Received: 18 July 2017 | Revised: 11 November 2017 | Accepted: 22 November 2017

DOI: 10.1002/prot.25432

RESEARCH ARTICLE

WILEY PROTEINS

Template-based quaternary structure prediction of proteins using enhanced profile–profile alignments

Tsukasa Nakamura[1,2] | Toshiyuki Oda[1] | Yoshinori Fukasawa[1] | Kentaro Tomii[1,2,3,4]

[1]Artificial Intelligence Research Center (AIRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo, 135-0064, Japan

[2]Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba, 277-8562, Japan

[3]Biotechnology Research Institute for Drug Discovery (BRD), National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo, 135-0064, Japan

[4]AIST-Tokyo Tech Real World Big-Data Computation Open Innovation Laboratory (RWBC-OIL), 2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8550, Japan

Correspondence
Kentaro Tomii, Artificial Intelligence Research Center (AIRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan.
Email: k-tomii@aist.go.jp

Funding information
Japan Society for the Promotion of Science, KAKENHI 17J06457; Platform Project for Supporting Drug Discovery and Life Science Research (Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS)) from Japan Agency for Medical Research and Development (AMED)

Abstract

Proteins often exist as their multimeric forms when they function as so-called biological assemblies consisting of the specific number and arrangement of protein subunits. Consequently, elucidating biological assemblies is necessary to improve understanding of protein function. Template-Based Modeling (TBM), based on known protein structures, has been used widely for protein structure prediction. Actually, TBM has become an increasingly useful approach in recent years because of the increased amounts of information related to protein amino acid sequences and three-dimensional structures. An apparently similar situation exists for biological assembly structure prediction as protein complex structures in the PDB increase, although the inference of biological assemblies is not a trivial task. Many methods using TBM, including ours, have been developed for protein structure prediction. Using enhanced profile–profile alignments, we participated in the 12th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP12), as the FONT team (Group # 480). Herein, we present experimental procedures and results of retrospective analyses using our approach for the Quaternary Structure Prediction category of CASP12. We performed profile–profile alignments of several types, based on FORTE, our profile–profile alignment algorithm, to identify suitable templates. Results show that these alignment results enable us to find templates in almost all possible cases. Moreover, we have come to understand the necessity of developing a model selection method that provides improved accuracy. Results also demonstrate that, to some extent, finding templates of protein complexes is useful even for MEDIUM and HARD assembly prediction.

KEYWORDS
biological assembly, community wide experiment, heterooligomers, homooligomers, protein complexes

1 | INTRODUCTION

Many proteins are known to function as complexes. Obtaining information about a quaternary structure, so-called biological assemblies formed using a protein in a living cell, is useful to estimate its function. The biological importance of protein assemblies is greatest, although protein complex structure prediction is still a demanding task when complex structures consisting of close homologous proteins are unavailable.[1] One reason for this difficulty is that quaternary structures are often not conserved during evolution.[2] For instance, regarding homooligomers, different quaternary structures are likely to be strongly associated with their specific functions.[3] However, recently, the amount of information related to the three-dimensional structure of the protein complex increases. Therefore, Template-Based Modeling (TBM), which has been used mainly for predicting the three-dimensional (3D) structures of protein monomers, is increasingly useful for predicting the 3D

Tsukasa Nakamura and Toshiyuki Oda contributed equally to this work.

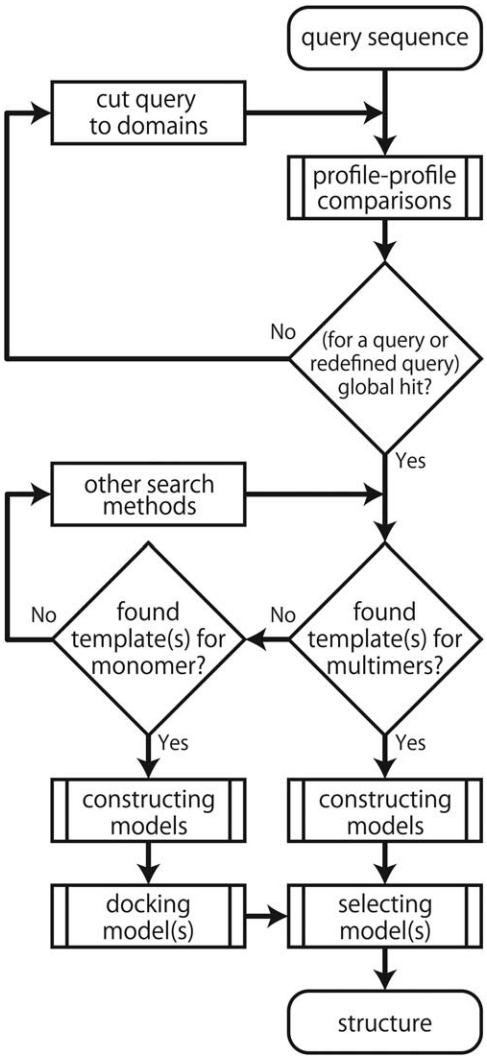

FIGURE 1 Schematic showing our prediction procedure

structures of the protein complexes.[4] Based on a TBM approach using our profile–profile alignment method, we participated in the CASP12 Quaternary Structure Prediction category, which provided a set of diverse protein complexes in terms of the number and form of its constituents. The prediction difficulty of target complexes varies a great deal depending on the availability of templates. Consequently, difficulties of three types, that is, EASY, MEDIUM, and HARD, are applied to the set of target complexes. According to the assessors' definition, there are quaternary structure template(s) for EASY targets, and are partial template(s) or template(s) with no sequence similarity for MEDIUM targets, but no adequate template exists for HARD targets.

The most fundamentally important step of TBM is the stage of template protein identification, for which various methods have been developed. In recent years, the profile–profile alignment method has been recognized as the most powerful method for template identification and for obtaining alignments between target and template proteins. We also developed our own profile–profile alignment method, FORTE,[5] and applied it to predictions of past CASP[6] and CAPRI[7] experiments, and of the TOM complex,[8] which is the translocase of the outer mitochondrial membrane. We have upgraded the method to construct profiles and have improved PSI-BLAST for use in profile construction. For CASP12, we used the revised PSI-BLAST,[9] called PSI-BLASTexB,[10] DELTA-BLAST,[11] and HHblits[12] to construct profiles of both targets and templates. In brief, PSI-BLASTexB is a revised implementation of PSI-BLAST based on the BLAST+ 2.3.0 package. We revised the source code of PSI-BLAST to obtain better PSSM(s) because the original PSI-BLAST was able to produce irregular scores for a gap-rich region. Using these profiles, profile–profile alignments were performed using FORTE. Results showed that these enable us to find templates in almost all possible cases. Nevertheless, we recognized the necessity of developing a model selection method that offers higher accuracy. To some degree, finding templates of a protein complex is useful even for MEDIUM and HARD assembly prediction. Herein, we present the experimental procedure and results of FONT (Group # 480). In addition, we describe retrospective analyses of our approach for the Quaternary Structure Prediction category of CASP12.

2 | MATERIALS AND METHODS

We predicted and constructed protein complexes for multimeric targets in CASP12 based on profile–profile alignment results. A schematic of our prediction procedure is presented in Figure 1. First, we applied template detection and alignment sampling using FORTE, our profile–profile alignment algorithm, with the scoring scheme based on the correlation coefficient between two profile columns.[5] It has been used for past CASP[6] and Critical Assessment of PRedicted Interactions (CAPRI)[7] experiments.

To identify appropriate templates and to obtain alignments between a query sequence and a template sequence, we conducted a series of profile–profile alignments that use sequence profiles of several forms by combining three sets of template libraries, five sequence-retrieval methods, position-specific matrices of two types, and scoring schemes of two types as described below. We developed the methods presented in Table 1 during the prediction season of CASP12. Consequently, some methods have been used only for a part of the set of CASP12 targets (see Supporting Information). For a retrospective analysis for the capability of template identification, we performed all possible types of profile–profile alignments using a partial sequence, corresponding to a domain that we assumed with results of the initial alignments, of a target protein (see below).

2.1 | Sequence retrieval and profile construction

To construct profiles for both a query protein and a template protein, we applied six methods by combining several tools for sequence retrieval and for construction of multiple alignment, and constructed and used position-specific matrices of two types: position-specific scoring matrix (PSSM) and position-specific residue probability (PSRP), as profiles (details are given in Supporting Information).

2.2 | Template libraries

We prepared three datasets as our template libraries for calculating profile–profile alignments. (i) We extracted a representative set of protein

TABLE 1 Summary of methods used for profile construction

Abbreviations	Query	Library	Profile construction (DB, # iterations)
PSI_PSSM	○	(ii)	(TM-align *only for library* +) PSI-BLASTexB (nr, 5)
DB_PSSM	○	(i)	DELTA-BLAST (CDD, 1)
SSM-PSI_PSSM	○(*)	(i)	SSEARCH (nr) + MAFFT + PSI-BLASTexB (nr, 1)
HH-PSI_PSSM	○	N/A	HHblits (up20, 3) + PSI-BLASTexB (nr, 1)
PSI_PSRP	○	(ii)	(TM-align *only for library* +) PSI-BLASTexB (nr, 5)
DB_PSRP	○	(i)	DELTA-BLAST (CDD, 1)
SSM-PSI_PSRP	N/A	(i)	SSEARCH (nr) + MAFFT + PSI-BLASTexB (nr, 1)
HH-PSI_PSRP	○	(i)	HHblits (up20, 3) + PSI-BLASTexB (nr, 1)
HH_PSRP	○	(iii)	HHblits (up20, 3)

The "Profile construction" column shows the methods (, databases, and number of iterations of search methods in parentheses) used in profile construction. "nr" and "CDD" respectively stand for the NCBI nr and conserved domain database. "up20" stands for HH-suite's uniprot20 database. In the "Abbreviations" column, PSI = PSI-BLASTexB, DB = DELTA-BLAST, SSM = SSEARCH + MAFFT, HH = HHblits, PSSM = position specific scoring matrix, PSRP = position specific residue's probability (see the text). In the "Query" column "○"denotes the procedure used in profile construction for query proteins. (*) SSM-PSI_PSSM was not used for constructing query profiles during the CASP12 experiments. Numbers (see Template libraries in the text) in the "Library" column represent the types of template libraries.

chains from PDB[13] using CD-HIT[14,15] (v4.6.3–2015-0515) with the threshold of 98% sequence identity. We used the 47,522 protein chains obtained on 5/10/2016 as template sequences. Those sequences were used for constructing profile libraries using three (A, B, and C; see Supporting Information) out of six sequence retrieval methods. In addition to this template library based on protein chains, we also used the following three libraries to exploit protein domain information. (ii) We generated a representative set of protein domains, removed redundancy by clustering domains with sequence identity of 40% using CD-HIT, based on the domain definition provided by the PDB. In all, we had 46,194 protein domains. The domain definition originates from the updated definition by SCOP[16] or protein domain parser (PDP).[17] We retrieved domain boundary information from the RCSB PDB and generated domain structures using BioJava.[18] To develop reliable profiles, we performed all-against-all structure comparison of 46,194 protein domains, found structurally similar pairs of protein domains, and obtained their pairwise alignments. We applied two criteria for defining similar pairs: (1) *P* values of FatCat[19] allowing 0 twists as .001 or fewer, and (2) TM-score of TM-align[20] that is 0.4 or higher. Pairwise alignments of protein domains satisfying these conditions were calculated using TM-align. Then, using PSI-BLASTexB[10] with NCBI's NRAA (D; see Supporting Information), they were compiled as a seed multiple sequence alignment (MSA) for constructing a profile of each protein domain. Here, the MSAs were obtained by stacking pairwise alignments of structurally similar proteins/domains produced by TM-align. (iii) We also prepared a representative set of protein domains and removed the redundancy by clustering domains with sequence identity of 98% using CD-HIT, based on the domain definition provided by SCOP. We constructed the profile library for these protein domains using HHblits[12] with its uniprot20 database (E; see Supporting Information).

2.3 | Scoring schemes of profile–profile alignment

We used FORTE, our profile–profile alignment algorithm, and used scoring schemes of two types for profile–profile alignments in this study. One is the original scoring scheme of FORTE, based on the correlation coefficient between two profile columns to be compared.[5] The other is the modified scoring scheme using sigmoid transformation of the original one as

$$s'_{ij} = (u-l)/\left(1+\exp\left(-\alpha\left(c_{ij}-t-m_i-m_j\right)\right)\right)+l,$$

where s'_{ij} stands for the modified similarity score for profile columns i and j to be compared, c_{ij} signifies the correlation coefficient for profile columns i and j, corresponding to the original similarity score, u and l respectively denote upper and lower bound to normalize scores, ranging from -1 to 1, and α (for steepness) and t are constants for defining the sigmoid function shape. Here, i represents an arbitrary position of the target profile; j denotes the position of the template profile. m_i and m_j respectively represent the mean values of correlation coefficients of columns i (for all j) and j (for all i). We used this modified score to adjust the abnormally high correlation coefficients in some positions (= columns) because of the poor profile values such as those presented in our study of PSI-BLASTexB.[10] The modified scoring scheme was used for 20 combinations of profile–profile alignments (four methods for query profiles and five methods for library profiles, see Figure 2). In both cases, the Z-scores of alignments were calculated using alignment scores and log-length correction, which is the same as that used by the original FORTE.

2.4 | 3D-model construction, evaluation, and selection

Based on alignments with templates and their Z-scores obtained using the methods described above, we built 3D-models of the target protein complexes using MODELLER[21] and Molecular Operating Environment (MOE)[22] in a case. We constructed 3D-models based on the higher-ranked templates, according to their Z-scores. As templates, we used higher-ranked proteins, in our libraries, registered in the oligomeric states in the PDB. Otherwise, we used close homologues (not in our libraries), which are registered in the oligomeric states in the PDB, of the proteins as templates because we used nonredundant set of

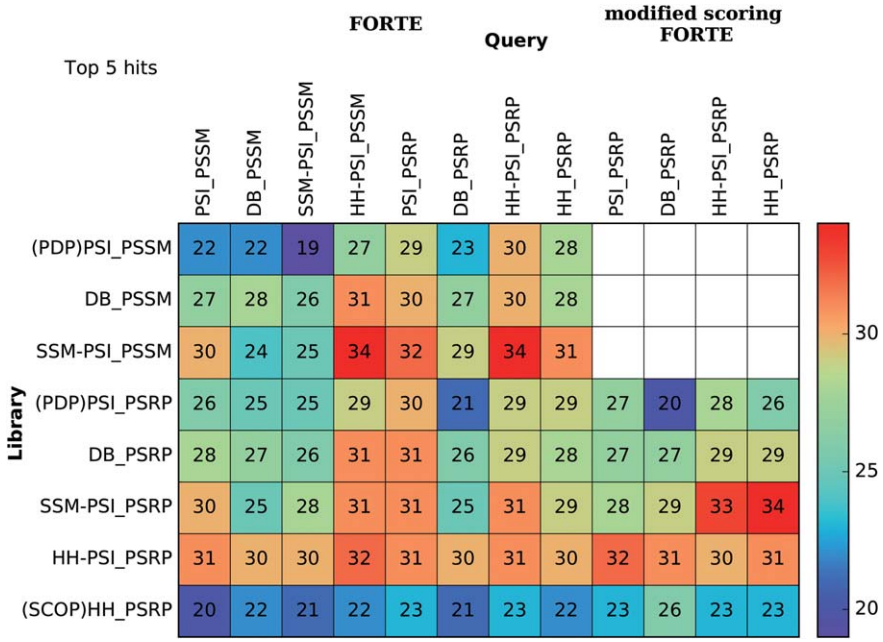

FIGURE 2 Numbers of target domains for which "correct" templates were detected. Each row corresponds to individual template libraries. Each column represents a type of query profile that we used. The modified scoring scheme was used for 20 combinations shown in the four rightmost columns. Numbers in cells show the numbers of target domains for which "correct" templates were detected among the top five hits by each combination. Colors of cells correspond to the numbers of target domains for which "correct" templates were detected. Warmer colors represent larger numbers; colder colors represent smaller numbers. The bar of the coloring schema is shown on the rightmost side

proteins as libraries. Moreover, we constructed 10 3D-models based on an alignment calculated using profile–profile alignments, and sorted the models in terms of the structural quality scores calculated using the Verify3D[23,24] and dDFIRE[25,26] programs. In the model selection step, the constructed models which show low-quality scores of Verify3D were removed. Subsequently, we selected 3D-models with the following criteria: (1) Prioritize templates with higher *Z*-scores, (2) Ranked templates based on results obtained using quality assessment methods. These procedures are executed mostly on an individual subunit basis. Then, to predict three-dimensional protein complex models, we observed oligomeric states of top candidates sorted by their structural quality scores to predict three-dimensional protein complex models. Many cases showed a similar arrangement of oligomeric states among top candidates for each target. We had no clue about oligomeric states for T0913. Therefore, we constructed protein complex models based on an individual subunit model using M-ZDOCK.[27] We usually submitted the model(s) with the highest score(s), but the orders of the submitting models were chosen by human intervention in some cases.

2.5 | Retrospective analysis of template identification

To verify and compare the performance of profile–profile alignment algorithms used for this study, we conducted a retrospective analysis for the capability of template identification. For this analysis, we defined a template with an LGA[28] value of 0.4 or more for a target domain as a "correct" one. This threshold is not so rigorous, but it has been used empirically.[7] Here, for simplicity and clarity, we used sequences of 44 protein domains, based on the CASP assessor definition, of multimeric targets in CASP12 as queries to ascertain whether a "correct" hit is obtained. The 44 domains used here had structurally similar domain(s), in terms of an LGA value of 0.4 or more, in the PDB before the expiration date of the targets. We regarded these 44 domains as those which were predictable using a TBM approach. Therefore, in this analysis, we did not include domains such as T0897-D1, which had no domain(s) with an LGA value of 0.4 or more in the PDB before the expiration date, and which were "true" free-modeling targets.

2.6 | Verification of the effects of profile–profile alignment results on assembly prediction

To elucidate the effects of monomer-based prediction results of profile–profile alignments on assembly prediction, we analyzed similarities between target complexes and template ones identified by profile–profile alignments. For this analysis, we measured the similarity between a target complex and a template one in terms of TM-scores calculated using MM-align,[29] which is an algorithm for structurally aligning multiple-chain protein complexes, and observed relations between TM-scores and Z-scores calculated using profile–profile alignments. TM-score is normalized using a length of the target multimer structure. We specifically examined the top five hits from all possible 84 types of profile–profile alignment methods (see below) as candidate structures.

3 | RESULTS

3.1 | Template identification based on profile–profile alignment results

We conducted a retrospective analysis to verify and compare the performance of profile–profile alignment algorithms used for this study.

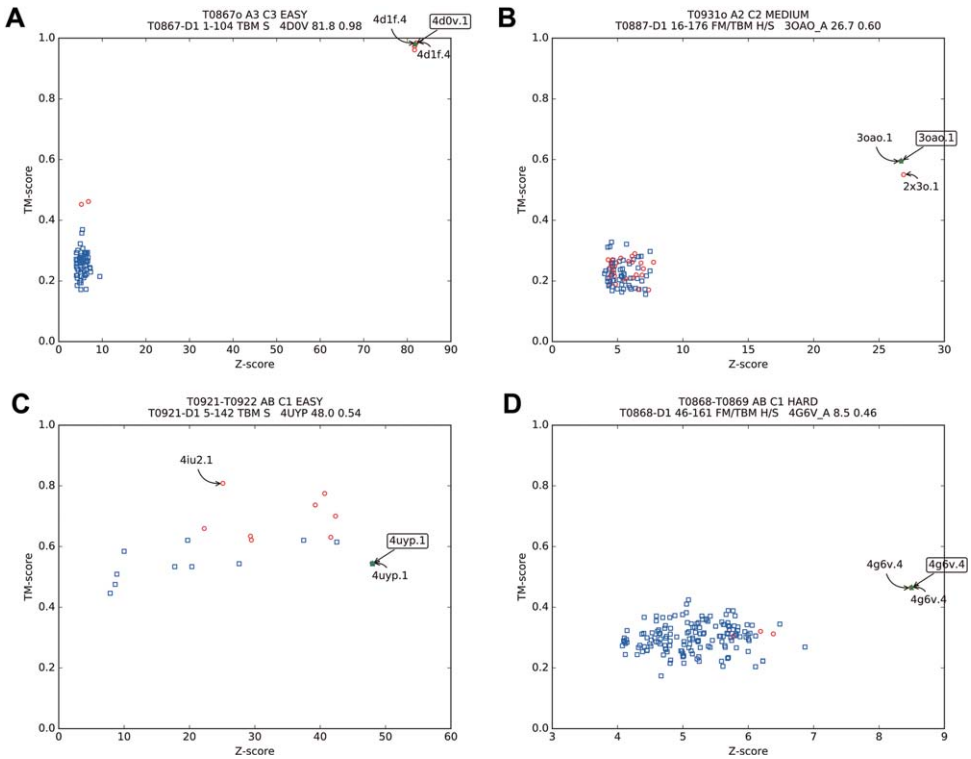

FIGURE 3 Plots of TM-scores vs. the highest Z-scores of templates. The horizontal axis shows Z-score of an alignment between a target domain sequence and a template sequence in PDB. We show the highest Z-score when the same template was identified within the top five hits using different profile–profile alignment methods. The vertical axis shows TM-scores calculated using MMalign between a target complex and a template complex in PDB. The red circle represents a template complex with stoichiometry that is the same as that of the target. Each blue square dot corresponds to a template structure that has different stoichiometry as the target structure. Green star with a rectangle label corresponds to a template structure that we used to construct a model in CASP12. Text above each figure shows the multimer target name, target stoichiometry, target symmetry, and target difficulty in the first line and the target domain name, domain range, domain difficulty classification, target type (Human/Server), template used to construct our model in the CASP term, Z-score of the template used, and the TM-score of the complex template used. Templates given the highest Z-score and the highest TM-score are annotated with a label. The label contains a PDB ID and a number, which represents the number of biological assembly defined in the PDB. We gave 0 for an asymmetric unit

For this analysis, we tested all possible 84 combinations of template libraries, sequence-retrieval methods, types of position-specific matrices, and scoring schemes, and surveyed the top five hits according their Z-scores, for each combination. We did not regard Z-scores of fewer than four as hits, even if they hit within the top five. It is noteworthy that we used only the combinations presented in Table 1, instead of 84 combinations, during the prediction season.

Figure 2 shows the number of target domains for which "correct" templates were detected using profile–profile alignments. Although the results vary in accordance with the combinations of methods, most combinations obtained "correct" hits among the top five hits in >27 (up to 37) cases. Results showed that we were able to detect templates with their LGA $>= 0.4$ for all targets when we consider the top five hits calculated from profile–profile alignments used for this study (Supporting Information Figure S1). This result demonstrates that the ability of the set of profile–profile alignments used for this study to search templates was sufficiently high for finding templates for these 44 domains, which were predictable by TBM, although the domain organization of a target protein was not given when a target sequence was released in CASP. It is noteworthy that most targets are single-domain targets, and that there are noticeable hits, on a domain basis, even for multi-domain targets. Therefore, we can readily recognize domains in a multi-domain target for many cases. It is also worth noting that the protein sequence and structure datasets used here were those before the expiration date of target proteins.

We can observe characteristics of different combinations of methods used for profile–profile alignments, although we realize that this is partly attributable to the difference of entries included in template libraries. According to the number of cases with "correct" hits among the top five hits, the sequence retrieval method C (HHblits + PSI-BLASTexB) for a query sequence is always equal or superior to the method E (HHblits) (see Supporting Information). Comparing results obtained using the two types of scoring scheme of FORTE reveals a slight difference between the original scoring scheme and the modified one. The modified scoring schemes are slightly better than the original one for several combinations of methods of profile construction and template libraries. However, the original scoring scheme is superior to the modified one for the combination of DELTA-BLAST and the template library, according to the number of cases with "correct" hits.

TABLE 2 QS-scores and TM-scores of our first models and baseline for EASY and MEDIUM targets

		QS-score		TM-score	
Target ID	Difficulty category	FONT (1st)	Baseline	MM-align	TM-score
T0861-T0862-T0870	MEDIUM	0.000	0.29	0.469	0.334
T0867	EASY	0.928	0.70	0.982	0.986
T0873	MEDIUM	0.548	0.32	0.484	0.492
T0880	MEDIUM	0.276	0.00	0.590	0.439
T0881	EASY	0.557	0.34	0.809	0.733
T0888	MEDIUM	0.422	0.00	0.820	0.713
T0893	EASY	0.472	0.04	0.419	0.411
T0906	EASY	0.815	0.73	-	-
T0909	EASY	0.391	0.02	0.764	0.359
T0917	EASY	0.658	0.10	0.867	0.860
T0921-T0922	EASY	0.065	0.02	0.655	0.553
T0931	MEDIUM	0.490	0.39	0.514	0.536

QS-scores of the first models of FONT and baseline QS-scores (A. Lafita, personal communication) for EASY and MEDIUM targets are shown. TM-scores, calculated with MM-align and TM-score, of our first models are also shown. Three (T0860, T0889, and T0903-T0904) targets that we missed the opportunity to submit are not shown. The TM-score of our first model for T0906 was not calculable because coordinate data of T0906 were unavailable.

3.2 | Relations between TM-scores and Z-scores

We analyzed relations between TM-scores calculated using MM-align and Z-scores calculated using profile–profile alignment methods to confirm the value of monomer-based prediction results obtained using these assembly prediction methods. For this analysis, we considered all possible permutations of subunit chains within the biological assemblies and also within the asymmetric units for template proteins from the PDB, and employed the highest TM-score obtained with all permutations using MM-align for each template to demonstrate values of top hits as complex templates. Figure 3, which contains typical examples extracted from Supporting Information Figure S2, presents plots of TM-scores of identified templates with the methods versus the highest Z-scores of templates for each target. Although, in total, the relations are not simple but rather complicated, the following lessons can be learnt. i) A prominent hit with the high Z-score indicates a good template for the multimeric form. Some EASY targets such as T0860 and T0889 show this type of distribution. Figure 3A (T0867) presents a typical example of this trend. Even for a MEDIUM target (T0931), this is the case to some extent (Figure 3B). In these cases, we readily decided to select the "correct" complex templates. However, ii) high Z-scores do not always guarantee good templates. This exceptional example is T0945, a HARD target, and this is consistent with the conventional observation that quaternary structures are often not conserved during evolution.[2] Therefore, we need exoteric method(s) or criteria to select adequate templates. In fact, iii) stoichiometry information of proteins can help to select "correct" complex templates. For instance, we were able to use "correct" complex template for an EASY target (T0921-T0922) as shown in Figure 3C if we concentrated on the complexes with the same stoichiometry as the target, although we failed to select "correct" complex template (see below). In addition, we found that iv) even for a prominent hit with the lower Z-scores, we can provide a moderate model based on the TBM approach (Figure 3D; see below).

It is noteworthy that the TM-scores shown here are for the ideal cases, that is, those are values for the "best" target-template alignments. In complex modeling, the quality of the alignment influences the prediction result. To illustrate this point, we show the QS-scores[30] and TM-scores, calculated by MM-align, between our first models and the actual complexes of targets in Table 2. In brief, QS-Score reflects the fraction of correctly modeled interface contacts. In terms of QS-scores, for EASY and MEDIUM targets, we were able to provide better 3D-models of target assemblies than their baseline, which are calculated performances with the QS-Score of top scoring sequence template (top HHSearch hit) by the assessor, except for the T0861-T0862-T0870 assembly and three (T0860, T0889, and T0903-T0904) targets, which we missed the opportunity to submit. To validate those values, we also show their TM-scores calculated by the TM-score,[31] which is also able to compare protein complexes. One can note small differences between an "ideal" TM-score and a TM-score of our first model for each target, especially for an EASY target. This point reflects the accuracy of alignments generated using our profile–profile alignment methods. As described above, our assembly prediction was underpinned strongly by the monomer-based prediction results of profile–profile alignments. Below, we describe what went right and wrong for some examples.

3.3 | Viral fibre head domains (T0880 and T0888)

Five target assemblies of viral fibre heads form homo trimers. Among them, there were two MEDIUM targets (T0880 and T0888) of fibre head trimers. We were able to obtain "correct" complex template(s) for these two Free Modeling (FM) targets among the top five hits (see Supporting Information Figure S1). More precisely, we were able to identify

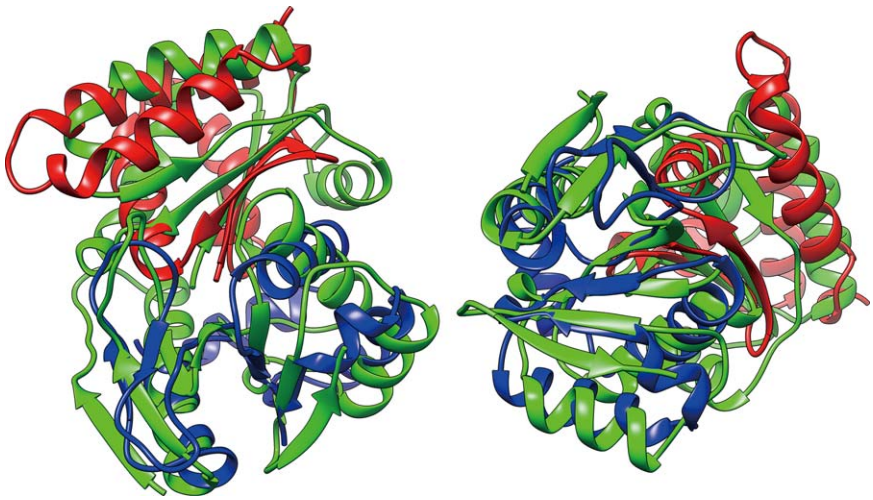

FIGURE 4 Comparison of target and template structures. The template structure (PDB ID: 4G6V[34]; green) was superimposed onto the target (T0868 (blue) and T0869 (red)) structure (PDB ID: 5J4A[35]) using UCSF Chimera.[36] Tentative top (right) and side (left) views are shown. RMSD Cα = 3.12 Å >90 amino acids between 4G6VA and 5J4AA

appropriate templates easily based on the consistent results of many profile–profile alignments for T0880, although our monomer model is partly good (GDT_TS = 63.89) for T0880-D1 and not so good (GDT_TS = 25.16) for T0880-D2. We used 1QIU, which is ranked 15th on the template list at the site of CASP12, as a template for T0880. Consequently, we were able to submit the model with the QS-score of 0.276 for T0880o. For T0888, we found very few similar sequences when we constructed its profiles. At the stage of selecting 3D-models among candidates, we were unable to find "correct" templates because of somewhat vague results of profile–profile alignments, which were attributable mainly to the poor contents of profiles for T0888. However, we were able to find a significant hit against the PDB using jackhmmer[32] and the full-length sequence using the full-length sequence of LAdV2 fibre 2 protein from UniProt.[33] We were able to use 4UE0 as a template. We also used the predicted secondary structure of the query sequence using PSIPRED to align the target sequence to a template. To obtain better alignment(s) between the target and template, we generated 300 alignments. First, we respectively divided the target and template sequences into nine fragments. Each pair of fragments roughly corresponds to a predicted and assigned secondary structure element, respectively, in the target and template proteins. Then we sampled alignments by shifting fragment pairs randomly, maintaining corresponding pairs. We built and evaluated 3D-models based on those alignments generated by shifting the fragment pairs. We submitted a 3D-model with the highest dDFIRE score among models based on 300 alignments. We constructed quaternary structure models and then verified them in the same way as standard procedure. As a result, we were able to submit the model with the QS-score of 0.422 for T0888o.

3.4 | T0868-T0869

For the case of T0868-T0869 (CdiA-CT/CdiI-SU1), a HARD target, we were able to identify a "correct" template, the 4G6V chain A (4G6VA), and construct a 3D-model of T0868 (GDT_TS = 53.02) based on the results of profile–profile alignments, although we failed to select an appropriate template for T0869 using our standard procedure during CASP12. We found, however, we could identify the "correct" template among top hits of several profile–profile alignment methods (see Supporting Information Figure S1). Although we used a poor model for T0869 (GDT_TS = 17.79), we found secondary structure elements similar to the N-terminal regions of both our model and the 4G6 V chain B (4G6VB), which forms a heterodimer with 4G6VA, and hypothesized that the patterns of protein–protein interaction of these proteins might be conserved, especially around the N-terminal regions of T0869. Then, we constructed the model based on the complex of 4G6V using similar secondary structure elements between our T0869 model and 4G6VB. We manually superimposed our T0869 model onto 4G6VB based on this similar arrangement of secondary structure elements. In this case, our TBM approach of protein complex was useful even for a HARD target of the Assembly category. We were able to submit the model with the QS-score of 0.114 for this complex. Indeed, we realized that the rough arrangement and orientation of two subunits have been conserved. Moreover, we infer from comparison of their structures that proteins constituting a heterodimer in 4G6V might be remote homologues of T0868-T0869 (Figure 4), although the topology of both N- and C-terminal regions is different between 4G6V and 5J4A (T0868-T0869).

3.5 | What went wrong

For the problem of T0921-T0922 (Coh5/Doc5), an EASY target, we identified multiple hits with high Z-scores. Among them, 4UYP and 4UYQ had mutually similar molecular arrangements, but they also had a complex structure with different orientation, which corresponds to a dual binding mode of cohesin–dockerin interactions, as shown in a recent study.[37] We were unable to find significant differences of *Z*-scores or structural quality scores for them, although we had 4DH2, which has a similar arrangement and orientation of two subunits with 4UYQ among top candidates. Because we submitted a complex model based on 4UYP, the orientation of subunits of our first model is not

correct (QS-score = 0.065), which indicates that room for improvement exists in selecting models using some novel method(s) other than Verify3D or dDFIRE. However, discerning these two complexes might be difficult because interactions at the interfaces are mutually similar as a result of the structural symmetry of dockerin. As described above, we should consider stoichiometry information of proteins for this target.

For a few HARD targets such as T0913 and T0945, we obtained prominent hits with the high Z-scores. Especially for T0945, we had hits with the same stoichiometry (Supporting Information Figure S2). However, our models are not correct (QS-scores = 0.005 for T0913, and 0.000 for T0945). These results might imply that quaternary structures are often not conserved during evolution.[2] However, the authors of T0945 assigned a monomer as its stoichiometry in PDB (5LEV). We suppose that further analysis should be made for this target.

4 | DISCUSSION

We participated in the first full-fledged Assembly category at CASP12 using enhanced profile–profile alignments. The target complexes have variety in terms of molecular size, symmetry group, and number of subunits in a complex, and reflect the entities in the PDB.

Profile–profile comparison is an effective method for template-based modeling (TBM) because of its power in similarity detection and its alignment accuracy. We performed template-based modeling for CASP12 targets using our updated and enhanced profile–profile comparison method with new profile construction pipelines. Because of an increase in the amount of information related to protein amino acid sequences and structures, TBM has become an extremely useful approach for protein structure prediction. Apparently, it represents a similar situation to that of protein complex structure prediction. As described above, we showed that TBM, based on profile–profile alignment methods, is useful for predicting protein complexes. For EASY and MEDIUM targets, a prominent hit with the high Z-score can indicate a good template, though high Z-scores do not always guarantee good templates. However, additional information about protein stoichiometry can help to select "correct" complex templates. We also acknowledge the necessity of improving the methods to identify "correct" complex templates based on the results of profile–profile alignments, especially for MEDIUM and HARD targets. In addition, we demonstrated the capability of finding similar interactions conserved between remotely related complexes for the case of T0868-T0869. However, we note that, of course, a TBM approach is only applicable to targets that already exist with similar structures in the PDB.

We have performed profile–profile alignments of many types by combining three template libraries, several sequence retrieval methods, position specific matrices of two types, and two scoring schemes for profile–profile comparison of a query profile with profiles in a library. Additionally, we widen the targets of retrospective analysis to 82 protein domains out of a total of 96 protein domains in CASP12. We found that most combinations listed "correct" hits among the top five hits in >50 (up to 65) cases (Supporting Information Figure S3), and that we were able to detect "correct" templates for all targets except one protein, T0918 (consisting of three domains). The 82 protein domains used here had similar protein domains with their LGA >= 0.4 in the PDB before the expiration date. Those results revealed that the use of only four combinations of profile–profile alignments was sufficient to identify "correct" templates for almost all targets, aside from two (T0859 and T0918) out of 82 target domains, when we consider the top five hits for each combination of profile–profile alignments (Supporting Information Figure S4). The two similar sets of four combinations of profile–profile alignments can cover 95% (78 out of 82), that is the highest coverage, of target domains. It is noteworthy that these two sets contain almost the same profile–profile alignment methods. Only a (slight) difference exists between the two sets of combinations, that is, SSM-PSI_PSSM (left) and PSI_PSSM (right). These might imply the superiority of contained methods compared with the other methods. We realized that combining varied but few profile–profile alignments is useful to enhance the capability of identifying a "correct" template(s) for a wide variety of targets. For instance, consideration of the top 13 hits revealed that the combination of profile–profile alignments of only three types was sufficient to identify a "correct" template(s) for almost any target, except for two (T0859 and T0918) (Supporting Information Figure S5). These results suggest that the combination of profile–profile alignment methods facilitates the ability for detecting appropriate templates, and that not using a holistic set of profile–profile alignments, but using a proper set of profile–profile alignments instead, is sufficient to find "correct" template(s) in the sense of template-based modeling.

5 | AVAILABILITY

FORTE and DELTA-FORTE are available for noncommercial use at http://forteprtl.cbrc.jp. PSI-BLASTexB can be downloaded from https://github.com/kyungtaekLIM/PSI-BLASTexB .

ACKNOWLEDGMENTS

The authors thank Aleix M. Lafita and the group of Prof. Guido Capitani (deceased) for their critical assessments. The authors thank Dr. Kyungtaek Lim and Dr. Kenichiro Imai for helpful discussions. The authors also thank Prof. Anna Tramontano (deceased) and all the CASP12 and CAPRI organizers and assessors, the structural biologists who provided targets, NCBI (BLAST) and Soding Lab (HHblits). This research is partially supported by the Platform Project for Supporting Drug Discovery and Life Science Research (Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS)) from the Japan Agency for Medical Research and Development (AMED) and partially supported by JSPS KAKENHI (Grant-in-Aid for JSPS Research Fellow) Grant Number 17J06457. The authors declare that they have no competing interests.

ORCID

Tsukasa Nakamura https://orcid.org/0000-0002-6312-3070
Kentaro Tomii http://orcid.org/0000-0002-4567-4768

REFERENCES

[1] Negroni J, Mosca R, Aloy P. Assessing the applicability of template-based protein docking in the twilight zone. *Structure*. 2014;22(9):1356–1362.

[2] Venkatakrishnan AJ, Levy ED, Teichmann SA. Homomeric protein complexes: evolution and assembly. *Biochem Soc Trans*. 2010;38(4):879–882.

[3] Hashimoto K, Nishi H, Bryant S, Panchenko AR. Caught in self-interaction: evolutionary and functional mechanisms of protein homooligomerization. *Phys Biol*. 2011;8(3):035007.

[4] Szilagyi A, Zhang Y. Template-based structure modeling of protein-protein interactions. *Curr Opin Struct Biol*. 2014;24:10–23.

[5] Tomii K, Akiyama Y. FORTE: a profile-profile comparison tool for protein fold recognition. *Bioinformatics*. 2004;20(4):594–595.

[6] Tomii K, Hirokawa T, Motono C. Protein structure prediction using a variety of profile libraries and 3D verification. *Proteins*. 2005;61(Suppl 7):114–121.

[7] Lensink MF, Velankar S, Kryshtafovych A, et al. Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment. *Proteins*. 2016;84(Suppl 1):323–348.

[8] Shiota T, Imai K, Qiu J, et al. Molecular architecture of the active mitochondrial protein gate. *Science*. 2015;349(6255):1544–1548.

[9] Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25(17):3389–3402.

[10] Oda T, Lim K, Tomii K. Simple adjustment of the sequence weight algorithm remarkably enhances PSI-BLAST performance. *BMC Bioinformatics*. 2017;18(1):288.

[11] Boratyn GM, Schaffer AA, Agarwala R, Altschul SF, Lipman DJ, Madden TL. Domain enhanced lookup time accelerated BLAST. *Biol Direct*. 2012;7:12.

[12] Remmert M, Biegert A, Hauser A, Soding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods*. 2011;9(2):173–175.

[13] Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. *Nucleic Acids Res*. 2000;28(1):235–242.

[14] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006;22(13):1658–1659.

[15] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012;28(23):3150–3152.

[16] Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*. 1995;247(4):536–540.

[17] Alexandrov N, Shindyalov I. PDP: protein domain parser. *Bioinformatics*. 2003;19(3):429–430.

[18] Prlic A, Yates A, Bliven SE, et al. BioJava: an open-source framework for bioinformatics in 2012. *Bioinformatics*. 2012;28(20):2693–2695.

[19] Ye Y, Godzik A. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*. 2003;19(Suppl 2):ii246–ii255.

[20] Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*. 2005;33(7):2302–2309.

[21] Webb B, Sali A. Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Bioinformatics*. 2016;54:5.6.1–5.6.37.

[22] Molecular Operating Environment (MOE), 2013.08. Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7; 2017.

[23] Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*. 1991;253(5016):164–170.

[24] Luthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. *Nature*. 1992;356(6364):83–85.

[25] Yang Y, Zhou Y. Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. *Protein Sci*. 2008;17(7):1212–1219.

[26] Yang Y, Zhou Y. Specific interactions for ab initio folding of protein terminal regions with secondary structures. *Proteins*. 2008;72(2):793–803.

[27] Pierce B, Tong W, Weng Z. M-ZDOCK: a grid-based approach for Cn symmetric multimer docking. *Bioinformatics*. 2005;21(8):1472–1478.

[28] Zemla A. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res*. 2003;31(13):3370–3374.

[29] Mukherjee S, Zhang Y. MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. *Nucleic Acids Res*. 2009;37(11):e83.

[30] Bertoni M, Kiefer F, Biasini M, Bordoli L, Schwede T. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Sci Rep*. 2017;7(1):10480.

[31] Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins*. 2004;57(4):702–710.

[32] Johnson LS, Eddy SR, Portugaly E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*. 2010;11:431.

[33] Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'donovan C, Redaschi N, Yeh LS. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*. 2004;32(Database issue):D115–D119.

[34] Morse RP, Nikolakakis KC, Willett JL, et al. Structural basis of toxicity and immunity in contact-dependent growth inhibition (CDI) systems. *Proc Natl Acad Sci U S A*. 2012;109(52):21480–21485.

[35] Johnson PM, Gucinski GC, Garza-Sanchez F, et al. Functional diversity of cytotoxic tRNase/immunity protein complexes from burkholderia pseudomallei. *J Biol Chem*. 2016;291(37):19387–19400.

[36] Pettersen EF, Goddard TD, Huang CC, et al. UCSF Chimera–a visualization system for exploratory research and analysis. *J Comput Chem*. 2004;25(13):1605–1612.

[37] Cameron K, Najmudin S, Alves VD, et al. Cell-surface attachment of bacterial multienzyme complexes involves highly dynamic protein-protein anchors. *J Biol Chem*. 2015;290(21):13578–13590.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Nakamura T, Oda T, Fukasawa Y, Tomii K. Template-based quaternary structure prediction of proteins using enhanced profile–profile alignments. *Proteins*. 2018;86:274–282. https://doi.org/10.1002/prot.25432