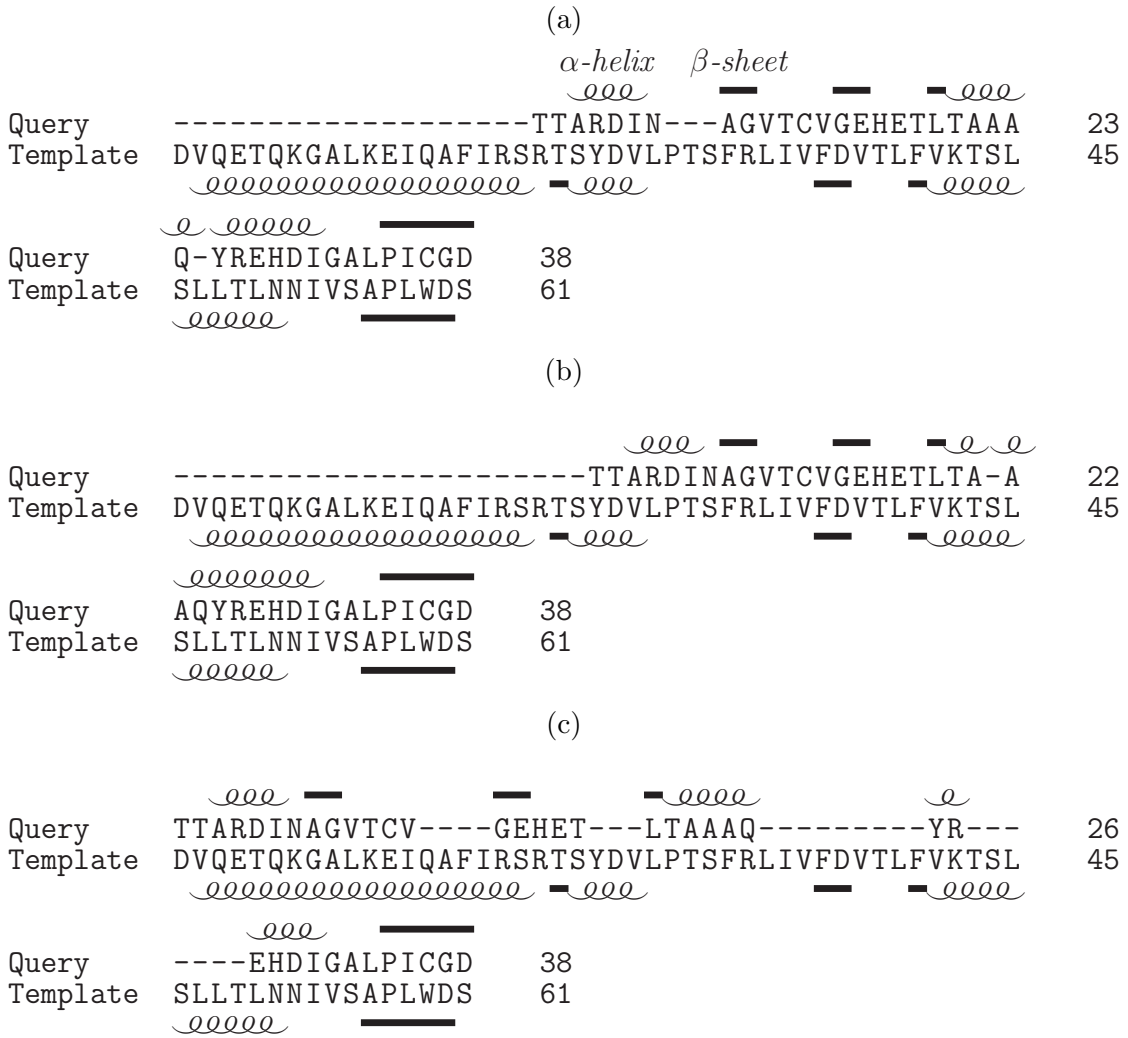


Figure 3.13.: Difference of superposition of secondary structure. Secondary structure was assigned by DSSP [80], [81]. (3.13a) Structural alignment (3.13b) Alignment by the proposed method (3.13c) Alignment by the PSSM distance with peripheral information

that the sequence alignment pattern at each position is learned and output as a score.

Figures 3.12b–3.12e are excerpts that show characteristics of our machine learning model. There were cases where the PSSM distance was close, and the proposed method's score was also high; figure 3.12d is an example. These regions were aligned between the query and template sequence, which was a reasonable result. The existing methods can be used for this part of the alignment.

Figure 3.12b is a case with a high score of the proposed method, although it is far away by PSSM distance with peripheral information. These regions were aligned between the query and template sequence without gaps. The amino acid pair of #104 in query sequence and #155 in the template sequence recorded a high score of the proposed method, but the PSSM distance was far. The normalized PSSM distance score was 0.533, and the normalized proposed score was 0.941 at the position. In this case, the PSSM distance between two hydrophobic amino acids (Valine and Leucine) was far, but the proposed method's score was high. Also, position (#105, #156) showed two hydrophilic amino acids (Asparagine and Serine) substitution, and (#103 and #154) showed two hydrophobic amino acids (Leucine and Valine). From the observation, our machine learning model learns the physicochemical amino acid similarity from training data. The diversity of the amino acids' frequency profile is low in query and template sequences, and the PSSM distance seems to be far.

On the other hand, there were some cases where PSSM distance was close, but the proposed scoring showed a low score. An example in figure 3.12e shows the case. For the position (#8, #8), the normalized PSSM distance score was 0.663, but the normalized proposed score was 0.205. These regions were not aligned between query and template sequence by our method, but alignment based on PSSM distance with peripheral information aligned these regions. The physicochemical properties of G (glycine) and D (aspartic acid), which occurred more often on the frequency profile of the query

sequence, are different. Also, fewer dominant amino acid species on the frequency profile and more amino acid species occur. These states may be learned as factors that make it difficult to give high scores.

Figure 3.12c shows the case of the region of random amino acid occurrence and biased amino acid occurrence; diversity of amino acids are different. These regions were also aligned between query and template sequence without gaps. Here, too, the PSSM distance between the two profiles was far, but the score was high as our algorithm. Significantly, the amino acid pairs of (query, template) sequences in (#10, #32) and (#13, #35) showed the phenomenon. At the position (#10, #32), the normalized PSSM distance score was 0.317, and the normalized proposed score was 0.566. At the position (#13, #35), they were 0.415 and 0.737, respectively. We have to rely on training data in these cases because it is difficult to estimate the regions' structural relationship only by the PSSMs at the positions. If one or both of the frequency profiles are not random-like but are dominated by individual amino acids, the proposed method seems to pick up the information well and give high scores.

The distribution and amino acid occurrence profiles of the scores were qualitatively analyzed. The proposed supervised machine learning model is thought to learn the sequence alignment pattern for each position from each position's amino acid diversity and frequency and output it as a score. Since the analysis was only a single example and only qualitative, quantitative analysis using large data sets remains a challenge for future work.

3.5. Conclusion

In this chapter, we proposed a new sequence alignment generation method that uses machine learning to accurately predict protein structures. Instead of a fixed substitution matrix, the proposed method predicts substitution scores at each residue pair. To apply

machine learning, we developed a method that converts pairwise alignments to numerical vectors of latent space, which enables us to employ a supervised machine learning algorithm for sequence prediction. The predicted scores are directly used to generate alignments, which are in turn used as input for homology modeling. We evaluated the model accuracy of our alignment generation method and found that it outperformed the state-of-the-art methods. We also investigated our method's ability to detect remote homologies; using AUC_{50} for comparison, our method did not perform better than other methods. However, we found that the proposed method generated relatively accurate 3D models compared with other methods.

Currently, our method requires a long execution time because of the kNN algorithm and dataset size. These factors caused us to reduce the amount of training data used because the model's execution time depends on the number of target proteins as well as protein size. It would be a natural extension of this work to employ faster kNN algorithms, including approximate schemes, because our method does not require precise solutions. The proposed feature vector design can be treated as two-dimensional; in the future, we will also consider the use of higher-performance models such as convolutional neural networks.

Chapter 4.

Sequence alignment generation for protein remote homologs

4.1. Introduction

In this chapter, we propose a new sequence alignment generation method for remote homologs detected by an intermediate sequence search (ISS) for use in homology modeling. Template structures are the structures of homologous proteins (homologs), and are often found by a homology search of protein structure databases, such as the PDB. If currently the available template are a good template structure and generates an accurate sequence alignment, homology modeling is the most practical structure prediction method. However, homology modeling requires homologous proteins with known structures to be used as templates. If the protein structure database does not have a homolog entry that closely resembles a query protein, classic sequence homology search algorithms, such as BLAST [34], fail to find a template. Thus, to detect remote (i.e., distantly related) homologs, more sensitive search methods are required. Sequence profile based on multiple sequence alignments, such as DELTA-BLAST [36], can detect remote homologs. In addition, HMM comparison methods, such as HHpred [40], have performed exceptionally well in structure prediction benchmarks [43], [44]. However, even when using the above mentioned sensitive homology search methods, the detection of remote homologs can fail due to insufficient search sensitivity.

To overcome this problem, the ISS method has been proposed to provide more dis-

tantly remote homology detection. The basic idea of ISS is the following: two sequences of remote homologous proteins, which do not have enough sequence identity or a close relationship evolutionally, can be related via another sequence whose characteristics and features are intermediate between the two remotely homologous proteins. If the match score between both the first and third sequences and the second and the third sequences is high, it can be concluded that the first and second sequences are related, even though their sequence similarity is low. In the ISS method, after searching for homologs of the query protein in the database, the results are used as new queries to detect more distantly related homologs by re-running the homology search. By identifying a connection via these intermediate sequences, the ISS method can detect relationships between the original query protein and remote homologs. The idea of the intermediate sequence search itself is not novel [63]. Decades ago, Entrez [64] provided intermediate sequence information. However, the naïve ISS procedure often provides many false positives [65] and requires significant computing resources to evaluate many homology searches. Recently, to overcome the computational demand and occurrence of false positives, approaches that utilize network or graph theory were proposed [66], [67]. In addition, machine learning-based intermediate sequence search methods have demonstrated good results [68], [69].

ISS is a useful technique for improving homology search sensitivity, and several studies have used this method for protein function prediction [63], [65], [82]. However, to our knowledge, there have been no examples of its use in protein structure prediction. The ISS method can detect remote homologs, but it does not generate any sequence alignments between the query and target proteins. As mentioned, homology modeling requires a template as well as a sequence alignment between the query and template proteins. Thus, to apply homology modeling to the result of homology detection via ISS, we have to generate a sequence alignment in a separate step. The simplest approach

to generate a sequence alignment is though the use of an algorithm, such as Smith-Waterman local alignment [51]. However, it is difficult to generate accurate sequence alignments between remote homologs, and an inaccurate sequence alignment often leads to low quality of predicted models in homology modeling. In essence, alignment quality is crucial to homology modeling. Thus, in order to apply ISS to homology modeling, we need a method to generate accurate sequence alignments specifically designed for ISS results.

Sequence alignment generation of remote homologs is a difficult task due to low sequence identity between the query and target sequences. On the other hand, for the case of ISS results, we can use additional intermediate sequence information that bridges the two sequences requiring alignment. Thus, we hypothesize that the intermediate sequences would help to generate more accurate sequence alignments. To evaluate the quality of the generated sequence alignments, we performed homology modeling based on the sequence alignments and measured their structure prediction accuracy. As a result, the proposed method showed better accuracy in comparison to a baseline method. Our method is expected to be valuable for distant homologs, and we also evaluate our method for these distant and difficult pairwise targets.

4.2. Materials and methods

For the homologs detection, we use the intermediate sequence search method. Currently, many ISS methods have been proposed to address early problems with the method [63], [65]–[69], [82]. However, the flow of the searches is basically the same. In this study, we implemented a basic ISS method. The ISS searches for homologs of the query protein in a protein structure or sequence database, and continuously uses the results as new queries to detect more distantly related homologs by re-running the homology search. Figure 4.1 shows the overview of the ISS method. Since our aim is homology modeling,

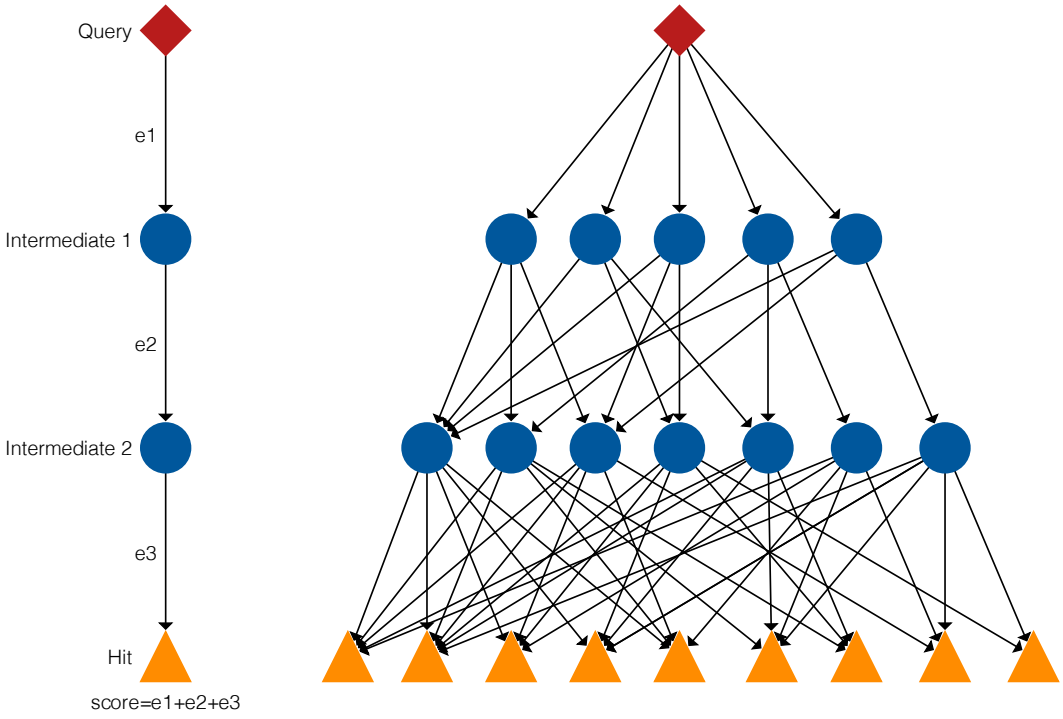


Figure 4.1.: Method overview for the intermediate sequence search. Red diamond, blue circles, and yellow triangles show the query, intermediate and hit sequences, respectively. First, we search homologs of a query sequence in the intermediate database and the resulting hits represent the first intermediates. Second, these hits are used as queries, such that the next series of find hits represent the second intermediates. Finally, these second intermediates are used as queries again, and the proposed method searches the final database. Hit score is calculated by the sum of each hit score.

the last search had to be done using a structural databases. Any arbitrary databases can be used for the intermediate databases, as they need not be structural databases. To achieve high sensitivity, the intermediate databases should be large; for example, NCBI nr [83] (over 180,000,000) or UniProt [76] (over 100,00,000 after 90 % sequence identity clustering). The number of intermediate searches is one of the parameters in this type of study, and, depending on the level of more sensitivity needed, this can be increased as more depth is required. Also, any search method can be applied for the ISS method. It is recommended that the tools used are fast as execution time often becomes long for numerous intermediate sequences, however, their sensitivity can be low because intermediate sequences should assure sensitivity.

4.2.1. Proposed alignment generation method

Although the detection performance of the ISS is high, this method often provides many false positives [65]. To overcome this problem, we implemented two improvements. First, our method uses a sub-region of the detected sequence as intermediate results to be used as the subsequent query, instead of using the whole sequence of the detected homolog. Many sequences in protein databases consist of multiple domains within one sequence; thus false positives are obtained because the domains, which are not related to the query sequence, are used as subsequent queries in intermediate searches. These domains are inappropriate for remote homology detection and cause many false positives during the ISS [63]. By narrowing the search region to a detected homology region, it is expected that the number of false positives will be reduced. Second, the proposed method assigns rankings to the final results set by the sum of similarities between intermediate sequences. The similarities are calculated during the ISS; for example, if DELTA-BLAST [36] or PSI-BLAST [35] is used, the similarity score will be the E-value or bit-score. Our method sums the similarity scores on the path from query to the final hits, and sorts them to

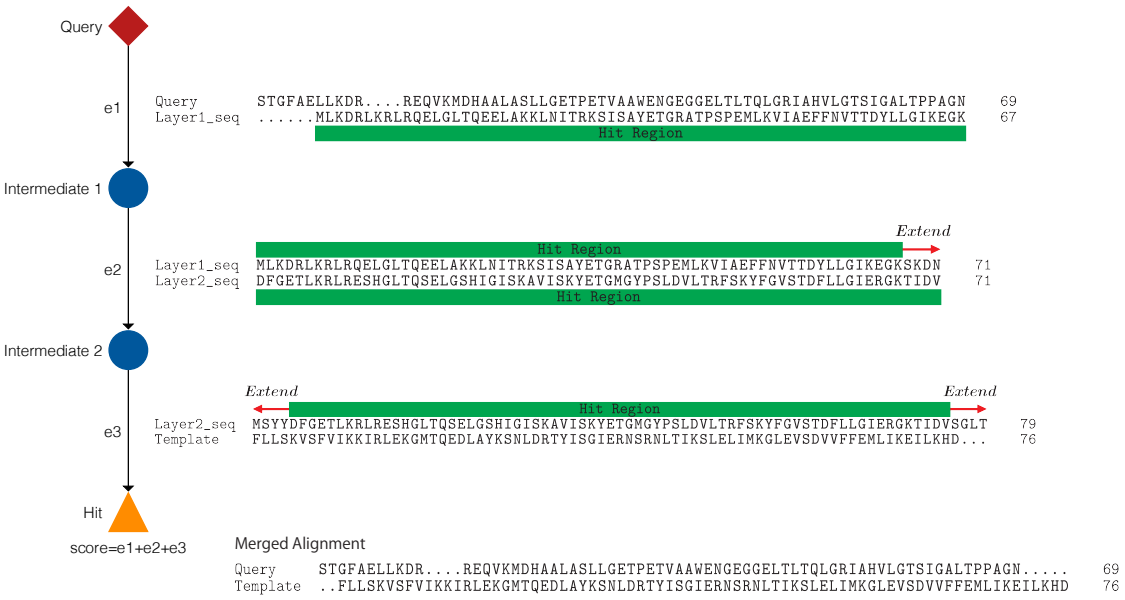


Figure 4.2.: Alignment generation with intermediate sequences. In each intermediate layer, a pairwise alignment is generated. The proposed method merges the pairwise alignments between intermediate sequences. Each sub-pairwise alignment is preserved, which means that the positions of residues in a pairwise alignment are preserved.

generate the final search results. If multiple paths exist between the query and the final hit, the path with the best one is selected. In this time, we used the Evalue as similarity score and selected a path of the smallest score as the best.

However, it is difficult to generate alignments between remote homologs because the sequence identity between them is often low. Even if some local alignments can be generated, the length of the alignment region is often too short for accurate homology modeling, such that prediction of the protein structure will not be accurate.

To overcome the problem, we use intermediate sequences detected by the ISS. This is reasonable because, in the ISS phase, the proposed method only uses aligned sequence regions, while other domains in a sequence are not used for the search. Figure 4.2 shows an overview of our alignment generation method. To extend the aligned region,

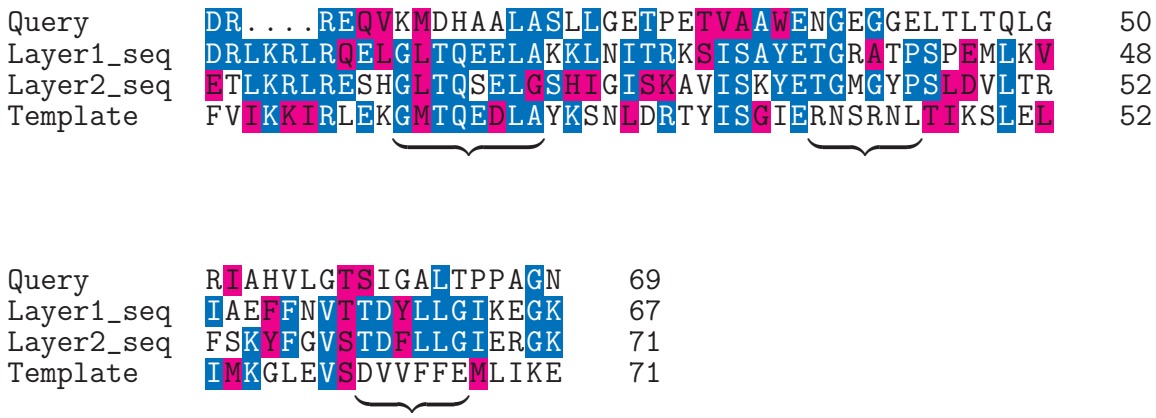


Figure 4.3.: Alignment generation of low a sequence identity region. Sequence similarity of the query and template sequences in the bracketed regions are low. However, the intermediate sub-sequences are similar, and the regions of low sequence similarity are aligned as a result. Blue and red shards identify identical and similar (positive BLOSUM62 score) residues, respectively.

hit regions are extended in intermediate layers as far as possible such that they do not include other domains too much. The length of the extension is one of the hyperparameters. The extended sub-sequence is used as a query in each intermediate search. In each intermediate layer, pairwise alignments are generated using the Smith-Waterman algorithm.

At the final phase of alignment generation, the proposed method merges pairwise alignments between intermediate sequences. During the merging procedure, no dynamic-programming-based multiple sequence alignment method is used. Each sub-pairwise alignment is preserved, which means that the positions of residues in a pairwise alignment are preserved. A pairwise alignment between the query protein and one of the final hits is then split out from the merged alignment.

Figure 4.3 shows an example of how intermediate sequences work. Usually, when the pairwise sequence identity of the query and template sequences is low, these regions are not aligned by naïve substitution matrix-based methods. However, if intermediate sub-

sequences exist, which are similar to the query or template sequence, similar sequence regions are aligned via these intermediate sequences. These intermediate-proxied alignment regions exist in the merged alignment, and they allow for the extension of aligned regions using remote homologous information.

This alignment method generally produces reasonable alignments, but sometimes generate obviously incorrect alignments because of a large shift of the aligned regions in intermediate search results and so on. Fortunately, we can detect most of these cases by checking the length of the aligned region. Therefore, in addition to merging multiple pairwise alignments, as described above, we also apply the Smith-Waterman algorithm to generate a pairwise alignment of the query and template. Then, the proposed method selects one that has a longer aligned region.

4.2.2. Materials

In this paper, we used the following datasets for evaluation: UniRef [76] was used as an intermediate sequence database, and Structural Classification of Proteins (SCOP) [71], [72] was used as the final database, which is the same standard database for evaluation as one used in the Chapter 3. The SCOP database classifies proteins by class, folds, superfamily (SF), family, and domain, based on manually curated function/structure classifications. Because the two databases contain redundant sequences, we used UniRef_{50}, which was a reduced by clustering sequences of 50% sequence identity; and SCOP_{95} was used as the final database, reduced by 95% sequence identity clustering. DELTA-BLAST was used for the intermediate search tool. For merging intermediate alignments, we used MAFFT's [84] alignment merge function. For evaluation, we selected 100 sequences for test data from SCOP_{40}, which is a SCOP database reduced by 40% sequence identity. Each of the 100 sequences were randomly selected, one from each of the top 100 superfamilies that were sorted according to the size of the superfamily. The selected 100 test

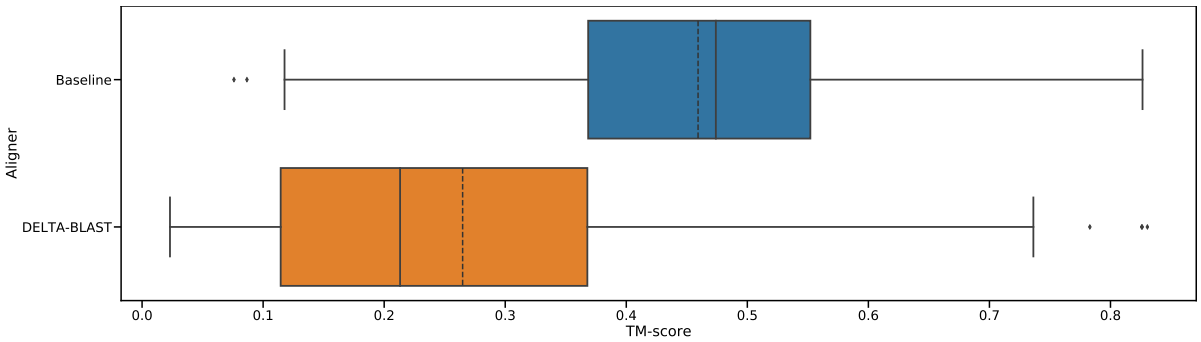


Figure 4.4.: TM-score distribution of hard targets, meaning hits that are not detected by DELTA-BLAST without the use of intermediates. The solid line represents the median, and the dashed line represents the mean. In these results, all templates are from the same superfamily as the query.

domains and their information are listed in appendix C.

4.2.3. Evaluation

To evaluate alignment quality for homology modeling, we generated structural models by homology modeling from alignments obtained using the proposed method. For homology modeling, we use the program MODELLER [31]. The TM-score [61] between native protein structure and the predicted one is used as a measure of structure prediction accuracy. The TM-score indicates global structure similarity by a regularized $(0, 1)$ value, and a TM-score $= 1$ means the predicted model corresponds to the native structure. We compared the TM-scores of predicted models obtained using the proposed method with two baseline methods. The baseline alignment methods used are the Smith-Waterman algorithm [51], and DELTA-BLAST.

4.3. Results

Figure 4.4 shows the model accuracy distribution of difficult targets, meaning pairs of query and templates that are not detected using DELTA-BLAST without ISS, but are detected by the proposed ISS method. The figure shows a comparison of the DELTA-

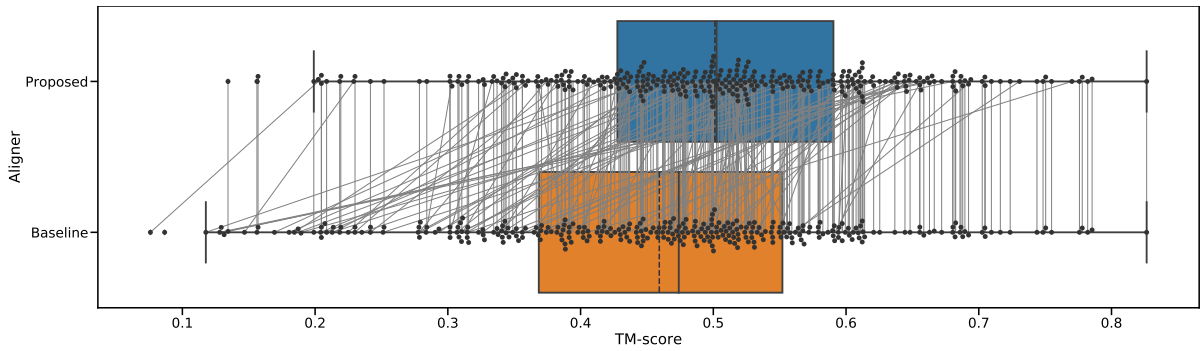
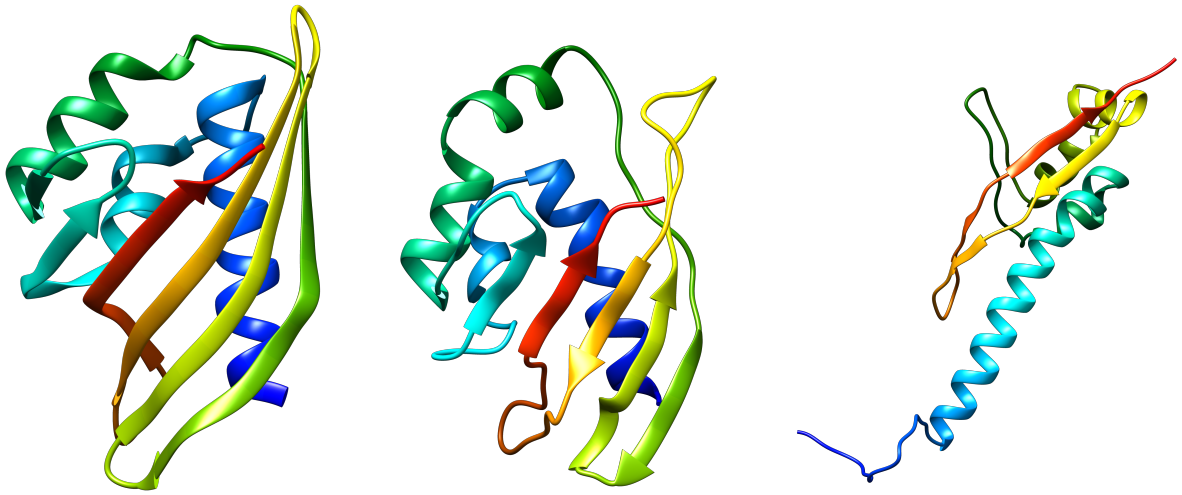


Figure 4.5.: TM-score distribution of hard targets, meaning hits that are not detected by DELTA-BLAST without intermediates. A solid line represents the median, and a dashed line represents the mean. Dots within the boxplot represents individual samples. Gray lines represent the same query and template pair. For the results shown, all templates are from the same superfamily as the query.

BLAST and Smith-Waterman algorithm as a baseline. For DELTA-BLAST, we changed the word score threshold to 1 (`-threshold 1`) to get longer aligned regions. On average, alignments using DELTA-BLAST did not generate more accurate models than alignments using the Smith-Waterman algorithm.

Figure 4.5 also shows the model accuracy distribution for a difficult target, which is a comparison of the proposed method with the Smith-Waterman algorithm as the baseline. Using these remote homologs, our method generated more accurate models than those generated using the Smith-Waterman alignment, with average TM-scores of 0.50 and 0.46, respectively. We tested the statistical significance using the related t-test. The p-value was 6.3×10^{-13}, and the average difference is significant ($p < 0.01$). Gray lines indicate the same query and template pair, and they reveal that many models have lower than average TM-scores, and thus improved accuracy.

Figure 4.6 and 4.7 show two examples of results. In these examples, we address alignment quality by comparing them with a structural alignment. In structural alignment, the structural difference between a target protein structure and a template protein structure is minimized; thus, sequence alignments generated by structural alignment are ideal



(a) Native structure (b) Proposed method (TM-score=0.73) (c) Smith–Waterman (TM-score=0.56)

(d) Alignment by proposed method

```
d1tp6a_          ................CAYRREIHHAHVAIRDWLA.GDSRADALDALMARFAEDFSMVTPHGVVLDKTALGELFRSKGGTRPGLRIEIDGESLLA     78
UniRef50_B5XV41  ................MNPYLQEVLDAHVLIERWLSQGEGSA...EALMTRFAAEFIMIPPGGEKMDYPAVSRFFHHAGATRPGLHIVVDQAKIIS     77
UniRef50_A0A2V8  ......................QEIRTLEHEWFDSYVRGN...RAAFDRIVADDAVMTYGNGKVGNKSEAIAEVKAPADASYSLTSDDVKVSVY     69
d2rfra1          MDDLTNLAARLRLLEDREEIRELIARYGPLADSGDAEALSE...LWVEDGEYAVVGFATAKGRAAIAALIDGQTHRALMADGCAHFLGPATVTVE     92

d1tp6a_          SGVDGATLAYRE......IQSDAAGRSERLSTVVLHRDDEGRLYWRHLQETFCG............     126
UniRef50_B5XV41  EWHDGAAVLYRE......SQTLADGSENVRWSTAIFQQAEGKMIWRHLQETRLG............     125
UniRef50_A0A2V8  GDTAIVTGRVTE......KGIFNGRSVNSQSRYTDVWVKRNGLWQVVAAQNTRLPQGPS.......     122
d2rfra1          GDTA..TARCHSVVFRCVSGTFGSHRV.SANRWT..FRRTPAGWRAVRRENALLDGSAAARALLQF     153
```

(e) Smith-Waterman alignment

```
d1tp6a_    CAYRREIHHAHVAIRDWLAGDSRADALDALMARFAEDFSMVTPHGVVLDKTALGELFRSKGGTRPGLRIEIDGESLLASGVDGATLAYREIQSDAAGRSERLSTV    105
d2rfra1    ..............MDDLTNLAARLRLLEDREEIRELIARYGPLADSGDAEALSELWVEDGEYAVVGFATAKGRAAIAALIDGQT..HRALMADGCAHFLGPATV     89

d1tp6a_    VLHRDDEGRLYWRHLQETFCG.....................................................     126
d2rfra1    TVEGDTATARCHSVVFRCVSGTFGSHRVSANRWTFRRTPAGWRAVRRENALLDGSAAARALLQF     153
```

(f) Structural alignment

```
d1tp6a_    ................CAYRREIHHAHVAIRDWLAGDSRADALDALMARFAEDFSMVTPHGVVLDK.TALGELFRSK.GGTRP.GLRIEIDGESLLASGVDGATL     86
d2rfra1    DDLTNLAARLRLLEDREEIRELIARYGPLADSG........DAEALSELWVEDGEYAVVGFATAKGRAAIAALIDGQTHRALADGCAHFLGPATV.TVEGDTATA     96

d1tp6a_    AYREIQSDAA....GR.SERLSTVVLHRDDEGRLYWRHLQETFCG............     126
d2rfra1    RCHSVVFRCVSGTFGSHRVSANRWTFRRTPA.GWRAVRRENALLDGSAAARALLQF     151
```

Figure 4.6.: Example 1: SCOP ID of query and template protein are `d1tp6a_` and `d2rfra1`, respectively.

for homology modeling. Often, the sequence alignments generated by homology detection methods are dissimilar to those generated by structural alignment, especially for remote homologs.

As shown in Figure 4.6, the TM-score of the proposed and Smith-Waterman methods are 0.73 and 0.56, respectively. This demonstrates that our method generated a more appropriate alignment for homology modeling. The TM-score of the query and template is 0.77. The alignment generated using intermediates is similar to the structural alignment. The aligned region of the pairwise Smith-Waterman alignment is right-shifted and narrower in comparison to that obtained using the proposed method. Therefore, the model generated from pairwise the Smith-Waterman alignment is different from the native structure. We have achieved an accuracy of more than TM-score 0.7 using template proteins that could not be discovered by DELTA-BLAST. The accuracy allows us to predict the important sites for the protein activity, discuss structurally important amino acid substitutions, and assign protein family.

In contrast, Figure 4.7 shows a slightly worse results than others. The TM-score of the proposed and Smith-Waterman methods were 0.50 and 0.56, respectively. The query and template's TM-score was 0.75, but the accuracy of the models was approximately 0.5. In this case, structural alignment by TM-align (Figure 4.7f) contains many small gap regions, and it isn't easy to make a similar alignment using algorithms based on affine-gap. A similar result will be obtained when the sequence identity is relatively low. Figure 4.5, three other models made the results worse, but the differences were approximately -0.01 only. We examined all results that made accuracy worse but could not find reasonable causes from the results.

(a) Native structure (b) Proposed method (TM-score=0.50) (c) Smith–Waterman (TM-score=0.56)

(d) Alignment by proposed method

```
d2pcsa1           ........................NGNGSIELKGTVEEVWSKLMDPSILSKCIMGCKSLELIGEDKYKADLQIGIAAVKGKYDAIIEVTDIKP     69
UniRef50_A0A2V8   .......................MKIEGTHELRAPRERVWQALVDPSVLQRCIPGCERLERTGEDSYAATLRTGVGAIKGVFQGSVRLEDMSA     70
UniRef50_A0A3D1   ........................QNGQCEIAAPRDLVWSALQNPDVLAASIPGCKSMRRIDATHYLASVQTKVGPVSASFEVQIDLQDIDP     68
d5i8fa1           MAAYTIVKEEESPIAPHRLFKALVLERHQVLVKAQPHVFKSGEIIEGDGGVGTVTKITFVDGHPLTYMLHKFDEIDAANFYCKYTLFEGDVLRDN     95

d2pcsa1           PYHYKLLVNGEGGPGFVNAEGVIDLTPINDECTQLTYTYSAEVGGKVAAIGQRMLGGVAKLLISDFFKKIQKEIAKS...    146
UniRef50_A0A2V8   PTHYRIVVDGKGQPGFLKGAGDLDLEE.RDGGTVVRYAGDVQVGGTLASVGQRMIQGAAKMMAAQFFTALEAEAQVEQG.    148
UniRef50_A0A3D1   PNRYTLSGEGKGVAGFAKGQAEVDLIE.AAQGTLLKYRLQATVGGKLAQVGSRLIDGTTRKLANEF..............    133
d5i8fa1           IEKVVYEVKLEAVGGGSKGKITVTYHP.KPGCTVNEEEVKIGEKKAYEFYK............................    145
```

(e) Smith-Waterman alignment

```
d2pcsa1    NGNGSIELKGTVEEVWSKLMDPSILSKCIMGCKSLELIGEDKYKADLQIGIAAVKGKYDAIIEVTDIKPPYHYKLLVNGEGGPGFVNAEGVIDLTPINDECTQLT    105
d5i8fa1    ...............................MAAYTIVKEEESPIAPHRLFKALVLERHQVLVKAQPHVFKSGEIIEGDGGVGTVTKITFVDGHPLTYMLHKFD     73

d2pcsa1    YTYSAEVGGKVAAIGQRMLGGVAKLLISDFFKKIQKEIAKS.....................................................    146
d5i8fa1    EIDAANFYCKYTLFEGDVLRDNIEKVVYEVKLEAVGGGSKGKITVTYHPKPGCTVNEEEVKIGEKKAYEFYKQVEEYLAANPEVFA    159
```

(f) Structural alignment

```
d2pcsa1    ....NGNGSIELKGTVEEVWSKL.MD.PSILSKCIM.GCKSLELIGE.....DKYKADLQIGIAAVKGKYDAIIEVTDIKPP.YHYKLLVNGEGGPGF.V.NAEG     90
d5i8fa1    MAAYTIVKEEESPIAPHRLFKALVLERHQVLVKAQPHVFKSGEIIEGDGGVGTVTKITFVDG....HPLTYMLHKFDEIDAANFYCKYTLFEGDVLRDNIEKVVY    101

d2pcsa1    VIDLTPINDECTQLTYTYSAEVGGKV..AAIGQRMLGGVAKLLISDFFKKIQKEIA.KS....    146
d5i8fa1    EVKLEA.VGGGSKGKITVTYHPK...PGCTVNE.EEVKIGEKKAYEFYKQVEEYLAANPEVFA    159
```

Figure 4.7.: Example 2: SCOP ID of query and template protein are `d2pcsa1` and `d5i8fa1`, respectively.

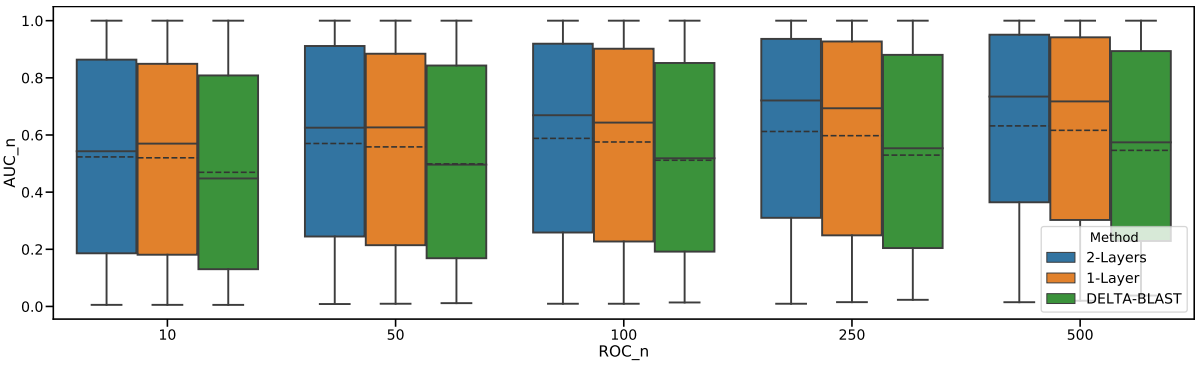


Figure 4.8.: AUC distribution of homology detection. In the boxplots, solid lines indicate medians; dashed lines indicate means. X and Y axes show the allowed count of false positives and AUC, respectively. 1-Layer and 2-Layers indicate a single intermediate sequence search and double intermediate search, respectively.

4.4. Discussion

4.4.1. Homology detection accuracy of intermediate sequence search

In this study, we implemented a simple intermediate sequence search to avoid any influence from specific algorithms. Thus, the homology detection accuracy of the search method was unclear. To verify the accuracy of our implementation, we performed an evaluation test using the SCOP database. For the evaluation of detection accuracy, we used a receiver operating characteristic (ROC) curve, because of the imbalance between the number of homologs and non-homologs. For quantitative analysis, we employed the area under the ROC curve (AUC) as the evaluation metrics [85]. Additionally, instead of using the original ROC and AUC, we used ROC_n and AUC_n according to previous methods [86]. These methods considered results only up to the nth false positive, and AUC_n was normalized by the number of false positives and cutoff value n. We define a true positive homology detection as results that are in the same superfamily as the query protein. By comparing the AUC of homology detection, we evaluated our method against DELTA-BLAST without ISS, as the baseline.

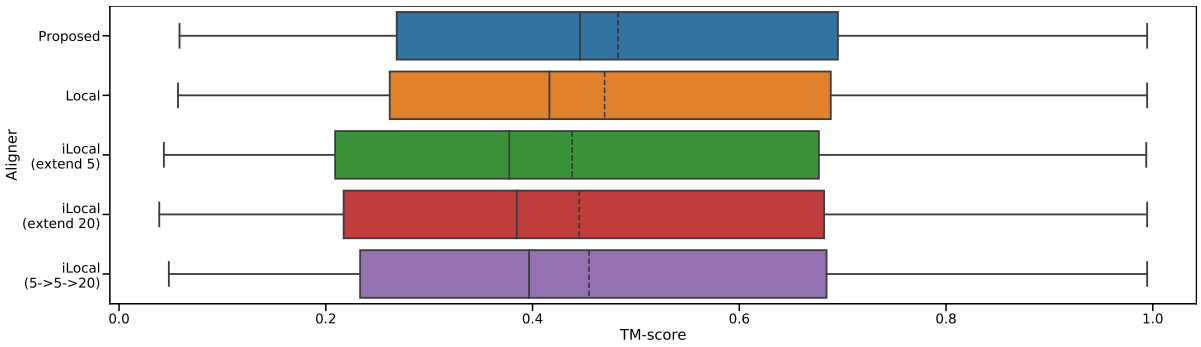


Figure 4.9.: TM-score distribution upon changing extension length. Local and iLocal indicate query-template pairwise Smith-Waterman without intermediates and Smith-Waterman with intermediates, respectively. Solid lines in boxplots indicate median, and dashed lines indicate mean.

Figure 4.8 shows the AUC distribution of the accuracy of homology detection. For all of the allowed number of false positives $n = (10, 50, 100, 250, 500)$, the proposed method with two intermediate layers overcomes the average AUC of DELTA-BLAST without intermediate layers. In the case of $n = 10$, the average AUC of the proposed method was 0.52 while that of the DELTA-BLAST was 0.46. In the case of $n = 500$, the average of our method reached a value of 0.63 while that of the BLAST approach increased to 0.55. As for the number of intermediate layers, the AUC of two intermediate layers is consistently higher than that of one layer. By increasing the allowed number of false positives (n), the AUC of both methods increased.

4.4.2. Model accuracy distribution by various expansion lengths

Figure 4.9 shows the model accuracy distribution upon changing the length of expansion of the hit region. We tried 5 and 20 for the length, and the combination of 5 in intermediate layers and 20 for the final, was used in the proposed method. Longer expansion length could generate more accurate models, and the proposed parameters, which are 5 for two intermediate layers and 20 for the final layer, show the best result. However, query-template pairwise Smith-Waterman alignment shows the highest TM-score

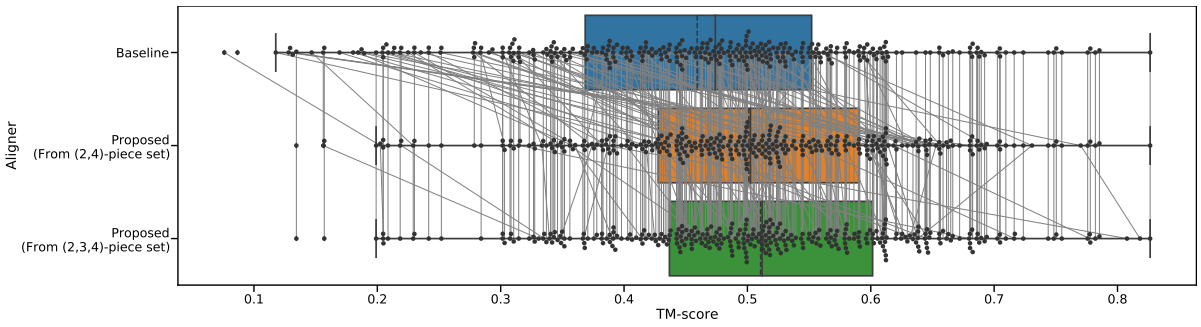


Figure 4.10.: TM-score distribution of hard targets, meaning hits that are not detected by DELTA-BLAST without intermediates. A solid line represents the median, and a dashed line represents the mean. Dots within the boxplot represents individual samples. Gray lines represent the same query and template pair. For the results shown, all templates are from the same superfamily as the query. Proposed (From (2,4)-piece set) shows the results from a pair of query and hit sequences or two intermediate sequences. Proposed (From (2,3,4)-piece set) shows the results from four different alignments: a pair of query and hit sequences, two intermediate sequences, the first intermediate sequence, and the second intermediate sequence.

average, yet. Finally, our method executes both of Smith-Waterman with intermediates and pairwise Smith-Waterman without intermediates. Using the proposed method, we selected the one that shows a wider aligned region and the results are shown in Figure 4.9.

4.4.3. Different combination of intermediate sequence alignments

In this study, two intermediate sequences were obtained during an intermediate sequence search. Our proposed method used all intermediate sequences for generating sequence alignment between query and template sequences. However, there are additional options to use different combination of intermediate sequences. Thus, in addition to the two alignment candidates, we considered a method that generated an alignment using only one of the intermediate sequences. In other words, we compared the alignment with four different alignments: a pair of query and hit sequences, two intermediate sequences, the first intermediate sequence, and the second intermediate sequence. The selection of

alignments was the same as in the proposed method, where we selected the one with the broadest coverage area for the query sequence.

The results of the evaluation are shown in figure 4.10. Selecting one of the four options showed improvement in both the mean and median. The results were better than those of the baseline method, but some of the alignments became worse. Currently, the coverage area was used in the selection (larger one was selected). However, the rule was too simple and improvement of this selection rule might improve it.

4.5. Conclusion

In this study, we developed an alignment generation algorithm suited for accurate homology modeling based on intermediate sequence search (ISS) for remote homology detection. Our method used the intermediate sequence search method to detect remote homologs and a sum of similarity score to assign rankings. In the alignment generation phase, we proposed a method that extended the hit region detected by the ISS, and used multiple pairwise alignments between intermediate sequences. In addition to alignment generation, we also applied the pairwise Smith-Waterman algorithm to query and the template sequences, and selected one alignment based on the length of the aligned region. We evaluated our method by comparing the AUC of homology detection for sensitivity and selectivity. We also evaluated the quality of the alignments by comparing the accuracy of template-based structural models generated from the alignments. As a result, the proposed method could detect homologs more accurately than DELTA-BLAST without intermediates. The evaluation of alignment quality based on the accuracy of structural models generated from the alignment, revealed that the proposed method generates more appropriate alignments for homology modeling, than those prepared without intermediate sequences. As for domains that are not detected by DELTA-BLAST, which were treated as difficult targets, model accuracy measured by TM-score improves by +0.04

on average, when compared using naïve dynamic programming-based alignment.

This study used a simple ISS model to generate alignments using intermediate sequences and evaluate the alignment quality. However, more intelligent ISS methods are available in existing studies and used in place of the simple ISS method used here, homology detection performance is expected to improve. Despite that, our alignment generation method can be successfully applied to the ISS. Thus, the evaluation of the more sophisticated ISS methods remains as one of future work. Also, because the intermediate layer uses a large sequence database in this research, it often outputs many similar sequences in the intermediate layers. One possibility is to cluster these similar sequences in the intermediate layer using multiple sequence alignment to generate an alignment that includes a lot of similarity information. However, we have not yet found a way to apply the multiple sequence alignment or the PSSMs generated from them to this method. Similarly, if intermediate sequences are similar in multiple pathways leading to the same hit sequence, they may generate multiple sequence alignments that provide important information on homology and structural similarity. The ISS using graph theory takes a similar approach, which may be helpful.

Chapter 5.

Discussion

5.1. Impact of model accuracy improvement for protein function estimation

As shown in our results, we achieved improved model accuracy in structural similarity to a native structure. However, it is difficult to judge whether this improvement is useful for advanced applications, such as protein function estimation. The protein shown in Figure 3.7 is fibronectin type III domain of integrin $\beta 4$, which makes a complex with plectin's actin-binding domain [87]. Thus, we applied a protein–protein docking to the modeled structures and a ligand structure from the complex structure (PDB ID: `4Q58`, Chain: A), using MEGADOCK 4.0 [88] with the default settings, to check the influence of model accuracy. Figure 5.1 shows the docking results using the model from the proposed method and that from HHsearch.

The best docking model selected from the top-10 models is shown. Where the docking calculation based on the HHsearch model failed to detect the correct binding position, the docking calculation based on the proposed method's model succeeded. In the HHsearch model, a loop region after 4^{th} β strand (the red circle in Figure 5.1) became longer than the correct structure because of this wrong alignment, causing a steric clash with

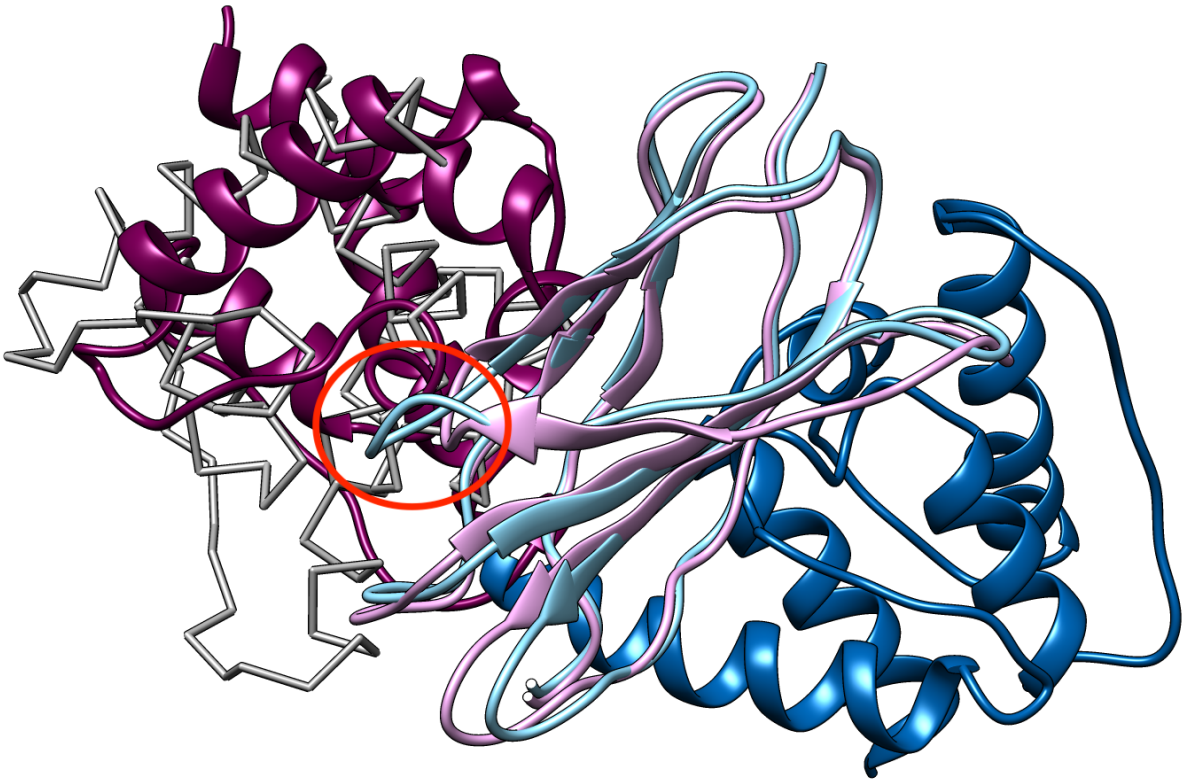


Figure 5.1.: Docking results using modeled structures. The light purple model is generated by the proposed method and the light blue model is from HHsearch. The dark purple model is a ligand structure (plectin) docked by MEGADOCK using the model from the proposed model (ninth model, LRMSD=8.8 Å), and the dark blue model is a ligand structure docked using the model from HHsearch (fourth model, LRMSD=23.8 Å). The light gray model drawn by C_α trace shows the correct position based on the native complex structure. The red circle shows a loop that HHsearch failed to model correctly.

the ligand structure. As a result, the docking calculation failed to detect the correct binding position and there was no docked result with ligand root mean squared deviation (LRMSD) < 10 Å within 3000 models output by MEGADOCK. This is simply one example, but it indicates that the model accuracy improvement achieved with the proposed method is sometimes effective in aid our understanding of the function of a protein.

5.2. Integration of proposed methods

In this section, we shows the integrated procedure of protein structure prediction that we developed and discusses their advantages and disadvantages from the results. The procedure is the following. The intermediate sequence search (ISS) is used for remote homology detection. Next, in the alignment generation phase, the machine learning-based method (described in chapter 3) is used. When the pair of query and template proteins are remote homologs, and the machine learning-based method could not generate alignment that length is enough for accurate homology modeling, an intermediate sequence search-based method (described in chapter 4) is used.

We define true positive hits as the domains in the same superfamily of the queries within the results from the ISS. In the search results, we define the easy target as the hit that is also found by DELTA-BLAST, and treat other hits as a difficult target, which can not be found by DELTA-BLAST but can be found by ISS. We predict protein structure by homology modeling with query and template proteins found by the ISS as true positives only in the top 10 hits. Then, we evaluate the quality of the alignments by comparing the accuracy of structure prediction from models. The test dataset is the same as the domains described in chapter 4. We selected 100 sequences for test data from SCOP_{40}, which is a SCOP database reduced by 40 % sequence identity. Each of the 100 sequences was randomly selected, one from each of the top 100 superfamilies that were sorted according to the superfamily size. In this evaluation, SCOP_{40} is used for the last layer of ISS to include remote homologs. Other parameters for the two proposed methods are the same as described in chapter 3 and chapter 4.

We define the easy target as the hit which is also found by DELTA-BLAST, and treat other hits as a difficult target, which cannot be found by DELTA-BLAST but can be found by ISS. Reviewing results of queries that have both difficult and easy targets, we found that some of them from ISS-based alignment outperform the result of the

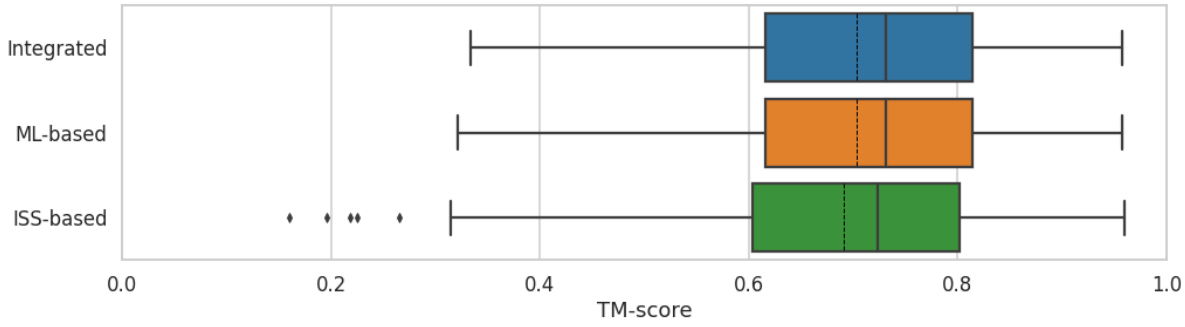


Figure 5.2.: Results of model accuracy score. X axis shows TM-score, and solid and dotted lines mean median and mean value, respectively.

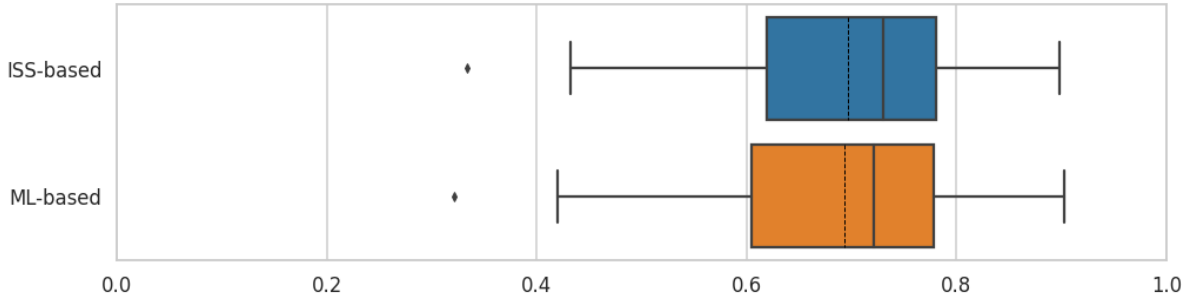


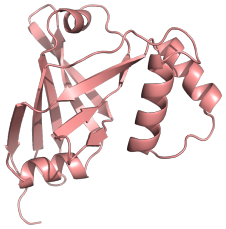
Figure 5.3.: Results of model accuracy score. X axis shows TM-score, and solid and dotted lines mean median and mean value, respectively. We select results from ISS-based alignments that comes from ISS only (not from Smith-Waterman). Also, among two alignments from the proposed methods, we select one that has longer aligned region.

machine leaning-based method. The results are shown in Figure 5.2. We tested the statistical significance of the differences using Wilcoxon signed-rank test. The p-values of "Integrated"−"ML-basd" and "ML-based"−"ISS-based" are 9.7×10^{-7} and 1.4×10^{-17}, respectively. ISS-based method is effective when generating alignment of remote homologs that are not found by existing alignment-based sequence search methods. We found that ISS-based method can be used for these difficult pair of query and template proteins that our machine learning-based alignment method cannot generate alignment for. In the clear case that the ISS-based method does not work well, generated alignment by ISS-based methods is shorter, the machine learning-based method can be applied and shows small improvements.

ISS-based alignment method can generate longer alignment in some cases, and that helps more accurate models generate. The ISS-based method tries Smith-Waterman alignment and ISS-based alignment and selects the one with a longer aligned region. Similarly, we select results from ISS-based alignments that comes from ISS (not Smith-Waterman). Also, among the two alignments from the proposed methods, we select one with longer aligned regions. The results is shown in Figure 5.3. By selecting an alignment that has longer aligned regions, ISS-based alignment outperforms the results of machine learning-based method.

Representative results from the integrated pipeline are shown in Figure 5.4 and 5.5. Figure 5.4 shows ISS-based method generates more suitable alignments for homology modeling; the scores of structure prediction accuracy (TM-score) of models generated from ISS-based alignment and ML-based alignment are 0.673 and 0.518, respectively. In this case, each residues in the target sequence are mapped to residues in the template sequence, but too many gaps are inserted unnaturally. Also, Figure 5.5 shows ML-based method generates more suitable alignments for homology modeling; the scores of structure prediction accuracy (TM-score) of models generated from ISS-based alignment and ML-based alignment are 0.353 and 0.578, respectively. In this case, ISS-based alignment contains too many gaps in the tail, and the gapped region of the target sequence is not mapped to the template sequence. In these unmaped region, homology modeling can not predict the structure because of lack of template structure information.

While developing the integrated procedure described in this chapter, our primary purpose of this evaluation is to show the advantage of the ISS-based alignment method compared with the machine learning-based method. However, there is no significant improvement. Ideally, more accuracy can be achieved because we see that higher accuracy is achieved when intentionally selecting the most accurate results.



(a) Native structure

(b) ISS-based (TM-score=0.673)

(c) ML-based (TM-score=0.518)

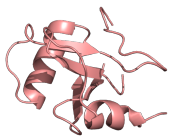
(d) Alignment by ISS-based

```
d1w9aa_  FDDKLLAVISGNSIGVLATIKHDGRPQLSNVQYHFDPRKLLIQVSIAEPR    50
d2fura1  YSDEDLVAMLDRNFTCTVSFIDGGIPYAIPMMLASEGKTIYL--HGSMKS    48

d1w9aa_  AKTRNLRRDPRASILV-DA------DDG------WSYAVAEGTAQLTPPA    87
d2fura1  RIYGILKTGQLIAISLLEINGIVLAKEIKNNSINYVSALIFGRPYEIDDT    98

d1w9aa_  AAPDD--DTVEALIALYR-NIAGE---HSDWDD-YRQAMVTDRRVLLTLP   130
d2fura1  EKKIEVFRLLTEKLVKGRWDNSIKPSYEDLNGVFVFAVKPETFSMKARTG   148

d1w9aa_  ISH-------VYGLPPGMR   142
d2fura1  PPHDTSTDDIWSGVLPI--   165
```

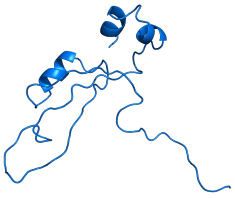
(e) Alignment by ML-based

```
d1w9aa_  ----FDDKLLAVISGNSIGVLATIKHDGRPQLSNVQYHFDPRKLLIQVSI    46
d2fura1  RASYSDEDLVAMLDRNFTCTVSFID-GGIPYAIPMMLASEG--KTIYLHG    47

d1w9aa_  AEPRAKTRNLRRDPRASILV-------DADDGWSYAVAEGTAQLTPPAAA    89
d2fura1  SMKSRIYGILKTGQLIAISLLEINGIVLAKEIKNNSINYVSALIFGRPYE    97

d1w9aa_  PDD-----DTVEALIA-LYRNIAG--EHSDWDDYRQAM------------   119
d2fura1  IDDTEKKIEVFRLLTEKLVKGRWDNSIKPSYEDLNGVFVFAVKPETFSMK   147

d1w9aa_  VTDRR-------------VLLTLPI--------SHVYGLPPGMR--   142
d2fura1  ARTGPPHDTSTDDIWSGVLPIQHTISEAGENAPEYVKSLYGKRIFI   193
```

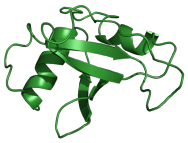
Figure 5.4.: In this case, ISS-based method generates more suitable alignments for homology modeling; the scores of structure prediction accuracy (TM-score) of models generated from ISS-based alignment and ML-based alignment are 0.673 and 0.518, respectively.



(a) Native structure



(b) ISS-based
(TM-score=0.353)



(c) ML-based
(TM-score=0.578)

(d) Alignment by ISS-based

```
d1k8rb_  CILRFIACNGQTRAVQSRGDYQKTLAI--ALKKFSLEDASKFIVCV----    44
d2c5lc1  IQQTLCKAKYSLSILSNPNPSDYVLL--EEVVKDTTNKKT--------TT    40

d1k8rb_  SQSSRIKLITEEEFKQICFNSSSPERDRLIIVPKEKPCPSFEDLRRSWEI    94
d2c5lc1  PKSSQRVLLDQ-ECVFQAQSKWKG-AGKFIL-KLKEQV------------    75

d1k8rb_  E    95
d2c5lc1  -    75
```

(e) Alignment by ML-based

```
d1k8rb_  ---CILRFIA--CNGQTRAVQS--RGDYQKTLAIALKKFS-----LEDAS    38
d2c5lc1  EESFFVQVHDVSPEQPRTVIKAPRVSTAQDVIQQTLCKAKYSLSILSNPN    50

d1k8rb_  KFIVCVSQS-----SRIKLITEEEFKQICFNSSSPERDRLIIVPKEKPCP    83
d2c5lc1  PSDYVLLEEVVKD---K------SSQRVLLDQE-----CVFQAQSKWKGA    86

d1k8rb_  SFEDLRRSWEIE    95
d2c5lc1  GKFILKLKEQV-    97
```

Figure 5.5.: In this case, ML-based method generates more suitable alignments for homology modeling; the scores of structure prediction accuracy (TM-score) of models generated from ISS-based alignment and ML-based alignment are 0.353 and 0.578, respectively.

Chapter 6.

Conclusion

This thesis aims to automatically generate computationally optimal sequence alignments for accurate homology modeling, even for distant homologs that are often detected by modern homology search tools. To achieve more accurate homology modeling, improving sequence alignment generation is an important issue. Therefore, there is a need for alignment generation methods for remote homologs. It aims to make more templates available and increase the number of candidate templates not available in traditional methods. We propose a new algorithm for generating appropriate pairwise alignments for more accurate homology modeling to achieve this goal. We focus on structural alignment-based and intermediate sequence-based as supervised machine learning. We validate the accuracy of the alignments generated by our method based on the accuracy of the predicted structural model. Our method generates more appropriate alignments for homology modeling, especially for remote homologs.

In chapter 3, we described a new sequence alignment generation method for accurate homology modeling that utilizes supervised machine learning. We evaluated the model accuracy of our alignment generation method and found that it outperformed the state-of-the-art methods. We also investigated our method's ability to detect remote

homologies; using AUC_{50} for comparison, our method did not perform better than other methods. However, we found that the proposed method generated relatively accurate 3D models compared with other methods.

In chapter 4, we proposed another alignment generation algorithm suited for accurate homology modeling based on intermediate sequence search (ISS) for remote homology detection. The evaluation of alignment quality based on the accuracy of structural models generated from the alignment, we revealed that the proposed method generates more appropriate alignments for homology modeling, than those prepared without intermediate sequences. As for domains that are not detected by DELTA-BLAST, which were treated as difficult targets, model accuracy measured by TM-score improves by +0.04 on average, when compared using naïve dynamic programming-based alignment.

In addition, in chapter 5, these two methods are integrated into a single pipeline. Firstly, the intermediate sequence search (ISS) is used for remote homology detection. Next, in the alignment generation phase, the machine learning-based method (chapter 3) is used. When the method could not generate alignment that length is enough for accurate homology modeling, an intermediate sequence search-based method is used. If the machine-learning-based method could not generate appropriate alignment, as the next candidate, this ISS-based method will be used. Reviewing results of queries that have both difficult and easy targets, some of them from ISS-based alignment overcomes the result of machine leaning-based. ISS-based method is suitable for the case that generating alignment of remote homologs and can be used for these difficult template–query pair that our machine learning-based alignment method can not generate alignment.

Chapter 7.

Future work

We show some problems that current homology modeling has: computationally optimal alignment generation for accurate homology modeling and suitable alignment generation of remote homologs. While is it true that the alignment adjustment proposed in this thesis can improve the accuracy model, other applications (for example, structure-based drug discovery) often require higher accuracy. We believe that this cannot be achieved by adjusting alignment alone. We need to use all possible information other than homology modeling, *de novo*-like optimization, and contact map prediction to contribute more. On the other hand, the structure optimization still requires a huge amount of combinatorial computation, which will take more time to resolve, although it is continuously evolving due to computers' advances. If improvements are made through future breakthroughs, predicting the three-dimensional structure of proteins will be significantly improved.

Regarding higher accurate structure prediction, we still have another problem. Our proposed methods are optimal only for domain-global structure, which means we do not care about local structure accuracy, such as the functional site of proteins or pocket for ligands. The model-global accuracy often affects function prediction, fold assignment, or computational prediction of protein-protein interaction described in chapter

5. Global structure optimization by alignment adjustment can still do things that can be contributed from computer science fields, like using more complex machine learning methods such as deep neural networks. However, some use-cases require highly accurate models and near-native local accuracies of an important region, such as structure-based drug design or small ligand docking simulation. The our machine learning-based (ML-based) method can create alignments by limiting alignment regions. Hence, if we know the important areas a priori, it can generate more locally accurate alignments. However, since the proposed ML-based method learns the global structural alignment of homologs, it may need a pocket structure or binding site. In any case, as well as alignment adjustment, we need computational structure optimization for highly accurate structure prediction.

Appendix A.

Summarized protocol proposed in chapter 3

A.1. Materials

Structural Classification of Proteins (SCOP) database The SCOP database classifies proteins by class, folds, superfamily (SF), family and domain based on manually curated function/structure classifications and contains redundant sequences. Thus, we used the $SCOP_{40}$ database instead, which contains only domains whose sequence identity is <40 % to avoid overfitting and reduce execution time.

UniRef [76] database For Position Specific Scoring Matrix (PSSM) generation, we used three-iteration PSI-BLAST [35] with the UniRef90 database.

A.2. Equipment

Computer >128 GiB RAM and >150 GiB free storage are recommended.

A.3. Software

PSI-BLAST To generate PSSM of an amino acid sequence

TM-align [61] To generate structural alignment of homologs

Implementation Available at `https://github.com/shuichiro-makigaki/exmachina`

Python 3.6 Required python packages are listed in the repository.

FLANN [89] k-Nearest Neighbor implementation

A.4. Procedure

A.4.1. Training

The primary purpose of the training phase is to generate k-NN model that will be used for substitution score prediction in the prediction and alignment generation phase.

1. Download SCOP_{40} database
2. Generate structural alignments of every domain pair in the same SF by TM-align
3. Select only pairs that the TM-score is ≥ 0.5
4. Generate a PSSM of the domain by three-iteration PSI-BLAST with the UniRef90 database
5. Generate training data and labels
   - As a hyper-parameter, window size is 5.
6. Reduce training dataset to 1/10 by random sampling
   - Because the original training dataset became too large to process within a reasonable computation time.
7. Save the training dataset and the labels as FLANN-acceptable data format

A.4.2. Prediction

The prediction phase consists of score prediction and alignment generation.

1. Prepare homologous two amino acid sequences
2. Generate PSSMs of each sequence by three-iteration PSI-BLAST with the UniRef90 database
3. Predict all substitution scores of each residue pairs
   - Query vector format is the same as the training phase, and the k-NN's classification scores are used for the substitution score directly.
   - As hyper-parameters, the window size is 5, and the number of the neighbor is 1000.
4. Save predicted substitution score matrix
5. Generate local sequence alignment by Smith-Waterman algorithm
   - During the dynamic-programming, the predicted substitution scores are used for score calculation.

A.5. Notes

- Proposed method implementation "ExMachina" is available in `https://github.com/shuichiro-makigaki/exmachina`.
- Usage documents and actual examples are also in the repository.

- Data for reproducing is available. How to download is described in the documents of the repository.

Appendix B.

Detailed dataset used in chapter 3

Domain IDs shown are the SCOP sid numbers.

Table B.1.: 14 domains from $SCOP_{40}$ as test data.

Class	Domains
a: All alpha proteins	d1wlqc_, d2axtu1, d2pqrb1, d3cr3a1, d2jn6a1
b: All beta proteins	d2zqna1, d1qg3a1, d2a5za1, d1wv3a1, d3etja1
c: Alpha and beta proteins (a/b)	d2w6ka1, d1wzca1, d1v7ra_, d2dsta1, d3ct6a1
d: Alpha and beta proteins (a+b)	d1y5ha3, d2nwua1, d1tvia_, d1t4ha_, d1th5a1
e: Multidomain proteins	d1ni9a_, d3cw9a1, d3beca2, d2qv7a1, d1wuil1
f: Membrane and cell surface proteins	d2axte1, d2axtd1, d2axtb1, d2axto1, d3dtub2
g: Small proteins	d2vy4a1, d3d9ta1, d2exfa1, d2ayja1, d3dplr1

Table B.2.: Seven domains from $SCOP_{40}$ used for hyperparameter optimization

Class	Domains
a: All alpha proteins	d2ij2a1
b: All beta proteins	d3d85d1
c: Alpha and beta proteins (a/b)	d3etja2
d: Alpha and beta proteins (a+b)	d2iiza1
e: Multidomain proteins	d1wuis1
f: Membrane and cell surface proteins	d3dhwa1
g: Small proteins	d3d4ub1

Table B.3.: Seven domains from $SCOP_{40}$ used for gap penalty optimization

Class	Domains
a: All alpha proteins	`d1tw9a1`
b: All beta proteins	`d3e5ua2`
c: Alpha and beta proteins (a/b)	`d1xria_`
d: Alpha and beta proteins (a+b)	`d2jmua1`
e: Multidomain proteins	`d2zd1b1`
f: Membrane and cell surface proteins	`d2zfga1`
g: Small proteins	`d2vuti1`

Appendix C.

Detailed test dataset used in chapter 4 and 5

Domain IDs shown are the SCOP sid numbers.

Table C.1.: Dom. and SF. shows SCOP domain ID and SuperFamily ID, respectively.

Dom.	SF.						
d1xg0c_	a.1.1	d1jjcb3	b.40.4	d2c42a1	c.36.1	d4c2va_	d.144.1
d3g0oa2	a.100.1	d1w9aa_	b.45.1	d1mkya1	c.37.1	d1k8rb_	d.15.1
d2zwua_	a.104.1	d4k60a_	b.47.1	d1ywfa1	c.45.1	d5le5a_	d.153.1
d3vyca_	a.118.1	d1mkea1	b.55.1	d3l9va_	c.47.1	d2e7ya_	d.157.1
d5j90a_	a.118.8	d2ux6a_	b.6.1	d1j24a_	c.52.1	d2nana_	d.169.1
d2xpwa2	a.121.1	d2ra6a_	b.60.1	d3mdqa2	c.55.1	d1tp6a_	d.17.4
d3b3hb_	a.25.1	d1e43a1	b.71.1	d1cz9a_	c.55.3	d2io8a2	d.3.1
d2d48a_	a.26.1	d5vf5a1	b.82.1	d1y0ya3	c.56.5	d1kw3b1	d.32.1
d1e29a_	a.3.1	d1gp6a_	b.82.2	d1hgxa_	c.61.1	d2nyca1	d.37.1
d2bnma1	a.35.1	d1wa3a_	c.1.10	d2g1pa_	c.66.1	d4a0yb_	d.38.1
d1qlsa_	a.39.1	d4j1oa2	c.1.11	d3f0ha1	c.67.1	d3toya1	d.54.1
d2oi8a1	a.4.1	d1geqa_	c.1.2	d1jyka1	c.68.1	d1tdja2	d.58.18
d4ejoa1	a.4.5	d4ddea1	c.1.8	d1qe3a_	c.69.1	d1tr0a_	d.58.4
d3q9va_	a.4.6	d2vhla2	c.1.9	d5c40a1	c.72.1	d2pe8a1	d.58.7
d1v2aa1	a.45.1	d1jl5a_	c.10.2	d4dnga_	c.82.1	d1q0qa3	d.81.1
d4pc3c1	a.5.2	d2ho4a_	c.108.1	d5uofa1	c.87.1	d1f5va_	d.90.1
d3edfa1	b.1.18	d4eu9a1	c.124.1	d3u7qa_	c.92.2	d2fpqa1	d.92.1
d3r4da2	b.1.1	d4qfea1	c.14.1	d4q6ba_	c.93.1	d2ysxa1	d.93.1
d2gysa4	b.1.2	d1xu9a1	c.2.1	d2fyia1	c.94.1	d1tvfa2	e.3.1
d1ddla_	b.121.4	d1n57a_	c.23.16	d4yuca1	c.95.1	d1quba1	g.18.1
d1zboa1	b.122.1	d3cu5a1	c.23.1	d2fiaa1	d.108.1	d2bz6l_	g.3.11
d2bbaa1	b.18.1	d5ljla1	c.23.5	d5c7qa_	d.113.1	d1pvza_	g.3.7
d1j1ta_	b.29.1	d1h3fa1	c.26.1	d2pcsa1	d.129.3	d2lcea1	g.37.1
d1wyxa_	b.34.2	d1zuna1	c.26.2	d1rwza2	d.131.1	d1rutx3	g.39.1
d5fb8c_	b.36.1	d1v59a1	c.3.1	d1z0wa_	d.14.1	d1weoa1	g.44.1

Appendix D.

Improvement of homology modeling by chimera alignment

D.1. Introduction

The structures of proteins determine their functions in organisms. Therefore, determining the structure of a protein provides important information that is valuable for various practical purposes in the biological sciences, such as virtual screening, function prediction, etc. However, while there are over 88 000 000 (July 2017) protein sequence entries in the Reference Sequence (RefSeq) [90] collection, there are only ∼130 000 entries in the Protein Data Bank (PDB) [91] as yet. There is a need for the prediction of a protein's structure from its amino acid sequence. Despite the demand, the prediction of protein structure is still a challenging problem. For small proteins, *ab initio* structure prediction methods can be used. However, for bigger proteins, *ab initio* methods are computationally difficult [92]. Hence, if we can find homologous proteins of a prediction target in PDB, we can employ template-based modeling methods. Template-based modeling is based on the premise that homologous proteins have similar structures. Therefore, the method requires an alignment between the target and template sequences. These alignments are usually retrieved by running template search tools. To make the alignments while searching homologous proteins, substitution matrices [34], sequence profiles [35], hidden Markov models (HMMs) [42], [46] techniques are often used. However, the sequence alignments generated by search tools vary depending on the algorithm used. Some of the alignments are incorrect and therefore unsuitable for use in template-based modeling.

Assuming that we could know the target structure before modeling, we could obtain a near-native predicted structure by template-based modeling with a structural alignment based on the shapes and three-dimensional conformations of the target and template proteins. The DALI [93] algorithm and the Combinatorial Extension (CE) method [94] were proposed in the 1990s, and the TM-align [61] technique was developed in the 2000s. As an example of the differences among sequence alignments performed by the structural alignment and non-structural alignment methods, figure D.1a shows the pairwise alignments of a target protein and a homologous protein sequence. The DALI algorithm was used for the structural alignment, and FFAS [38], HHPRED [46], and SPARKS-X [39] were used as template search tools. Figure D.1b–D.1f and table D.1g

show the predicted models predicted by these alignments and their accuracies. As a measure of model accuracy, we used the template modeling score (TM-score) [47], which extends the approaches used in the Global Distance Test (GDT) [48] and MaxSub [49]. The TM-score does not depend on protein size, and it has a strong correlation with the quality of each final full-length model. All alignments by template search tools differed from the structural alignment, and a model based on the structural alignment was the nearest match to the a native structure. In addition, we observed the interesting phenomenon that a structural alignment is essentially composed of sub-strings from each template search tool, and small gaps are inserted much more frequently.

In this study, we propose a novel method to generate more accurate sequence alignments when using template-based modeling. The approaches employed in template-based modeling are often grouped into two categories according to the number of template proteins used; namely, single-template and multiple-template methods. The use of multiple templates is often effective for improving the accuracy of modeling. However, single-template modeling is more widely used because it requires only a single template protein. Thus, in this study, we concentrated on improving the accuracy of single-template modeling. Although our proposed method generates multiple predicted models using multiple pairwise alignments, it requires only a single template. The proposed method merges multiple pairwise alignments generated by multiple template search algorithms for a single template, and then constructs a multiple sequence alignment from those alignments. The multiple sequence alignment can be treated as a profile and converted to a HMM. The model can emit residues following probability and generate pairwise alignments between the target and template sequences. The generated alignments are composed of several parts of different algorithm-based alignments analogously to a chimera, which is a mythical monstrous creature composed of the parts of multiple animals. We can generate an arbitrary number of alignments from the model, which means that an arbitrary number of models is generated. As a result, some of the generated models are more accurate than those produced from the "original" pairwise alignments (that is, the alignments from the template search tools used for making HMMs). However, we have to select the most accurate model from many candidate models. To select the final model, we defined a model assessment score based on the joint probability of the HMM and the existing model quality assessment score. Finally, we assessed our method by comparing the quality of the models constructed by our methods with those of the "original" alignments from the template search tools.

D.2. Methods

Our basic idea is that the sequence alignments produced by multiple template search algorithms are not totally incorrect, but are partially correct and have suitable parts for template-based modeling. By merging such partially correct alignments, we can obtain new alignments that generate more accurate models.

To implement our idea, we: (1) retrieve various alignments of template and target sequences from multiple template search tools; (2) merge them into a multiple sequence

Structural alignment (DALI)

```
QQQEATLAIRPVGQ--GIGMPD-GFSVWHHLDANGIRFKSITPQ-KDGLLIKFDSTAQGAAAKEVLGRALPHGYIIALLE
--RTLMTFVSVTGNPTREESDTITKLWQTSLWNNHIQA-ERYMVDDNRAIFLFKDGTQAWDAKDFLI-EQERCKGVTIEN
```

Alignment 1 (FFAS)

```
QQQEATLAIRPVGQGIGMPDGFSV----WHHLDANGIRFKSITPQKDGLLIKFDSTAQGAAAKEVLGRALPHGYIIALLE
---RTLMTFVSVTGNPTREESDTITKLWQTSLWNNHIQAERYMVDDNRAIFLFKDGTQAWDAKDFLIE------------
```

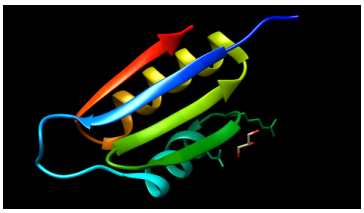
Alignment 2 (HHPRED)

```
QQQEATLAIRPVGQ-GIGMPDGFSVWHHLDANGIRFKSITPQKDGLLIKFDSTAQGAAAKEVLGRALPHGYIIALLE
------VSVTGNPTREESDTITKLWQTSLWNNHIQAERYMVDDNRAIFLFKDGTQAWDAKDFL--------------
```

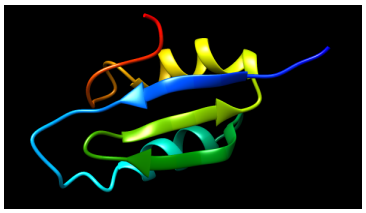
Alignment 3 (SPARKS-X)

```
QQQEATLAIRPVGQGIG--MPDGFSVWH-HLDANGIRFKSITPQKDGLLIKFDSTAQGAAAKEVLGRALPH-GYIIALLE
--RTLMTFVSVTGNPTREESDTITKLWQTSLWNNHIQAERYMVDDNRAIFLFKD---GTQAWDAKDFLIEQERCKGVTIE
```

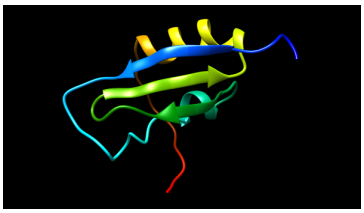
(a) Alignments from the structural alignment (DALI) and template search tools (FFAS, HHPRED, and SPARKS-X). In each alignment, the first line is a target and the second is a template sequence. Colored residues are shared in alignment in DALI.



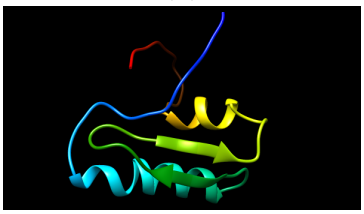
(b)



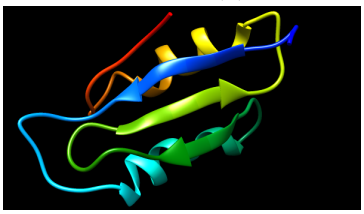
(c)



(d)



(e)



(f)

(D.1b) Native	TM-score
(D.1c) DALI	0.710
(D.1d) FFAS	0.476
(D.1e) HHPRED	0.474
(D.1f) SPARKS-X	0.524

(g)

Figure D.1.: An example of alignments that are not suitable for template-based modeling. The target protein is PDBID: 4PWU, and the template protein is PDBID: 3OFE. (All the images were rendered by UCSF Chimera [70].)

alignment and convert it to a HMM; (3) generate N pairwise alignments from the HMM by random sampling; (4) make models from all the alignments; (5) estimate the quality of each model; and (6) select a final model. Figure D.2 illustrates an overview.

D.2.1. Template Search

Template search is the most important phase in template-based modeling, and many methods have been developed during the past couple of decades. First, we search templates into multiple template search tools and collect results from them. The template candidates found by multiple search tools are sorted based on their frequency, and the same ranked templates are re-ranked based on their mean rank value. The protein at the top of the ranking is selected as the template. In this way, a template protein and the alignments between the target and template sequences are decided.

Template search tools often return multiple search results for a single template candidate when different sub-structures in a template protein can be used for the template. After deciding the template protein, to classify sub-structure matches, template sequences in alignments are clustered by CD-HIT [95]. Only sequences in the biggest cluster are added to the multiple sequence alignment.

D.2.2. Making a HMM

To obtain a consensus among the sequence alignments produced by search tools, we merge them into a multiple sequence alignment (MSA) and convert it to HMM. Many tools to make an MSA have been developed and are available such as Clustal Omega [96], MAFFT [97], and T-Coffee [98]. In this study, however, we aim to construct an MSA of the same protein. Thus, we use a simple algorithm to merge pairwise alignments and make an MSA. If an alignment being added to the MSA has gaps in a target sequence, we insert the same number of gaps to all sequences in the MSA. This algorithm does not use any substitution matrices and generates a relatively "sparse" alignment. However, the generated alignment can reflect all used alignments, and is a reasonable method to obtain a consensus.

We found that certain template search tools outperformed other tools on many targets. Therefore, we treat the results from these high-performing template search tools as preferential alignments. To enlarge the effect of these tools' results in a profile HMM, we add the alignments produced by these tools to each multiple alignment twice. Which tools should be added is one of the hyperparameters, and we determined it by a 3-fold cross-validation. The details of this step are described in the Results section.

Next, this multiple sequence alignment is treated as a profile and used to make a HMM. Regarding the HMM profiles that best reflect MSAs, HMMER [42] and HHPRED are well known. In this study, we employed a similar process to these algorithms. Additionally, emission nodes in the HMM have a residue position as well as as emission probability. During the next sampling phase, the profile HMM with a residue position ID can generate pairwise alignments that keep the same order of residues as the original target and template sequences.

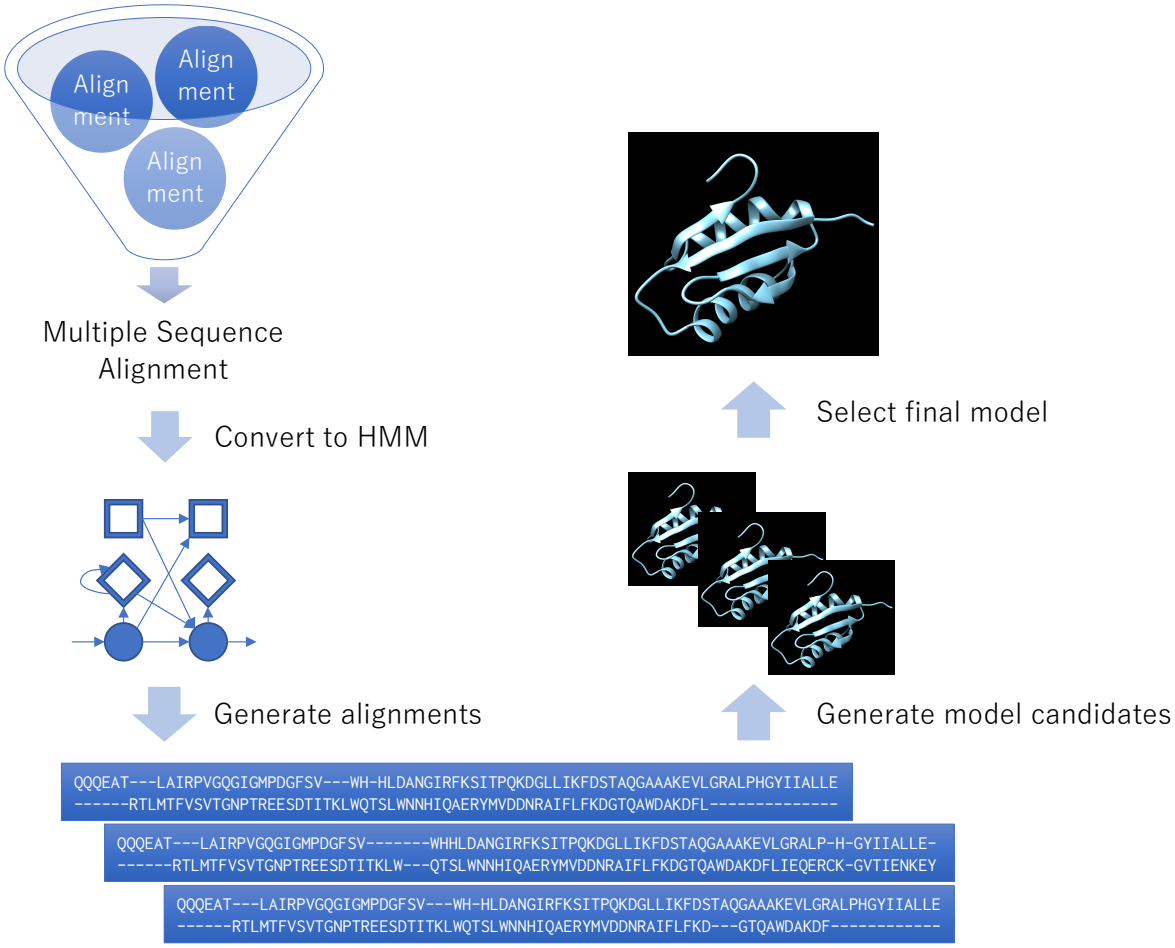


Figure D.2.: Overview of proposed method

D.2.3. Alignment Sampling and Model Candidate Generation

For template-based modeling, we need to generate alignments between a template and a target sequence. We use the profile HMM with residue position ID to generate pairwise alignments by concatenating residues and gaps that are observed from the HMM. By generating a random probability before state transition, we move to the next node according to the random probability. At each hidden node, we treat it as a gap in the target sequence or a template sequence. At each emission node, we generate a random probability again, and the HMM emits one template residue following the random probability. Only if the residue position ID being emitted now is bigger than the proceding emitted residue ID, the residue is emitted. To keep the same order as in the template sequence, skipped position IDs between the current and proceding residues are filled with the skipped residues of each template sequence and gaps are added to the target sequence accordingly. Each pairwise alignment will be used for template-based modeling.

D.2.4. Final Model Selection

Models generated in the steps described above include both near-native and inaccurate models. Therefore, we need to estimate the model accuracy and select the best model. Many algorithms have been developed for the assessment of model quality, and it is still a challenging problem [99]. In this study, we use Pcons [100] and ProQ2 [101] for model quality assessment. Because we generate many model candidates from different alignments and have to select one of them, it is reasonable to use the consensus methods such as Pcons. In addition to the scores, we also use a joint probability of the alignment from HMM to select the best model from consensus clusters. The final quality assessment score function in this study is defined by

$$S_{\mathrm{P}} + \frac{1}{1 + e^{-\alpha S_{\mathrm{J}}}} \tag{D.1}$$

where S_{J} is the Z score of $\log J$ in which J is the joint probability of the alignment, and S_{P} is

$$S_{\mathrm{P}} = (1 - \beta) S_{\mathrm{Pcons}} + \beta S_{\mathrm{ProQ2}} \tag{D.2}$$

where S_{Pcons} is the Pcons score and S_{ProQ2} is a global score of ProQ2. $\alpha \in [0, 1]$ and $\beta \in \mathbb{R}_{++}$ are free parameters. We define them by cross-validation as described in the Results section. In addition, we tried to use other functions such as a simple linear combination for that purpose but found that the sigmoid-style score performed best.

D.3. Results

D.3.1. Datasets

We used a subset of a target dataset provided by Critical Assessment of Methods of Protein Structure Prediction round (CASP) 11 [102] (`http://predictioncenter.org/casp11/`). In CASP 11, the targets were classified into three categories for evaluation based on the prediction difficulty: template-based modeling (TM) for an easy target, TM-hard for a midium-difficulty target, and free modeling (FM) for a difficult target after the final submission phase [103]. For FM targets, it is difficult to apply template-based modeling because there is no template in the PDB. Thus, we used only single-domain targets in the TM and TM-hard target categories, and finally, 40 targets were selected. In this dataset, the maximum length of sequences is 458, the minimum is 67, and the average is 207.

D.3.2. Determination of Preferential Alignments and Parameter Optimization

In the proposed method, the sequence alignments of several relatively accurate alignment tools are merged twice to produce a multiple sequence alignment. To determine

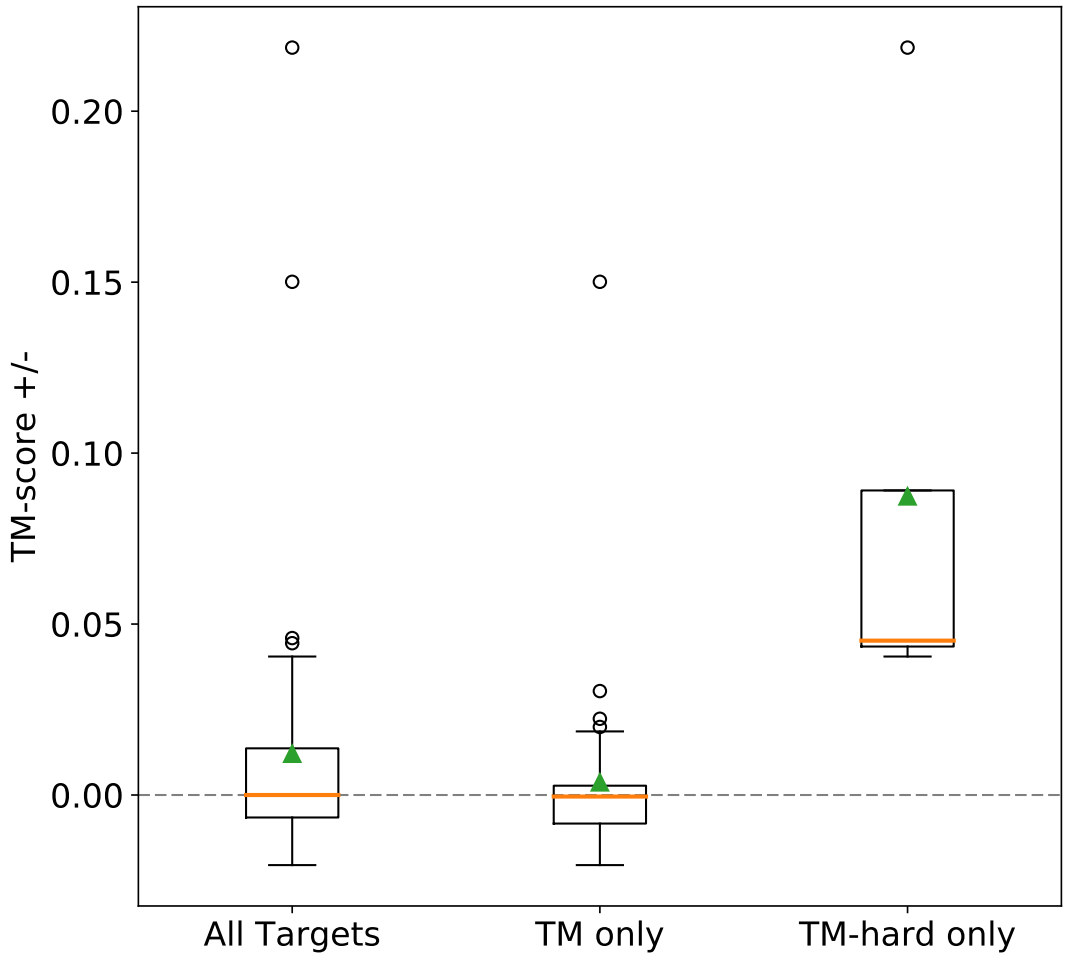


Figure D.3.: Boxplots that show differences of TM-score between the final model from proposed method and a model selected by Pcons and ProQ2 from meta-server results. Red lines and green triangle show median and mean, respectively.

which tools achieve a relatevely high model accuracy, we selected the tools by 3-fold cross-validation. In this study, we used the LOMETS server [104], which searches templates using multiple template search tools including FFAS-3D, HHsearch, MUSTER, pGenTHREADER, SPARKS-X, etc. It is suitable for our purpose because it applies several state-of-the-art methods simultaneously. At each fold, the results from each search tool were sorted by their average TM-score. Results from tools that exceeded the average TM-score were treated as preferential alignment tools. As a result of the 3-fold cross-validation, the tools that were commonly selected as preferential among folds were SPARKS-X, HHSERCH, FFAS, and MUSTER [105].

We also optimized α to 8.3 by the same 3-fold cross validation as described above. β is 0.2, which was the best shown by Ray's study [101]. The number of samples that are generated from HMM (N) is a user-defined parameter, and we set this parameter 500 in this study. For template-based modeling, we used MODELLER [31] with the default parameters.

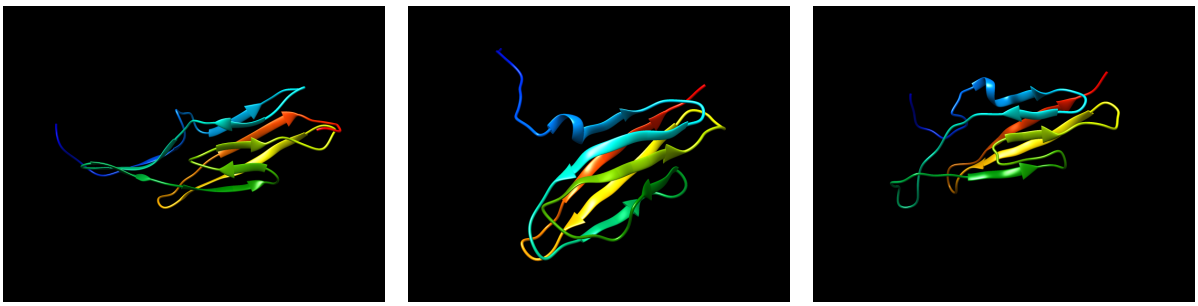
D.3.3. Evaluation of Prediction Accuracy

For evaluating the effectiveness of the proposed method, we compared the accuracy of the models selected by the proposed method to that of the models selected by Pcons and ProQ2 [defined in (D.2)] from models based on alignments generated by template search tools. We used the TM-score as a measure of model accuracy. To keep the same situation as in CASP11, protein structures released after CASP 11 were filtered out from the templates used in this evaluation.

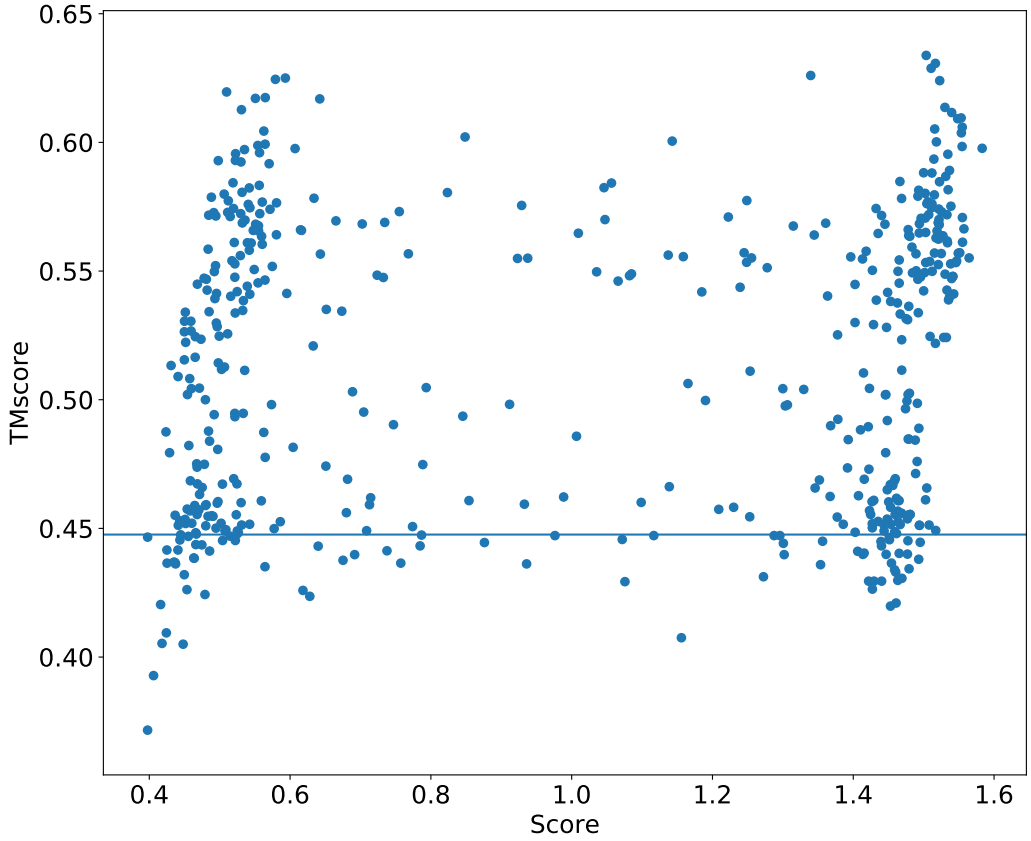
Figure D.3 shows the differences in model accuracy between the proposed method and the baseline method. The proposed method improved the model accuracy by +0.012 on average for all targets. We tested the statistical significance of this improvement using a one-sided t-test. The p-value was 0.044 and is significant ($p < 0.05$). The accuracies of 6 of 40 targets were not changed ($\pm$ 0), which explains why the mean value is better than 0, but the median stays around 0. For the results of TM-hard category targets, the average improvement was +0.087, which is better than that for all targets. However, there were only 4 TM-hard targets in this evaluation, and we could not reject the null hypothesis by t-test ($p = 0.070$). For some targets, the TM-score decreased by a maximum of -0.021. We discuss the reason in the Discussion section.

D.4. Discussion

Figure D.4d and D.5d are scatter plots of the TM-score and the proposed model assessment score. Figure D.4d was generated from an example of models where the proposed method worked well and improved model accuracy. The TM-score had a weak correlation with the proposed model assessment score. The proposed method could generate models that were more accurate than those that were selected from models of meta-server results by score function [defined in (D.2)]. By contrast, the models shown in figure D.5d were generated from one of the models for which the proposed method did

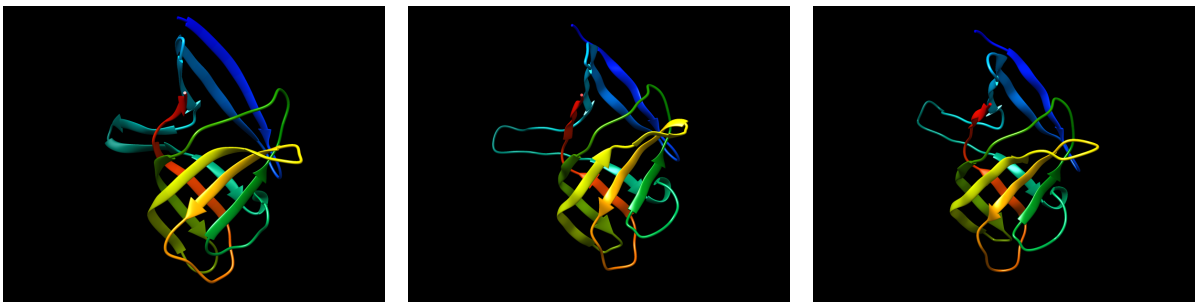


(a) (b) (c)

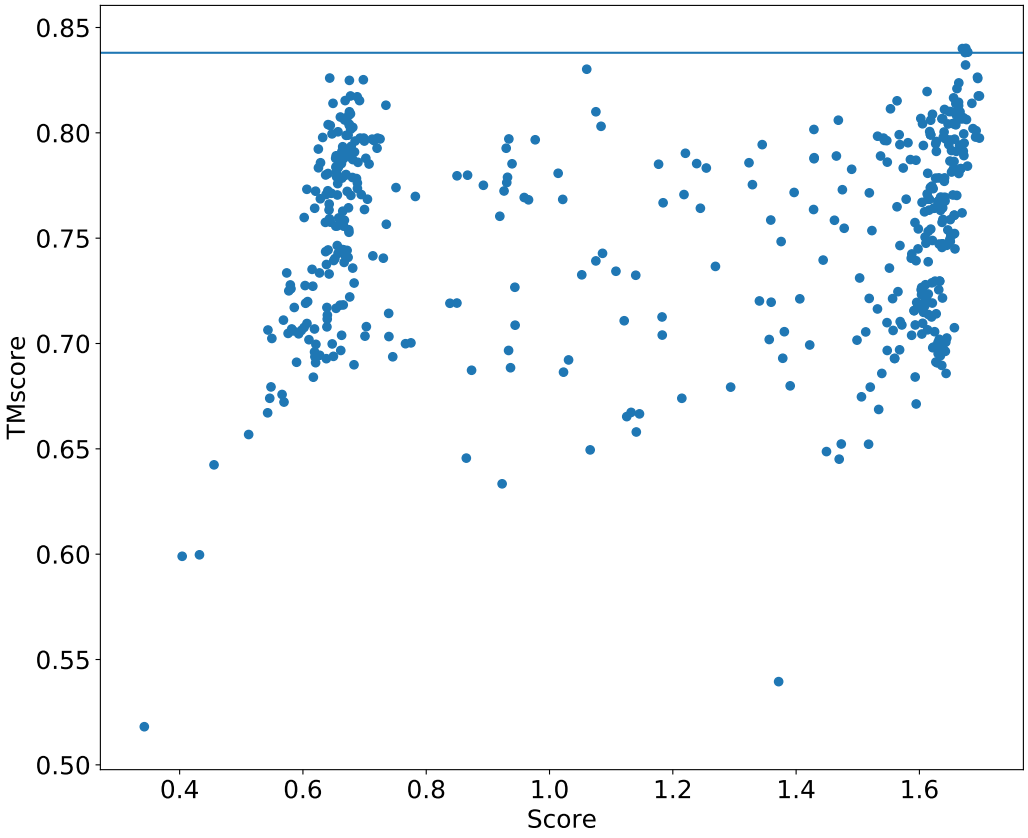


(d)

Figure D.4.: Results from models where the proposed method worked well (Native: 2MQC, Template: 3MZR): figure D.4a, D.4b, and D.4c show a native structure, the best model selected by (D.2) from template search tools, and the best model by the proposed method, respectively. Figure D.4d is a scatter plot of the proposed score against the TM-score. The blue line shows the TM-score of figure D.4b.



(a) (b) (c)



(d)

Figure D.5.: Results from models where the proposed method did not work well (Native: 5FJL,Template: 3ZPE): figure D.5a, D.5b, and D.5c show a native structure, the worst model generated from the original alignments that were generated by the template search server, and the worst model generated by the proposed method, respectively. Figure D.5d shows scatter plots of the proposed score against the TM-score. The blue line shows the TM-score of figure D.5b. In this case, baseline selection could detect a near-native model. Meanwhile, our proposed selection method could not detect it.

not work well. In this case, few models had a high accuracy and the proposed score function could not detect one of them because the score function [defined in (D.2)] selected the most accurate model from the meta-server results. Therefore, although the TM-score also had a weak correlation with the proposed model assessment score and our method could select the near-best model, the absolute model accuracy decreased as a result. To further assess the ability of our proposed method for TM-hard targets, we should add more targets and evaluate the models.

We used the joint probability of HMM in the score function. Even if we did not use it in the score function, there would be no significant changes in TM-score. However, without the joint probability, the results would become unstable, and the deviation of the changes in the accuracies would become larger. For instance, the difference in the prediction accuracy in the worst case becomes -0.042 in the TM-score. From figure D.3, these results are worse than those obtained when using the joint probability of HMM. We think, from the results, that adding the joint probability of HMM helps Pcons and ProQ2 when they could not detect accurate models.

D.5. Conclusion and Future Work

In this paper, we have proposed a new single-template-based method for modeling protein structure. To integrate the alignments generated by state-of-the-art template search algorithms, we merge the alignments and make a profile HMM based on them. This HMM can generate alignments by random sampling and can be applied to template-based modeling to make model candidates using MODELLER. To select the final model from the candidate models, we use the joint probability of the HMM and a model quality score. We evaluated the proposed method based on the TM-score. The proposed method improved the prediction accuracies compared with the results for the original alignments.

In this research, we concentrated on improving single-template modeling. However, the use of multiple templates often improves prediction accuracy. Thus, we aim to focus on extending the method to multiple template modeling in our future work. In addition, the question of how we should treat the case of repeated structures in a template or target structure remains problematic. When a template protein has repeated sub-structures, the target sequence matches to one of the template sub-structures. In this study, this special case is outside the scope by clustering sequence alignments, and our method cannot use information regarding different sub-structure matches. Additionally, when the target structure contains repetitive elements, a target sub-sequence may not be aligned with the corresponding residues of a given template protein, and modeling the sub-sequence will fail. However, we think that this case can be resolved by using multiple templates, and finding a solution for this case is also one of the aims of our future work.

Bibliography

[1] wwPDB consortium, "Protein Data Bank: the single global archive for 3D macromolecular structure data," *Nucleic Acids Research*, vol. 47, no. D1, pp. D520–D528, Oct. 2018. DOI: 10.1093/nar/gky949.

[2] D. Baker and A. Sali, "Protein Structure Prediction and Structural Genomics," *Science*, vol. 294, no. 5540, pp. 93–96, 2001, ISSN: 0036-8075. DOI: 10.1126/science.1065659.

[3] A. Hillisch, L. F. Pineda, and R. Hilgenfeld, "Utility of homology models in the drug discovery process," *Drug Discovery Today*, vol. 9, no. 15, pp. 659–669, 2004, ISSN: 1359-6446. DOI: 10.1016/s1359-6446(04)03196-4.

[4] P. M. Colman, J. N. Varghese, and W. G. Laver, "Structure of the catalytic and antigenic sites in influenza virus neuraminidase," *Nature*, vol. 303, no. 5912, pp. 41–44, 1983, ISSN: 0028-0836. DOI: 10.1038/303041a0.

[5] M. v. Itzstein, "The war against influenza: discovery and development of sialidase inhibitors," *Nature Reviews Drug Discovery*, vol. 6, no. 12, pp. 967–974, 2007, ISSN: 1474-1776. DOI: 10.1038/nrd2400.

[6] T. Schindler, W. Bornmann, P. Pellicena, W. T. Miller, B. Clarkson, and J. Kuriyan, "Structural Mechanism for STI-571 Inhibition of Abelson Tyrosine Kinase," *Science*, vol. 289, no. 5486, pp. 1938–1942, 2000, ISSN: 0036-8075. DOI: 10.1126/science.289.5486.1938.

[7] S. Cowan-Jacob, G. Fendrich, A. Floersheimer, P. Furet, J. Liebetanz, G. Rummel, P. Rheinberger, M. Centeleghe, D. Fabbro, and P. Manley, "Structural biology contributions to the discovery of drugs to treat chronic myelogenous leukaemia," *Acta Crystallographica Section D: Biological Crystallography*, vol. 63, no. 1, pp. 80–93, 2007, ISSN: 0907-4449. DOI: 10.1107/s0907444906047287.

[8] N. S. Pagadala, K. Syed, and J. Tuszynski, "Software for molecular docking: a review," *Biophysical Reviews*, vol. 9, no. 2, pp. 91–102, 2017, ISSN: 1867-2450. DOI: 10.1007/s12551-016-0247-1.

[9] R. Taylor, P. Jewsbury, and J. Essex, "A review of protein-small molecule docking methods," *Journal of Computer-Aided Molecular Design*, vol. 16, no. 3, pp. 151–166, 2002, ISSN: 0920-654X. DOI: 10.1023/a:1020155510718.

[10] D. E. Scott, A. R. Bayly, C. Abell, and J. Skidmore, "Small molecules, big targets: drug discovery faces the protein–protein interaction challenge," *Nature Reviews Drug Discovery*, vol. 15, no. 8, pp. 533–550, 2016, ISSN: 1474-1776. DOI: 10.1038/nrd.2016.29.

[11] Q. C. Zhang, D. Petrey, L. Deng, L. Qiang, Y. Shi, C. A. Thu, B. Bisikirska, C. Lefebvre, D. Accili, T. Hunter, T. Maniatis, A. Califano, and B. Honig, "Structure-based prediction of protein–protein interactions on a genome-wide scale," *Nature*, vol. 490, no. 7421, pp. 556–560, 2012, ISSN: 0028-0836. DOI: 10.1038/nature11503.

[12] B. G. Pierce, K. Wiehe, H. Hwang, B.-H. Kim, T. Vreven, and Z. Weng, "ZDOCK server: interactive docking prediction of protein–protein complexes and symmetric multimers," *Bioinformatics*, vol. 30, no. 12, pp. 1771–1773, Feb. 2014, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu097.

[13] M. Ohue, Y. Matsuzaki, N. Uchikoga, T. Ishida, and Y. Akiyama, "MEGADOCK: An All-to-All Protein-Protein Interaction Prediction System Using Tertiary Structure Data," *Protein & Peptide Letters*, vol. 21, no. 8, pp. 766–778, 2013, ISSN: 0929-8665. DOI: 10.2174/09298665113209990050.

[14] Y. Matsuzaki, M. Ohue, N. Uchikoga, and Y. Akiyama, "Protein-protein Interaction Network Prediction by Using Rigid-Body Docking Tools: Application to Bacterial Chemotaxis," *Protein & Peptide Letters*, vol. 21, no. 8, pp. 790–798, 2013, ISSN: 0929-8665. DOI: 10.2174/09298665113209990066.

[15] H. Naveed and J. J. Han, "Structure-based protein-protein interaction networks and drug design," *Quantitative Biology*, vol. 1, no. 3, pp. 183–191, 2013, ISSN: 2095-4689. DOI: 10.1007/s40484-013-0018-y.

[16] P. F. Gherardini and M. Helmer-Citterich, "Structure-based function prediction: approaches and applications," *Briefings in Functional Genomics*, vol. 7, no. 4, pp. 291–302, 2008, ISSN: 2041-2649. DOI: 10.1093/bfgp/eln030.

[17] T. I. Zarembinski, L.-W. Hung, H.-J. Mueller-Dieckmann, K.-K. Kim, H. Yokota, R. Kim, and S.-H. Kim, "Structure-based assignment of the biochemical function of a hypothetical protein: A test case of structural genomics," *Proceedings of the National Academy of Sciences*, vol. 95, no. 26, pp. 15 189–15 193, 1998, ISSN: 0027-8424. DOI: 10.1073/pnas.95.26.15189.

[18] J. D. Watson, R. A. Laskowski, and J. M. Thornton, "Predicting protein function from sequence and structural data," *Current Opinion in Structural Biology*, vol. 15, no. 3, pp. 275–284, 2005, ISSN: 0959-440X. DOI: 10.1016/j.sbi.2005.04.003.

[19] Y. Loewenstein, D. Raimondo, O. C. Redfern, J. Watson, D. Frishman, M. Linial, C. Orengo, J. Thornton, and A. Tramontano, "Protein function annotation by homology-based inference," *Genome Biology*, vol. 10, no. 2, p. 207, 2009, ISSN: 1465-6906. DOI: 10.1186/gb-2009-10-2-207.

[20] R. C. Stevens, "The cost and value of three-dimensional protein structure," *Drug Discovery World*, vol. 4, no. 3, pp. 35–48, 2003.

[21] A. Dove, "Structural biology shapes up," *Science*, vol. 353, no. 6296, pp. 306–308, 2016, ISSN: 0036-8075. DOI: 10.1126/science.353.6296.306.

[22] A. Andreeva, D. Howorth, C. Chothia, E. Kulesha, and A. G. Murzin, "SCOP2 prototype: a new approach to protein structure mining," *Nucleic Acids Research*, vol. 42, no. D1, pp. D310–D314, 2014, ISSN: 0305-1048. DOI: 10.1093/nar/gkt1242.

[23] A. Andreeva, E. Kulesha, J. Gough, and A. G. Murzin, "The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures," *Nucleic Acids Research*, vol. 48, no. D1, pp. D376–D382, 2019, ISSN: 0305-1048. DOI: 10.1093/nar/gkz1064.

[24] F. DiMaio, A. Leaver-Fay, P. Bradley, D. Baker, and I. André, "Modeling Symmetric Macromolecular Structures in Rosetta3," *PLoS ONE*, vol. 6, no. 6, e20450, 2011. DOI: 10.1371/journal.pone.0020450.

[25] D. Xu and Y. Zhang, "Ab initio protein structure assembly using continuous structure fragments and optimized knowledge - based force field," *Proteins: Structure, Function, and Bioinformatics*, vol. 80, no. 7, pp. 1715–1735, 2012, ISSN: 1097-0134. DOI: 10.1002/prot.24065.

[26] Xu, Dong and Zhang, Yang, "Toward optimal fragment generations for ab initio protein structure assembly," *Proteins: Structure, Function, and Bioinformatics*, vol. 81, no. 2, pp. 229–239, 2013, ISSN: 1097-0134. DOI: 10.1002/prot.24179.

[27] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, and D. Hassabis, "Improved protein structure prediction using potentials from deep learning," *Nature*, vol. 577, no. 7792, pp. 706–710, 2020, ISSN: 0028-0836. DOI: 10.1038/s41586-019-1923-7.

[28] Senior, Andrew W. and Evans, Richard and Jumper, John and Kirkpatrick, James and Sifre, Laurent and Green, Tim and Qin, Chongli and Žídek, Augustin and Nelson, Alexander W. R. and Bridgland, Alex and Penedones, Hugo and Petersen, Stig and Simonyan, Karen and Crossan, Steve and Kohli, Pushmeet and Jones, David T. and Silver, David and Kavukcuoglu, Koray and Hassabis, Demis, "Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13)," *Proteins: Structure, Function, and Bioinformatics*, vol. 87, no. 12, pp. 1141–1148, 2019, ISSN: 0887-3585. DOI: 10.1002/prot.25834.

[29] L. A. Abriata, G. E. Tamò, and M. D. Peraro, "A further leap of improvement in tertiary structure prediction in CASP13 prompts new routes for future assessments," *Proteins: Structure, Function, and Bioinformatics*, vol. 87, no. 12, pp. 1100–1112, 2019, ISSN: 0887-3585. DOI: 10.1002/prot.25787.

[30] T. I. Croll, M. D. Sammito, A. Kryshtafovych, and R. J. Read, "Evaluation of template - based modeling in CASP13," *Proteins: Structure, Function, and Bioinformatics*, vol. 87, no. 12, pp. 1113–1127, 2019, ISSN: 0887-3585. DOI: 10.1002/prot.25800.

[31] A. Šali and T. L. Blundell, "Comparative protein modelling by satisfaction of spatial restraints," *Journal of Molecular Biology*, vol. 234, no. 3, pp. 779–815, 1993. DOI: 10.1006/jmbi.1993.1626.

[32] A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. T. Heer, T. A. P. de Beer, C. Rempfer, L. Bordoli, R. Lepore, and T. Schwede, "Swiss-model: Homology modelling of protein structures and complexes," *Nucleic Acids Research*, vol. 46, no. W1, W296–W303, May 2018, ISSN: 0305-1048. DOI: 10.1093/nar/gky427.

[33] W. R. Pearson and D. J. Lipman, "Improved tools for biological sequence comparison," *Proceedings of the National Academy of Sciences*, vol. 85, no. 8, pp. 2444–2448, 1988. DOI: 10.1073/pnas.85.8.2444.

[34] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990. DOI: 10.1016/S0022-2836(05)80360-2.

[35] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, Sep. 1997. DOI: 10.1093/nar/25.17.3389.

[36] G. M. Boratyn, A. A. Schäffer, R. Agarwala, S. F. Altschul, D. J. Lipman, and T. L. Madden, "Domain enhanced lookup time accelerated BLAST," *Biology Direct*, vol. 7, no. 1, p. 12, Apr. 2012. DOI: 10.1186/1745-6150-7-12.

[37] K. Tomii and Y. Akiyama, "FORTE: a profile–profile comparison tool for protein fold recognition," *Bioinformatics*, vol. 20, no. 4, pp. 594–595, Mar. 2004. DOI: 10.1093/bioinformatics/btg474.

[38] D. Xu, L. Jaroszewski, Z. Li, and A. Godzik, "Ffas-3d: Improving fold recognition by including optimized structural features and template re-ranking," *Bioinformatics*, vol. 30, no. 5, pp. 660–667, 2014, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt578.

[39] Y. Yang, E. Faraggi, H. Zhao, and Y. Zhou, "Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates," *Bioinformatics*, vol. 27, no. 15, pp. 2076–2082, Nov. 2011, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr350.

[40] L. Zimmermann, A. Stephens, S.-Z. Nam, D. Rau, J. Kübler, M. Lozajic, F. Gabler, J. Söding, A. N. Lupas, and V. Alva, "A completely reimplemented mpi bioinformatics toolkit with a new hhpred server at its core," *Journal of Molecular Biology*, vol. 430, no. 15, pp. 2237–2243, 2018, Computation Resources for Molecular Biology. DOI: 10.1016/j.jmb.2017.12.007.

[41] K. Karplus, C. Barrett, and R. Hughey, "Hidden Markov models for detecting remote protein homologies.," *Bioinformatics*, vol. 14, no. 10, pp. 846–856, Nov. 1998, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/14.10.846.

[42] R. D. Finn, J. Clements, W. Arndt, B. L. Miller, T. J. Wheeler, F. Schreiber, A. Bateman, and S. R. Eddy, "Hmmer web server: 2015 update," *Nucleic Acids Research*, vol. 43, no. W1, W30–W38, 2015, ISSN: 0305-1048. DOI: 10.1093/nar/gkv397.

[43] A. Hildebrand, M. Remmert, A. Biegert, and J. Söding, "Fast and accurate automatic structure prediction with hhpred," *Proteins: Structure, Function, and Bioinformatics*, vol. 77, no. S9, pp. 128–132, 2009. DOI: 10.1002/prot.22499.

[44] A. Meier and J. Söding, "Automatic Prediction of Protein 3D Structures by Probabilistic Multi-template Homology Modeling," *PLoS Computational Biology*, vol. 11, no. 10, pp. 1–20, 2015. DOI: 10.1371/journal.pcbi.1004343.

[45] L. Holm, "DALI and the persistence of protein shape," *Protein Science*, vol. 29, no. 1, pp. 128–140, 2020, ISSN: 0961-8368. DOI: 10.1002/pro.3749.

[46] V. Alva, S.-Z. Nam, J. Söding, and A. N. Lupas, "The mpi bioinformatics toolkit as an integrative platform for advanced protein sequence and structure analysis," *Nucleic Acids Research*, vol. 44, no. W1, W410–W415, 2016, ISSN: 0305-1048. DOI: 10.1093/nar/gkw348.

[47] Y. Zhang and J. Skolnick, "Scoring function for automated assessment of protein structure template quality," *Proteins: Structure, Function, and Bioinformatics*, vol. 57, no. 4, pp. 702–710, 2004. DOI: 10.1002/prot.20264.

[48] A. Zemla, "Lga: A method for finding 3d similarities in protein structures," *Nucleic acids research*, vol. 31, no. 13, pp. 3370–4, Mar. 2003, ISSN: 0305-1048. DOI: 10.1093/nar/gkg571.

[49] N. Siew, A. Elofsson, L. Rychlewski, and D. Fischer, "Maxsub: An automated measure for the assessment of protein structure prediction quality," *Bioinformatics*, vol. 16, no. 9, pp. 776–785, 2000. DOI: 10.1093/bioinformatics/16.9.776.

[50] J. Kopp, L. Bordoli, J. N. Battey, F. Kiefer, and T. Schwede, "Assessment of casp7 predictions for template-based modeling targets," *Proteins: Structure, Function, and Bioinformatics*, vol. 69, no. S8, pp. 38–56, 2007. DOI: 10.1002/prot.21753.

[51] T. Smith and M. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, 1981. DOI: 10.1016/0022-2836(81)90087-5.

[52] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the National Academy of Sciences*, vol. 89, no. 22, pp. 10 915–10 919, 1992, ISSN: 0027-8424. DOI: 10.1073/pnas.89.22.10915.

[53] J. Söding, "Protein homology detection by HMM-HMM comparison," *Bioinformatics*, vol. 21, no. 7, pp. 951–960, 2005. DOI: 10.1093/bioinformatics/bti125.

[54] R. Cao, D. Bhattacharya, J. Hou, and J. Cheng, "Deepqa: Improving the estimation of single protein model quality with deep belief networks," *BMC Bioinformatics*, vol. 17, no. 1, p. 495, Dec. 2016. DOI: 10.1186/s12859-016-1405-y.

[55] J. Lyons, A. Dehzangi, R. Heffernan, A. Sharma, K. Paliwal, A. Sattar, Y. Zhou, and Y. Yang, "Predicting backbone c_α angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network," *Journal of Computational Chemistry*, vol. 35, no. 28, pp. 2040–2046, 2014. DOI: 10.1002/jcc.23718.

[56] B. Manavalan and J. Lee, "SVMQA: support–vector-machine-based protein single-model quality assessment," *Bioinformatics*, vol. 33, no. 16, pp. 2496–2503, Apr. 2017, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx222.

[57] S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu, "Accurate de novo prediction of protein contact map by ultra-deep learning model," *PLOS Computational Biology*, vol. 13, no. 1, pp. 1–34, Jan. 2017. DOI: 10.1371/journal.pcbi.1005324.

[58] S. Wang, J. Peng, J. Ma, and J. Xu, "Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields," *Scientific Reports*, vol. 6, no. 1, srep18962, Jan. 2016. DOI: 10.1038/srep18962.

[59] L. Wei and Q. Zou, "Recent progress in machine learning-based methods for protein fold recognition," *International Journal of Molecular Sciences*, vol. 17, no. 12, pp. 1–13, 2016. DOI: 10.3390/ijms17122118.

[60] D. Lupyan, A. Leo-Macias, and A. R. Ortiz, "A new progressive-iterative algorithm for multiple structure alignment," *Bioinformatics*, vol. 21, no. 15, pp. 3255–3263, 2005, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti527.

[61] Y. Zhang and J. Skolnick, "TM-align: a protein structure alignment algorithm based on the TM-score," *Nucleic Acids Research*, vol. 33, no. 7, pp. 2302–2309, Jan. 2005. DOI: 10.1093/nar/gki524.

[62] X. Liu, K. Fan, and W. Wang, "The number of protein folds and their distribution over families in nature," *Proteins: Structure, Function, and Bioinformatics*, vol. 54, no. 3, pp. 491–499, 2004, ISSN: 1097-0134. DOI: 10.1002/prot.10514.

[63] J. Park, S. A. Teichmann, T. Hubbard, and C. Chothia, "Intermediate sequences increase the detection of homology between sequences," *Journal of Molecular Biology*, vol. 273, no. 1, pp. 349–354, 1997, ISSN: 0022-2836. DOI: 10.1006/jmbi.1997.1288.

[64] G. D. Schuler, J. A. Epstein, H. Ohkawa, and J. A. Kans, "Entrez: Molecular biology database and retrieval system," in *Computer Methods for Macromolecular Sequence Analysis*, ser. Methods in Enzymology, vol. 266, Academic Press, 1996, pp. 141–162. DOI: 10.1016/S0076-6879(96)66012-1.

[65] A. A. Salamov, M. Suwa, C. A. Orengo, and M. B. Swindells, "Combining sensitive database searches with multiple intermediates to detect distant homologues," *Protein Engineering, Design and Selection*, vol. 12, no. 2, pp. 95–100, Feb. 1999. DOI: 10.1093/protein/12.2.95.

[66] B. Liu, S. Jiang, and Q. Zou, "HITS-PR-HHblits: protein remote homology detection by combining PageRank and Hyperlink-Induced Topic Search," *Briefings in Bioinformatics*, Nov. 2018, ISSN: 1477-4054. DOI: 10.1093/bib/bby104.

[67] P. Pipenbacher, A. Schliep, S. Schneckener, A. Schönhuth, D. Schomburg, and R. Schrader, "ProClust: improved clustering of protein sequences with an extended graph-based approach," *Bioinformatics*, vol. 18, no. suppl_2, S182–S191, Oct. 2002, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/18.supp_2.S182.

[68] J. Weston, A. Elisseeff, D. Zhou, C. S. Leslie, and W. S. Noble, "Protein ranking: From local to global structure in the protein similarity network," *Proceedings of the National Academy of Sciences*, vol. 101, no. 17, pp. 6559–6563, 2004, ISSN: 0027-8424. DOI: 10.1073/pnas.0308067101.

[69] J. Weston, R. Kuang, C. Leslie, and W. S. Noble, "Protein ranking by semi-supervised network propagation," in *BMC bioinformatics*, Springer, vol. 7, 2006, S10. DOI: 10.1186/1471-2105-7-S1-S10.

[70] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, "Ucsf chimera–a visualization system for exploratory research and analysis," *Journal of Computational Chemistry*, vol. 25, no. 13, pp. 1605–1612, 2004. DOI: 10.1002/jcc.20084.

[71] N. K. Fox, S. E. Brenner, and J.-M. Chandonia, "SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures," *Nucleic Acids Research*, vol. 42, no. D1, pp. D304–D309, Dec. 2013, ISSN: 0305-1048. DOI: 10.1093/nar/gkt1240.

[72] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "Scop: A structural classification of proteins database for the investigation of sequences and structures," *Journal of Molecular Biology*, vol. 247, no. 4, pp. 536–540, 1995. DOI: 10.1016/S0022-2836(05)80134-2.

[73] A. Hijikata, K. Yura, T. Noguti, and M. Go, "Revisiting gap locations in amino acid sequence alignments and a proposal for a method to improve them by introducing solvent accessibility," *Proteins: Structure, Function, and Bioinformatics*, vol. 79, no. 6, pp. 1868–1877, 2011. DOI: 10.1002/prot.23011.

[74] L. Rychlewski, W. Li, L. Jaroszewski, and A. Godzik, "Comparison of sequence profiles. strategies for structural predictions using sequence information," *Protein Science*, vol. 9, no. 2, pp. 232–241, 2008. DOI: 10.1110/ps.9.2.232.

[75] J. Xu and Y. Zhang, "How significant is a protein structure similarity with TM-score = 0.5?" *Bioinformatics*, vol. 26, no. 7, pp. 889–895, Feb. 2010. DOI: 10.1093/bioinformatics/btq066.

[76] T. U. Consortium, "UniProt: the universal protein knowledgebase," *Nucleic Acids Research*, vol. 45, no. D1, pp. D158–D169, Nov. 2016. DOI: 10.1093/nar/gkw1099.

[77] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, Jan. 2008. DOI: 10.1007/s10115-007-0114-2.

[78] A. Marchler-Bauer, S. Lu, J. B. Anderson, F. Chitsaz, M. K. Derbyshire, C. DeWeese-Scott, J. H. Fong, L. Y. Geer, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, J. D. Jackson, Z. Ke, C. J. Lanczycki, F. Lu, G. H. Marchler, M. Mullokandov, M. V. Omelchenko, C. L. Robertson, J. S. Song, N. Thanki, R. A. Yamashita, D. Zhang, N. Zhang, C. Zheng, and S. H. Bryant, "CDD: a Conserved Domain Database for the functional annotation of proteins," *Nucleic Acids Research*, vol. 39, no. suppl_1, pp. D225–D229, Nov. 2010. DOI: 10.1093/nar/gkq1189.

[79] M. Mirdita, L. von den Driesch, C. Galiez, M. J. Martin, J. Söding, and M. Steinegger, "Uniclust databases of clustered and deeply annotated protein sequences and alignments," *Nucleic Acids Research*, vol. 45, no. D1, pp. D170–D176, Nov. 2017. DOI: 10.1093/nar/gkw1081.

[80] W. G. Touw, C. Baakman, J. Black, T. A. H. te Beek, E. Krieger, R. P. Joosten, and G. Vriend, "A series of PDB-related databanks for everyday needs," *Nucleic Acids Research*, vol. 43, no. D1, pp. D364–D368, Oct. 2014, ISSN: 0305-1048. DOI: 10.1093/nar/gku1028.

[81] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983. DOI: 10.1002/bip.360221211.

[82] B. John and A. Sali, "Detection of homologous proteins by an intermediate sequence search," *Protein Science*, vol. 13, no. 1, pp. 54–62, 2004. DOI: 10.1110/ps.03335004.

[83] N. R. Coordinators, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research*, vol. 46, no. D1, pp. D8–D13, Nov. 2017, ISSN: 0305-1048. DOI: 10.1093/nar/gkx1095.

[84] K. Katoh, K. Misawa, K.-i. Kuma, and T. Miyata, "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform," *Nucleic Acids Research*, vol. 30, no. 14, pp. 3059–3066, Jul. 2002. DOI: 10.1093/nar/gkf436.

[85] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve.," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982, PMID: 7063747. DOI: 10.1148/radiology.143.1.7063747.

[86] M. Gribskov and N. L. Robinson, "Use of receiver operating characteristic (roc) analysis to evaluate sequence matching," *Computers & Chemistry*, vol. 20, no. 1, pp. 25–33, 1996, ISSN: 0097-8485. DOI: 10.1016/S0097-8485(96)80004-0.

[87] J.-G. Song, J. Kostan, F. Drepper, B. Knapp, E. de Almeida Ribeiro, P. V. Konarev, I. Grishkovskaya, G. Wiche, M. Gregor, D. I. Svergun, B. Warscheid, and K. Djinović-Carugo, "Structural Insights into Ca^{2+}-Calmodulin Regulation of Plectin 1a-Integrin $\beta 4$ Interaction in Hemidesmosomes," *Structure*, vol. 23, pp. 558–570, 2015. DOI: 10.1016/j.str.2015.01.011.

[88] M. Ohue, T. Shimoda, S. Suzuki, Y. Matsuzaki, T. Ishida, and Y. Akiyama, "MEGADOCK 4.0: an ultra–high-performance protein–protein docking software for heterogeneous supercomputers," *Bioinformatics*, vol. 30, pp. 3281–3283, 2014. DOI: 10.1093/bioinformatics/btu532.

[89] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," *VISAPP (1)*, vol. 2, no. 331-340, p. 2, 2009.

[90] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, and K. D. Pruitt, "Reference sequence (refseq) database at ncbi: Current status, taxonomic expansion, and functional annotation," *Nucleic Acids Research*, vol. 44, no. D1, pp. D733–D745, 2016, ISSN: 0305-1048. DOI: 10.1093/nar/gkv1189.

[91] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic acids research*, vol. 28, no. 1, pp. 235–42, 2000, ISSN: 0305-1048. DOI: 10.1093/nar/28.1.235.

[92] L. N. Kinch, W. Li, B. Monastyrskyy, A. Kryshtafovych, and N. V. Grishin, "Evaluation of free modeling targets in casp11 and roll," *Proteins: Structure, Function, and Bioinformatics*, vol. 84, no. S1, pp. 51–66, 2016, ISSN: 1097-0134. DOI: 10.1002/prot.24973.

[93] L. Holm and C. Sander, "Dali: A network tool for protein structure comparison," *Trends in Biochemical Sciences*, vol. 20, no. 11, pp. 478–480, 1995, ISSN: 0968-0004. DOI: 10.1016/S0968-0004(00)89105-7.

[94] I. N. Shindyalov and P. E. Bourne, "Protein structure alignment by incremental combinatorial extension (ce) of the optimal path," *Protein Engineering, Design and Selection*, vol. 11, no. 9, pp. 739–747, 1998, ISSN: 1741-0126. DOI: 10.1093/protein/11.9.739.

[95] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "Cd-hit: Accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, Dec. 2012, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts565.

[96] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson, and D. G. Higgins, "Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega," *Molecular Systems Biology*, vol. 7, no. 1, p. 539, Nov. 2011, ISSN: 1744-4292. DOI: 10.1038/msb.2011.75.

[97] K. Katoh and D. M. Standley, "Mafft multiple sequence alignment software version 7: Improvements in performance and usability," *Molecular Biology and Evolution*, vol. 30, no. 4, pp. 772–780, 2013, ISSN: 0737-4038. DOI: 10.1093/molbev/mst010.

[98] C. Notredame, D. G. Higgins, and J. Heringa, "T-coffee: A novel method for fast and accurate multiple sequence alignment," *Journal of Molecular Biology*, vol. 302, no. 1, pp. 205–217, 2000, ISSN: 0022-2836. DOI: 10.1006/jmbi.2000.4042.

[99] A. Kryshtafovych, A. Barbato, B. Monastyrskyy, K. Fidelis, T. Schwede, and A. Tramontano, "Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in casp11," *Proteins: Structure, Function, and Bioinformatics*, vol. 84, no. S1, pp. 349–369, 2016, ISSN: 1097-0134. DOI: 10.1002/prot.24919.

[100] J. Lundström, L. Rychlewski, J. Bujnicki, and A. Elofsson, "Pcons: A neural-network-based consensus predictor that improves fold recognition," *Protein Science*, vol. 10, no. 11, pp. 2354–2362, Jan. 2001, ISSN: 1469-896X. DOI: 10.1110/ps.08501.

[101] A. Ray, E. Lindahl, and B. Wallner, "Improved model quality assessment using proq2," *BMC Bioinformatics*, vol. 13, no. 1, pp. 1–12, Dec. 2012, ISSN: 1471-2105. DOI: 10.1186/1471-2105-13-224.

[102] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano, "Critical assessment of methods of protein structure prediction: Progress and new directions in round xi," *Proteins: Structure, Function, and Bioinformatics*, vol. 84, no. S1, pp. 4–14, 2016, ISSN: 1097-0134. DOI: 10.1002/prot.25064.

[103] L. N. Kinch, W. Li, R. D. Schaeffer, R. L. Dunbrack, B. Monastyrskyy, A. Kryshtafovych, and N. V. Grishin, "Casp 11 target classification," *Proteins: Structure, Function, and Bioinformatics*, vol. 84, no. S1, pp. 20–33, 2016, ISSN: 1097-0134. DOI: 10.1002/prot.24982.

[104] S. Wu and Y. Zhang, "Lomets: A local meta-threading-server for protein structure prediction," *Nucleic Acids Research*, vol. 35, no. 10, pp. 3375–3382, Jul. 2007, ISSN: 0305-1048. DOI: 10.1093/nar/gkm251.

[105] Wu, Sitao and Zhang, Yang, "Muster: Improving protein sequence profile–profile alignments by using multiple sources of structure information," *Proteins: Structure, Function, and Bioinformatics*, vol. 72, no. 2, pp. 547–556, Aug. 2008, ISSN: 1097-0134. DOI: 10.1002/prot.21945.