博士論文

Development of methods for large-scale bioinformatics analysis of protein sequences and structures and their applications

（蛋白質配列及び立体構造情報の大規模解析手法の開発とその応用）

中村　司

Tsukasa Nakamura

Preface

In this thesis, the contents are based on the following published articles.

Nakamura T, Yamada K D, Tomii K, Katoh K: **Parallelization of MAFFT for large-scale multiple sequence alignments.** *Bioinformatics*, Oxford University Press, 34(14): 2490–2492, 2018.

Nakamura T†, Oda T†, Fukasawa Y, Tomii K: **Template-based quaternary structure prediction of proteins using enhanced profile-profile alignments.** *Proteins: Structure, Function, and Bioinformatics*, Wiley, 86(S1): 274–282, 2017. († co-first author)

Nakamura T, Tomii K: **Different similarity measures effects on protein ligand-binding pocket comparison using a reduced vector representation of pockets.** *Biophysics and Physicobiology,* The Biophysical Society of Japan, 13:139–147, 2016.

Nakamura T, Tomii K: **Protein ligand-binding site comparison by a reduced vector representation derived from multidimensional scaling of generalized description of binding sites.** *Methods*, Elsevier, 93:35–40, 2016.

Abstract

Proteins serve various functions in living cells. In order to understand the functions of proteins, like when, where, with what, how and why they operate, protein structures can help us to gain deeper insight about them. The number of known protein sequences is increasing rapidly, through the widespread use of next-generation sequencing, but the number of known protein structures is only gradually increasing, because of the difficulty of conducting experiments. Because of this situation, the use of computational methods to understand proteins is becoming ever more important.

Many proteins are known to function as complexes. Therefore, predicting quaternary structure is useful for understanding their functions. In the first chapter of this thesis, we present the details of our prediction methods for protein quaternary structure, our prediction results and the retrospective analysis of our prediction method through the 12th community wide experiment on the critical assessment of techniques for protein structure prediction. We used a template-based modeling method to predict quaternary structures and assessed the validity of this method.

In the second chapter, we present the parallelized multiple sequence alignment software, MAFFT-MPI, which we developed, and the results of an efficiency analysis of this software. We also present the results of an accuracy analysis of multiple sequence alignments produced by this software using the prediction of the secondary structure of proteins and the contact prediction of proteins. We demonstrate that we could achieve the highest accuracy among existing methods within a practical timeframe. Furthermore, we reveal that there is no decrease in the accuracy of multiple sequence alignments themselves produced by this software even though the number of sequences increases.

In the third chapter, we present a method for comparing ligand-binding pockets in proteins. In this method, one pocket is represented by one reduced vector. Using our novel representation, the similarity between ligand-binding pockets can be performed efficiently by merely calculating the inner product of about 200-dimensional vectors. The novel method exhibits higher performance in detection of similar binding pockets than metrics currently used in existing alignment-free methods and an accurate alignment-dependent method. We also investigated the effects of modifications in the expansion and revision of edge classes for improving the ability to detect similar binding pockets, using two datasets. The computational times required for calculating the similarity of randomly selected pocket pairs suggests that this novel method can identify similarities faster than the other currently-used methods. Our novel method is expected to be useful for the large-scale comparison of binding pockets to infer the ligands and functions of proteins.

Table of Contents

Chapter 1

The methods for quaternary structure prediction and verification of their effects

1.1 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 (Negroni *et al.*, 2014). One reason for this difficulty is that quaternary structures are often not conserved during evolution (Venkatakrishnan *et al.*, 2010). For instance, regarding homo-oligomers, different quaternary structures are likely to be strongly associated with their specific functions (Hashimoto *et al.*, 2011). 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 structures of the protein complexes (Szilagyi and Zhang, 2014).

Based on a TBM approach using our profile-profile alignment method, we participated in the 12th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP12). During CASP experiment period, participants are provided amino acid sequences and predict their protein structures. These predicted structures are compared with the actual structure which is determined experimentally. Through this experiment, methods for predicting protein structure are assessed in a blinded manner, which means none of the participants know the actual structure when making their predictions.

The Assembly category is an assessment category in CASP12. CASP12 held the first full-fledged Assembly category in CASP history. Along with the popularization of experimental methods which can reveal protein complex structures, like cryo-EM, the importance of predicting complex structures is increasingly being recognized even though a complex structure comprises monomer structures which are also unknow. We achieved 1st place among 25 groups participated in this category. There are other assessment categories, such as the Contact Prediction category assess methods that predict three-dimensional contacts in protein structure and the Refinement category that assesses how well methods can refine a provided starting structure. The targets in the Assembly category consist of amino acid sequences derived from diverse protein complexes in terms of the number and form of their 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 (Tomii and Akiyama, 2004), and applied it to predictions of past CASP (Tomii *et al.*, 2005) and CAPRI (Lensink *et al.*, 2016) experiments, and of the TOM complex (Shiota *et al.*, 2015), 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 (Altschul *et al.*, 1997), called PSI-BLASTexB (Oda *et al.*, 2017), DELTA-BLAST (Boratyn *et al.*, 2012), and HHblits (Remmert *et al.*, 2012) 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 enhanced 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 our group: FONT (Group #480). In addition, we describe retrospective analyses of our approach for the Quaternary Structure Prediction category of CASP12.

1.2 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 1-1. 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 (Tomii and Akiyama, 2004). It has been used for past CASP and Critical Assessment of PRedicted Interactions (CAPRI) 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 1-1 during the prediction season of CASP12. Consequently, some methods have been used only for a part of the set of CASP12 targets. 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.

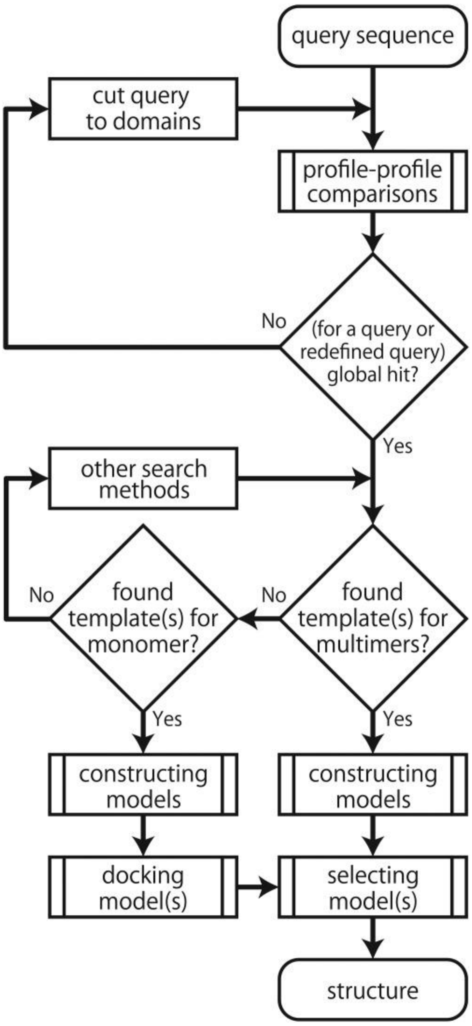


Figure 1-1 Schematic showing our prediction procedure.

Table 1-1 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 1.2.2 Template libraries) in the "Library" column represent the types of template libraries.

1.2.1 Sequence retrieval and profile construction

To construct profiles for both a query protein and a template protein, we applied five methods as described below (A to E) by combining several tools for sequence retrieval and for construction of multiple alignment. Then we constructed using PSI-BLASTexB and used position-specific matrices of two types: position-specific scoring matrix (PSSM) and position-specific residue probability (PSRP), as profiles.

A: SSEARCH + PSI-BLASTexB

In this method, first, we used SSEARCH (36.3.8d) (Pearson, 1996), which is an implementation of the Smith–Waterman algorithm, to obtain similar protein sequences with a novel sensitive matrix we have developed, MIQS (Yamada and Tomii, 2014). We optimized gap penalties against the NCBI's NRAA database downloaded from NCBI-FTP site (ftp://ftp.ncbi.nlm.nih.gov/blast/db/) on 5/12/2016. Then, we constructed a multiple sequence alignment (MSA) using the MAFFT (v7.245) (Katoh and Standley, 2013) automatic selection of alignment strategy using MIQS with the sequences collected in the prior step. When too many (more than 45,000) similar sequences were found in the database, we used CD-HIT (Li and Godzik, 2006; Fu *et al.*, 2012) with the threshold of 95% sequence identity and with a reduced number of sequences in an MSA.

Then, to construct a profile, we conducted a PSI-BLASTexB search against the NCBI's NRAA database using the constructed MSA as a seed MSA with no iteration. For the implementation of PSI-BLASTexB, we have improved the search sensitivity of PSI-BLAST through reduction of the effects of narrow-width blocks on the sequence weight calculation by considering a minimum block width (MBW), as described in our earlier report (Oda *et al.*, 2017). We set MBW as 13 in the implementation. Although this method was used only for proteins in the template libraries during the CASP12 experiment, we also used this method for query sequences (= target domains of CASP12) to test and compare our strategies applied for this study. We used PSSMs and weighted observed residue frequencies at each position as PSRPs, calculated using PSI-BLASTexB, as profiles.

B: DELTA-BLAST

We conducted a DELTA-BLAST search with one iteration against the NCBI's Conserved Domain Database (CDD) downloaded from NCBI-FTP site (ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd/) at 5/6/2016 to construct a profile because the DELTA-BLAST performance often saturates by the first iteration. We used PSSMs produced by DELTA-BLAST as profiles for both target and template proteins. We used this procedure, designated as DELTA-FORTE, for the CAPRI round 30 experiment (Lensink *et al.*, 2016).

We also produced and used weighted observed residue frequencies at each position as profiles for both target and template proteins. Because DELTA-BLAST does not output weighted residue frequencies at each position, as PSI-BLAST does, we calculated the weighted frequencies of an MSA based on the result of DELTA-BLAST. The result of DELTA-BLAST includes pairwise alignments between a query sequence and a representative sequence of domains in the database. We produced an MSA by merging MSAs in CDD of detected domains with the simple pile-up strategy (i.e., MSAs of detected domains are merged using only columns aligned with residues in a query sequence) according to pairwise alignments between a query sequence and a representative sequence of detected domains. When the number of sequences in the merged MSA was 6,000 of more, CD-HIT was used to reduce the number of sequences in the MSA by removing redundancy with the threshold of 90% sequence identity. When the number of sequences remained 6,000 and more later in the procedure, CD-HIT was used again with the threshold of 80% sequence identity.

As a PSRP, we calculated the weighted frequencies of the merged MSA using position-specific sequence weights. The sequence weights are computed using the procedure proposed by Henikoff and Henikoff (Henikoff and Henikoff, 1991), although we did not use blocks covering whole alignments but used blocks covering amino acids less than 11aa distant from the positions of interest. Gaps were regarded as the 21st amino acids.

C: HHblits + PSI-BLASTexB

For this method, we first performed an HHblits (2.0.15) search to find similar protein sequences and to obtain an MSA using them, against its uniprot20 database (uniprot20_2016_02) downloaded from the HH-suite site (http://wwwuser.gwdg.de/~compbiol/data/hhsuite/databases/hhsuite_dbs/) with the default setting, except for option "-n 3". In cases for which we found 6,000 and more similar sequences, we used CD-HIT with the threshold of 90% sequence identity with reduction of the number of sequences in the MSA.

Then, to construct a profile, we performed a PSI-BLASTexB search against the NCBI's NRAA database using the constructed MSA as a seed MSA with no iteration, as described above. We used this method to obtain profiles for both target and template proteins. As PSRPs, we used both PSSMs and weighted observed residue frequencies at each position, calculated using PSI-BLASTexB.

D: PSI-BLASTexB

To construct a profile only for target proteins, we also performed a PSI-BLASTexB search against the NCBI's NRAA database with (up to) five iterations. As PSRPs, we used profiles of two types, PSSMs and weighted observed residue frequencies at each position, calculated using PSI-BLASTexB.

E: HHblits

To obtain weighted residue frequencies for each residue type at each position as PSRPs, we also conducted an HHblits search against the uniprot20 database with three iterations, or until convergence in fewer than three iterations. The weighted frequencies were calculated in the manner described for method B based on the MSA produced by HHblits, as described for method C. We prepared PSRPs for target proteins and the template library (iii) (see Table 1).

1.2.2 Template libraries

We prepared three datasets as our template libraries for calculating profile-profile alignments.

(i) We extracted a representative set of protein chains from PDB (Berman *et al.*, 2000) using CD-HIT (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) out of five sequence retrieval methods.

In addition to this template library based on protein chains, we also used the following two 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 (Murzin *et al.*, 1995) or protein domain parser (PDP) (Alexandrov and Shindyalov, 2003). We retrieved domain boundary information from the RCSB PDB and generated domain structures using BioJava (Prlic *et al.*, 2012). 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 (Ye and Godzik, 2003) allowing 0 twists as .001 or fewer, and (2) TM-score of TM-align (Zhang and Skolnick, 2005) that is 0.4 or higher. Pairwise alignments of protein domains satisfying these conditions were calculated using TM-align. Then, using PSI-BLASTexB with NCBI's NRAA (D in the previous section), 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 with its uniprot20 database (E in the previous section).

1.2.3 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. The other is the modified scoring scheme using sigmoid transformation of the original one as

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

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. 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 1-2). 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.

1.2.4 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 (Webb and Sali, 2016) and Molecular Operating Environment (MOE) (Chemical Computing Group ULC, 2017) 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 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 (Bowie *et al.*, 1991; Lüthy *et al.*, 1992) and dDFIRE (Yang and Zhou, 2008a, 2008b) 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 (Pierce *et al.*, 2005). 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.

1.2.5 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 (Zemla, 2003) 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 (Lensink *et al.*, 2016).

Here, for simplicity and clarity, we used sequences of 44 protein domains (see Appendix Figure A1), 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.

1.2.6 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 (Mukherjee and Zhang, 2009), 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.

1.3 Results

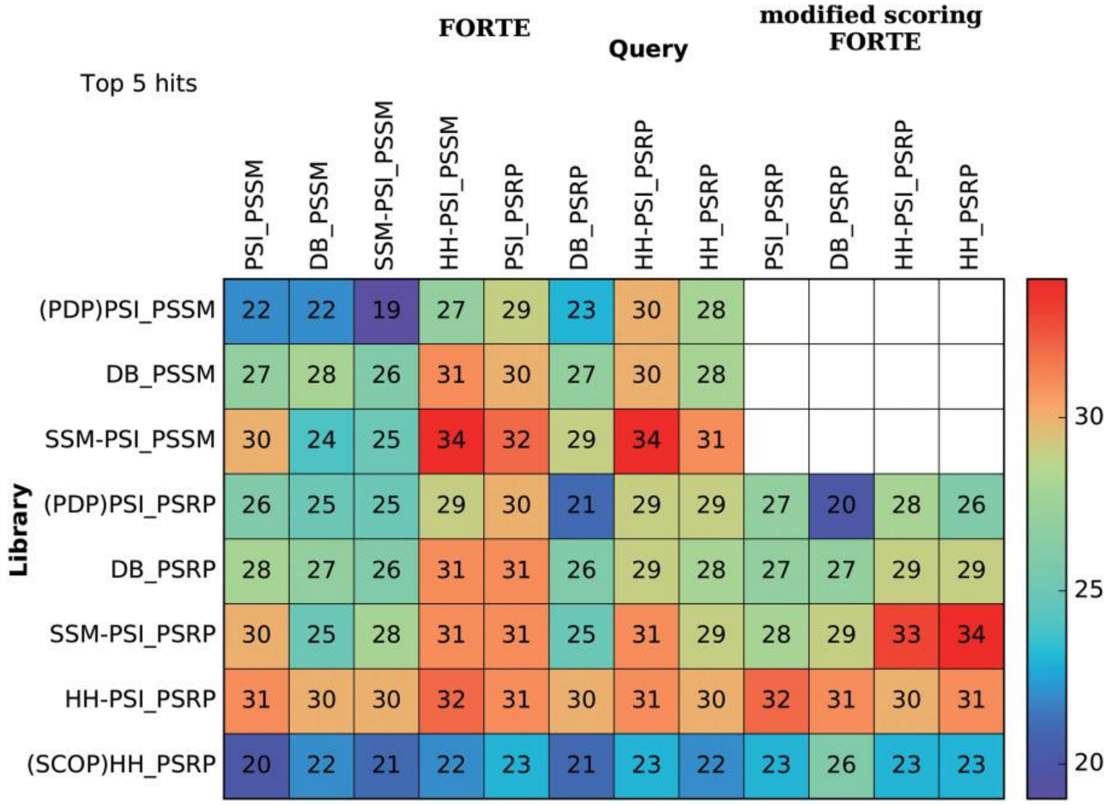
1.3.1 Template identification based on profile-profile alignment results

We conducted a retrospective analysis to verify and compare the performance of profile-profile alignment algorithms used for this study. For this analysis, we tested all possible 84 (= 8x8+5x4) 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 1-1, instead of 84 combinations, during the prediction season.

Figure 1-2 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 34) 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 (Appendix Figure A1). 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.

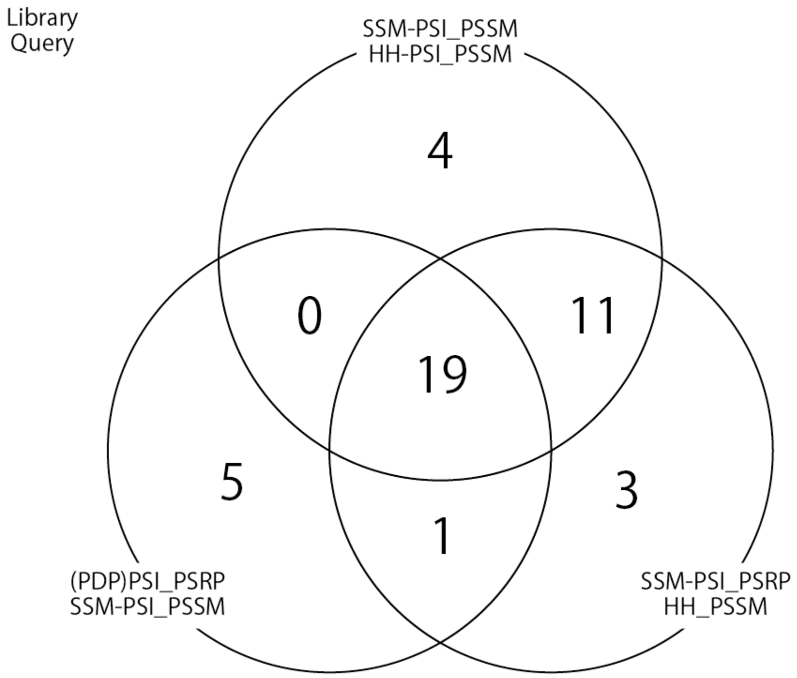
We can observe characteristics of different combinations of methods used for profile-profile alignments, although we realize that this is partly attributable to the difference of entries included in template libraries. According to the number of cases with “correct” hits among the top five hits, the sequence retrieval method C (HHblits + PSI-BLASTexB) for a query sequence is always equal or superior to the method E (HHblits) under the original scoring scheme. 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 (PDP)PSI-BLASTexB and the template library, according to the number of cases with “correct” hits.

The results in Figure 1-2 reveal that the use of only three combinations of profile-profile alignments was sufficient to identify the “correct” templates for almost all targets except for one (T0859-D1) of the 44 domains of multimeric targets when we consider the top five hits which have Z-score 4.0 or more for each combination of profile-profile alignments. The two sets of three combinations of profile-profile alignments can cover 43 of the 44 target domains (Figure 1-3). It is noteworthy that these two sets contain the same profile construction method, namely (PDP)PSI-BLASTexB_PSRP. This might imply the importance of including the profile construction method that has been derived from the protein domain sequences. If we use four combinations of profile-profile alignments, the eight sets of four combinations can cover all 44 domains (see T0929o (= T0859-D1) in Appendix Figure A1). It is also noteworthy that the all eight sets of four combinations include (PDP)PSI-BLASTexB_PSRP.



Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	22	22	19	27	29	23	30	28				
DB_PSSM	27	28	26	31	30	27	30	28				
SSM-PSI_PSSM	30	24	25	34	32	29	34	31				
(PDP)PSI_PSRP	26	25	25	29	30	21	29	29	27	20	28	26
DB_PSRP	28	27	26	31	31	26	29	28	27	27	29	29
SSM-PSI_PSRP	30	25	28	31	31	25	31	29	28	29	33	34
HH-PSI_PSRP	31	30	30	32	31	30	31	30	32	31	30	31
(SCOP)HH_PSRP	20	22	21	22	23	21	23	22	23	26	23	23

Figure 1-2 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.



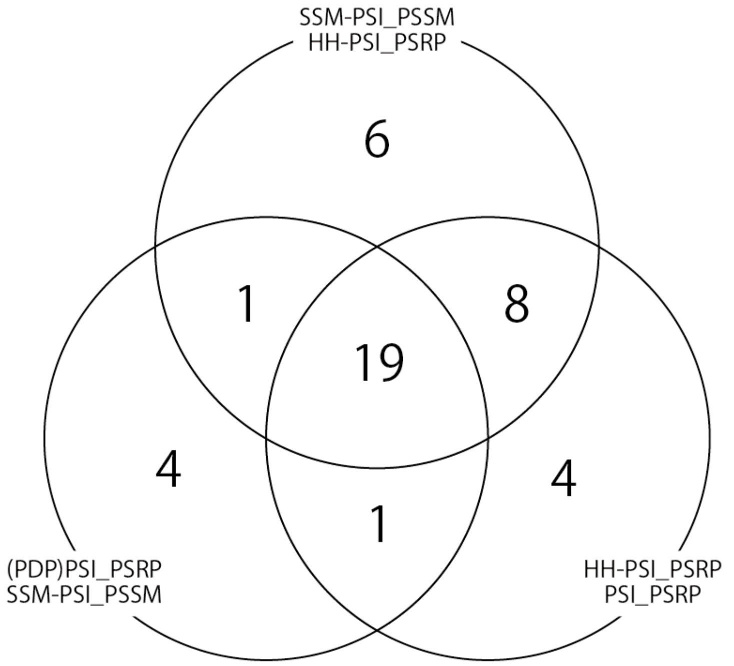


Figure 1-3 Venn diagrams which represents numbers of target domains for which “correct” templates were detected under consideration of the top five hits in the retrospective analysis to 44 protein domains of multimeric targets in CASP12.

1.3.2 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 1-4, which contains typical examples extracted from Appendix Figure A2, 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 1-4A (T0867) presents a typical example of this trend. Even for a MEDIUM target (T0931), this is the case to some extent (Figure 1-4B). 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 (Venkatakrishnan *et al.*, 2010). 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 1-4C if we concentrated on the complexes with the same stoichiometry as the target, although we failed to select "correct" complex template (see 1.3.4 T0868-T0869). 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 1-4D; see 1.3.6 What went wrong).

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 (Bertoni *et al.*, 2017) and TM-scores, calculated by MM-align, between our first models and the actual complexes of targets in Table 1-2. 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 (Zhang and Skolnick, 2004), 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.

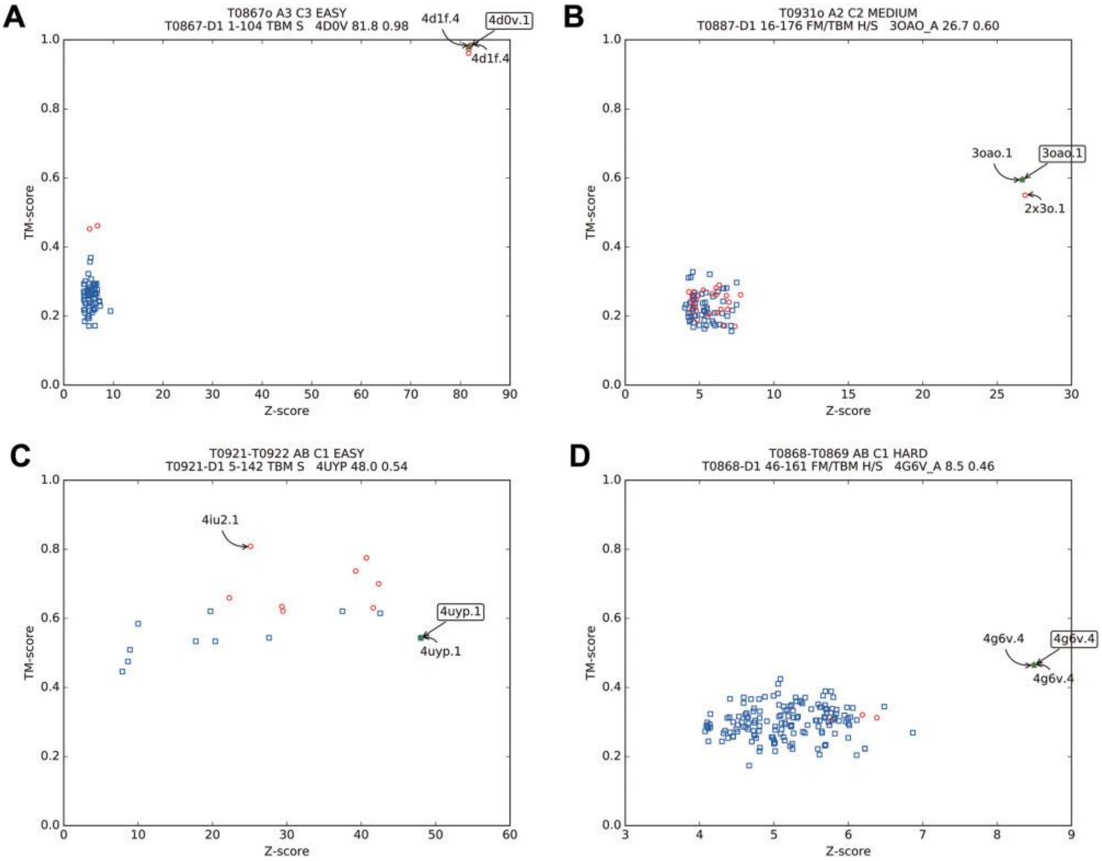


Figure 1-4 Plots of TM-scores vs. the highest Z-scores of templates. The horizontal axis shows Z-score of an alignment between a target domain sequence and a template sequence in PDB. We show the highest Z-score when the same template was identified within the top five hits using different profile-profile 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 serial number of biological assembly defined in the PDB. We gave 0 for an asymmetric unit.

Table 1-2 QS-scores and TM-scores of our first models and baseline for EASY and MEDIUM targets

Target ID	Difficulty category	QS-score		TM-score	
		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

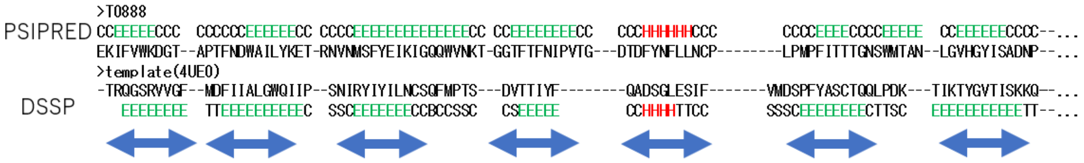
QS-scores of the first models of FONT and baseline QS-scores (A. Lafita, personal communication) for EASY and MEDIUM targets are shown. TM-scores, calculated with MM-align and TM-score, of our first models are also shown. Three (T0860, T0889, and T0903-T0904) targets that we missed the opportunity to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

1.3.3 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 Appendix Figure A1). More precisely, we were able to identify 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 (Johnson *et al.*, 2010) and the full-length sequence using the full-length sequence of LAdV2 fibre 2 protein from UniProt (Apweiler *et al.*, 2004). 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 (Figure 1-5). 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.



```
>T0888
CCEEEEECCC  CCCCCCEEEEEECC CCCCEEEEEEEEEEEEEECC CCEEEEEEEECC   CCCHHHHHHCCC        CCCCEEEECCCCEEEEE  CCEEEEEECCCC--...
EKIFVWKDGT--APTFNDWAILYKET-RNVNMSFYEIKIGQQWVNKT-GGTFTFNIPVTG---DTDFYNFLLNCP--------LPMPFITTTGNSWMTAN--LGVHGYISADNP--...
>template(4UE0)
-TRQGSRVVGF--MDFIIALGWQIIP--SNIRYIYILNCSQFMPTS---DVTTIYF--------QADSGLESIF-------VMDSPFYASCTQQLPDK---TIKTYGVTISKKQ--...
   EEEEEEEE  TTEEEEEEEEEEC  SSCEEEEEEECCBCCSSC   CSEEEEE        CCHHHHTTCC       SSSCEEEEEEEECTTSC   EEEEEEEEEEETT--...
```

Figure 1-5 Schematic diagram how we generated alignments for T0888. The predicted secondary structure of the query sequence using PSIPRED and the secondary structure of the template structure are used to align sequences. We divided into nine fragments (two rest fragments were abbreviated in this figure).

1.3.4 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) (Morse *et al.*, 2012), 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 Appendix Figure A1).

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 4G6V 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 (Figure 1-6).

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 1-7), although the topology of both N- and C-terminal regions is different between 4G6V and 5J4A (T0868-T0869) (Johnson *et al.*, 2016).

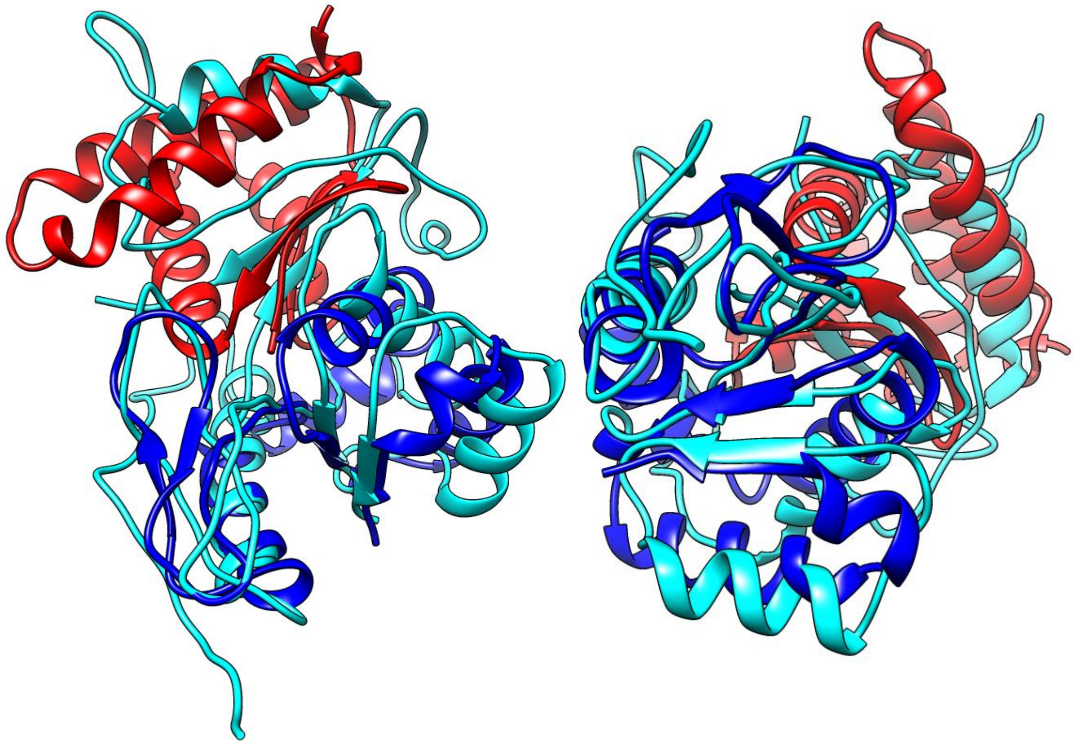


Figure 1-6 Comparison of target and predicted structures. The predicted structure (cyan) was superimposed onto the target (T0868 (blue) and T0869 (red)) structure (PDB ID: 5J4A) using UCSF Chimera. Tentative top (right) and side (left) views are shown.

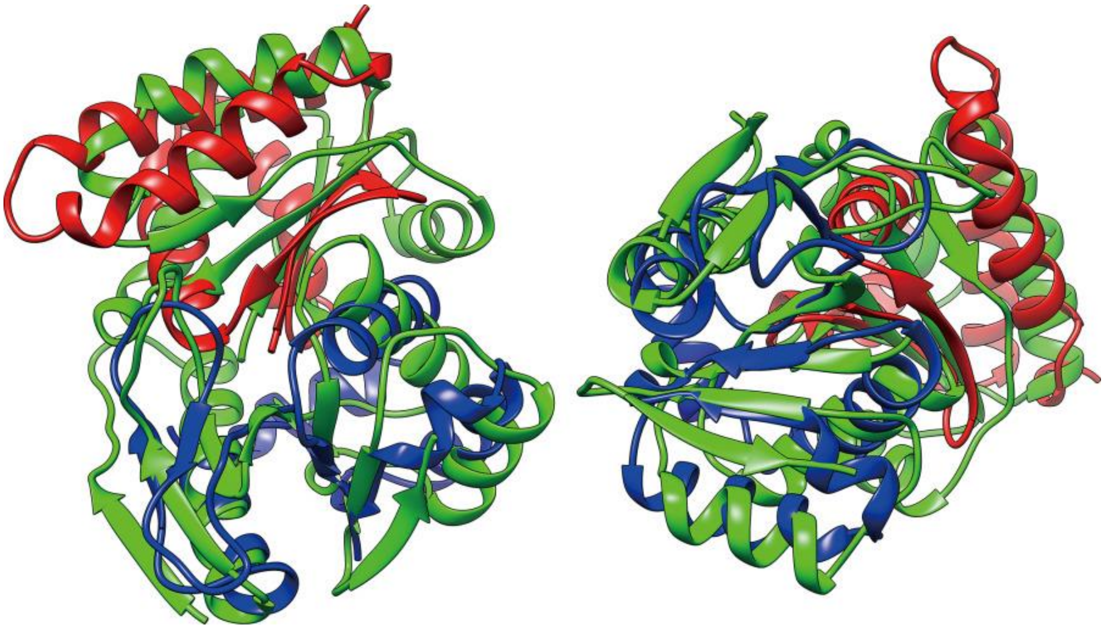


Figure 1-7 Comparison of target and template structures. The template structure (PDB ID: 4G6V; green) was superimposed onto the target (T0868 (blue) and T0869 (red)) structure (PDB ID: 5J4A) using UCSF Chimera. Tentative top (right) and side (left) views are shown. RMSD Cα = 3.12 Å >90 amino acids between 4G6VA and 5J4AA.

1.3.5 Some EASY targets

QS-scores of our first models about targets: T0881, T0893, T0909 and T0917 are shown in Table 1-2. Our models of them are ranked at the 1st, 1st, 1st and 4th places respectively among the first models of all participating groups. However, based on Figure 1-9, it seems that there might still be room for improvement by choosing a "better" template structure. Regarding T0881, Figure 1-9A, we had not chosen 2IUM but instead chose 2VTW. Both have the same stoichiometry (A3) as T0881. We speculate that this is because we could not put a high enough quality score, calculated using Verify3D and dDFIRE, to overturn the slight superiority of 2VTW in Z-score. Regarding T0893, Figure 1-9B,C, because we had not cut the query sequence to domains, we missed the opportunity to select the template, 2C2A. Regarding T0909, Figure 1-9D, we did not choose 3SUC, instead choosing 2X6W. We speculate that this is because we could not decide which to submit, based on quality score and 2X6W had slightly longer alignment length than 3SUC. Then, we tried to overcut the query sequence to the domains, but it did not make us choose 3SUC. Regarding T0917, Figure 1-9E, because only a part of the profile construction methods were used for this target in CASP12 experiment period, we missed 1VLJ.

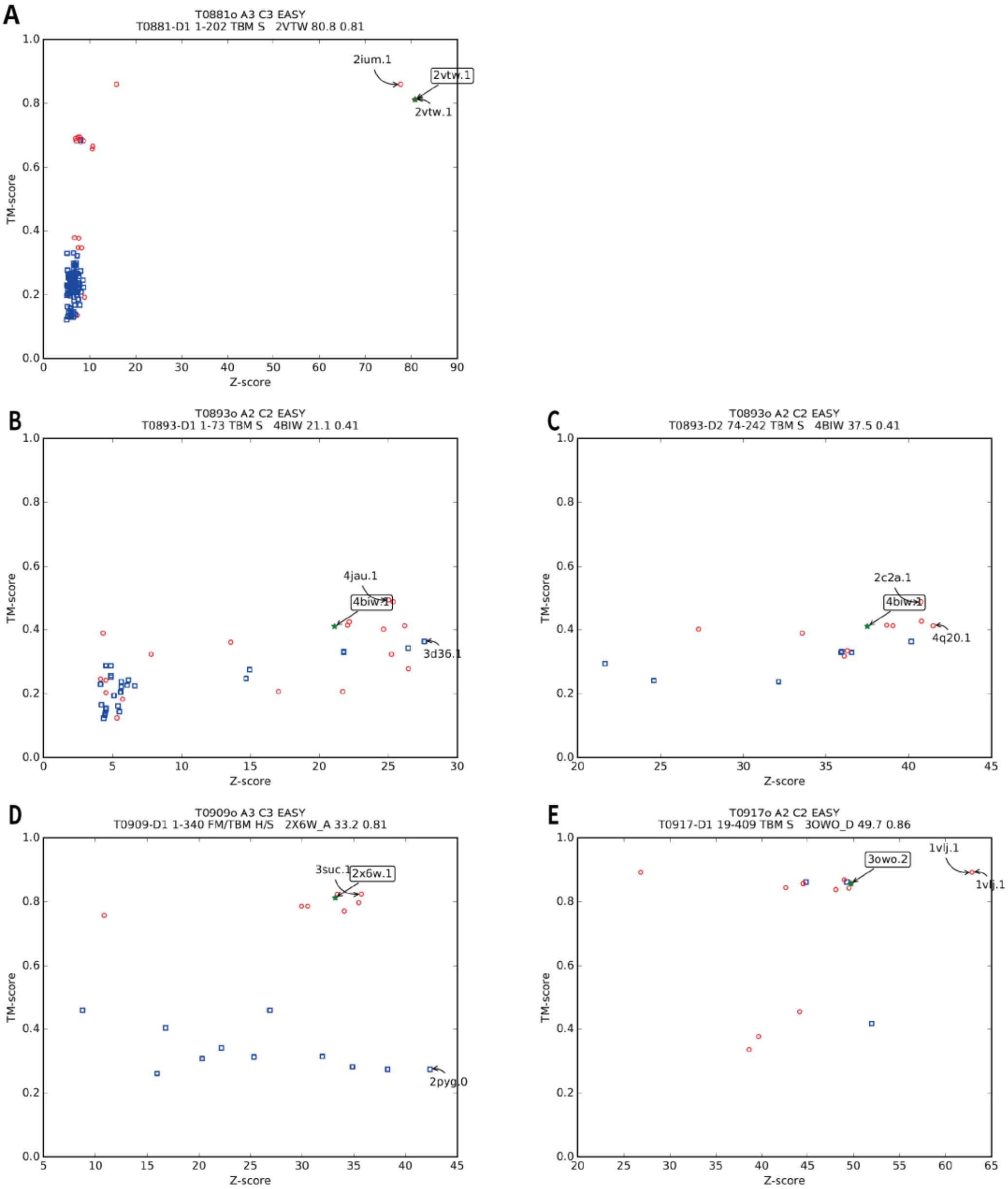


Figure 1-8 Plots of TM-scores vs. the highest Z-scores of templates. (about details, see the footnote of Figure 1-4) (A) T0881, (B-C) T0893, (D) T0909, (E) T0917.

1.3.6 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 (Cameron *et al.*, 2015). 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 (Appendix Figure A2). 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 (Venkatakrishnan *et al.*, 2010). 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.

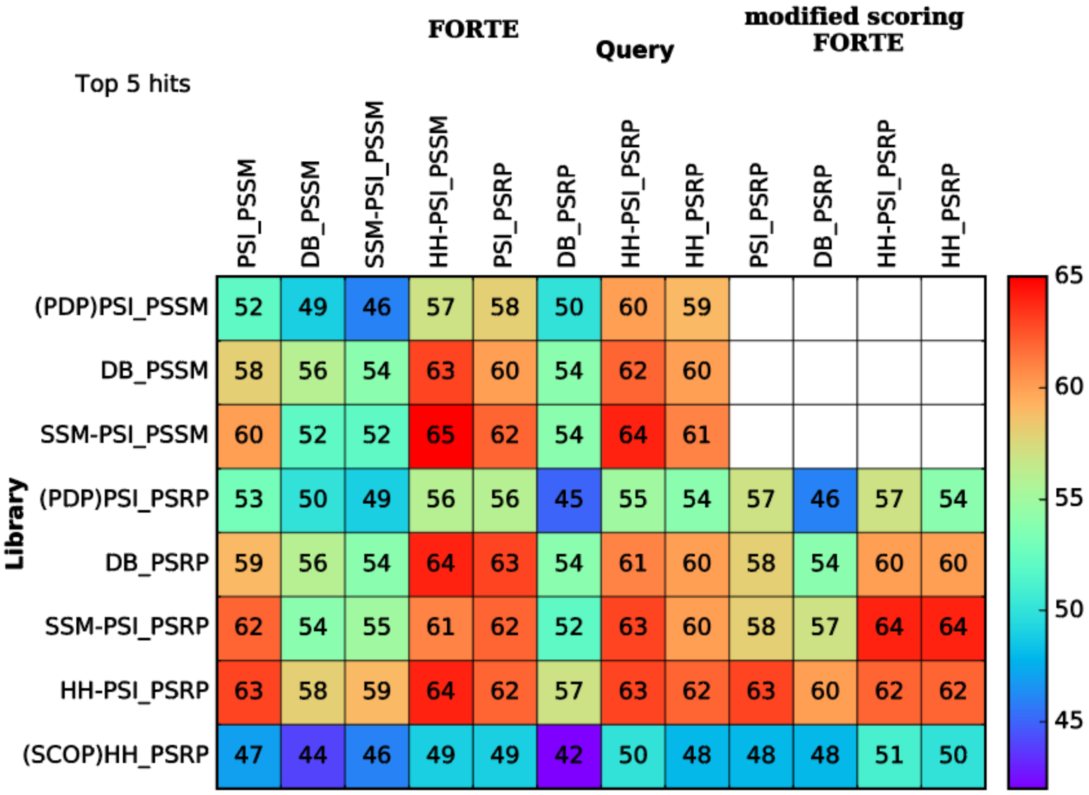
1.4 Discussion

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 (Figure 1-9), 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 (Figure 1-10).

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 (top) and PSI_PSSM (bottom). 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) (Figure 1-11). 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.



	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	52	49	46	57	58	50	60	59				
DB_PSSM	58	56	54	63	60	54	62	60				
SSM-PSI_PSSM	60	52	52	65	62	54	64	61				
(PDP)PSI_PSRP	53	50	49	56	56	45	55	54	57	46	57	54
DB_PSRP	59	56	54	64	63	54	61	60	58	54	60	60
SSM-PSI_PSRP	62	54	55	61	62	52	63	60	58	57	64	64
HH-PSI_PSRP	63	58	59	64	62	57	63	62	63	60	62	62
(SCOP)HH_PSRP	47	44	46	49	49	42	50	48	48	48	51	50

Figure 1-9 Numbers of target domains for which “correct” templates were detected in the retrospective analysis to 82 protein domains out of a total of 96 protein domains in CASP12. 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.

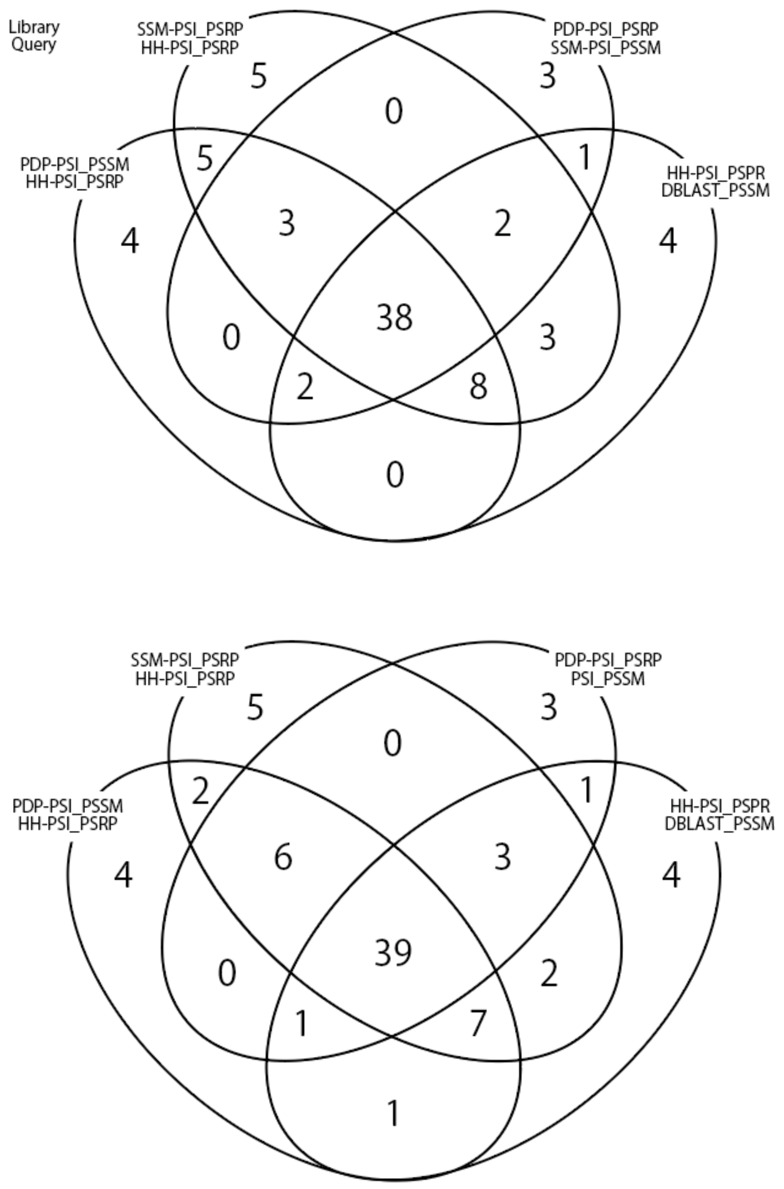


Figure 1-10 Venn diagrams of numbers of target domains for which "correct" templates were detected under consideration of the top five hits in the retrospective analysis to 82 protein domains out of a total of 96 protein domains in CASP12.

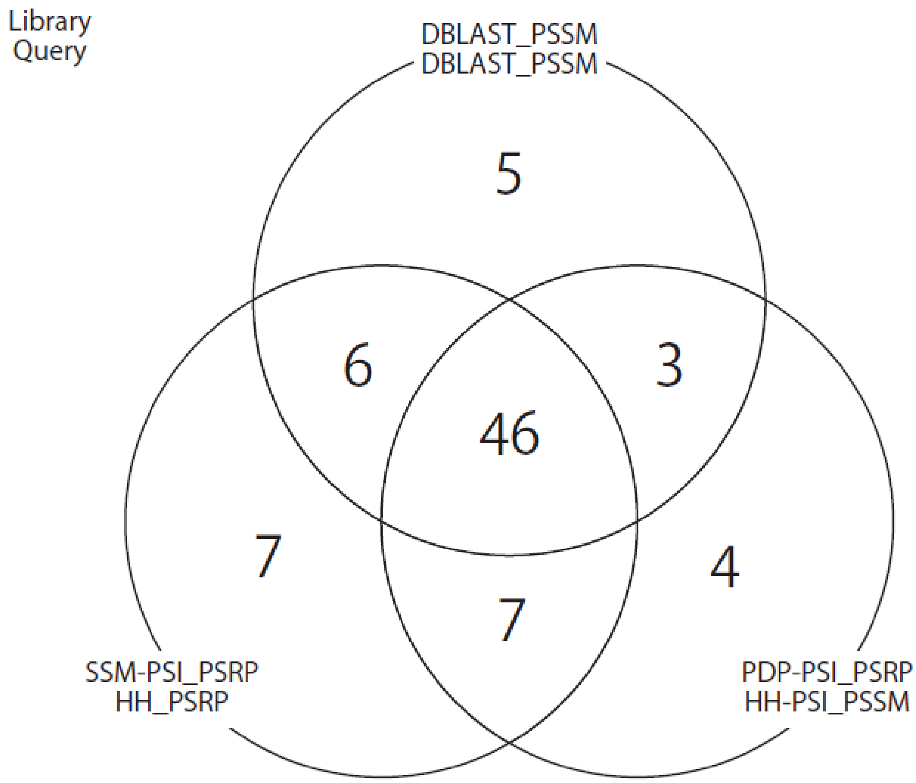


Figure 1-11 A Venn diagram which represents numbers of target domains for which “correct” templates were detected under consideration of the top 13 hits in the retrospective analysis to 82 protein domains out of a total of 96 protein domains in CASP12.

Chapter 2

Large-scale parallelization for construction of MSA and performance comparison with other methods

2.1 Introduction

A large number of biological sequences from widely divergent organisms are becoming available. Accordingly, the need for multiple alignments of large numbers of sequences is increasing for various kinds of sequence analysis. The G-INS-1 option of MAFFT was recently reported to have higher accuracy than other methods for large multiple sequence alignments (MSAs) in independent benchmarks (Le *et al.*, 2017; Yamada *et al.*, 2016). However, this method was impractical for actual analyses, requiring large computational resources in both space and time to perform all-to-all pairwise alignments by dynamic programming (DP) (Needleman and Wunsch, 1970), which are used for a guide tree and a scoring function similar to COFFEE (Notredame *et al.*, 1998).

Here, we introduce a scalable variant, G-large-INS-1, which has equivalent accuracy to G-INS-1 and is applicable to 50000 or more sequences. Our strategies to reduce computational costs are (i) parallelization across multiple machines and/or processor cores using MPI and Pthreads to increase speed and (ii) the use of a high-speed shared filesystem, which is becoming common for processing big data. An MPI-based parallelization of another high-accuracy MSA method, MSAProbs, was recently released (Gonzalez-Dominguez *et al.*, 2016), but it cannot be applied to thousands of sequences. The present update of MAFFT is designed to satisfy the need for accurately aligning large numbers of sequences but is not applicable to long genomic sequences since the length dependence of the computational cost is unchanged. The G-large-INS-1 option is available in MAFFT versions 7.355 or later and the online service (Katoh *et al.*, 2017).

2.2 Materials and methods

2.2.1 Pairwise alignments

In both G-INS-1 and G-large-INS-1, all-to-all pairwise alignments are computed with DP. In G-INS-1, those alignments are computed by multiple processor cores in a single machine using POSIX threads (Pthreads) and stored in RAM. In G-large-INS-1, Message Passing Interface (MPI) can be used to distribute the tasks to multiple processes on different machines (referred to as the MPI version hereafter) and the resulting alignments are stored in temporary files on a filesystem shared by the machines. The pairwise alignments are used to build a guide tree and for objective score in the subsequent progressive alignment step. For this step, in the current implementation, G-large-INS-1 uses multiple cores in a single machine to load the temporary files. Variants with pairwise local alignments (L-INS-1 and L-large-INS-1) are also available and are expected to work better for sequences with long flanking regions with no homology. However, we used only MSA problems with global homology in this research, and thus the difference in accuracy due to different pairwise alignment algorithms (G- or L-) was small here.

2.2.2 Guide tree

MAFFT uses a guide tree with a UPGMA-like method by default. In this method, when merging two clusters, the distance between the new merged cluster and another cluster is set to a weighted average of the average distance and the minimum distance of sequence pairs between the new merged cluster and another cluster, as noted in (Yamada *et al.*, 2016). Instead, G-large-INS-1 uses the stepwise addition strategy, which is often used to build an initial tree in phylogeny inference programs. There is no guarantee that the distance between the merged cluster and another cluster is the minimum or average one. The resulting tree also depends highly on input order.

2.2.3 Pthreads version

For small-scale shared-memory systems with up to approximately 20 cores, a Pthreads version of G-large-INS-1 (without MPI) is also available. For larger systems with more cores, even with shared memory, the MPI version has a higher efficiency for technical reasons. A hybrid mode with both MPI and Pthreads is also selectable as necessary. The calculation of these two versions is identical, apart from the methods for parallelization.

2.3 Results

2.3.1 Performance comparison with other methods

Accuracy of G-large-INS-1 was compared with that of conventional G-INS-1 using different benchmarks, QuanTest (see 2.3.4) (Le *et al.*, 2017) (Figure 2-1a), HomFam (Sievers *et al.*, 2011), OXFam (Raghava *et al.*, 2003; Yamada *et al.*, 2016) and ContTest (Fox *et al.*, 2016) (Table 2-1). Both methods ran with different input orders and/or minor variations in pairwise alignment and guide tree in order to assess instability of accuracy scores (Boyce *et al.*, 2015). In all cases, the difference between G-large-INS-1 (red lines in Figure 2-1a) and G-INS-1 (blue lines) was small.

We also confirmed that G-large-INS-1 and G-INS-1 outperformed the other methods: MSAProbs-MPI; Clustal Omega; Kalign; MAFFT-FFT-NS-2; and Muscle, when the number of sequences is large (200 or more), as expected, based on the original QuanTest results, presented in Figure 1 of (Le *et al.*, 2017). Also, G-large-INS-1 and G-INS-1 slightly outperformed QuickProbs2, when the number of sequences is much larger (1000 or more).

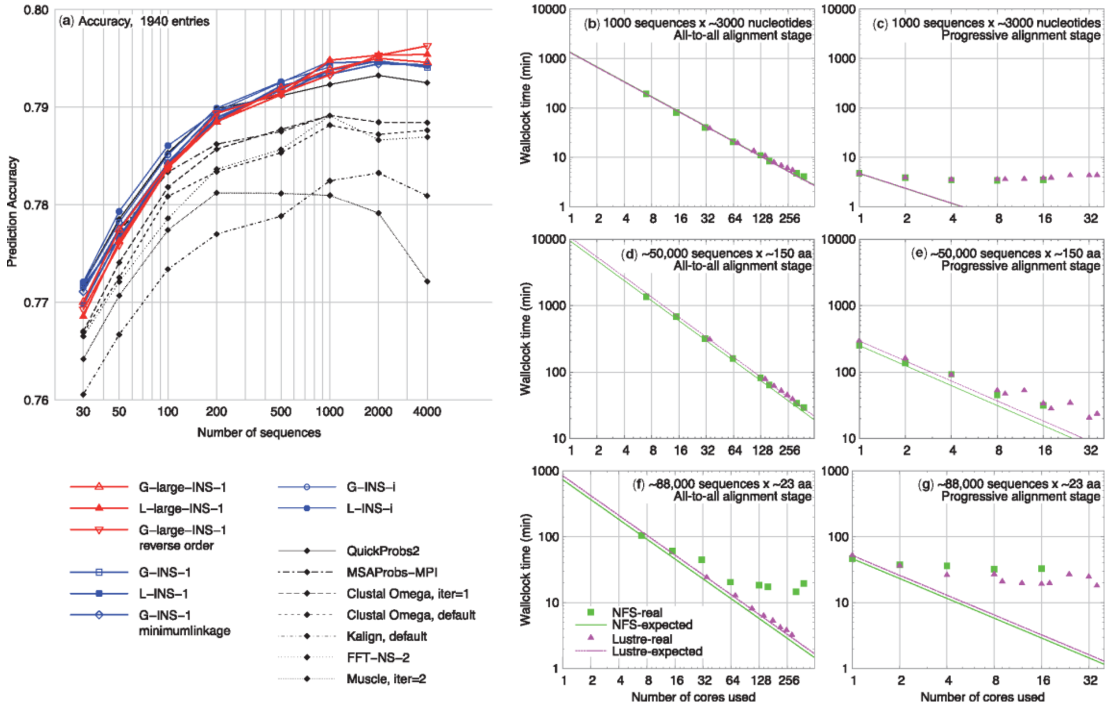


Figure 2-1 (a) QuanTest. Accuracy of protein secondary structure prediction based on various sizes of MSAs by G-large-INS-1 (red bold lines), G-INS-1 (version 7.245; blue bold lines) and other popular methods. We used 1940 (out of 2265) entries so that JPred4 can be consistently applied to the MSAs by all methods. (b)–(g), Parallelization efficiency of all-to-all alignment stage (b, d and f) and progressive stage (c, e and g) when applying G-large-INS-1 to LSU rRNA (b, c) sdr (d, e) and zf-CCHH (f, g). Green squares and magenta triangles are the computational time on NFS and Lustre filesystem, respectively. Lines are the expected time based on the cases using seven cores [NFS; green solid lines in (b), (d) and (f)], 35 cores [Lustre; magenta dotted lines in (b), (d) and (f)] and single core (c, e and g), assuming a perfect efficiency. The calculations with NFS (green) were performed on a heterogeneous cluster system (each node has 16–20 cores of Intel Xeon E5-2660 v3 2.6 GHz, E5-2680 2.7 GHz and E5-2670 v2 2.50 GHz with 64–128GB RAM). The calculations with the Lustre filesystem (magenta) were performed on Intel Xeon E5-2695 v4 2.10 GHz 36 cores with 256GB RAM per node using Lustre version 2.5.42.

2.3.2 Computational cost

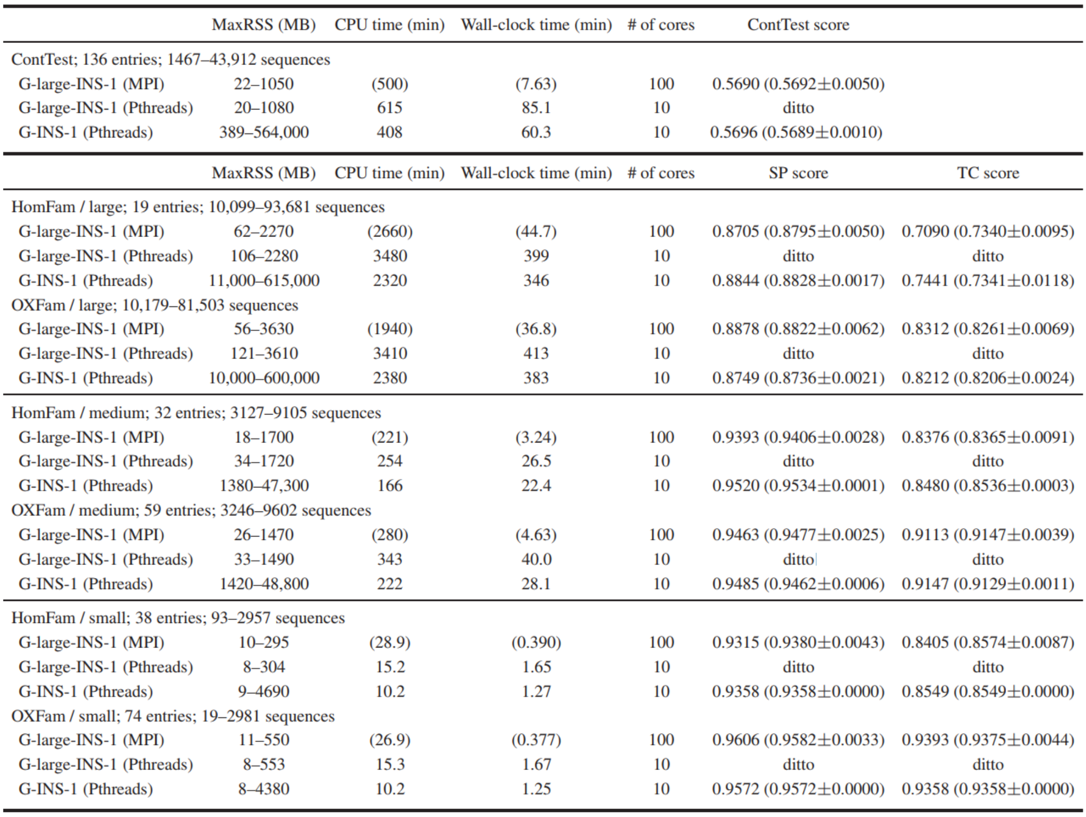
Large amounts of RAM are required if conventional tools for high-quality MSAs are applied to a large number of sequences. For example, MAFFT-L-INS-i and MSAProbs-MPI used at most 9.23GB and 74.8GB for a subset of 1000 sequences in QuanTest. For a larger subset (4000 sequences), MAFFT-G-INS-1 and QuickProbs2 (Gudys and Deorowicz, 2017) used at most 26.0 GB and 411 GB RAM, respectively. In contrast, G-large-INS-1 used only 5.72GB at most, for the subset of 4000 sequences. Memory usage for larger problems (up to ~90 000 sequences) is shown in Table 2-1, which suggests that this advantage increases with the number of sequences. Note that G-large-INS-1 uses files to save temporary data and thus requires a high-speed filesystem when the input sequences are very short, as discussed below.

Parallelization efficiency in three examples is shown in Figure 2-1(b–g), separately for two stages: (i) the all-to-all alignment stage (b, d and f) and (ii) the progressive alignment stage (c, e and g).

For LSU rRNA sequences (b, 1521–4102 bases, 1000 sequences randomly selected from the SEED alignment in Silva (Gloeckner *et al.*, 2017) and protein sequences with usual lengths (d, 21–297 amino acids, 50157 sequences, the 'sdr' family taken from HomFam), the wall-clock time for the all-to-all alignment stage decreased almost linearly with the number of cores used for the calculation. However, for a dataset with very short sequences (f, 12–35 amino acids, 88345 sequences, the 'zf-CCHH' family taken from HomFam), the efficiency differs depending on filesystem: high in Lustre (shown with magenta triangles) but low in NFS (shown with green squares). This difference is due to the balance between calculation and disk operations. As noted earlier, a considerable amount of temporary data is written in parallel into the filesystem: approximately 218 MB, 100 GB and 142 GB for LSU rRNA, 'sdr' and 'zf-CCHH', respectively, in the examples shown here. Overhead due to these disk operations is almost negligible in the former two cases but not in the latter case, where alignment of ~23 amino acids takes only a short time in comparison with the time to write the temporary data to disk using NFS.

Figure 2-1c, e and g suggest that the wall-clock time of the progressive stage varies for each run and does not linearly decrease, but usually this is not a speed-limiting step. CPU time and wall-clock time for various problems are shown in Table 2-1.

Table 2-1 Comparison of computational costs (RAM usage, CPU time and wall-clock time) and accuracy scores of the G-INS-1 (version 7.291) and G-large-INS-1 options using HomFam, OXFam and ContTest



	MaxRSS (MB)	CPU time (min)	Wall-clock time (min)	# of cores	ContTest score	
ContTest; 136 entries; 1467–43,912 sequences						
G-large-INS-1 (MPI)	22–1050	(500)	(7.63)	100	0.5690 (0.5692±0.0050)	
G-large-INS-1 (Pthreads)	20–1080	615	85.1	10	ditto	
G-INS-1 (Pthreads)	389–564,000	408	60.3	10	0.5696 (0.5689±0.0010)	
	MaxRSS (MB)	CPU time (min)	Wall-clock time (min)	# of cores	SP score	TC score
HomFam / large; 19 entries; 10,099–93,681 sequences						
G-large-INS-1 (MPI)	62–2270	(2660)	(44.7)	100	0.8705 (0.8795±0.0050)	0.7090 (0.7340±0.0095)
G-large-INS-1 (Pthreads)	106–2280	3480	399	10	ditto	ditto
G-INS-1 (Pthreads)	11,000–615,000	2320	346	10	0.8844 (0.8828±0.0017)	0.7441 (0.7341±0.0118)
OXFam / large; 10,179–81,503 sequences						
G-large-INS-1 (MPI)	56–3630	(1940)	(36.8)	100	0.8878 (0.8822±0.0062)	0.8312 (0.8261±0.0069)
G-large-INS-1 (Pthreads)	121–3610	3410	413	10	ditto	ditto
G-INS-1 (Pthreads)	10,000–600,000	2380	383	10	0.8749 (0.8736±0.0021)	0.8212 (0.8206±0.0024)
HomFam / medium; 32 entries; 3127–9105 sequences						
G-large-INS-1 (MPI)	18–1700	(221)	(3.24)	100	0.9393 (0.9406±0.0028)	0.8376 (0.8365±0.0091)
G-large-INS-1 (Pthreads)	34–1720	254	26.5	10	ditto	ditto
G-INS-1 (Pthreads)	1380–47,300	166	22.4	10	0.9520 (0.9534±0.0001)	0.8480 (0.8536±0.0003)
OXFam / medium; 59 entries; 3246–9602 sequences						
G-large-INS-1 (MPI)	26–1470	(280)	(4.63)	100	0.9463 (0.9477±0.0025)	0.9113 (0.9147±0.0039)
G-large-INS-1 (Pthreads)	33–1490	343	40.0	10	ditto	ditto
G-INS-1 (Pthreads)	1420–48,800	222	28.1	10	0.9485 (0.9462±0.0006)	0.9147 (0.9129±0.0011)
HomFam / small; 38 entries; 93–2957 sequences						
G-large-INS-1 (MPI)	10–295	(28.9)	(0.390)	100	0.9315 (0.9380±0.0043)	0.8405 (0.8574±0.0087)
G-large-INS-1 (Pthreads)	8–304	15.2	1.65	10	ditto	ditto
G-INS-1 (Pthreads)	9–4690	10.2	1.27	10	0.9358 (0.9358±0.0000)	0.8549 (0.8549±0.0000)
OXFam / small; 74 entries; 19–2981 sequences						
G-large-INS-1 (MPI)	11–550	(26.9)	(0.377)	100	0.9606 (0.9582±0.0033)	0.9393 (0.9375±0.0044)
G-large-INS-1 (Pthreads)	8–553	15.3	1.67	10	ditto	ditto
G-INS-1 (Pthreads)	8–4380	10.2	1.25	10	0.9572 (0.9572±0.0000)	0.9358 (0.9358±0.0000)

For RAM usage (maximum resident set size; MaxRSS), the maximum and minimum values in each subset are shown. In the MPI version, the MaxRSS values of the most memory consuming process of the 100 was selected in each problem. CPU time and wall-clock time were averaged for each subset. The calculation of the Pthreads version was performed on Intel Xeon E7-4870 2.4GHz with 2TB RAM using the Lustre (version 2.5.41) filesystem. The calculation of the MPI version was performed using 10 machines in a heterogeneous cluster system (see the footnote of Figure 2-1), for which calculation times are shown in parentheses. The last columns show the ContTest, SP and TC benchmark scores with the input order in the data of https://mafft.sb.ecei.tohoku.ac.jp/, followed by average score ± sample standard deviation, with 20 (“small” subsets of HomFam and OXFam) or 10 (the others) randomized orders, in parentheses. The FastSP program was used to compute the SP and TC scores.

2.3.3 Guide tree

G-large-INS-1 uses the stepwise addition strategy to build a guide tree while G-INS-1 uses a guide tree with a UPGMA-like method. Probably as a result of this difference, in Figure 2-3, the conventional guide trees (G-INS-1, blue) slightly outperformed the trees using the stepwise addition strategy (G-large-INS-1, red) when the number of sequences is small (from 30 to 100). However, for a larger number of sequences (200 or more), little difference was observed. This is consistent with an earlier report (Boyce *et al.*, 2014), which found that the importance of accurate guide trees decreases with the number of sequences, at least when used for protein structural analyses.

The resulting tree with the stepwise addition strategy depends highly on input order. To investigate the instability of alignment accuracy (Boyce *et al.*, 2015), we performed additional tests with shuffled sequence order, 10–20 times, for HomFam and OXFam and ContTest (Table 1-1), which suggest the effect of the input order is generally larger in G-large-INS-1 than in G-INS-1. It was difficult to repeat QuanTest with shuffled orders, as it took a large amount of computational time. Instead, we ran G-large-INS1 with the reverse sequence order and G-INS-1 with a non-default guide tree (Figure 2-1a), in order to check the instability of benchmark scores to some extent.

2.3.4 Separately estimating alignment accuracy in QuanTest

QuanTest utilizes secondary structure prediction accuracy (SSPA) to measure alignment quality. Figure 2-1a suggests that SSPA increases with the number of sequences, but two different factors, alignment accuracy and structure-prediction accuracy, are combined in this observation. The alignment accuracy was separately estimated by the procedure described in (Le *et al.*, 2017): (i) Subalignments of the same set of 200 sequences were extracted from the alignments of 500, 1000, 2000 and 4000 sequences. (ii) Secondary structure was predicted by JPred4 (Drozdetskiy *et al.*, 2015) for each subalignment (Figure 2-2). The difference in SSPA should purely reflect the difference in alignment, because the same set of sequences were used in these alignments. It is known (Sievers *et al.*, 2013; Le *et al.*, 2017) that the alignment accuracy of approximate methods, such as Clustal Omega and MAFFT-FFT-NS-2, decreases with the increase of sequences, shown as black dashed lines in Figure 2-3. By contrast, the alignment accuracy of G-INS-1 and G-large-INS-1 (red and blue dashed lines in Figure 2-3) does not decrease with the increase of sequences.

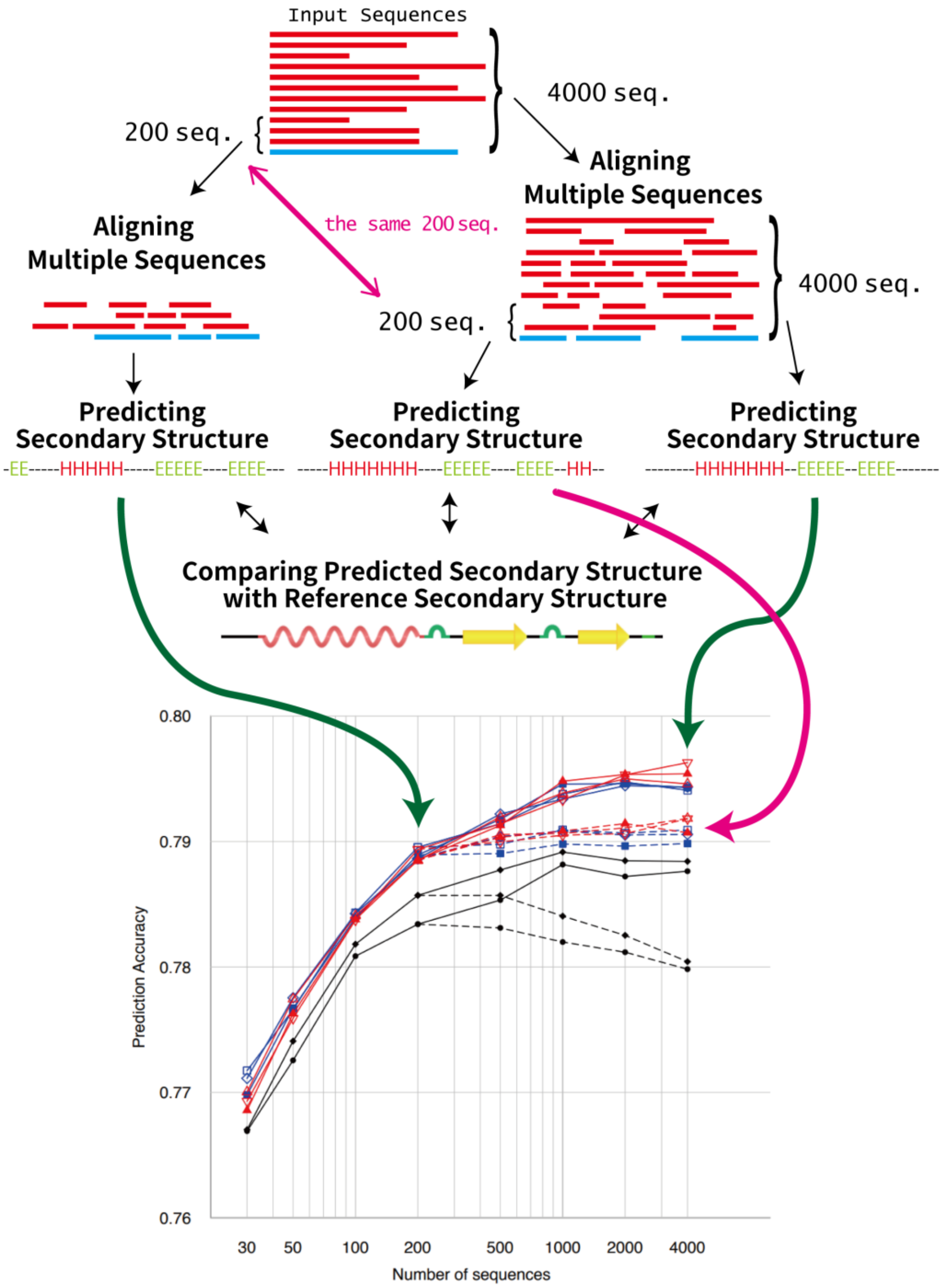


Figure 2-2 Schematic diagram of QuanTest and the correspondences between the procedures and the results in Figure 2-3. The green arrows represent use of full alignment (solid line in the result figure). The magenta arrow represents use of sub alignments of 200 sequences included in the full alignments (dashed line in the result figure).

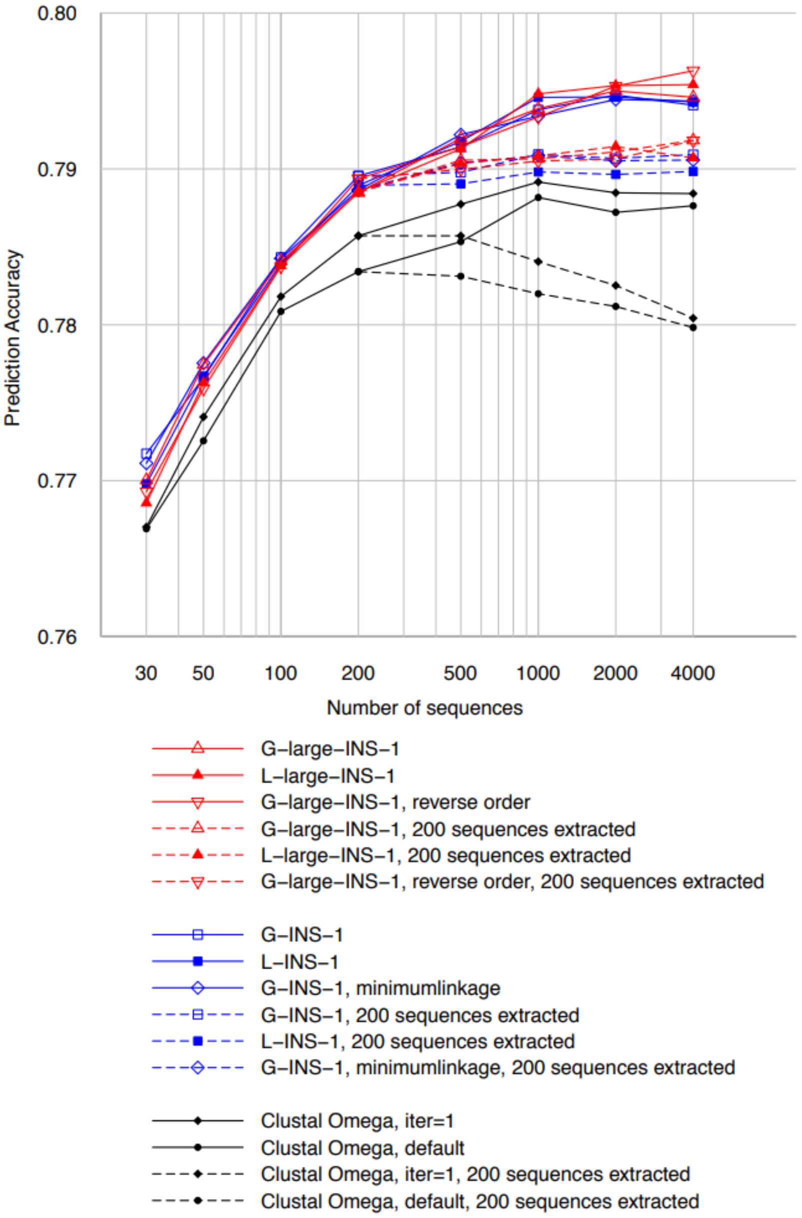


Figure 2-3 Effect of data size on alignment accuracy only (dashed lines) and total effect on alignment accuracy and secondary-structure prediction accuracy (solid lines) in QuanTest. The same 1940 entries as Figure 2-1 were used.

2.4 Discussion

Until now, it was necessary to use highly approximate methods, such as the FFT-NS-2 option of MAFFT or the progressive option of Clustal Omega, in order to construct large MSAs. In terms of the MSA itself, the accuracy of these methods tends to decrease along with the increase in the number of sequences. This was first pointed out by (Sievers *et al.*, 2013) and confirmed by (Le *et al.*, 2017). The increase in accuracy observed in Figure 2-1a for more than 200 sequences is due to the prediction phase not due to the alignment phase (black dashed lines in Figure 2-3). As a result, it was difficult to know how many sequences should be included in an MSA. With more sequences, the MSA has richer comparative information, but the alignment quality is expected to decrease. The optimal balance between these two factors may differ by case. In contrast, the accuracy of G-large-INS-1 and G-INS-1 (red and blue dashed lines in Figure 2-3) was robust to data size in this test. The number of sequences to include in the MSA can now be determined simply based on the computational resources available and the requirements for the downstream analysis.

Chapter 3

Development of methods for an effective reduced vector representation of ligand-binding pockets of proteins

3.1 Introduction

Proteins often execute their functions in a cell through interactions with their ligands. Interactions between proteins and small molecules are particularly important, especially with relation to metabolism, drug discovery, and drug repositioning. Consequently, comparing and/or classifying ligand-binding pockets of small molecules can facilitate the functional elucidation of proteins. For instance, performing classification about known protein structures provides insight into the structural features of proteins and advances our understanding of functions and structures. With the recent increase of a database of known protein structure, the Protein Data Bank (PDB), we have huge amounts of structural information for approximately 350,000 known and 6.2 million unknown (estimated using a pocket identification program) ligand-binding pockets (Ito *et al.*, 2015). Consequently, comprehensive comparison and classification of both known and predicted protein ligand-binding pockets provide important insights into predicting ligands and drug discovery. For such a comprehensive analysis, a fast pocket comparison method is extremely useful. Indeed, various approaches have already been proposed (Konc and Janezic, 2014).

Pocket comparison methods are divisible into two classes, i.e., alignment-dependent and alignment-free methods (Gao and Skolnick, 2013a). Although alignment-dependent methods for pocket comparison perform structural alignment of binding residues, alignment-free methods are independent of the structural alignment. Such methods often use descriptors that represent binding residues in a pocket. Both methods have been developed to be applicable to large-scale comparison of binding pockets. For instance, Gao and Skolnick developed a fast alignment-dependent method called APoc (Gao and Skolnick, 2013b). They argued that alignment-dependent methods are “generally more accurate, albeit slower than alignment-free methods." However, although alignment-free methods are unable to provide information of matched residues, they are efficient in terms of computational time. Consequently, these methods enable comparison of known binding pockets with numerous predicted ligand-binding pockets estimated using a pocket detection program. Furthermore, alignment-free methods are compatible with analysis of “flexible” binding pockets, and are readily applicable to binding pockets comprising multiple protein chains. It remains difficult to apply alignment-dependent methods directly to such pockets.

Considering these reasons, we assume that alignment-free methods can particularly contribute to protein-ligand interaction prediction and can enable the prediction of protein function. Therefore, we developed an alignment-free method that enables us to perform exhaustive comparison of both known and predicted ligand-binding pockets of 1,000,000 order (Ito *et al.*, 2012), and to develop a database called PoSSuM that includes the comparison results.

However, some possible caveats must be associated with those methods. First, these methods have no abundant ability of expression because they cannot increase the number of types of labels for counting the occurrence frequency of triangles because the occurrence frequency vector becomes sparse and because it is difficult to calculate the similarity properly. Second, the error by which the similarity between similar triangles is regarded as completely dissimilar occurs when a ligand binding pocket is converted into triangles because the similarity between triangles is not considered beyond the bin.

To overcome these difficulties, we defined the similarity among all triangle types which is producible under our labeling method, and developed a method to represent a ligand binding site with a reduced vector using multidimensional scaling (MDS). In other words, the vector representation of a ligand binding site is obtained using the linear combination of the occurrence frequency of triangles using the coordinates of triangles in a metric space. Using this method, one can calculate the similarity between two ligand binding pockets merely by calculating the inner product of two reduced vectors.

We also sought to revise the way of integrating the similarity between triangle types to improve the discriminative ability of our method. We introduced the new converting matrix $\boldsymbol{X}'$ by making the double centered matrix

D nonnegative, as calculated from the similarity matrix ***S*** between triangle types. We confirmed there is a slight increase in the discriminative ability because of this change, but we need further analysis to conclude that this performance improvement arises from this change only.

The novel method presented herein exhibits higher performance at the detecting similarity between binding pockets than existing alignment-free methods. It outperforms the fast sequence order-independent structural comparison method, the Alignment of Pockets (APoc) (Gao and Skolnick, 2013b), which necessitates solving the optimization problem. We also confirmed, using the *TOUGH-M1* dataset, that the novel method outperforms SiteEngine (Shulman-Peleg *et al.*, 2004), which measures similarity between pockets using geometric hashing and matching of triangles of centers of physico-chemical properties, and Graph-based Local Structure Alignment (G-LoSA) (Lee and Im, 2012), which measures similarity between pockets with iterative maximum clique search and solving the linear sum assignment problem. The results of this study suggest that this novel method is faster than our previous method (Ito *et al.*, 2012) and the other methods.

3.2 Materials and Methods

This study was undertaken to enhance our original method (Ito *et al.*, 2012) for finding similar ligand-binding pockets using all triangles consisting of three amino acids in the pockets and similarities between triangle types. The entire procedure for converting structural information of a pocket into a reduced vector representation using multidimensional scaling (MDS) is the following.

3.2.1 Enumerating possible triangle types

In this study, each binding pocket is described by an ensemble of all triangles consisting of three Cα atoms of amino acids in the pocket, in contrast to our previous study which considered a several types of triangle types at a time. Each triangle vertex is labeled with one amino acid of 20 types. We treated modified residues such as selenomethionine as naturally occurring amino acids (e.g. methionine). Regarding the triangle edges, first of all, we used the same definition as that presented in our previous report. We considered only triangles with edges, i.e. $C\alpha$-$C\alpha$ distances of residue pairs, within 13.6 Å. Moreover, we classified edges into five classes at 2.2 Å intervals and labeled them Roman numerals (I, II, III, IV, and V) (hereinafter designated as '5 edge'). Secondly, we extended $C\alpha$-$C\alpha$ distances to be considered, ranging from 1.0 Å to 15.8 Å, and added a class of edges. Edges were classified into 6 classes at intervals of 2.2 Å. We labeled them Roman numerals (I, II, III, IV, V, and VI) in ascending order (hereinafter '6 edge interval set α'). Lastly, we also investigated the effects of revising the interval distances which affect the classes of edges. We modified the intervals of 6 edge classes from 2.2 Å each to 1.0, 4.0, 6.36, 8.72, 11.08, 13.44, and 15.8 Å (hereinafter, '6 edge interval set β').

Superposition of one triangle to another triangle can be done in six ways. We regarded triangles as identical if they have the same label, with regard to their vertices and edges, considering all six ways of superposition. Given these conditions, we listed all possible triangle types. Therefore, we found triangles of 171,700 types under the condition that the classes of edge labels are 5 labels and 295,240 types under the 6 labels.

3.2.2 Definition of similarity between two triangle types

In our novel method, for two triangle types of p and q, we defined the similarity s_{pq} consisting of two terms, which respectively measure mutual physicochemical and geometrical similarity, as the following form:

$$s_{pq} \equiv \max \begin{bmatrix} r(m_{\mathrm{AD}} + m_{\mathrm{BE}} + m_{\mathrm{CF}}) + (1-r)(-1)\big(f(\mathrm{AB}, \mathrm{DE}) + f(\mathrm{BC}, \mathrm{EF}) + f(\mathrm{CA}, \mathrm{FD})\big) \\ : 6 \text{ way superposition} \end{bmatrix} \quad (3\text{-}1)$$

Here, m_{XY} represents a physicochemical similarity between two amino acids X and Y, defined with an amino acid substitution matrix. For this study, we used the PAM50 matrix (Dayhoff and Schwartz, 1978), which is not rounded after a decimal point, because we assumed that residues consisting of a ligand-binding pocket are conservative for substituting amino acids. Also, the not-rounded PAM50 matrix yielded better performance than the rounded (data not shown). In addition, A, B, and C respectively denote the vertices of the triangle type p; D, E, and F respectively denote those of triangle type q. r is a weighting factor, ranging from 0 to 1, for physicochemical and geometrical similarity terms in this equation. AB, BC, and CA denote the edges of the triangle type p; DE, EF, and FD denote those of triangle type q. The function f, which represents the geometrical dissimilarity between two edges X and Y, is defined as

f(edge X, edge Y) ≡ | value of the class for edge X − value of the class for edge Y |. In this definition, the value of a class is given according to the assigned numerals for a class. For example, the function f gives 4 when edge X belongs to class I and edge Y belongs to class V. Then, f is summed up for three edges and multiplied by -1 for converting dissimilarity to similarity (see eq. (3-1)). We regarded the maximum value of s_{pq} for all possible ways of superposition of triangle types considering rotation and reflection as similarity s_{pq} for two triangle types p and q.

3.2.3 Multidimensional Scaling (MDS)

To perform MDS, we calculated the similarities for all possible pairs of 295,240 or 171,700 triangle types based on the similarity definition presented above. The number of triangle types is hereinafter designated as N. We were able to obtain a similarity matrix $\boldsymbol{S}$ between triangle types as a square matrix of order N. We assumed a model by which the similarity between triangle types corresponds to the inner product, and used MDS to obtain the coordinates of each triangle type in a high-dimensional space. The procedures used for this study are summarized briefly as follows. First centering is performed over the previously described similarity matrix $\boldsymbol{S}$ to obtain the double centered matrix $\boldsymbol{D}$ because we want to obtain, eventually, those coordinates which have zero mean. The element of the double centered matrix $\boldsymbol{D}$ is obtainable as

$$\boldsymbol{D}_{ij} = \boldsymbol{S}_{ij} - \frac{1}{N}\sum_{a=1}^{N} \boldsymbol{S}_{ia} - \frac{1}{N}\sum_{a=1}^{N} \boldsymbol{S}_{aj} + \frac{1}{N^2}\sum_{a=1}^{N}\sum_{b=1}^{N} \boldsymbol{S}_{ab}\,. \tag{3-2}$$

Then, the eigenvalue decomposition of $\boldsymbol{D}$ is performed, thereby yielding eigenvalue vector $\boldsymbol{\lambda}$ and a matrix of eigenvector $\boldsymbol{Z}$. Using $\boldsymbol{\lambda}$ and $\boldsymbol{Z}$, the coordinates of triangle types are then found using the following formula:

$$\boldsymbol{X} = \left(\sqrt{\boldsymbol{\lambda}_1}\boldsymbol{z}_1, \sqrt{\boldsymbol{\lambda}_2}\boldsymbol{z}_2, \sqrt{\boldsymbol{\lambda}_3}\boldsymbol{z}_3, \dots, \sqrt{\boldsymbol{\lambda}_l}\boldsymbol{z}_l\right). \tag{3-3}$$

We used the randomized algorithm (Halko *et al.*, 2011) to compute large-scale singular value decomposition for eigenvalue decomposition of $\boldsymbol{D}$ because the order of $\boldsymbol{D}$ is huge, and because the only necessary eigenvalues are those with a large absolute value. We designate this $\boldsymbol{X}$ as with negatives (*wn).*

3.2.4 Introducing the new converting matrix $\boldsymbol{X}'$

We introduced the new converting matrix $\boldsymbol{X}'$ by making the double centered matrix $\boldsymbol{D}$ nonnegative. First, we calculated $\boldsymbol{D}'$ by replacing all negative elements of matrix $\boldsymbol{D}$ with zeros, as follows;

$$\boldsymbol{D}' = \left(d'_{ij}\right);\ d'_{ij} = \max\left(0, d_{ij}\right). \tag{3-4}$$

Then, we performed eigenvalue decomposition of $\boldsymbol{D}'$ in as described in Section 3.2.3 and obtained $\boldsymbol{X}'$. We designate this $\boldsymbol{X}'$ as greater than or equal to 0 (*gt0*).

3.2.5 Convert pockets into reduced vector representations

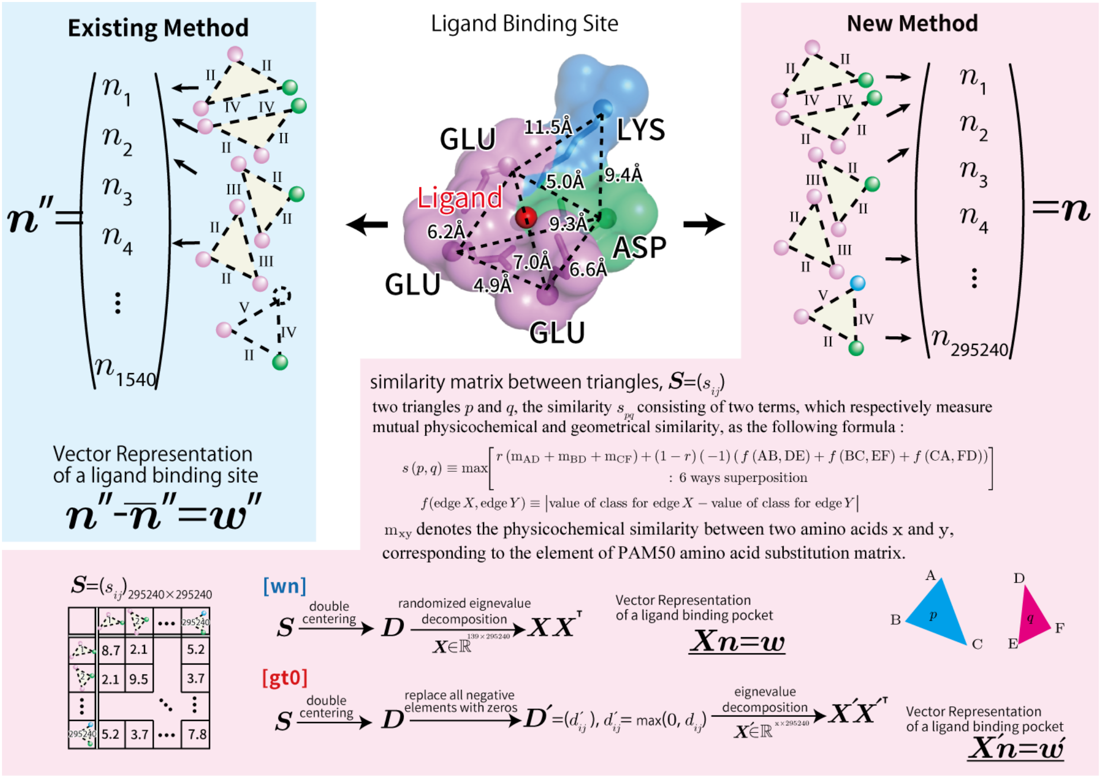
We defined $\boldsymbol{n}$ as a N-dimensional vector based on the occurrence frequencies of triangle types at a ligand-binding pocket. All triangles that occur in a pocket with edge lengths of 1.0 Å to 13.6 or15.8 Å are classified as one of N triangle types. Using $\boldsymbol{X}$ (or $\boldsymbol{X}'$) described in the previous section, we found the following.

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{n} = \left(\sqrt{\boldsymbol{\lambda}_1}\boldsymbol{z}_1, \sqrt{\boldsymbol{\lambda}_2}\boldsymbol{z}_2, \sqrt{\boldsymbol{\lambda}_3}\boldsymbol{z}_3, \dots, \sqrt{\boldsymbol{\lambda}_l}\boldsymbol{z}_l\right)^{\mathrm{T}}(n_1, n_2, n_3, \dots, n_N)^{\mathrm{T}}$$

$$= (\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \dots, \boldsymbol{x}_N)(n_1, n_2, n_3, \dots, n_N)^{\mathrm{T}}$$

$$= (w_1, w_2, w_3, \dots, w_l)^{\mathrm{T}} = \boldsymbol{w}$$

In the equations above, n_i represents the number of the i-th triangle in the list of triangle types. $\boldsymbol{w}$ stands for a vector representing a pocket (Figure 3-1). To represent a pocket with a reduced vector based on the MDS result, we used the lowest number of dimensions that satisfy a certain extent of cumulative contribution ratio, which was calculated using positive eigenvalues only. For this study, we set the criteria of the cumulative contribution ratio as 0.98. Thus, the number of dimensions changes according the change of parameters, which are r in equation (3-1) and the number of the classes of edge labels (Figure 3-2). We define similarity between two pockets i and j as a cosine distance between $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$. Therefore, the similarity can be found easily by calculating the inner product between normalized $\boldsymbol{w}_i$ and normalized $\boldsymbol{w}_j$. This procedure can be regarded as calculation of the weighted arithmetic mean over $\boldsymbol{X}$ (or $\boldsymbol{X}'$) weighted by $\boldsymbol{n}$.


Existing Method
$\boldsymbol{n}''=$
n_1
n_2
n_3
n_4
n_{1540}
II
IV
III
V
Vector Representation
of a ligand binding site
$\boldsymbol{n}''-\overline{\boldsymbol{n}}''=\boldsymbol{w}''$
Ligand Binding Site
GLU
LYS
Ligand
ASP
GLU
GLU
11.5Å
5.0Å
9.4Å
6.2Å
9.3Å
7.0Å
6.6Å
4.9Å
New Method
n_1
n_2
n_3
n_4
n_{295240}
$=\boldsymbol{n}$
$\boldsymbol{S}=(s_{ij})_{295240\times295240}$
1
2
295240
8.7
2.1
5.2
2.1
9.5
3.7
5.2
3.7
7.8
[wn]
$\boldsymbol{S} \xrightarrow[]{\text{double centering}} \boldsymbol{D} \xrightarrow[\boldsymbol{X}\in\mathbb{R}^{139\times295240}]{\text{randomized eignevalue decomposition}} \boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}$
Vector Representation
of a ligand binding pocket
$\boldsymbol{Xn}=\boldsymbol{w}$
[gt0]
$\boldsymbol{S} \xrightarrow[]{\text{double centering}} \boldsymbol{D} \xrightarrow[]{\text{replace all negative elements with zeros}} \boldsymbol{D}'=(d'_{ij}),\ d'_{ij}=\max(0, d_{ij}) \xrightarrow[\boldsymbol{X}'\in\mathbb{R}^{x\times295240}]{\text{eignevalue decomposition}} \boldsymbol{X}'\boldsymbol{X}'^{\mathsf{T}}$
Vector Representation
of a ligand binding pocket
$\boldsymbol{X}'\boldsymbol{n}=\boldsymbol{w}'$
A
B
C
p
D
E
F
q

similarity matrix between triangles, $\boldsymbol{S}=(s_{ij})$

two triangles p and q, the similarity s_{pq} consisting of two terms, which respectively measure mutual physicochemical and geometrical similarity, as the following formula :

$$s\left(p,q\right)\equiv\max\begin{bmatrix} r\left(\mathrm{m_{AD}}+\mathrm{m_{BD}}+\mathrm{m_{CF}}\right)+\left(1-r\right)\left(-1\right)\left(f\left(\mathrm{AB},\mathrm{DE}\right)+f\left(\mathrm{BC},\mathrm{EF}\right)+f\left(\mathrm{CA},\mathrm{FD}\right)\right) \\ :\ 6\text{ ways superposition}\end{bmatrix}$$

$$f(\text{edge}\,X,\text{edge}\,Y)\equiv\left|\text{value of class for edge}\,X-\text{value of class for edge}\,Y\right|$$

$\mathrm{m_{xy}}$ denotes the physicochemical similarity between two amino acids x and y, corresponding to the element of PAM50 amino acid substitution matrix.

Figure 3-1 Schematic diagram of a vector representation of a ligand-binding pocket. Structural and amino acid information of a ligand-binding pocket are converted into a vector.

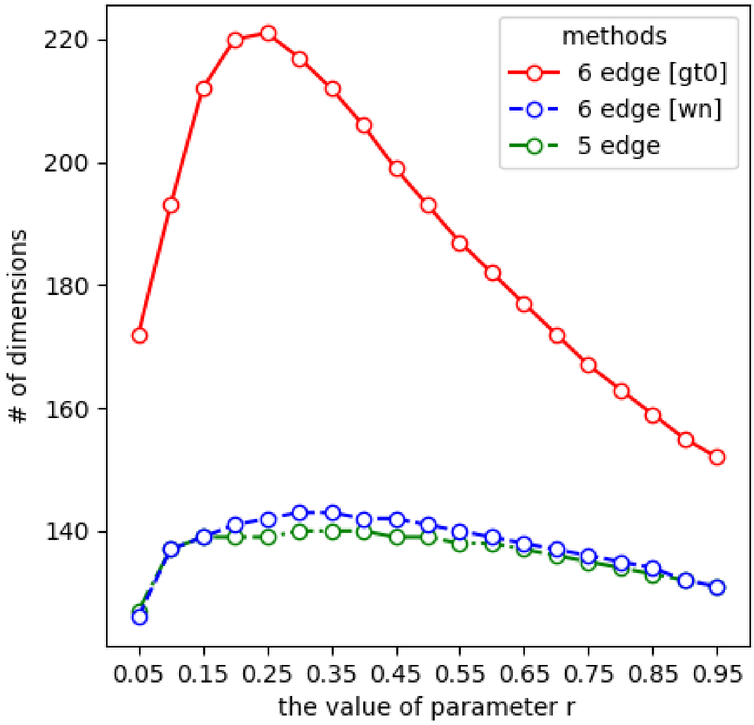


Figure 3-2 Relation between the value of weighting factor r and the number of dimensions. The X axis shows r, and the Y axis shows the dimensions used to represent a binding pocket.

3.2.6 Datasets

We examined three datasets for this study.

i) For *Ito138*, we used a relatively hard dataset based on our previous study (Ito *et al.*, 2012) to optimize the weighting factor r in equation (3-1). The dataset comprises pocket pairs that share the same types of small molecules in proteins with different global structures. This dataset originally comprised 167 known small molecules such as ATP, NAD, and FAD, binding pockets. Nevertheless, we noticed that some binding pockets are inadequate because of the small numbers (< 5) of triangles derived from their binding pockets. For that reason, we used 138 binding pockets (Appendix Table A1) among the 167 binding pockets to calculate the optimized r for our novel method. Then we compared the method with existing ones. We designate this dataset as *Ito138*.

ii) Regarding *APocS3*, we applied the Subject and Control dataset used in a report about APoc (Gao and Skolnick, 2013b) to evaluate our novel method with optimized r and to compare the performance of the method to that of APoc. This dataset is relatively easy because the Subject dataset comprises pocket pairs that share the same or similar types of ligands. While *APocS3* originally comprised 38,066 pairs each in Subject and Control dataset, nevertheless, we noted that some of binding pocket are inadequate because of small number of binding residues, based on the default setting of APoc which is that "minimal number of pocket residue is 10". We omitted pocket pairs which could not be handled by APoc with default setting, and then confirmed that we reproduced the same ROC curves (Fig. 3B in (Gao and Skolnick, 2013b)). For that reason, we used 37,956/26,527 pairs for Subject/Control dataset in this study. Coordinate files of binding pockets were obtained from the APoc website (http://cssb.biology.gatech.edu/APoc). We designate this dataset as *APocS3*.

iii) We used *APocS3_LIGSITE* to compare our novel method with APoc. This dataset includes pairs of predicted pockets generated by LIGSITE (Huang and Schroeder, 2006) based on the pocket pairs in *APocS3*. In comparison with *APocS3*, pocket pair numbers were reduced to 34,511/17,408 pairs of the Subject/Control datasets because the binding residues of some pockets could not be predicted correctly.

iv) We used *TOUGH-M1* to compare our novel method with APoc, SiteEngine (Shulman-Peleg *et al.*, 2005) , and G-LoSA (Lee and Im, 2017). This dataset was presented in a study (Govindaraj and Brylinski, 2018). This dataset includes pairs of predicted pockets generated by Fpocket 2.0 (Le Guilloux *et al.*, 2009). This dataset is composed of 505,116/ 556,810 pairs of Subject(Positive)/Control(Negative) datasets.

3.2.7 Performance analysis

For the *Ito138* dataset, we performed all-against-all 9,453 ($= \binom{138}{2}$) comparisons. In this dataset, a positive example was defined if a pocket pair shares the same ligand; otherwise a pair was regarded as a negative example.

Regarding *APocS3/APocS3_LIGSITE/TOUGH-M1*, for example *APocS3*, we performed comparisons of the 64,483 (= 37,956 + 26,527) binding pocket pairs defined in are earlier report about APoc for the Subject and Control datasets. When a pocket pair with the same/similar ligand gives a similarity score higher than a threshold value, it was regarded as a true positive (TP). Otherwise it was regarded as a false negative (FN). However, if a pair with ligands that are not the same/similar gives a similarity score that is less than the threshold value, then it was classified as a true negative (TN). Otherwise, it was classified a false positive (FP). A true positive rate (TPR) is calculated using TP/(TP+FN). A false positive rate (FPR) is given as FP/(FP+TN). The receiver operating characteristic (ROC) curve is used to present results. ROC is a curve based on a true positive rate against a false positive rate at various thresholds. The area under the ROC curve (AUC) is used to evaluate and compare performances.

3.3 Results and Discussion

We investigated the effects of change of the scheme to calculate the similarity between pockets with novel the similarity definition. In addition, we investigated the effects of modifications in the following three points, i.e., the expansion of edge classes, the revision of intervals of edge classes and the introduction of new converting matrix "$\boldsymbol{X'}$". Then we compared our novel method with existing methods.

3.3.1 Effects of novel similarity measure scheme between two ligand binding pockets

First, we evaluated the effectiveness of changing the scheme to calculate the similarity between two ligand binding pockets. For evaluation, a novel method was used with the number of edge classes set as five classes. The method of classifying them is the same as that of our previous method. We tested the weighting factor in every 0.05 sampling from 0.05 to 0.95 to define the optimized value of r using the *Ito138* dataset. Plots of the weighting factor vs. AUC are presented in Figure 3-3a as "5 edge". According to this result, the best AUC is obtained with r = 0.15. Therefore, we used this value to evaluate the effectiveness of the novel similarity definition.

Figure 3-3b presents the ROC curves, i.e., plots of TPR vs. FPR for this evaluation and shows the novel similarity definition outperforms both FuzCav-like (Weill and Rognan, 2010) and our previous method, PoSSuM-like, in terms of AUC. We identified the main reason behind the superiority of the novel definition.

Figure 3-3c presents actual similarity values for all-against-all 9,453 pairs as a heat map to compare the novel similarity-based method (shown at the lower left) with PoSSuM-like (shown at the upper right). Each square in the graph corresponds to one similarity of a pair of pockets. We found that the novel method assigns lower similarity to pockets, especially to those which bind to HEM or SF4 with pockets which bind to the other ligands. Similarly, the discriminate power of GDP binding pockets from other pockets by the novel method is slightly better than that of the PoSSuM-like method.

Next, we evaluated the effectiveness using *APocS3*. Figure 3-4 presents the ROC curves for this evaluation. It shows that the novel similarity definition also outperforms APoc.

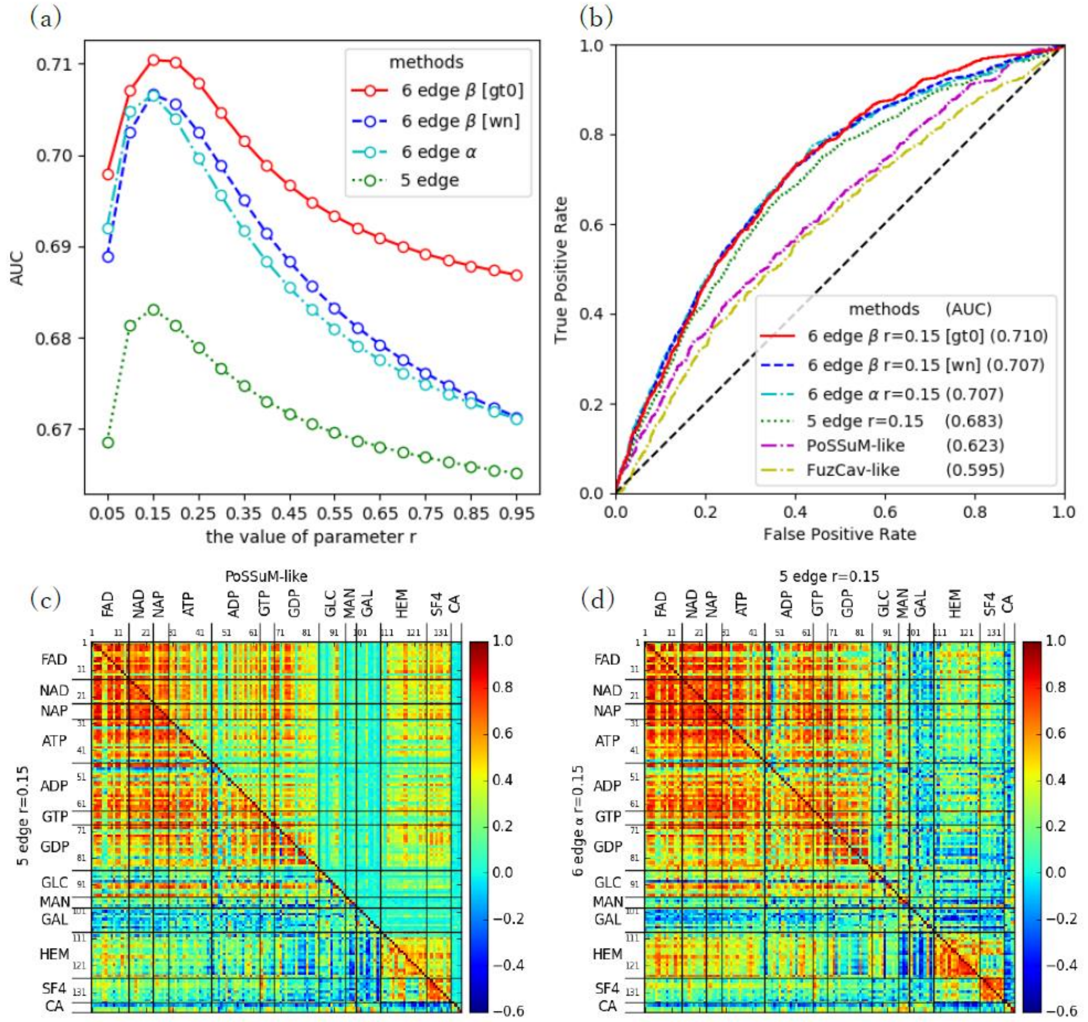


Figure 3-3 Benchmark results with *Ito138*. (a) Results of our novel methods, including "5 edge", "6 edge (with interval set) *α*", "6 edge (with interval set) β [*wn*]" and "6 edge (with interval set) β [*gt0*]", with various values (0.05–0.95 with the sampling interval of 0.05) of the weighting factor *r* are shown. The X axis indicates *r*, and the Y axis indicates AUC values. (b) ROC curves of our novel methods, FuzCav-like and PoSSuM-like are shown. The X axis shows FPR. The Y axis shows TPR. (c) Heat maps to compare the '5 edge class' method (lower left) with the PoSSuM-like method (upper right) are shown. The color of each square in the map represents a similarity value for a pocket pair. Ligand abbreviations placed by axes correspond to the ligand to which a pocket binds. (d) Heat maps to compare the '6 edge *α*' method (lower left) with the '5 edge class' method (upper right) are shown.

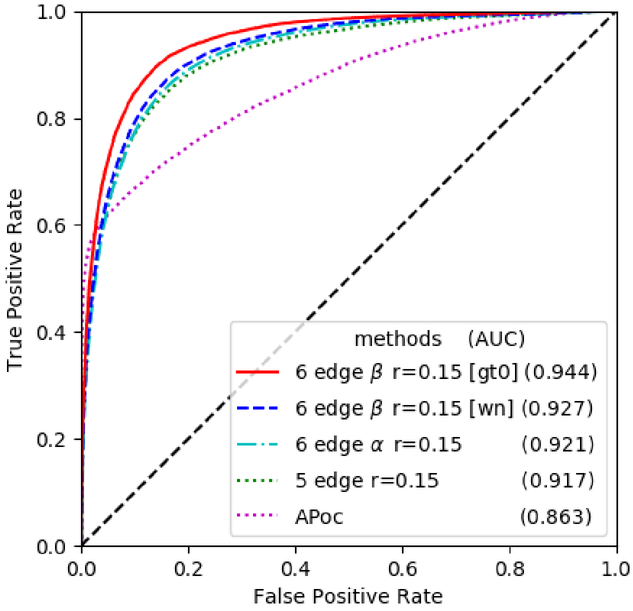


Figure 3-4 Benchmark results with *APocS3*. ROC curves of our novel methods and APoc are shown.

3.3.2 Effects of increasing the number of edge classes

Next, we increased the number of edge classes from 5 to 6, according to the expansion of Cα-Cα distances of residue pairs, ranging from 1.0 Å to 15.8 Å, used as the triangle edge. Edges were classified into 6 classes at intervals of 2.2 Å (interval set α). The effects of this modification were evaluated using *Ito138*. Figure 3-3a and b show that 6 edge classes outperform 5 edge classes, which suggests that addition of edge classes engenders better ability to recognize similar binding pockets.

In Figure 3-3d, as heat maps, we compared individual result obtained using the '6 edge α' method (shown in the lower left) with it using the '5 edge class' method (shown in the upper right). In this case we also found that discrimination of HEM binding pockets and SF4 binding pockets from other binding pockets is improved in the '6 edge α' method compared with the '5 edge class' method. We suppose that discrimination of HEM binding pockets and SF4 binding pockets became better because the maximum value of edge length changed to 15.8 Å. The HEM binding pocket is commonly large. The distance between some binding residues which face each other through HEM is about 15 Å.

We also evaluated the effectiveness of our novel method using an easy dataset: *APocS3*. Figure 3-4 shows that the '6 edge α' method slightly outperforms the '5 edge class' method.

3.3.3 Effects of the intervals of class of edge labels

Next, we examined the procedure used to decide the intervals and modified the intervals of 6 edge classes from 2.2 Å each to 1.0, 4.0, 6.36, 8.72, 11.08, 13.44, and 15.8 Å (interval set β). The setting of every 2.2 Å interval is the same as that of our previous method. The first interval was set to 4.8 Å, which distance originated from the FuzCav method. However, we investigated the frequency of the edge length of triangles taken from all pockets in the *Ito138* dataset (Figure 3-5). In the figure, the green line shows the frequencies of all edge lengths. The red line shows the frequency of edge lengths which comprise two adjacent residues in a chain. According to this figure, almost all edges shorter than about 4.0 Å comprise adjacent residues. Thus, we considered it natural to set the first interval as 4.0 Å based on the difference of chemical characteristics between adjacent residues, or lack thereof.

The effectiveness of this modification was evaluated using *Ito138*. Figure 3-3a and b show that this modification was not so influential, in terms of AUC values, to our novel method on *Ito138*, probably because *Ito138* is comprised of pockets that are too diverse to be affected by this improvement. On the other hand, using *APocS3,* Figure 3-4 shows that the '6 edge β' method performs better than the '6 edge α' method.

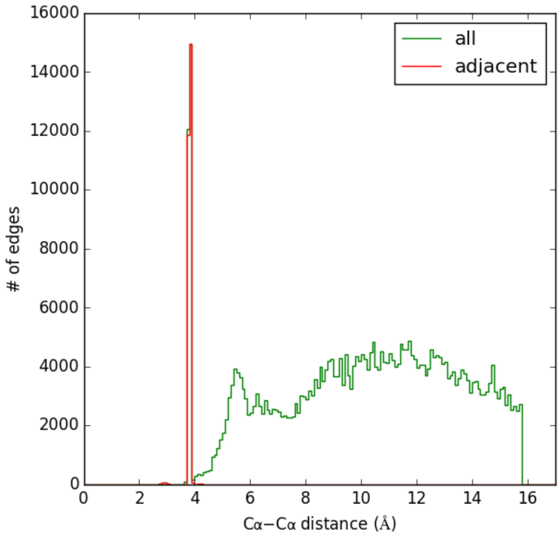


Figure 3-5 Distribution of the edge length (Cα-Cα distance) of triangles taken from all pockets in *Ito138*.

3.3.4 Effects of introducing the new converting matrix X'

Next, we evaluated the effectiveness of introducing the new converting matrix X' to calculate the reduced vector representation of a pocket. The effectiveness of this modification was evaluated using *Ito138*. Figure 3-3a and b show that *gt0* slightly outperforms *wn* in terms of AUC values. We also evaluated the effectiveness of this modification using *APocS3*. Figure 3-4 shows that *gt0* outperforms *wn* as assessed by AUC.

However, we need further analysis to conclude that this performance improvement arises from introducing the new converting matrix only. This performance improvement may possibly come from the changes of the number of dimensions of reduce vector representation of pockets (Figure 3-2). The '6 edge β *gt0*' method using about 200-dimensions, while the '6 edge β *wn*' method using about 140-dimensions. While, of course, changing from *wn* scheme to *gt0* scheme cause the changes of the number of dimensions, which means introducing the new converting matrix cause the changes, to eliminate the possibility that the performance improvement mainly caused by the difference of the numbers of dimensions, we need further analysis with fixing the number of dimensions of reduced vector by the‘6 edge β *gt0*’ method to about 140-dimensions.

Based on the above results, we compared performances with other methods using the ‘6 edge β *gt0*’method as our novel method.

3.3.5 Performance comparison with *APocS3_LIGSITE*

We conducted farther analysis using datasets of predicted pockets. First, we compared the '6 edge β *gt0*' method with APoc using the *APocS3_LIGSITE* dataset. Figure 3-6 shows that, in the low-FPR region, APoc showed higher performance. However, in the region higher than 20% FPR, the '6 edge β *gt0*' method showed higher performance than that of APoc. Additionally, the AUC value of the '6 edge β *gt0*' method was higher than that of APoc.

There is a small difference in contents between *APocS3_LIGSITE* and the "APoc dataset" used in the study (Govindaraj and Brylinski, 2018). For example *APocS3_LIGSITE* is composed of 34,511/17,408 pairs of the Subject/Control datasets, while the "APoc dataset" is composed of 34,970/20,744 pairs. If we compare the AUC values directly, APoc has an AUC of 0.82, G-LoSA has 0.77 and SiteEngine has 0.60. We can confirm that our novel method outperforms the other methods.

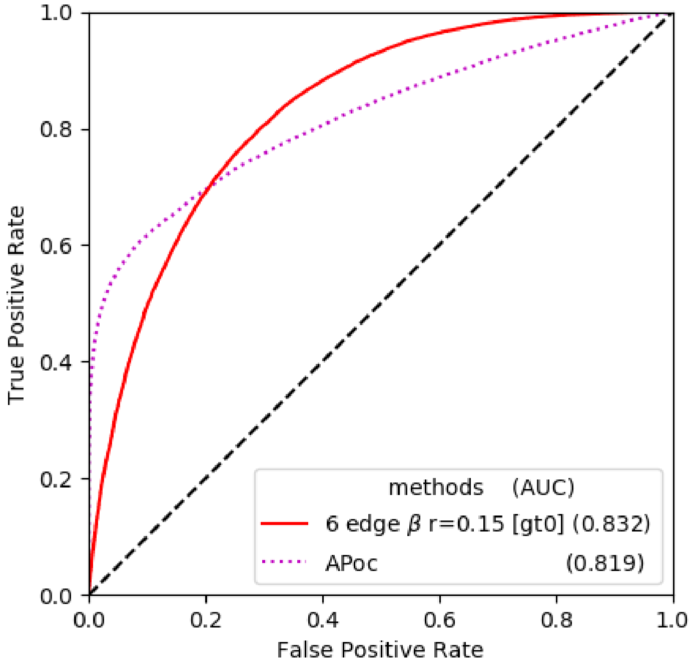


Figure 3-6 Benchmark results with *APocS3_LIGSITE*. ROC curves of our novel method and APoc are shown.

3.3.6 Performance comparison with *TOUGH-M1*

Finally, we compared the '6 edge β *gt0*' method with the other methods using the *TOUGH-M1* dataset. Figure 3-7 shows that the '6 edge β *gt0*' method performed better than the other methods. (cf. Table 1 and Fig.9 in (Govindaraj and Brylinski, 2018))

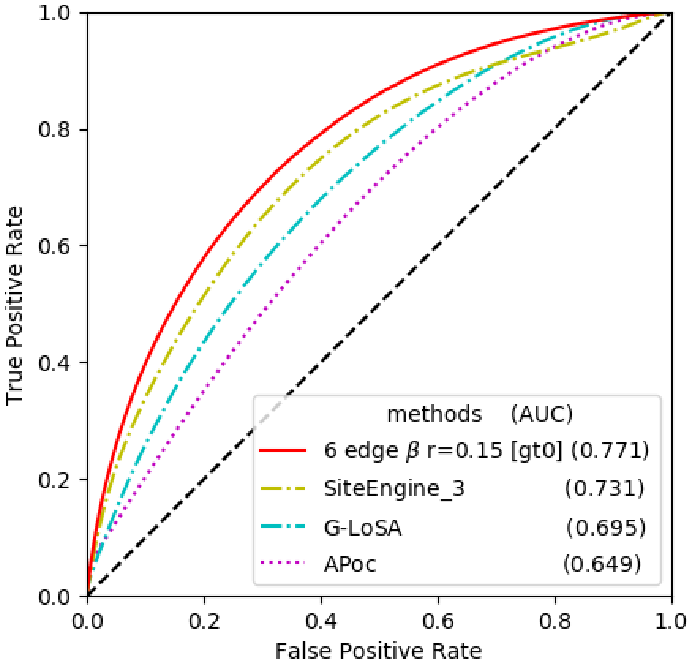


Figure 3-7 Benchmark results with *TOUGH-M1*. ROC curves of our novel method (*gt0*) , SiteEngine, G-LoSA and APoc are shown.

3.3.7 Computational time for calculating similarity of randomly selected pocket pairs

Figure 3-8 shows estimated computational time of APoc and our novel method. Both X axis and Y axis are shown in logarithmic scale. For comparing randomly selected one pair of pockets, while APoc took 0.036 second, our novel method took 10^{-6} second (less than 0.1 second per pocket as preparation process). We speculate that we can achieve 6.5 million pockets (350,000 known plus 6.2 million estimated pockets) comparison using this novel method in few days when we employ a hundred multithreads/processes.

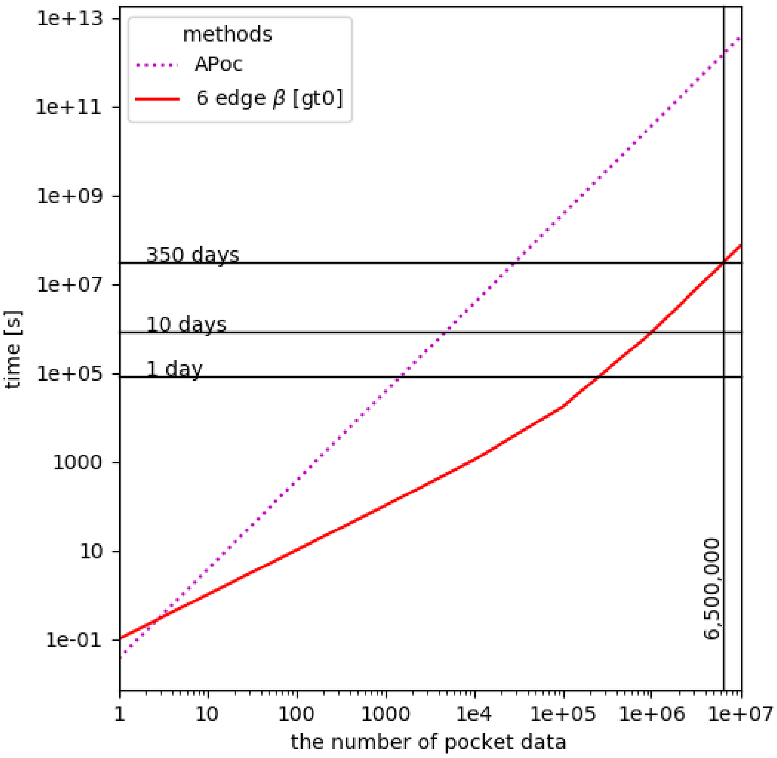


Figure 3-8 Estimated computational time of APoc and our novel method. Both X axis and Y axis are shown in logarithmic scale.

3.3.8 Expedient examples of our method

We present examples that demonstrate the usefulness of '6 edge β *gt0'* method. First, we show an example from *Ito138* dataset. The interferon-inducible p47 resistance GTPases from mouse (PDBID: 1TQ4 (Ghosh *et al.*, 2004)) and the alpha1,3-fucosyltransferase with GDP from H. pylori (2NZX (Sun *et al.*, 2007)) have the same ligand: GDP, though the two proteins possess different global structures; P loop containing nucleoside triphosphate hydrolases fold (1TQ4) and UDP-Glycosyltransferase/glycogen phosphorylase fold (2NZX). Our novel method gave 0.921 as the similarity score for this pocket pair (the higher the similarity score, the more likely the pair is composed of pockets to which the same/similar ligand bind). It is noteworthy that APoc gave 0.772 as the p value for this pair (the lower the p value, the more likely the pair is composed of pockets to which the same/similar ligand bind).

We present two more examples from the *APocS3* dataset. The aldehyde dehydrogenase from rat (1AD3 (Liu *et al.*, 1997); ALDH-like fold) and the 17-beta-hydroxysteroid dehydrogenase type 4 from human (1ZBQ; NAD(P)-binding Rossmann fold) have the same ligand: NAD. As discussed in the Discussion and Conclusion sections in the paper about APoc, this is an example of dissimilar pockets with different ligand conformations. APoc assigned this pair of pockets a p value of 0.417, even though our novel method produced a similarity score of 0.896. We regard this fact as demonstrating the effect of usefulness of our alignment-free method, which can accommodate the pocket conformation change associated with the ligand conformation change. Furthermore, the asparagine synthetase from E. coli (12AS (Nakatsu *et al.*, 1998)) and the electron transfer flavoprotein from human (1EFV (Roberts *et al.*, 1996)) have the same ligand: AMP. Similarly, as discussed in the paper related to APoc, this is an example of dissimilar pockets and similar ligand conformations. Whereas APoc gave 0.212 as a p value for this pocket pair, our novel method showed a similarity score of 0.814. We regard this feature as demonstrating the usefulness of this alignment-free method, which can vaguely represent the circumstances related to a ligand.

3.4 Conclusions

Based on our previous method, for improving the ability to detect similar ligand-binding pockets, we changed similarity measures of pockets. We observed the effectiveness of the change of scheme to calculate the similarity between pockets with two different datasets, *Ito138* and *APocS3*. We also found that the modifications in expansion, revision of edge classes and the introduction of new converting matrix "$\boldsymbol{X}'$" are effective. These results should be considered for future development of pocket comparison methods. The method proposed herein showed higher detection performance of similar binding pockets than the other methods, even if datasets are composed of predicted pockets. Because of its succinct representation, our novel method is expected to be useful for large-scale comparison of binding pockets to infer ligands and functions of proteins.

Acknowledgements

I would like to express my sincere gratitude and appreciation to my advisor, Dr. Kentaro Tomii, for his continuous help and guidance during the period of my Master’s and Doctoral courses. He provided me with a lot of valuable feedback, opportunities to express my ideas, and the chance for educational development.

I would also like to show my appreciation to my Ph. D. thesis committee members, Professor Koji Tsuda, Professor Asai Kiyoshi, Associate Professor Hisanori Kiryu and Professor Akio Kitao.

I would like to express my sincere appreciation to Mr. Toshiyuki Oda, Dr. Yoshinori Fukasawa, Associate Professor Kazunori D. Yamada and Associate Professor Kazutaka Katoh for their collaborative efforts towards the studies presented in this thesis.

I would like to thank former and current members of the Tomii laboratory: Dr. Kenichiro Imai, Dr. Nozomi Nagano, Dr. Yu Yamamori, Dr. Yuko Tsuchiya, Dr. Ikeda Masami, Dr. Kyungtaek Lim, Dr. Jun-ichi Ito, Mr. Hiroyuki Fukuda and Ms. Ming Yang, for their suggestions regarding my research.

I would like to thank the support by Grant-in-Aid for JSPS Research Fellow and computational resources provided by the Chimera cluster system at CBRC/AIRC-AIST, the NIG supercomputer at ROIS National Institute of Genetics, and the Reedbush supercomputer system in the Information Technology Center, The University of Tokyo.

References

Alexandrov,N. and Shindyalov,I. (2003) PDP: protein domain parser. *Bioinformatics*, **19**, 429–430.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Apweiler,R. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bertoni,M. *et al.* (2017) Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Sci. Rep.*, **7**.

Boratyn,G.M. *et al.* (2012) Domain enhanced lookup time accelerated BLAST. *Biol. Direct*, **7**.

Bowie,J.U. *et al.* (1991) A method to identify protein sequences that fold into a known three-dimensional stucture. *Science (80-.).*, **253**, 164–170.

Boyce,K. *et al.* (2015) Instability in progressive multiple sequence alignment algorithms. *Algorithms Mol. Biol.*, **10**.

Boyce,K. *et al.* (2014) Simple chained guide trees give high-quality protein multiple sequence alignments. *Proc. Natl. Acad. Sci.*, **111**, 10556–10561.

Cameron,K. *et al.* (2015) Cell-surface Attachment of Bacterial Multienzyme Complexes Involves Highly Dynamic Protein-Protein Anchors. *J. Biol. Chem.*, **290**, 13578–13590.

Chemical Computing Group ULC (2017) Molecular Operating Environment (MOE), 2013.08.

Dayhoff,M. and Schwartz,R. (1978) A Model of Evolutionary Change in Proteins. *Atlas protein Seq. Struct.*, 345–352.

Drozdetskiy,A. *et al.* (2015) JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.*, **43**, W389–W394.

Fox,G. *et al.* (2016) Using de novo protein structure predictions to measure the quality of very large multiple sequence alignments. *Bioinformatics*, **32**, 814–820.

Fu,L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.

Gao,M. and Skolnick,J. (2013a) A Comprehensive Survey of Small-Molecule

Binding Pockets in Proteins. *Plos Comput. Biol.*, **9**.

Gao,M. and Skolnick,J. (2013b) APoc: large-scale identification of similar protein pockets. *Bioinformatics*, **29**, 597–604.

Ghosh,A. *et al.* (2004) Crystal structure of IIGP1: A paradigm for interferon-inducible p47 resistance GTPases. *Mol. Cell*, **15**, 727–739.

Gloeckner,F.O. *et al.* (2017) 25 years of serving the community with ribosomal RNA gene reference databases and tools. *J. Biotechnol.*, **261**, 169–176.

Gonzalez-Dominguez,J. *et al.* (2016) MSAProbs-MPI: parallel multiple sequence aligner for distributed-memory systems. *Bioinformatics*, **32**, 3826–3828.

Govindaraj,R.G. and Brylinski,M. (2018) Comparative assessment of strategies to identify similar ligand-binding pockets in proteins. *BMC Bioinformatics*.

Gudys,A. and Deorowicz,S. (2017) QuickProbs 2: Towards rapid construction of high-quality alignments of large protein families. *Sci. Rep.*, **7**.

Le Guilloux,V. *et al.* (2009) Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics*.

Halko,N. *et al.* (2011) Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *Siam Rev.*, **53**, 217–288.

Hashimoto,K. *et al.* (2011) Caught in self-interaction: evolutionary and functional mechanisms of protein homooligomerization. *Phys. Biol.*, **8**.

Henikoff,S. and Henikoff,J.G. (1991) Automated assembly of protein blocks for database searching. *Nucleic Acids Res.*, **19**, 6565–6572.

Huang,B. and Schroeder,M. (2006) LIGSITE(csc): predicting ligand binding sites using the Connolly surface and degree of conservation. *Bmc Struct. Biol.*, **6**.

Ito,J.-I. *et al.* (2012) PDB-scale analysis of known and putative ligand-binding sites with structural sketches. *Proteins-Structure Funct. Bioinforma.*, **80**, 747–763.

Ito,J.I. *et al.* (2015) PoSSuM v.2.0: Data update and a new function for investigating ligand analogs and target proteins of small-molecule drugs. *Nucleic Acids Res.*, **43**, D392–D398.

Johnson,L.S. *et al.* (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, **11**.

Johnson,P.M. *et al.* (2016) Functional Diversity of Cytotoxic tRNase/Immunity

Protein Complexes from Burkholderia pseudomallei. *J. Biol. Chem.*, **291**, 19387–19400.

Katoh,K. *et al.* (2017) MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.*

Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.

Konc,J. and Janezic,D. (2014) Binding site comparison for function prediction and pharmaceutical discovery. *Curr. Opin. Struct. Biol.*, **25**, 34–39.

Le,Q. *et al.* (2017) Protein multiple sequence alignment benchmarking through secondary structure prediction. *Bioinformatics*, **33**, 1331–1337.

Lee,H.S. and Im,W. (2017) G-LoSA for prediction of protein-ligand binding sites and structures. In, *Methods in Molecular Biology*.

Lee,H.S. and Im,W. (2012) Identification of ligand templates using local structure alignment for structure-based drug design. *J. Chem. Inf. Model.*, **52**, 2784–2795.

Lensink,M.F. *et al.* (2016) Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment. *Proteins-Structure Funct. Bioinforma.*, **84**, 323–348.

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

Liu,Z.J. *et al.* (1997) The first structure of an aldehyde dehydrogenase reveals novel interactions between NAD and the Rossmann fold. *Nat. Struct. Biol.*, **4**, 317–326.

Lüthy,R. *et al.* (1992) Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83–85.

Morse,R.P. *et al.* (2012) Structural basis of toxicity and immunity in contact-dependent growth inhibition (CDI) systems. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, 21480–21485.

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

Murzin,A.G. *et al.* (1995) Scop - a Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *J. Mol. Biol.*, **247**, 536–540.

Nakatsu,T. *et al.* (1998) Crystal structure of asparagine synthetase reveals a close evolutionary relationship to class II aminoacyl-tRNA synthetase. *Nat. Struct. Biol.*, **5**, 15–19.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Negroni,J. *et al.* (2014) Assessing the Applicability of Template-Based Protein Docking in the Twilight Zone. *Structure*, **22**, 1356–1362.

Notredame,C. *et al.* (1998) COFFEE: An objective function for multiple sequence alignments. *Bioinformatics*, **14**, 407–422.

Oda,T. *et al.* (2017) Simple adjustment of the sequence weight algorithm remarkably enhances PSI-BLAST performance. *BMC Bioinformatics*, **18**.

Pearson,W.R. (1996) Effective protein sequence comparison. In, *Methods in enzymology.*, pp. 227–258.

Pierce,B. *et al.* (2005) M-ZDOCK: a grid-based approach for C-n symmetric multimer docking. *Bioinformatics*, **21**, 1472–1478.

Prlic,A. *et al.* (2012) BioJava: an open-source framework for bioinformatics in 2012. *Bioinformatics*, **28**, 2693–2695.

Raghava,G.P.S. *et al.* (2003) OXBench: A benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics*, **4**.

Remmert,M. *et al.* (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Roberts,D.L. *et al.* (1996) Three-dimensional structure of human electron transfer flavoprotein to 2.1-A resolution. *Proc. Natl. Acad. Sci. U. S. A.*, **93**, 14355–60.

Shiota,T. *et al.* (2015) Molecular architecture of the active mitochondrial protein gate. *Science (80-.).*, **349**, 1544–1548.

Shulman-Peleg,A. *et al.* (2004) Recognition of functional sites in protein structures. *J. Mol. Biol.*, **339**, 607–633.

Shulman-Peleg,A. *et al.* (2005) SiteEngines: Recognition and comparison of binding sites and protein-protein interfaces. *Nucleic Acids Res.*

Sievers,F. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**.

Sievers,F. *et al.* (2013) Making automated multiple alignments of very large numbers of protein sequences. *Bioinformatics*, **29**, 989–995.

Sun,H.Y. *et al.* (2007) Structure and mechanism of Helicobacter pylori fucosyltransferase: A basis for lipopolysaccharide variation and inhibitor design. *J. Biol. Chem.*, **282**, 9973–9982.

Szilagyi,A. and Zhang,Y. (2014) Template-based structure modeling of protein-protein interactions. *Curr. Opin. Struct. Biol.*, **24**, 10–23.

Tomii,K. *et al.* (2005) Protein structure prediction using a variety of profile libraries and 3D verification. *Proteins-Structure Funct. Bioinforma.*, **61**, 114–121.

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

Venkatakrishnan,A.J. *et al.* (2010) Homomeric protein complexes: evolution and assembly. *Biochem. Soc. Trans.*, **38**, 879–882.

Webb,B. and Sali,A. (2016) Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinforma.*

Weill,N. and Rognan,D. (2010) Alignment-Free Ultra-High-Throughput Comparison of Druggable Protein-Ligand Binding Sites. *J. Chem. Inf. Model.*, **50**, 123–135.

Yamada,K. and Tomii,K. (2014) Revisiting amino acid substitution matrices for identifying distantly related proteins. *Bioinformatics*, **30**, 317–325.

Yamada,K.D. *et al.* (2016) Application of the MAFFT sequence alignment program to large data-reexamination of the usefulness of chained guide trees. *Bioinformatics*, **32**, 3246–3251.

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

Yang,Y. and Zhou,Y. (2008b) Specific interactions for ab initio folding of protein terminal regions with secondary structures. *Proteins-Structure Funct. Bioinforma.*, **72**, 793–803.

Ye,Y. and Godzik,A. (2003) Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, **19**, II246-II255.

Zemla,A. (2003) LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.*, **31**, 3370–3374.

Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins-Structure Funct. Bioinforma.*, **57**, 702–710.

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

Appendices

T0860o
T0860-D1
TBM S
1-136
max GDTTS:81.8

Query

Library	FORTE								modified scoring FORTE			
	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	6.2	6.2	6.2	11.4 1	5.9 1	4.7 5	16.3 1	13.6 1				
DB_PSSM	8.9 1	8.9 1	8.9 1	20.9 1	10.0 1	8.5 1	29.4 1	25.0 1				
SSM-PSI_PSSM	7.4	7.4	7.4	6.4 1	6.7	6.6	9.8 1	8.5 1				
(PDP)PSI_PSRP	5.0	5.0	5.0	9.2 1	7.5	4.7	19.6 1	7.5 1	6.7	5.2	30.0 1	25.3 1
DB_PSRP	7.9 1	7.9 1	7.9 1	23.8 1	7.6 2	6.6 1	33.2 1	23.6 1	6.0 12	7.9 1	55.3 1	48.7 1
SSM-PSI_PSRP	5.3	5.3	5.3	9.3 1	6.3	4.0 10	14.8 1	11.0 1	7.1	5.4	21.0 1	17.8 1
HH-PSI_PSRP	19.5 1	19.5 1	19.5 1	32.9 1	47.0 1	21.1 1	55.2 1	33.3 1	73.4 1	45.6 1	90.8 1	73.3 1
(SCOP)HH_PSRP	4.9	4.9	4.9	4.3	6.6	4.6	4.7	4.6	8.8	6.8	6.9	6.4

T0861o-T0862o-T0870o
T0861-D1
TBM S
2-313
max GDTTS:99.04

Query

Library	FORTE								modified scoring FORTE			
	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	46.3 1	44.5 1	44.9 1	41.6 1	40.5 1	39.2 1	36.4 1	37.7 1				
DB_PSSM	35.6 1	48.4 1	37.0 1	33.9 1	29.4 1	36.6 1	27.6 1	28.6 1				
SSM-PSI_PSSM	33.8 1	34.5 1	38.0 1	30.3 1	30.3 1	30.4 1	26.8 1	27.8 1				
(PDP)PSI_PSRP	39.9 1	25.4 1	40.3 1	38.6 1	23.8 1	22.9 1	18.7 1	20.0 1	50.9 1	53.0 1	44.1 1	44.1 1
DB_PSRP	35.1 1	38.0 1	38.0 1	35.3 1	26.5 1	34.3 1	23.5 1	24.5 1	42.9 1	56.4 1	38.4 1	38.4 1
SSM-PSI_PSRP	33.3 1	30.1 1	37.1 1	31.8 1	26.4 1	27.3 1	22.6 1	23.5 1	41.7 1	44.9 1	35.7 1	35.7 1
HH-PSI_PSRP	31.2 1	27.3 1	31.6 1	35.0 1	24.7 1	24.8 1	24.3 1	25.5 1	36.3 1	38.7 1	35.6 1	35.6 1
(SCOP)HH_PSRP	36.6 1	32.2 1	36.4 1	40.8 1	29.2 1	29.2 1	28.9 1	30.1 1	41.9 1	44.8 1	40.7 1	40.7 1

T0861o-T0862o-T0870o
T0862-D1
FM H/S
139-239
max GDTTS:61.56

Query

Library	FORTE								modified scoring FORTE			
	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	4.9 12	5.6 3	4.8 7	4.6 14	4.8 18	5.1 5	5.0 4	4.9 2				
DB_PSSM	7.2 13	6.0 4	7.0 6	7.4 2	7.6 4	6.3 3	6.9 2	6.9 2				
SSM-PSI_PSSM	6.6 1	6.5 1	6.6 2	6.1 2	6.2 2	6.0 2	6.0 2	5.9 2				
(PDP)PSI_PSRP	4.4 11	4.2 7	4.4 7	4.8 1	4.7 2	4.6 1	4.9 1	4.8 1	6.1 1	4.6 11	5.4 6	6.0 3
DB_PSRP	8.3 1	9.1 1	8.6 1	7.1 2	6.9 4	5.4 3	5.9 2	6.0 3	5.2 3	5.9 1	4.9 7	4.7 9
SSM-PSI_PSRP	5.7 1	7.6 1	6.1 1	5.5 1	5.0 2	5.0 2	5.0 2	4.7 7	4.4 3	5.4 2	4.8 1	5.2 1
HH-PSI_PSRP	5.9 1	6.8 1	6.3 1	6.5 1	5.9 1	4.8 1	6.3 1	6.2 1	4.8 1	6.3 1	4.2 16	4.0 16
(SCOP)HH_PSRP	5.5 1	5.8 1	5.8 1	6.6 1	5.9 1	4.7 1	6.2 1	6.2 1	4.9 1	5.2 1	4.2 2	4.2 2

T0861o-T0862o-T0870o
T0870-D1
FM H/S
2-124
max GDTTS:51.63

Query

Library	FORTE								modified scoring FORTE			
	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	4.4 10	6.7	5.5	5.2	4.8 3	5.6	4.7 3	4.8				
DB_PSSM	5.0 8	6.6 1	5.4 2	6.3	5.7	5.4 9	6.3	6.5				
SSM-PSI_PSSM	4.8	5.1 13	5.2	5.1	5.1	4.6 13	5.7	4.4 17				
(PDP)PSI_PSRP	0.0	4.7 1	4.2 1	4.1 1	4.2 1	5.1 1	4.3 1	4.5 1	6.6	5.5	5.9	5.8
DB_PSRP	5.5	6.3	5.8	6.4	5.5	6.0	6.0	6.0	4.9	4.8 10	5.1	5.2
SSM-PSI_PSRP	4.1 16	5.3	4.6 12	5.0	4.4	4.6	4.2	4.2	4.9	4.4 6	4.6	4.6
HH-PSI_PSRP	4.9	5.4 1	4.8	4.7	4.3	4.4 5	4.4	4.4	4.6	4.9	4.1	4.4
(SCOP)HH_PSRP	5.2	5.0	5.0	5.1	4.7	4.5	4.7	4.8	4.8	4.3	4.3	4.3

T0866o
T0866-D1
FM H/S
38-141
max GDTTS:80.77

Query

Library	FORTE								modified scoring FORTE			
	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	4.9	4.7	4.6	4.7	5.3	4.6	5.1	5.2				
DB_PSSM	4.6 19	6.0	5.0	5.0 3	6.2	5.5 10	5.2 12	5.3 9				
SSM-PSI_PSSM	6.1 6	5.3 20	5.8 6	5.9 2	6.4	5.9 4	5.6 12	5.8 9				
(PDP)PSI_PSRP	5.2	4.9	5.2	4.9	4.5	4.4	4.1	4.2	5.3	6.2	5.8	5.5
DB_PSRP	5.4	4.5 13	5.2	4.6 4	5.2	4.7 7	4.0 18	4.2 14	6.2	5.9 6	5.2 20	6.5
SSM-PSI_PSRP	4.6 9	5.3	4.6 18	4.8 5	4.9	5.3 1	4.4 3	4.6 3	5.5 9	7.8 1	6.3 3	6.6 3
HH-PSI_PSRP	4.4 13	4.1 12	4.5 2	4.5 2	4.0 8	5.0 2	4.2 4	4.4 2	6.2 5	8.7 1	7.1 3	7.2 3
(SCOP)HH_PSRP	4.4	4.1	4.4	4.2	4.4	4.1	4.1	0.0	6.2	6.7	6.4	5.9

T0867o
T0867-D1
TBM S
1-104
max GDTTS:97.6

Query

Library	FORTE								modified scoring FORTE			
	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	12.2 1	8.3 1	12.3 1	12.6 1	26.3 1	10.7 1	26.9 1	17.8 1				
DB_PSSM	28.9 1	19.7 1	29.3 1	29.1 1	51.0 1	23.2 1	50.8 1	37.2 1				
SSM-PSI_PSSM	14.5 1	12.1 1	14.8 1	14.4 1	29.1 1	14.7 1	28.0 1	18.7 1				
(PDP)PSI_PSRP	9.7 1	4.6 1	9.8 1	10.1 1	31.0 1	4.2 6	31.3 1	10.1 1	41.1 1	15.0 1	42.5 1	32.9 1
DB_PSRP	32.5 1	20.7 1	32.9 1	32.7 1	54.2 1	20.1 1	53.9 1	34.1 1	71.1 1	32.7 1	73.1 1	60.7 1
SSM-PSI_PSRP	17.5 1	13.7 1	17.8 1	17.5 1	30.8 1	15.6 1	30.0 1	20.1 1	40.2 1	23.9 1	39.9 1	32.9 1
HH-PSI_PSRP	30.3 1	27.0 1	30.4 1	30.3 1	65.1 1	27.9 1	61.7 1	33.0 1	81.3 1	54.1 1	82.1 1	61.5 1
(SCOP)HH_PSRP	4.2	4.2	4.3	4.1	6.1	0.0	6.4	4.1	9.0	5.3	9.4	6.9

T0868-T0869
T0868-D1
FM/TBM H/S
46-161
max GDTTS:86.64

Query

Library	FORTE								modified scoring FORTE			
	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	5.8	4.9 17	4.8 9	5.0 2	5.3	5.4	6.3 1	5.6 1				
DB_PSSM	6.1	6.5	6.3	5.2	6.9	5.6	6.3	6.2				
SSM-PSI_PSSM	5.2 2	6.4	6.0	5.1 5	4.9 7	5.7	5.4 5	5.4 3				
(PDP)PSI_PSRP	0.0	0.0	0.0	4.1	4.3	4.4	4.0 13	4.5	6.0	4.5 16	8.5 1	6.6 1
DB_PSRP	6.2	6.2	5.5	4.8	6.0	4.5	5.1 9	5.0 5	4.7 19	5.7 1	7.4 1	6.4 1
SSM-PSI_PSRP	5.1	6.2	4.2 7	5.0	5.0	4.1 4	4.7	4.6	6.0	6.1 1	6.7 1	6.1 1
HH-PSI_PSRP	4.2 10	4.8	4.1 14	4.7	5.2	4.4	4.6 6	5.0	4.2 14	4.4 10	8.2 1	7.0 1
(SCOP)HH_PSRP	4.4	4.2	4.2	4.1	4.8	4.3	5.7	4.9	5.2	4.6 8	4.6 6	4.4 6

T0868-T0869
T0869-D1
FM H/S
3-106
max GDTTS:52.4

Query

Library	FORTE								modified scoring FORTE			
	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	5.2 8	6.1	4.8 16	4.9 15	5.7 1	4.7 16	4.9 7	4.7 7				
DB_PSSM	5.8	5.7 5	5.4 3	6.7	5.2 7	5.3 5	5.7	5.5				
SSM-PSI_PSSM	5.5 8	6.3 3	5.0 11	5.5 5	4.9 7	5.5 1	5.5 3	4.7 11				
(PDP)PSI_PSRP	4.2	4.0 2	0.0	4.5	4.2 5	4.3 1	4.1 9	4.3 3	6.1 2	4.6 15	7.8	6.9
DB_PSRP	7.2	5.3 3	5.1 16	7.3	6.4	4.6 14	7.1	6.4	6.3	7.6	7.7	7.6
SSM-PSI_PSRP	4.5 16	6.2	5.6	5.4	4.5 16	4.2 12	6.3	5.2	5.3	6.3	6.1	5.8
HH-PSI_PSRP	5.4	4.6 4	5.3	6.6	5.1	4.7	6.2	5.7	5.4	5.1	5.7	6.1
(SCOP)HH_PSRP	5.6	4.4	5.1	6.4	5.1	4.2	6.1	5.4	4.2 17	4.9	6.6	6.7

T0873o
T0873-D1
TBM S
16-501
max GDTTS:83.5

Query

	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	47.5 1	43.8 1	47.5 1	43.3 1	47.8 1	47.3 1	41.7 1	42.6 1				
DB_PSSM	55.1 1	61.3 1	54.8 1	51.8 1	51.8 1	56.9 1	45.8 1	47.5 1				
SSM-PSI_PSSM	56.3 1	50.8 1	59.0 1	54.5 1	54.0 1	51.9 1	48.9 1	50.3 1				
(PDP)PSI_PSRP	39.9 1	30.2 1	38.6 1	30.6 1	25.8 1	24.7 1	20.6 1	22.6 1	71.6 1	69.5 1	63.4 1	62.4 1
DB_PSRP	61.2 1	59.4 1	61.1 1	55.3 1	45.7 1	49.9 1	39.6 1	41.6 1	87.6 1	92.5 1	78.3 1	77.1 1
SSM-PSI_PSRP	54.1 1	46.4 1	56.0 1	49.8 1	42.4 1	40.9 1	37.8 1	39.5 1	73.6 1	71.6 1	69.5 1	68.0 1
HH-PSI_PSRP	53.2 1	44.3 1	53.7 1	56.1 1	41.9 1	39.5 1	41.5 1	43.8 1	73.2 1	71.9 1	69.9 1	68.7 1
(SCOP)HH_PSRP	47.5 1	39.9 1	45.0 1	43.2 1	38.0 1	36.6 1	33.4 1	34.8 1	66.9 1	65.5 1	60.3 1	58.9 1

T0875o
T0875-D1
FM/TBM H/S
1-122
max GDTTS:44.18

Query

	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	17.4 1	6.0	5.8	15.5 1	17.5 1	5.6	14.7 1	14.8 1				
DB_PSSM	24.7 1	7.4	6.8 2	22.6 1	27.5 1	6.4	25.7 1	25.7 1				
SSM-PSI_PSSM	32.6 1	7.3	6.1 1	30.5 1	28.0 1	6.2	26.7 1	27.4 1				
(PDP)PSI_PSRP	20.0 1	4.3	4.2	16.0 1	10.6 1	4.7	8.0 1	8.4 1	35.2 1	6.7	27.5 1	27.6 1
DB_PSRP	31.9 1	6.1	6.8 1	28.7 1	24.8 1	5.7	20.4 1	21.2 1	43.9 1	5.4 13	38.5 1	38.0 1
SSM-PSI_PSRP	37.5 1	6.3	9.1 1	33.3 1	29.4 1	5.5	25.7 1	26.3 1	44.9 1	6.3 4	40.7 1	39.6 1
HH-PSI_PSRP	32.0 1	5.4	7.6 1	30.3 1	22.8 1	4.5	20.1 1	20.7 1	40.9 1	5.8 2	38.3 1	37.8 1
(SCOP)HH_PSRP	6.8	4.8	5.2	6.7	4.7	0.0	0.0	0.0	7.5	5.4	9.1	8.8

T0880o
T0880-D1
FM H/S
1-36
max GDTTS:63.89

Query

	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	5.0 1	5.0 1	5.0 1	5.0 1	6.4 1	5.5 1	6.4 1	5.5 1				
DB_PSSM	7.0 1	7.0 1	7.0 1	7.0 1	9.2 1	7.2 1	9.2 1	7.2 1				
SSM-PSI_PSSM	7.2 2	7.2 2	7.2 2	7.2 2	8.2 1	7.7 1	8.2 1	7.7 1				
(PDP)PSI_PSRP	6.7 1	6.7 1	6.7 1	6.7 1	7.8 2	4.3	7.8 2	4.3	7.6 1	6.1 1	7.6 1	6.1 1
DB_PSRP	6.9 1	6.9 1	6.9 1	6.9 1	8.3 1	6.2 3	8.3 1	6.2 3	8.4 1	8.5 1	8.4 1	8.5 1
SSM-PSI_PSRP	6.0 2	6.0 2	6.0 2	6.0 2	7.2 1	6.1 3	7.2 1	6.1 3	7.0 2	7.4 1	7.0 2	7.4 1
HH-PSI_PSRP	6.3 1	6.3 1	6.3 1	6.3 1	7.7 1	5.8 1	7.7 1	5.8 1	8.6 1	7.2 1	8.6 1	7.2 1
(SCOP)HH_PSRP	5.8 1	5.8 1	5.8 1	5.8 1	6.4 1	5.3 1	6.4 1	5.3 1	6.1 1	5.9 1	6.1 1	5.9 1

T0880o
T0880-D2
FM H/S
37-193
max GDTTS:39.81

Query

	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	5.9	5.9	5.9	5.9	5.8 3	5.6	5.8 3	5.6				
DB_PSSM	6.9	6.9	6.9	6.9	6.7 2	6.1	6.7 2	6.1				
SSM-PSI_PSSM	8.2	8.2	8.2	8.2	7.1 1	7.2	7.1 1	7.2				
(PDP)PSI_PSRP	5.6	5.6	5.6	5.6	6.9 3	4.8	6.9 3	4.8	5.8 3	4.1 9	5.8 3	4.1 9
DB_PSRP	6.1	6.1	6.1	6.1	6.7 1	4.8	6.7 1	4.8	7.7 2	6.0 3	7.7 2	6.0 3
SSM-PSI_PSRP	4.8 18	4.8 18	4.8 18	4.8 18	7.5 2	5.4	7.5 2	5.4	8.2 7	6.1 1	8.2 7	6.1 1
HH-PSI_PSRP	5.3	5.3	5.3	5.3	6.3 1	4.4	6.3 1	4.4	7.4 9	4.6 18	7.4 9	4.6 18
(SCOP)HH_PSRP	5.4	5.4	5.4	5.4	6.8 1	4.1	6.8 1	4.1	7.8 2	4.9 5	7.8 2	4.9 5

T0881o
T0881-D1
TBM S
1-202
max GDTTS:69.43

Query

	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	5.8	6.3	5.4	10.9 1	6.0 1	6.7	12.0 1	11.1 1				
DB_PSSM	6.8 2	6.6 5	7.2 1	14.5 1	7.1 1	6.0 5	16.3 1	15.9 1				
SSM-PSI_PSSM	8.0 1	8.1 1	8.4 1	23.1 1	10.0 1	8.1 1	24.1 1	22.5 1				
(PDP)PSI_PSRP	6.6	7.0	7.3	10.9 1	7.3 2	5.3	11.3 1	8.6 1	6.0	5.7	15.8 1	13.3 1
DB_PSRP	6.0	6.1	5.8 11	16.1 1	7.0 1	5.3 2	17.4 1	16.6 1	7.9 3	5.8 11	35.0 1	34.4 1
SSM-PSI_PSRP	6.8	6.1	6.7	21.9 1	8.8 1	6.3	26.6 1	23.7 1	12.1 1	7.3 1	59.6 1	54.4 1
HH-PSI_PSRP	13.2 1	11.6 1	13.9 1	29.2 1	25.0 1	12.1 1	35.4 1	32.1 1	47.0 1	24.6 1	80.8 1	74.5 1
(SCOP)HH_PSRP	6.0	6.6	6.2	5.8 6	7.4	6.7	6.0 10	6.6	6.0 19	8.8	6.5 5	5.7 19

T0884-T0885
T0884-D1
FM/TBM S
2-72
max GDTTS:65.84

Query

	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	4.9	4.7 9	4.1 12	11.3 1	4.4 8	4.2 9	16.0 1	14.9 1				
DB_PSSM	4.9 5	5.7 2	4.9 4	9.3 1	5.9 3	5.4 5	15.3 1	14.3 1				
SSM-PSI_PSSM	5.2 1	5.3 15	4.9 8	12.8 1	6.3 2	5.0 6	20.0 1	19.4 1				
(PDP)PSI_PSRP	4.0	4.3	0.0	10.0 1	5.4 1	0.0	14.4 1	14.3 1	5.5 1	4.9 10	32.2 1	29.1 1
DB_PSRP	5.9	6.4	5.9	9.8 1	6.0 1	5.0 9	15.9 1	15.3 1	7.0	6.4 2	24.8 1	22.8 1
SSM-PSI_PSRP	5.7 2	4.5 13	5.8 3	14.9 1	5.7 8	5.2	25.1 1	23.6 1	6.4 6	5.3 12	39.5 1	35.7 1
HH-PSI_PSRP	4.1 17	4.2 4	4.1 14	13.9 1	5.2 10	4.9 1	21.3 1	20.3 1	5.8 11	5.9 2	41.2 1	35.8 1
(SCOP)HH_PSRP	5.3	4.1 5	5.5	15.0 1	4.7 6	4.2	21.9 1	20.8 1	5.2 9	5.6 1	39.8 1	37.0 1

T0884-T0885
T0885-D1
TBM S
2-115
max GDTTS:87.94

Query

	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	6.8 1	5.3 5	6.6 1	5.8 1	7.5 1	5.8	6.3 1	6.0 1				
DB_PSSM	7.2 1	5.9 3	7.5 1	7.3 1	7.9 1	5.3 6	7.8 1	7.6 1				
SSM-PSI_PSSM	7.0 1	4.8 12	7.0 1	7.0 1	8.4 1	5.9 2	7.7 1	7.6 1				
(PDP)PSI_PSRP	5.4 1	4.6 1	5.6 1	5.0 1	5.0 1	5.0 1	4.8 1	4.9 1	9.2 1	4.9 9	7.7 1	7.2 1
DB_PSRP	8.1 1	6.1	8.8 1	8.0 1	8.5 1	5.0 11	8.2 1	7.9 1	10.3 1	5.6 2	9.5 1	9.2 1
SSM-PSI_PSRP	6.5 1	5.5	6.7 1	6.5 1	7.3 1	5.0 3	7.0 1	6.8 1	8.9 1	4.8 8	7.7 1	7.5 1
HH-PSI_PSRP	6.7 1	4.6 12	7.2 1	6.6 1	7.3 1	5.3 1	7.0 1	6.7 1	8.4 1	6.5 1	7.8 1	7.5 1
(SCOP)HH_PSRP	6.8 1	4.5 9	7.5 1	6.7 1	7.5 1	4.6 20	7.0 1	6.7 1	8.5 1	5.4 1	7.8 1	7.5 1

T0888o
T0888-D1
FM H/S
1-121
max GDTTS:52.69

Query

	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	5.6	5.6	5.6	5.6	6.4	5.4	6.4	5.4				
DB_PSSM	7.0	7.0	7.0	7.0	7.4 3	6.5	7.4 3	6.5				
SSM-PSI_PSSM	8.1	8.1	8.1	8.1	7.2	6.4	7.2	6.4				
(PDP)PSI_PSRP	4.4 3	4.4 3	4.4 3	4.4 3	5.8	5.2	5.8	5.2	7.4	4.2 16	7.4	4.2 16
DB_PSRP	5.1	5.1	5.1	5.1	6.3 3	4.8 16	6.3 3	4.8 16	7.8	5.0 13	7.8	5.0 13
SSM-PSI_PSRP	5.7	5.7	5.7	5.7	6.9	5.8	6.9	5.8	8.8	7.3	8.8	7.3
HH-PSI_PSRP	4.7	4.7	4.7	4.7	7.9	5.1	7.9	5.1	7.9	7.1	7.9	7.1
(SCOP)HH_PSRP	4.1 13	4.1 13	4.1 13	4.1 13	6.4	4.7	6.4	4.7	7.4	4.9 6	7.4	4.9 6

T0889o
T0889-D1
TBM S
4-242
max GDTTS:87.55

Query

	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	35.1 1	29.6 1	32.1 1	29.3 1	31.6 1	25.4 1	25.1 1	25.7 1				
DB_PSSM	28.1 1	25.9 1	27.9 1	26.1 1	25.1 1	20.5 1	20.2 1	20.9 1				
SSM-PSI_PSSM	25.6 1	21.1 1	23.9 1	22.0 1	24.4 1	18.7 1	18.8 1	19.5 1				
(PDP)PSI_PSRP	26.6 1	18.5 1	26.7 1	24.8 1	23.4 1	14.3 1	12.7 1	13.5 1	46.8 1	32.4 1	29.9 1	30.0 1
DB_PSRP	26.1 1	21.7 1	25.8 1	24.7 1	23.4 1	18.5 1	17.2 1	17.7 1	37.9 1	29.4 1	27.2 1	27.2 1
SSM-PSI_PSRP	23.9 1	18.7 1	23.3 1	22.5 1	22.2 1	16.7 1	16.6 1	17.1 1	37.4 1	27.0 1	25.7 1	25.8 1
HH-PSI_PSRP	21.7 1	17.5 1	20.9 1	20.8 1	20.2 1	15.5 1	15.6 1	16.1 1	30.5 1	22.2 1	22.0 1	22.0 1
(SCOP)HH_PSRP	24.3 1	19.7 1	23.7 1	23.2 1	22.7 1	17.7 1	18.0 1	18.4 1	33.3 1	25.0 1	24.5 1	24.4 1

T0893o
T0893-D1
TBM S
1-73
max GDTTS:78.08

Query

	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	7.6 1	7.9 1	6.6 1	6.8 1	10.3 1	8.7 1	8.6 1	8.4 1				
DB_PSSM	16.9 1	19.2 1	14.1 1	12.8 1	18.7 1	16.9 1	13.7 1	13.7 1				
SSM-PSI_PSSM	17.6 1	14.9 1	15.2 1	13.7 1	19.3 1	15.0 1	14.1 1	14.1 1				
(PDP)PSI_PSRP	5.8 1	4.7 1	5.2 1	4.8 1	5.0 1	4.5 1	4.2 1	4.3 1	14.9 1	14.1 1	12.1 1	11.8 1
DB_PSRP	17.9 1	16.0 1	14.5 1	13.4 1	18.6 1	15.7 1	12.4 1	12.4 1	26.4 1	26.1 1	19.4 1	19.1 1
SSM-PSI_PSRP	17.4 1	13.7 1	15.5 1	13.5 1	18.7 1	13.9 1	13.2 1	13.3 1	27.6 1	25.0 1	21.4 1	21.2 1
HH-PSI_PSRP	15.8 1	12.9 1	14.1 1	12.6 1	16.0 1	12.7 1	12.0 1	12.0 1	24.9 1	21.1 1	18.0 1	17.8 1
(SCOP)HH_PSRP	17.6 1	14.2 1	14.9 1	14.4 1	17.5 1	13.7 1	12.9 1	13.1 1	25.3 1	22.0 1	17.8 1	17.6 1

T0893o
T0893-D2
TBM S
74-242
max GDTTS:87.28

Query

	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	35.4 1	30.9 1	33.8 1	31.9 1	28.6 1	26.2 1	23.8 1	24.8 1				
DB_PSSM	29.9 1	29.1 1	28.9 1	26.2 1	24.4 1	23.0 1	19.4 1	20.2 1				
SSM-PSI_PSSM	34.3 1	26.2 1	32.4 1	29.7 1	27.5 1	22.5 1	20.6 1	21.7 1				
(PDP)PSI_PSRP	25.9 1	16.3 1	27.0 1	24.0 1	16.0 1	12.3 1	10.8 1	11.5 1	41.5 1	37.4 1	33.6 1	33.6 1
DB_PSRP	29.4 1	24.1 1	29.6 1	26.0 1	24.0 1	21.0 1	17.2 1	18.0 1	40.6 1	36.3 1	31.6 1	32.0 1
SSM-PSI_PSRP	30.3 1	22.8 1	28.9 1	26.8 1	27.1 1	20.9 1	19.7 1	20.6 1	41.4 1	32.5 1	32.4 1	32.3 1
HH-PSI_PSRP	25.7 1	18.9 1	26.8 1	24.5 1	20.9 1	16.8 1	16.4 1	17.1 1	35.3 1	28.3 1	28.5 1	28.7 1
(SCOP)HH_PSRP	28.1 1	21.0 1	28.7 1	25.9 1	23.0 1	19.6 1	17.9 1	18.8 1	39.5 1	32.5 1	31.9 1	32.0 1

T0894-T0895
T0894-D1
FM H/S
182-270
max GDTTS:63.48

Query

	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	5.6	6.2	5.8	4.7	5.3	5.6	5.0	5.2				
DB_PSSM	5.2	6.3	5.3	4.7 7	6.3	6.4	6.9	6.9				
SSM-PSI_PSSM	5.9	5.9 9	5.5	5.3 2	6.8	6.5	5.7	5.5				
(PDP)PSI_PSRP	4.6	4.1 16	4.3	4.7	4.1 12	4.2	4.6 1	4.5 2	5.6 6	6.4	6.6	5.6 7
DB_PSRP	5.4	6.4	5.1	4.7 3	6.1 1	6.7	4.6 15	4.8 12	6.6 1	7.0	6.2 1	7.3 1
SSM-PSI_PSRP	5.7	5.9	5.5	4.9	4.9 16	5.6	5.7	5.7	6.0 6	6.0	5.5 19	5.5 13
HH-PSI_PSRP	5.0	5.0 2	4.6 9	4.2 4	5.7	5.0	4.6	4.7	5.7 3	7.4	6.3	5.3 14
(SCOP)HH_PSRP	4.6	4.5	4.4	4.0	5.2	4.4	4.2	4.2	5.9 3	7.3	5.2 11	5.5 8

T0894-T0895
T0894-D2
FM/TBM H/S
271-324
max GDTTS:78.7

Query

	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	5.5 1	5.9 1	5.5 1	5.0 3	6.3 1	5.1 1	5.2 1	5.5 1				
DB_PSSM	6.5 1	5.7 1	6.0 1	6.1 1	7.5 1	7.6 1	6.1 1	6.1 1				
SSM-PSI_PSSM	6.6 1	6.1 1	6.5 1	6.8 1	6.6 1	6.5 1	6.3 2	6.5 2				
(PDP)PSI_PSRP	4.4 1	5.1 1	4.3 1	0.0	5.2 1	5.0 1	4.6 1	4.0 1	8.6 1	6.6 1	8.4 1	8.7 1
DB_PSRP	6.6 1	6.3 1	5.7 1	6.1 1	6.8 1	6.6 1	7.3 1	7.3 1	8.3 2	7.0 1	9.4 1	9.6 1
SSM-PSI_PSRP	7.0 1	6.1 1	6.3 1	7.0 1	8.1 1	6.7 1	6.9 2	6.8 2	7.8 1	6.8 1	7.8 1	7.7 2
HH-PSI_PSRP	5.9 1	5.6 1	6.1 1	5.7 1	6.8 1	5.5 1	7.0 1	7.0 1	7.5 1	6.2 1	7.9 1	8.5 1
(SCOP)HH_PSRP	5.2 1	5.1 1	5.8 1	5.2 1	6.7 1	6.0 1	7.4 1	6.5 1	8.2 1	7.3 1	8.4 1	7.8 1

T0894-T0895
T0895-D1
TBM H/S
1-120
max GDTTS:75.42

Query

	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	6.9 1	8.8 1	8.2 1	7.5 1	6.4 1	7.5 1	6.9 1	7.0 1				
DB_PSSM	7.1 1	8.1 1	7.6 1	7.5 1	6.2 2	6.7 1	6.6 1	6.6 1				
SSM-PSI_PSSM	6.5 1	8.7 1	7.0 1	6.8 1	6.1 1	7.0 1	6.2 1	6.3 1				
(PDP)PSI_PSRP	4.9	4.8	5.0	4.5	4.2	4.6	4.3	4.3	6.4	6.6	7.2	7.0
DB_PSRP	5.9 1	6.2 1	6.4 1	6.1 2	4.7 19	5.0 4	4.8 17	5.0 11	4.3 7	4.6 18	4.5 10	4.6 6
SSM-PSI_PSRP	5.2	6.0	5.8	5.4	4.8	5.7	4.7	4.7	5.2	5.7	5.5	5.4
HH-PSI_PSRP	5.4	5.5	4.7 15	5.5	5.0	5.4	5.2	5.2	4.2	5.2	4.7	4.4
(SCOP)HH_PSRP	5.6	5.8	5.3	5.8	5.1	5.7	5.2	5.3	4.5	5.7	4.7	4.4

T0897-T0898
T0897-D2
FM H/S
162-285
max GDTTS:54.64

Query

	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	5.3	5.1	4.6	4.8	5.4	4.8	5.0	4.8				
DB_PSSM	5.3	5.9	5.1	5.1	7.1	5.5	6.1	6.1				
SSM-PSI_PSSM	6.7 1	6.7	5.2 12	6.6	6.0 11	6.5	7.7	7.8				
(PDP)PSI_PSRP	5.2	4.9	4.6	4.5	5.2	4.4	5.8	5.2	5.8	6.8	5.7	5.9
DB_PSRP	5.7	5.4	5.2	4.9	5.0	5.7	5.1	4.8	6.9	7.5	6.4	6.0
SSM-PSI_PSRP	6.7 1	6.1	5.1 4	5.5	6.1 1	6.1	4.9 11	4.7 17	9.6 1	6.8	6.9 2	7.0 2
HH-PSI_PSRP	5.3 1	5.1	5.1	4.8	5.1 1	5.4	5.6	5.4	9.4 1	7.0	6.4	6.0
(SCOP)HH_PSRP	4.5	4.5	4.5	4.3	4.1	4.1	4.7	4.8	6.0	5.9	6.6	6.2

T0897-T0898
T0898-D1
FM H/S
4-109
max GDTTS:42.69

Query

	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	6.3	6.4	6.3	5.2	7.0	6.1	5.6	5.5				
DB_PSSM	6.1 7	6.1 8	6.1 7	6.8	6.6	7.0	7.1	7.4				
SSM-PSI_PSSM	6.7	6.7	6.7	6.4	6.5	5.7	6.0	5.8				
(PDP)PSI_PSRP	4.6 3	4.6 3	4.6 3	5.0	6.0 10	4.7 11	4.9 16	5.4	7.0	5.1 3	5.6 11	7.7
DB_PSRP	6.8	6.8	6.8	7.2	6.7	5.8	6.2	6.7	8.4	4.9 7	6.4	6.8
SSM-PSI_PSRP	5.0 15	5.1 7	5.0 15	6.5	6.5	5.9	6.2	6.5	8.0	5.7	5.7	6.0
HH-PSI_PSRP	5.7	5.7	5.7	6.0	5.8	5.2	5.6	5.6	8.2	5.5	5.7	5.5
(SCOP)HH_PSRP	6.0	6.0	6.0	5.8	6.4	5.2	5.3	5.2	8.2	4.6	5.7	5.6

T0897-T0898
T0898-D2
FM/TBM H/S
110-164
max GDTTS:75.45

	FORTE								modified scoring FORTE			
	Query											
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	5.4 6	5.4 6	5.4 6	12.1 11	7.8 3	5.5 8	11.9 11	11.9 11				
DB_PSSM	5.7 11	5.7 11	5.7 11	21.9	7.8 1	6.3 2	23.1	23.2				
SSM-PSI_PSSM	5.7 7	5.7 7	5.7 7	22.8	8.5 1	6.7 1	21.1 13	21.3 14				
(PDP)PSI_PSRP	5.7 4	5.7 4	5.7 4	15.4 12	7.5 5	4.3	8.5 11	8.8 12	8.4 2	7.9 1	26.6 11	26.1 12
DB_PSRP	5.7 5	5.7 5	5.7 5	25.6	7.6 1	5.9 2	23.4	23.6	9.8 2	7.3 1	41.2	41.1
SSM-PSI_PSRP	5.5 3	5.5 3	5.5 3	24.0 20	8.9 1	7.1 2	24.1	24.3	11.2 1	8.4 1	42.5	42.5
HH-PSI_PSRP	7.4 1	7.4 1	7.4 1	25.7	8.0 4	8.3 1	21.5	21.7	9.6 2	9.1 1	39.4	39.1
(SCOP)HH_PSRP	6.8 1	6.8 1	6.8 1	19.3 12	9.3 1	7.1 1	16.4 12	16.5 12	10.6 1	8.0 2	29.9 11	29.7 11

T0903o-T0904o
T0903-D1
TBM S
15-350
max GDTTS:97.38

	FORTE								modified scoring FORTE			
	Query											
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	17.0 1	20.3 1	17.5 2	14.9 2	19.8 1	19.3 1	16.3 2	16.4 2				
DB_PSSM	27.4 1	30.6 1	28.8 1	27.5 1	21.6 1	24.6 1	21.1 1	21.6 1				
SSM-PSI_PSSM	27.6 1	26.5 1	31.7 1	28.2 1	24.3 1	21.4 1	22.1 1	22.9 1				
(PDP)PSI_PSRP	14.6 1	15.0 1	16.0 2	13.6 2	8.5 1	8.5 1	8.1 1	8.3 1	28.2 2	30.8 1	26.7 2	26.1 2
DB_PSRP	24.9 1	28.5 1	28.1 1	25.2 1	19.4 1	23.4 1	19.6 1	19.8 1	34.1 1	45.9 1	36.0 1	35.2 1
SSM-PSI_PSRP	25.4 1	23.3 1	30.1 1	26.7 1	19.5 1	18.7 1	19.3 1	19.7 1	35.2 1	34.0 1	37.6 1	37.0 1
HH-PSI_PSRP	21.8 1	21.5 1	24.3 1	24.1 1	18.2 1	17.3 1	19.3 1	19.7 1	29.9 1	28.8 1	34.6 1	34.2 1
(SCOP)HH_PSRP	23.9 2	24.5 1	23.0 2	22.0 2	19.6 2	18.9 2	17.6 2	18.0 2	33.7 2	32.4 1	31.8 2	31.1 2

T0903o-T0904o
T0904-D1
FM H/S
61-311
max GDTTS:43.62

	FORTE								modified scoring FORTE			
	Query											
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	10.0	7.8	10.3	9.5	9.9	7.1	9.6	9.4				
DB_PSSM	8.0 1	5.7 20	6.1 10	6.3 4	7.5 9	6.8	6.6 11	6.4 11				
SSM-PSI_PSSM	7.7 3	5.9 13	9.0	6.2 9	7.5 1	5.2 20	6.5 5	6.2 9				
(PDP)PSI_PSRP	7.2 1	5.4 4	6.1 2	6.3 1	6.7 1	5.5 4	6.5 2	6.3 2	8.7 1	6.1	6.2 2	6.0 10
DB_PSRP	9.5 1	6.0 5	8.7 5	8.1 4	8.2 1	5.6 5	7.9 1	7.6 1	11.2 1	4.6 20	8.7 1	8.2 1
SSM-PSI_PSRP	8.1 1	6.2 3	6.6 6	6.7 3	7.3 1	5.8	6.6 1	6.4 1	10.0 1	5.1 2	8.6 1	7.9 1
HH-PSI_PSRP	8.8 1	6.0 5	7.6 5	7.3 3	7.8 1	5.4 1	7.2 1	7.0 1	10.9 1	4.9 2	8.4 1	7.9 1
(SCOP)HH_PSRP	10.1 1	6.9 1	8.5 1	8.7 1	8.5 1	6.0 2	7.9 1	7.6 1	12.8 1	4.4 3	9.0 1	8.2 1

T0906o
T0906-D1
TBM S
2-353
max GDTTS:95.19

	FORTE								modified scoring FORTE			
	Query											
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	8.5	9.1	8.7	8.4	7.6	7.7	7.8	7.8				
DB_PSSM	53.9 1	55.4 1	53.9 1	51.5 1	70.1 1	72.3 1	70.3 1	68.9 1				
SSM-PSI_PSSM	59.9 1	52.9 1	59.9 1	54.8 1	68.5 1	66.9 1	67.6 1	66.4 1				
(PDP)PSI_PSRP	6.5	6.8	6.3	6.1	6.0	6.2	6.1	6.1	6.4	6.3	7.0	7.0
DB_PSRP	60.2 1	62.7 1	60.9 1	56.3 1	68.4 1	72.0 1	67.9 1	66.6 1	102.7 1	107.0 1	103.8 1	100.9 1
SSM-PSI_PSRP	53.7 1	51.4 1	54.5 1	49.6 1	62.7 1	61.2 1	61.9 1	60.3 1	93.9 1	96.7 1	96.8 1	92.5 1
HH-PSI_PSRP	51.8 1	50.3 1	52.2 1	50.5 1	62.6 1	62.7 1	65.5 1	64.1 1	86.4 1	89.9 1	90.5 1	87.3 1
(SCOP)HH_PSRP	67.4 1	65.3 1	67.7 1	64.3 1	79.1 1	79.7 1	83.8 1	81.7 1	107.6 1	114.2 1	114.8 1	111.6 1

T0909o
T0909-D1
FM/TBM H/S
1-340
max GDTTS:62.01

	FORTE								modified scoring FORTE			
	Query											
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	25.3 1	11.8 1	9.3 1	22.4 1	18.1 1	12.8 1	15.1 1	16.0 1				
DB_PSSM	21.6 1	20.6 1	10.3 1	16.7 1	14.7 1	17.8 1	13.1 1	13.9 1				
SSM-PSI_PSSM	24.0 1	13.9 1	10.0 1	20.5 1	17.2 1	13.9 1	15.5 1	16.5 1				
(PDP)PSI_PSRP	25.6 1	15.7 1	12.8 1	23.6 1	14.7 1	12.2 1	13.5 1	15.2 1	38.3 1	21.6 1	31.8 1	32.2 1
DB_PSRP	23.9 1	21.7 1	12.3 1	20.7 1	15.9 1	18.2 1	13.7 1	14.5 1	37.5 1	31.3 1	26.8 1	26.9 1
SSM-PSI_PSRP	25.0 1	16.1 1	12.5 1	21.3 1	17.9 1	14.6 1	15.6 1	16.6 1	42.4 1	24.7 1	29.3 1	29.9 1
HH-PSI_PSRP	23.5 1	16.5 1	12.2 1	21.7 1	16.3 1	13.6 1	15.1 1	16.1 1	39.7 1	23.9 1	29.2 1	29.9 1
(SCOP)HH_PSRP	27.6 1	19.2 1	14.5 1	26.9 1	19.8 1	14.8 1	17.8 1	19.3 1	43.9 1	22.5 1	33.0 1	34.2 1

T0912o
T0912-D1
TBM H/S
24-113,299-622
max GDTTS:66.42

	FORTE								modified scoring FORTE			
	Query											
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	31.5 1	12.1 1	25.0 1	30.5 1	23.1 1	13.7 1	21.2 1	22.3 1				
DB_PSSM	24.1 1	17.6 1	16.1 1	21.2 1	16.0 1	13.3 1	14.2 1	15.1 1				
SSM-PSI_PSSM	27.3 1	14.7 1	21.9 1	25.6 1	19.2 1	14.7 1	17.2 1	18.2 1				
(PDP)PSI_PSRP	36.1 1	17.3 1	24.4 1	32.6 1	17.7 1	13.2 1	16.2 1	18.9 1	55.6 1	26.2 1	49.1 1	48.1 1
DB_PSRP	27.9 1	18.7 1	21.0 1	26.5 1	17.0 1	14.1 1	15.4 1	16.5 1	45.3 1	32.2 1	39.4 1	39.0 1
SSM-PSI_PSRP	29.5 1	16.2 1	24.8 1	28.1 1	17.6 1	13.4 1	16.1 1	17.2 1	46.8 1	26.5 1	41.3 1	41.0 1
HH-PSI_PSRP	29.6 1	16.1 1	24.1 1	29.8 1	17.5 1	13.6 1	16.5 1	17.9 1	46.0 1	30.2 1	44.3 1	44.0 1
(SCOP)HH_PSRP	35.7 1	19.8 1	27.8 1	35.1 1	22.7 1	16.2 1	21.5 1	23.4 1	54.3 1	34.2 1	50.5 1	50.6 1

T0912o
T0912-D2
FM/TBM H/S
114-154,258-299
max GDTTS:77.71

	FORTE								modified scoring FORTE			
	Query											
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	5.0 2	4.8 13	5.5	5.1 1	6.1 1	5.3 4	5.0 2	5.3 1				
DB_PSSM	5.1 4	5.5 14	6.1	5.2 4	5.5 9	6.5	6.3 1	6.4 1				
SSM-PSI_PSSM	5.7 11	4.9 19	5.1 16	11.0 1	7.3 1	5.1 3	10.5 1	10.6 1				
(PDP)PSI_PSRP	5.1 1	4.5	4.7	4.8 3	4.4 3	0.0	0.0	4.3 1	6.6	7.0	6.9	6.7
DB_PSRP	5.1 15	5.3 3	5.7	4.9 20	6.0	6.0	5.6	5.9	7.2	6.3	7.6	7.9
SSM-PSI_PSRP	6.6 1	4.7 5	5.2 3	12.1 1	7.3 1	4.9	9.8 1	10.2 1	13.6 1	5.9	21.3 1	20.7 1
HH-PSI_PSRP	7.3 1	5.1	4.7 6	10.4 1	9.4 1	5.2	11.1 1	12.0 1	15.9 1	5.7 1	24.7 1	24.6 1
(SCOP)HH_PSRP	4.0 8	4.8	4.0 9	4.7	4.6 19	4.6 6	5.3	5.3	6.4	4.9 4	7.1	6.9

T0912o
T0912-D3
FM H/S
155-257
max GDTTS:41.99

	FORTE								modified scoring FORTE			
	Query											
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	5.5	5.7	5.7	6.3	5.6	5.3	5.4	4.2 20				
DB_PSSM	6.0	6.6	5.9	5.4	6.3	6.1	6.0	5.6				
SSM-PSI_PSSM	5.7	6.5	6.2	6.4	6.6	5.8	6.3	6.5				
(PDP)PSI_PSRP	4.5	4.1	4.4	4.2	4.5	4.1	4.4	4.1	6.0	5.8	5.1 5	6.5
DB_PSRP	5.2	5.4	5.8	5.5	6.0	6.5	4.9 14	6.0	6.6	7.3	5.7 10	7.2
SSM-PSI_PSRP	6.1	5.1	5.3	5.5	6.9	4.4	5.2 13	4.9 15	6.3 6	7.3	6.3 3	6.4 2
HH-PSI_PSRP	5.7	6.1	5.7	4.2 12	4.7 13	5.4	5.1	5.1	7.1 2	6.7	6.3 2	6.1 2
(SCOP)HH_PSRP	4.0	4.9	4.7	4.5	4.8	4.8	4.5	4.0	5.9	6.6	4.8 14	6.6

T0913o
T0913-D1
TBM H/S
49-386
max GDTTS:68.56

Query

	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	40.0 1	26.1 1	33.8 1	32.7 1	34.8 1	25.8 1	31.1 1	32.0 1				
DB_PSSM	27.9 1	38.4 1	26.0 1	23.1 1	22.9 1	32.6 1	22.3 1	22.8 1				
SSM-PSI_PSSM	24.0 1	25.8 1	25.7 1	24.1 1	21.3 1	22.3 1	21.4 1	22.7 1				
(PDP)PSI_PSRP	32.2 1	19.0 1	26.6 1	26.8 1	18.6 1	15.2 1	15.4 1	17.0 1	43.2 1	37.9 1	35.9 1	36.0 1
DB_PSRP	26.8 1	33.0 1	24.8 1	23.7 1	17.0 1	26.3 1	15.3 1	16.5 1	31.9 1	45.6 1	27.2 1	26.9 1
SSM-PSI_PSRP	23.8 1	22.8 1	23.8 1	22.5 1	16.8 1	18.8 1	16.6 1	17.5 1	30.6 1	34.4 1	26.8 1	26.4 1
HH-PSI_PSRP	24.9 1	20.5 1	22.7 1	28.1 1	18.0 1	17.1 1	20.4 1	21.6 1	29.0 1	28.7 1	30.1 1	29.5 1
(SCOP)HH_PSRP	29.7 1	25.7 1	26.9 1	32.0 1	21.9 1	21.6 1	24.8 1	26.1 1	35.3 1	35.0 1	35.2 1	34.6 1

T0914-T0915
T0915-D1
FM H/S
6-159
max GDTTS:52.44

Query

	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	6.0	7.3	5.9	6.0	6.1	7.1	6.2	6.0				
DB_PSSM	6.7	7.3	7.2	6.4	5.7	6.1	6.0	6.8				
SSM-PSI_PSSM	6.0	6.6	6.3	5.7	6.5	5.4 5	6.0	5.9				
(PDP)PSI_PSRP	4.8 1	4.3	4.9 1	4.6 1	5.9 1	5.0 1	5.8 1	5.5 1	6.8	5.7	5.4 12	6.9
DB_PSRP	5.7	5.9	5.5	6.1	6.0	5.2	5.8	5.9	5.8	6.0	5.8	6.2
SSM-PSI_PSRP	5.7	5.5	4.7 13	5.8	6.0	4.7	4.6 7	4.4 11	5.7	5.4	5.8	5.7
HH-PSI_PSRP	4.9	4.9 12	4.8 11	5.1	5.5	4.8 14	4.7 16	4.6 12	4.6 7	6.7	4.3 15	7.1
(SCOP)HH_PSRP	5.3	5.8	5.4	5.6	5.7	5.5	5.7	5.6	4.9	5.1	5.4	5.2

T0917o
T0917-D1
TBM S
19-409
max GDTTS:85.68

Query

	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	51.1 1	52.1 1	51.3 1	48.7 1	48.0 1	48.2 1	44.0 1	45.8 1				
DB_PSSM	35.4 1	39.3 1	38.0 1	33.5 1	29.6 1	32.0 1	27.1 1	28.4 1				
SSM-PSI_PSSM	34.1 1	35.7 1	35.8 1	32.5 1	31.2 1	31.7 1	27.7 1	29.2 1				
(PDP)PSI_PSRP	39.0 1	28.4 1	38.2 1	34.4 1	25.0 1	25.0 1	20.8 1	22.6 1	60.4 1	62.9 1	56.5 1	55.9 1
DB_PSRP	35.7 1	35.3 1	39.1 1	33.9 1	27.0 1	31.1 1	23.9 1	25.4 1	44.0 1	49.7 1	41.9 1	41.2 1
SSM-PSI_PSRP	32.9 1	31.0 1	34.7 1	30.6 1	25.7 1	27.1 1	22.3 1	23.7 1	40.9 1	43.2 1	38.3 1	37.5 1
HH-PSI_PSRP	28.9 1	27.4 1	30.0 1	30.9 1	23.9 1	24.7 1	23.2 1	24.6 1	34.6 1	36.7 1	35.3 1	35.2 1
(SCOP)HH_PSRP	35.6 1	33.9 1	36.3 1	38.1 1	30.7 1	31.6 1	30.2 1	31.7 1	42.8 1	44.8 1	44.3 1	43.8 1

T0921-T0922
T0921-D1
TBM S
5-142
max GDTTS:70.65

Query

	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	11.0 1	10.0 1	7.0 1	7.1 1	12.1 1	11.8 1	8.8 1	8.4 1				
DB_PSSM	18.8 1	23.1 1	12.9 1	12.2 1	20.8 1	26.4 1	15.1 1	14.7 1				
SSM-PSI_PSSM	19.9 1	20.6 1	13.8 1	14.0 1	20.7 1	21.2 1	15.6 1	15.2 1				
(PDP)PSI_PSRP	11.4 1	12.0 1	9.2 1	8.4 1	11.4 1	12.5 1	9.3 1	8.0 1	25.1 1	24.6 1	13.6 1	12.8 1
DB_PSRP	18.5 1	23.3 1	13.0 1	12.6 1	20.6 1	28.6 1	14.1 1	13.1 1	33.2 1	48.0 1	19.8 1	17.9 1
SSM-PSI_PSRP	21.5 1	22.3 1	13.3 1	13.9 1	25.2 1	27.0 1	16.7 1	15.4 1	38.2 1	42.5 1	23.8 1	22.5 1
HH-PSI_PSRP	15.4 1	15.4 1	10.8 1	11.4 1	17.8 1	18.1 1	14.2 1	12.8 1	28.0 1	30.4 1	19.2 1	17.4 1
(SCOP)HH_PSRP	15.4 1	15.6 1	10.9 1	11.5 1	17.7 1	18.2 1	14.6 1	13.1 1	28.4 1	29.3 1	19.4 1	17.9 1

T0921-T0922
T0922-D1
TBM S
23-96
max GDTTS:83.78

Query

	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	25.5 1	18.3 1	13.1 1	22.9 1	23.6 1	20.1 1	19.3 1	20.4 1				
DB_PSSM	27.0 1	25.5 1	17.4 1	26.7 1	27.6 1	27.9 1	24.8 1	26.0 1				
SSM-PSI_PSSM	29.3 1	19.2 1	18.0 1	26.9 1	26.9 1	22.2 1	22.6 1	23.9 1				
(PDP)PSI_PSRP	18.5 1	11.1 1	10.3 1	16.4 1	13.0 1	11.5 1	9.5 1	10.7 1	37.7 1	34.8 1	36.6 1	37.0 1
DB_PSRP	27.9 1	23.1 1	19.6 1	26.8 1	26.3 1	27.1 1	22.4 1	24.0 1	36.6 1	41.2 1	38.7 1	38.3 1
SSM-PSI_PSRP	28.5 1	20.0 1	19.7 1	26.7 1	27.8 1	24.1 1	24.1 1	25.9 1	39.3 1	37.3 1	39.5 1	40.2 1
HH-PSI_PSRP	26.5 1	18.1 1	17.6 1	25.8 1	24.7 1	22.6 1	21.3 1	23.1 1	36.8 1	35.5 1	39.1 1	40.1 1
(SCOP)HH_PSRP	26.0 1	18.1 1	17.6 1	24.7 1	23.1 1	22.2 1	19.3 1	20.8 1	36.8 1	35.6 1	38.8 1	39.6 1

T0929o
T0859-D1
FM H/S
4-132
max GDTTS:28.32

Query

	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	6.2	6.4	6.2	6.4	6.5	5.8	6.4	5.8				
DB_PSSM	6.8	7.1	6.8	7.1	6.7	6.3	6.8	6.3				
SSM-PSI_PSSM	6.8	7.0	6.8	7.0	6.4	6.6	6.4	6.6				
(PDP)PSI_PSRP	4.2	4.3	4.2	4.3	7.6	4.8	7.3	4.8	7.0	5.2	6.8	5.2
DB_PSRP	6.3	6.1	6.3	6.1	7.0	5.0	7.0	5.0	8.6	7.3	8.4	7.3
SSM-PSI_PSRP	6.7	6.4	6.7	6.4	7.7	5.0	7.6	5.0	6.7	6.1	6.9	6.1
HH-PSI_PSRP	5.7	5.9	5.7	5.9	6.6	4.2	6.3	4.2	10.0	5.7	9.6	5.7
(SCOP)HH_PSRP	4.4	4.6	4.4	4.6	6.1	0.0	5.7	0.0	5.3 15	4.3 5	7.0	4.3 5

T0931o
T0887-D1
FM/TBM H/S
16-176
max GDTTS:57.45

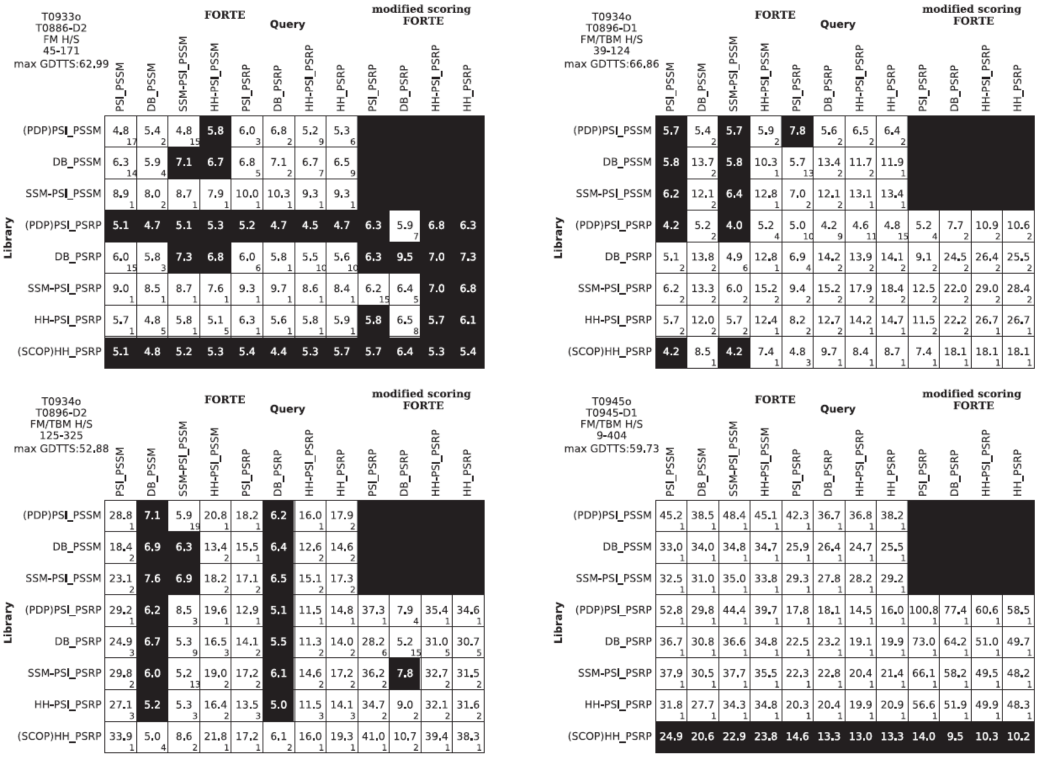
Query

	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	7.1	6.2	7.1	6.8	6.5	5.9	6.3	6.1				
DB_PSSM	12.4 1	14.4 1	11.9 1	12.0 1	11.9 1	13.5 1	11.7 1	11.9 1				
SSM-PSI_PSSM	16.9 1	11.7 1	16.5 1	15.9 1	15.9 1	11.3 1	15.0 1	15.3 1				
(PDP)PSI_PSRP	4.8	4.2	4.8	4.9	4.5	4.8	4.6	4.7	6.9	5.9	6.5	6.1
DB_PSRP	11.2 1	12.4 1	11.5 1	10.3 1	10.0 1	12.0 1	9.2 1	9.8 1	16.3 1	18.2 1	13.8 1	14.0 1
SSM-PSI_PSRP	16.6 1	11.3 1	16.9 1	15.3 1	15.6 1	10.3 1	14.4 1	15.0 1	26.9 1	16.6 1	22.7 1	22.4 1
HH-PSI_PSRP	15.2 1	11.3 1	15.3 1	15.3 1	14.0 1	10.7 1	14.4 1	14.7 1	20.2 1	14.9 1	20.1 1	19.3 1
(SCOP)HH_PSRP	6.0	5.7	6.1	5.8	5.1	5.0	4.6	4.7	4.3	4.4	0.0	4.1

T0933o
T0886-D1
FM H/S
21-44,172-216
max GDTTS:71.01

Query

	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	5.3 1	5.0 1	5.2 1	4.9 3	5.7 3	5.5 1	5.5 1	5.3 2				
DB_PSSM	4.9 7	5.5	5.3 9	4.9 4	5.6 11	5.5 13	6.1 7	5.6 12				
SSM-PSI_PSSM	6.2 1	4.9 14	6.4 2	5.4 3	7.4 1	5.9 7	6.7 1	6.4 1				
(PDP)PSI_PSRP	4.4 2	4.9 1	4.9 1	4.8 1	4.5 3	5.4 2	4.8 1	4.7 1	8.9 1	6.3 11	8.0 1	7.5 1
DB_PSRP	5.6 3	5.3 14	6.1 1	5.7 3	5.7 11	5.6 19	6.2 7	5.5 16	7.1	5.9 1	5.7 17	5.8 19
SSM-PSI_PSRP	6.3 1	5.5 3	5.8 3	5.6 3	7.4 3	6.5 8	6.6 8	6.1 12	7.0	7.2 2	6.5 2	6.3 2
HH-PSI_PSRP	4.9 5	5.2 1	5.7 1	5.7 1	5.7 10	6.1 3	5.9 7	5.6 8	7.1 17	7.3 1	6.5 10	6.7 4
(SCOP)HH_PSRP	4.5 8	4.6 3	5.2 1	4.8 3	5.8 1	5.9 1	5.7 2	5.3 2	8.6	5.9 2	6.0 17	6.3 10



T0933o
T0886-D2
FM H/S
45-171
max GDTTS:62.99

Query

	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	4.8 17	5.4 2	4.8 15	5.8	6.0 3	6.8 2	5.2 9	5.3 6				
DB_PSSM	6.3 14	5.9 4	7.1	6.7	6.8 5	7.1 2	6.7 7	6.5 9				
SSM-PSI_PSSM	8.9 1	8.0 2	8.7 1	7.9 1	10.0 1	10.3 1	9.3 1	9.3 1				
(PDP)PSI_PSRP	5.1	4.7	5.1	5.3	5.2	4.7	4.5	4.7	6.3	5.9 7	6.8	6.3
DB_PSRP	6.0 15	5.8 3	7.3	6.8	6.0 6	5.8 1	5.5 10	5.6 10	6.3	9.5	7.0	7.3
SSM-PSI_PSRP	9.0 1	8.5 1	8.7 1	7.6 1	9.3 1	9.7 1	8.6 1	8.4 1	6.2 15	6.4 5	7.0	6.8
HH-PSI_PSRP	5.7 1	4.8 5	5.8 1	5.1 5	6.3 1	5.6 1	5.8 1	5.9 1	5.8	6.5 8	5.7	6.1
(SCOP)HH_PSRP	5.1	4.8	5.2	5.3	5.4	4.4	5.3	5.7	5.7	6.4	5.3	5.4

T0934o
T0896-D1
FM/TBM H/S
39-124
max GDTTS:66.86

Query

	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	5.7	5.4 2	5.7	5.9 2	7.8	5.6 2	6.5 2	6.4 2				
DB_PSSM	5.8	13.7 2	5.8	10.3 1	5.7 13	13.4 2	11.7 2	11.9 1				
SSM-PSI_PSSM	6.2	12.1 2	6.4	12.8 1	7.0 2	12.1 2	13.1 1	13.4 1				
(PDP)PSI_PSRP	4.2	5.2 2	4.0	5.2 4	5.0 10	4.2 9	4.6 11	4.8 15	5.2 4	7.7 2	10.9 2	10.6 2
DB_PSRP	5.1 2	13.8 2	4.9 6	12.8 1	6.9 4	14.2 2	13.9 2	14.1 2	9.1 2	24.5 2	26.4 2	25.5 2
SSM-PSI_PSRP	6.2 2	13.3 2	6.0 2	15.2 2	9.4 2	15.2 2	17.9 2	18.4 2	12.5 2	22.0 2	29.0 2	28.4 2
HH-PSI_PSRP	5.7 2	12.0 2	5.7 2	12.4 1	8.2 2	12.7 2	14.2 1	14.7 1	11.5 2	22.2 2	26.7 1	26.7 1
(SCOP)HH_PSRP	4.2	8.5 1	4.2	7.4 1	4.8 3	9.7 1	8.4 1	8.7 1	7.4 1	18.1 1	18.1 1	18.1 1

T0934o
T0896-D2
FM/TBM H/S
125-325
max GDTTS:52.88

Query

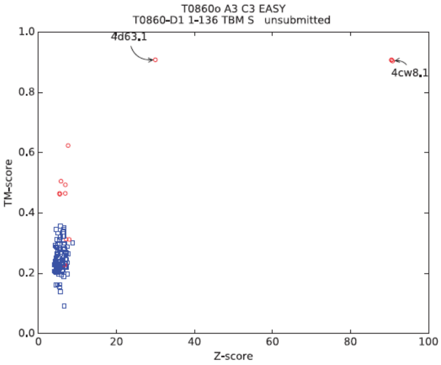
	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	28.8 1	7.1	5.9 19	20.8 1	18.2 1	6.2	16.0 1	17.9 2				
DB_PSSM	18.4 2	6.9	6.3	13.4 2	15.5 1	6.4	12.6 2	14.6 2				
SSM-PSI_PSSM	23.1 2	7.6	6.9	18.2 2	17.1 2	6.5	15.1 2	17.3 2				
(PDP)PSI_PSRP	29.2 1	6.2	8.5 3	19.6 1	12.9 1	5.1	11.5 1	14.8 1	37.3 1	7.9 4	35.4 1	34.6 1
DB_PSRP	24.9 3	6.7	5.3 9	16.5 3	14.1 2	5.5	11.3 2	14.0 2	28.2 6	5.2 15	31.0 5	30.7 5
SSM-PSI_PSRP	29.8 2	6.0	5.2 13	19.0 2	17.2 2	6.1	14.6 2	17.2 2	36.2 2	7.8	32.7 2	31.5 2
HH-PSI_PSRP	27.1 3	5.2	5.3 3	16.4 2	13.5 3	5.0	11.5 3	14.1 3	34.7 2	9.0 2	32.1 2	31.6 2
(SCOP)HH_PSRP	33.9 1	5.0 4	8.6 2	21.8 1	17.2 1	6.1 2	16.0 1	19.3 1	41.0 1	10.7 2	39.4 1	38.3 1

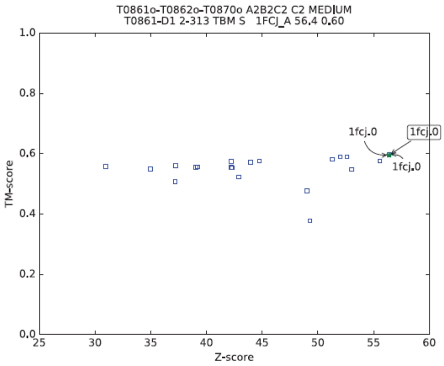
T0945o
T0945-D1
FM/TBM H/S
9-404
max GDTTS:59.73

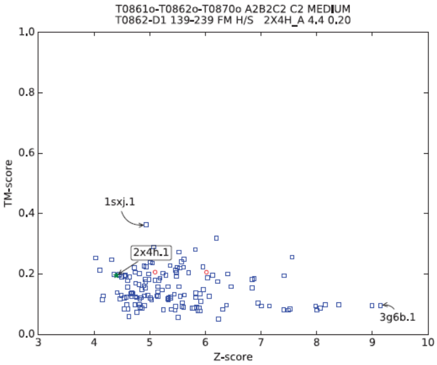
Query

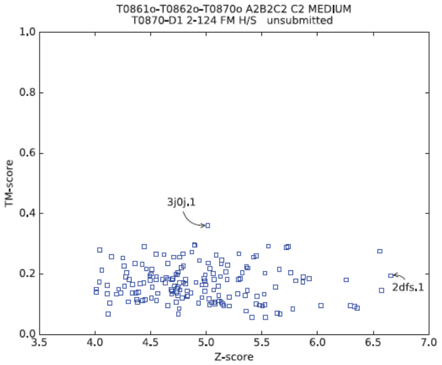
	FORTE								modified scoring FORTE			
Library	PSI_PSSM	DB_PSSM	SSM-PSI_PSSM	HH-PSI_PSSM	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP	PSI_PSRP	DB_PSRP	HH-PSI_PSRP	HH_PSRP
(PDP)PSI_PSSM	45.2 1	38.5 1	48.4 1	45.1 1	42.3 1	36.7 1	36.8 1	38.2 1				
DB_PSSM	33.0 1	34.0 1	34.8 1	34.7 1	25.9 1	26.4 1	24.7 1	25.5 1				
SSM-PSI_PSSM	32.5 1	31.0 1	35.0 1	33.8 1	29.3 1	27.8 1	28.2 1	29.2 1				
(PDP)PSI_PSRP	52.8 1	29.8 1	44.4 1	39.7 1	17.8 1	18.1 1	14.5 1	16.0 1	100.8 1	77.4 1	60.6 1	58.5 1
DB_PSRP	36.7 1	30.8 1	36.6 1	34.8 1	22.5 1	23.2 1	19.1 1	19.9 1	73.0 1	64.2 1	51.0 1	49.7 1
SSM-PSI_PSRP	37.9 1	30.5 1	37.7 1	35.5 1	22.3 1	22.8 1	20.4 1	21.4 1	66.1 1	58.2 1	49.5 1	48.2 1
HH-PSI_PSRP	31.8 1	27.7 1	34.3 1	34.8 1	20.3 1	20.4 1	19.9 1	20.9 1	56.6 1	51.9 1	49.9 1	48.3 1
(SCOP)HH_PSRP	24.9	20.6	22.9	23.8	14.6	13.3	13.0	13.3	14.0	9.5	10.3	10.2

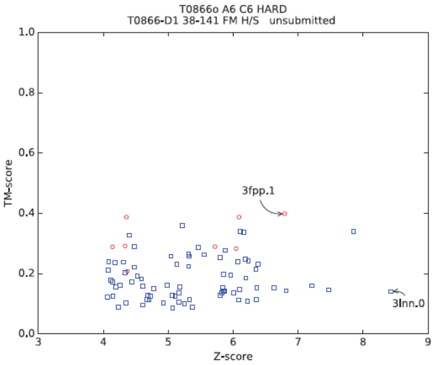
Figure A1 Each row corresponds to individual template libraries. Each column represents a type of query profile that we used. Values in cells show Z-scores which “correct” templates were detected by each combination.

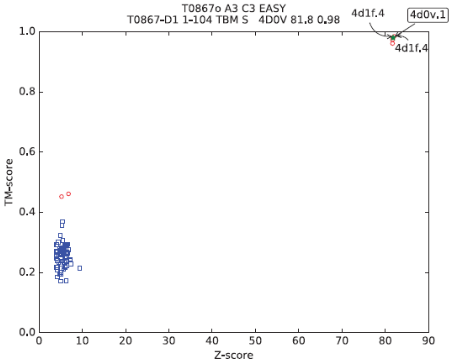

T0860o A3 C3 EASY
T0860-D1 1-136 TBM S   unsubmitted

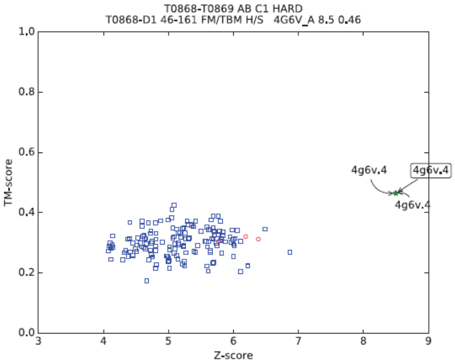

T0861o-T0862o-T0870o A2B2C2 C2 MEDIUM
T0861-D1 2-313 TBM S   1FCJ_A 56.4 0.60

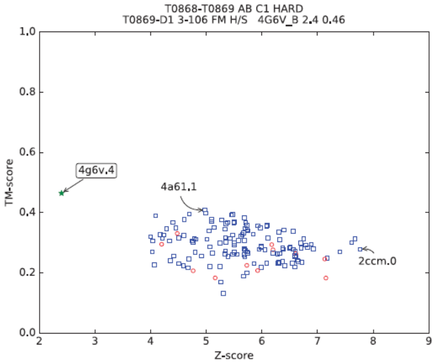

T0861o-T0862o-T0870o A2B2C2 C2 MEDIUM
T0862-D1 139-239 FM H/S   2X4H_A 4.4 0.20

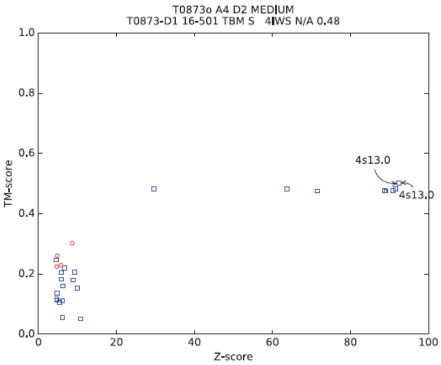

T0861o-T0862o-T0870o A2B2C2 C2 MEDIUM
T0870-D1 2-124 FM H/S   unsubmitted

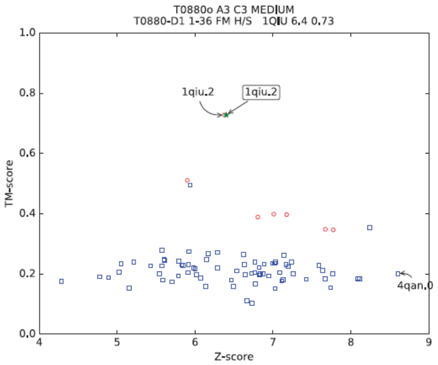

T0866o A6 C6 HARD
T0866-D1 38-141 FM H/S   unsubmitted

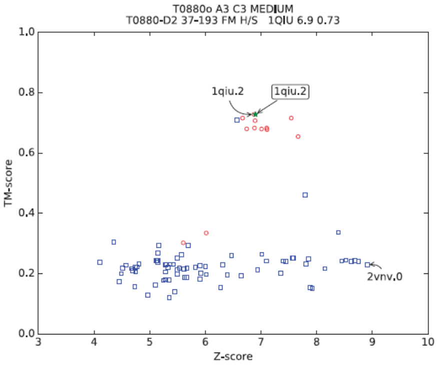

T0867o A3 C3 EASY
T0867-D1 1-104 TBM S   4D0V 81.8 0.98

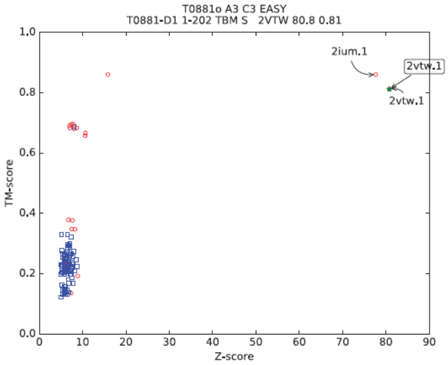

T0868-T0869 AB C1 HARD
T0868-D1 46-161 FM/TBM H/S   4G6V_A 8.5 0.46

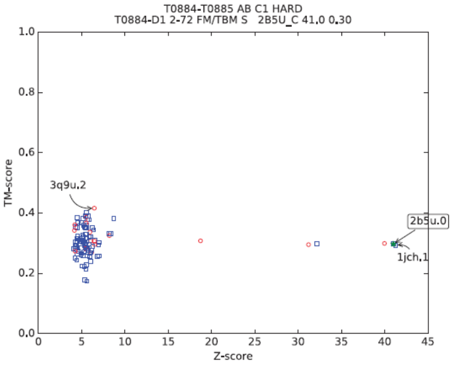

T0868-T0869 AB C1 HARD
T0869-D1 3-106 FM H/S   4G6V_B 2.4 0.46

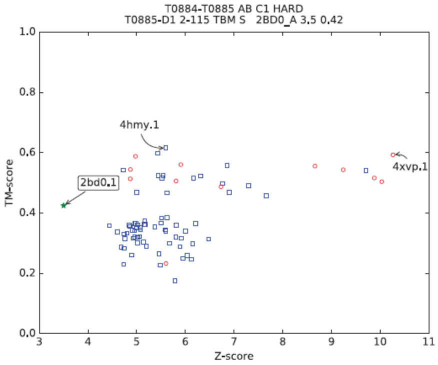

T0873o A4 D2 MEDIUM
T0873-D1 16-501 TBM S   4IWS N/A 0.48
TM-score
Z-score
4s13.0
4s13.0
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100

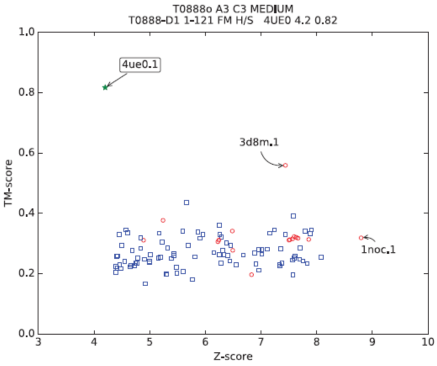

T0880o A3 C3 MEDIUM
T0880-D1 1-36 FM H/S   1QIU 6.4 0.73
TM-score
Z-score
1qiu.2
1qiu.2
4qan.0
0.0
0.2
0.4
0.6
0.8
1.0
4
5
6
7
8
9

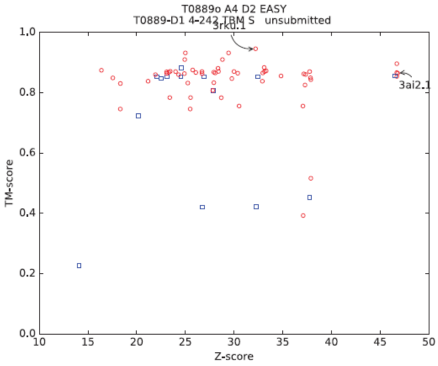

T0880o A3 C3 MEDIUM
T0880-D2 37-193 FM H/S   1QIU 6.9 0.73
TM-score
Z-score
1qiu.2
1qiu.2
2vnv.0
0.0
0.2
0.4
0.6
0.8
1.0
3
4
5
6
7
8
9
10

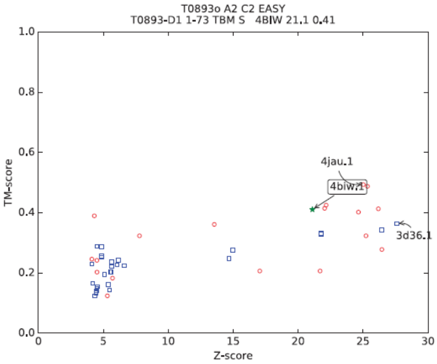

T0881o A3 C3 EASY
T0881-D1 1-202 TBM S   2VTW 80.8 0.81
TM-score
Z-score
2ium.1
2vtw.1
2vtw.1
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
70
80
90

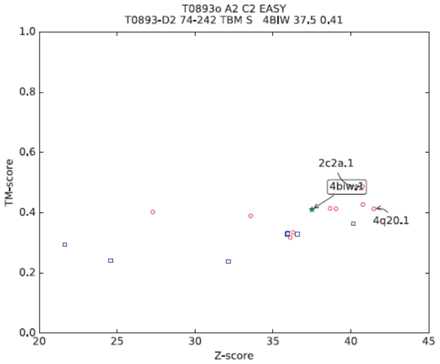

T0884-T0885 AB C1 HARD
T0884-D1 2-72 FM/TBM S   2B5U_C 41.0 0.30
TM-score
Z-score
3q9u.2
2b5u.0
1jch.1
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
35
40
45

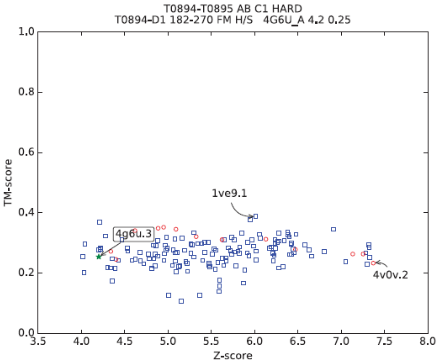

T0884-T0885 AB C1 HARD
T0885-D1 2-115 TBM S   2BD0_A 3.5 0.42
TM-score
Z-score
4hmy.1
2bd0.1
4xvp.1
0.0
0.2
0.4
0.6
0.8
1.0
3
4
5
6
7
8
9
10
11

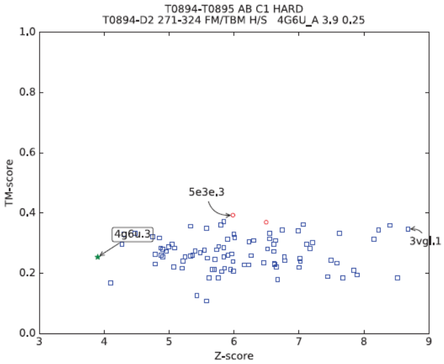

T0888o A3 C3 MEDIUM
T0888-D1 1-121 FM H/S   4UE0 4.2 0.82
TM-score
Z-score
4ue0.1
3d8m.1
1noc.1
0.0
0.2
0.4
0.6
0.8
1.0
3
4
5
6
7
8
9
10

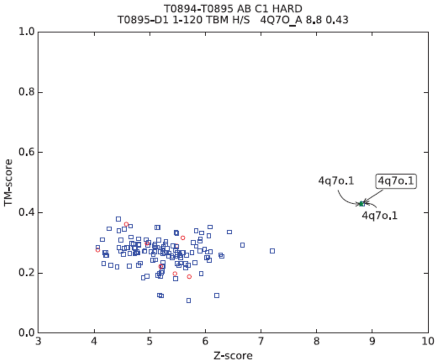

T0889o A4 D2 EASY
T0889-D1 4-242 TBM S   unsubmitted
TM-score
Z-score
3rku.1
3ai2.1
0.0
0.2
0.4
0.6
0.8
1.0
10
15
20
25
30
35
40
45
50

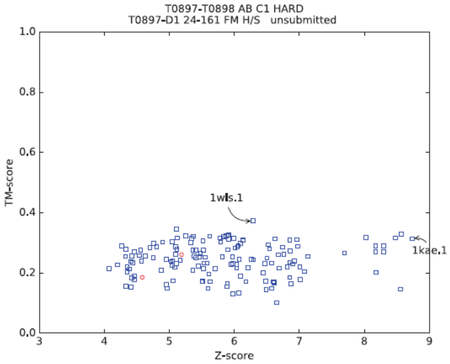

T0893o A2 C2 EASY
T0893-D1 1-73 TBM S 4BIW 21.1 0.41
TM-score
Z-score
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
4jau.1
4biw.1
3d36.1

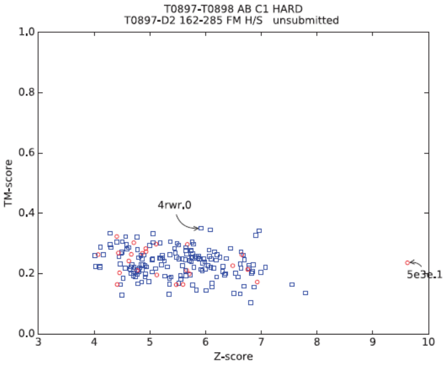

T0893o A2 C2 EASY
T0893-D2 74-242 TBM S 4BIW 37.5 0.41
TM-score
Z-score
1.0
0.8
0.6
0.4
0.2
0.0
20
25
30
35
40
45
2c2a.1
4biw.1
4q20.1

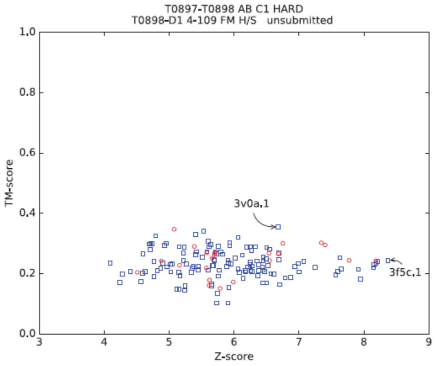

T0894-T0895 AB C1 HARD
T0894-D1 182-270 FM H/S 4G6U_A 4.2 0.25
TM-score
Z-score
1.0
0.8
0.6
0.4
0.2
0.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
1ve9.1
4g6u.3
4v0v.2


T0894-T0895 AB C1 HARD
T0894-D2 271-324 FM/TBM H/S 4G6U_A 3.9 0.25
TM-score
Z-score
1.0
0.8
0.6
0.4
0.2
0.0
3
4
5
6
7
8
9
5e3e.3
4g6u.3
3vgl.1

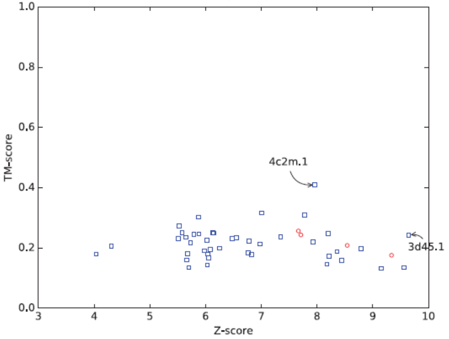

T0894-T0895 AB C1 HARD
T0895-D1 1-120 TBM H/S 4Q7O_A 8.8 0.43
TM-score
Z-score
1.0
0.8
0.6
0.4
0.2
0.0
3
4
5
6
7
8
9
10
4q7o.1
4q7o.1
4q7o.1


T0897-T0898 AB C1 HARD
T0897-D1 24-161 FM H/S unsubmitted
TM-score
Z-score
1.0
0.8
0.6
0.4
0.2
0.0
3
4
5
6
7
8
9
1wls.1
1kae.1

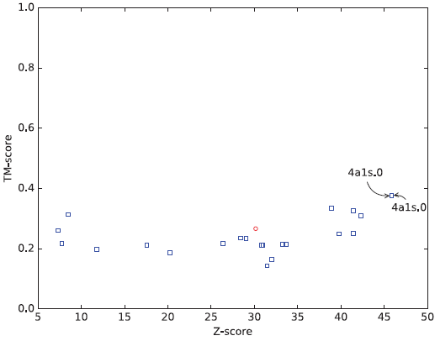

T0897-T0898 AB C1 HARD
T0897-D2 162-285 FM H/S unsubmitted
TM-score
Z-score
1.0
0.8
0.6
0.4
0.2
0.0
3
4
5
6
7
8
9
10
4rwr.0
5e3e.1


T0897-T0898 AB C1 HARD
T0898-D1 4-109 FM H/S unsubmitted
TM-score
Z-score
1.0
0.8
0.6
0.4
0.2
0.0
3
4
5
6
7
8
9
3v0a.1
3f5c.1

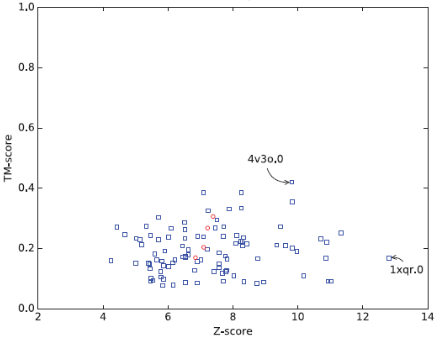
T0897-T0898 AB C1 HARD
T0898-D2 110-164 FM/TBM H/S   unsubmitted



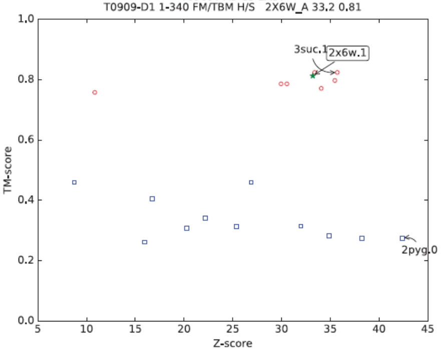
T0903o-T0904o A2B2 C2 MEDIUM
T0903-D1 15-350 TBM S   unsubmitted



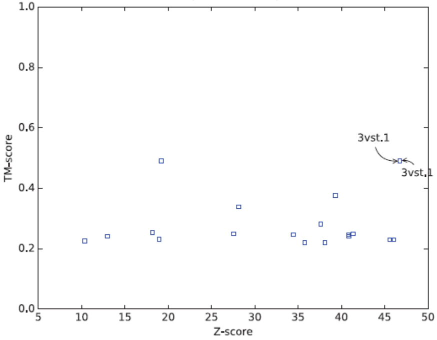
T0903o-T0904o A2B2 C2 MEDIUM
T0904-D1 61-311 FM H/S   unsubmitted



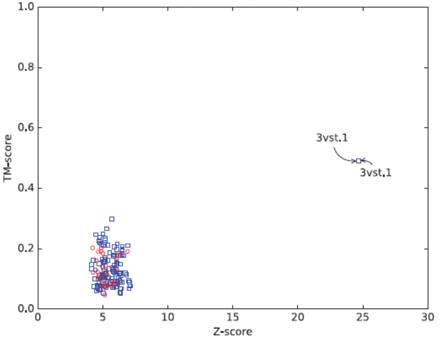
T0909o A3 C3 EASY
T0909-D1 1-340 FM/TBM H/S   2X6W_A 33.2 0.81



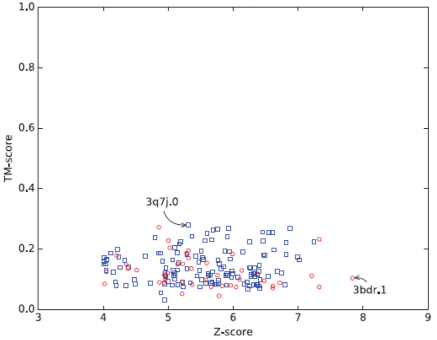
T0912o A2 C2 HARD
T0912-D1 24-113,299-622 TBM H/S   unsubmitted



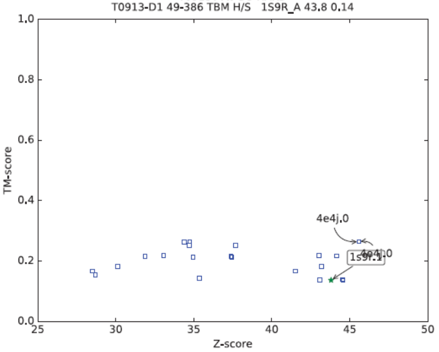
T0912o A2 C2 HARD
T0912-D2 114-154,258-299 FM/TBM H/S   unsubmitted

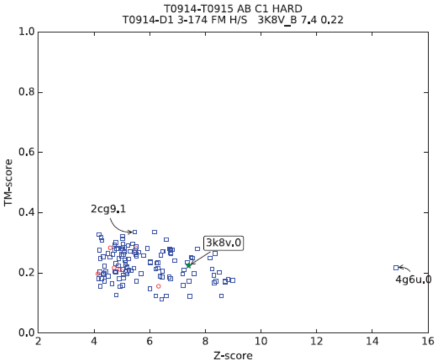


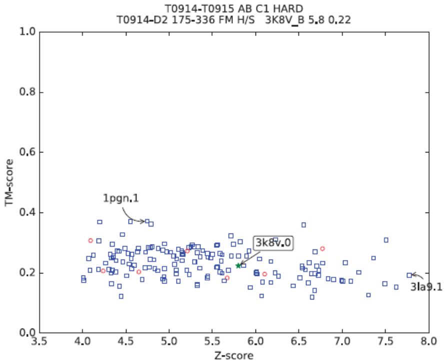
T0912o A2 C2 HARD
T0912-D3 155-257 FM H/S   unsubmitted

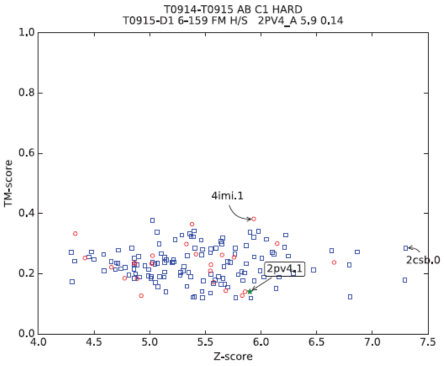


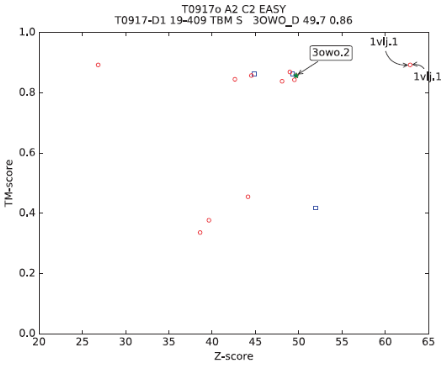
T0913o A6 D3 HARD
T0913-D1 49-386 TBM H/S   1S9R_A 43.8 0.14

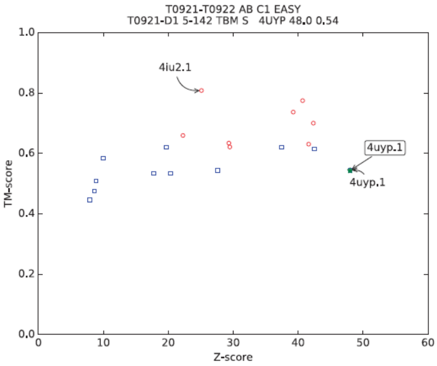


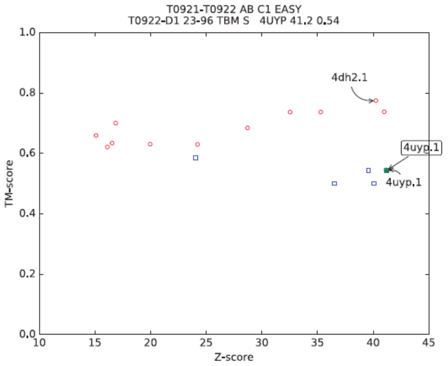

T0914-T0915 AB C1 HARD
T0914-D1 3-174 FM H/S   3K8V_B 7.4 0.22
TM-score
Z-score
1.0
0.8
0.6
0.4
0.2
0.0
2
4
6
8
10
12
14
16
2cg9.1
3k8v.0
4g6u.0

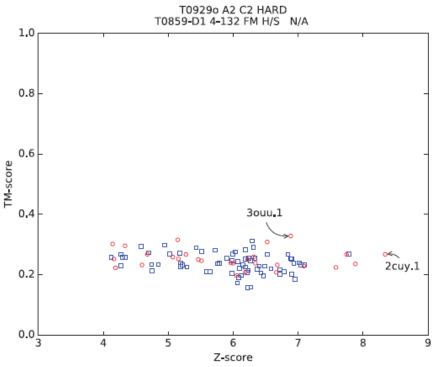

T0914-T0915 AB C1 HARD
T0914-D2 175-336 FM H/S   3K8V_B 5.8 0.22
TM-score
Z-score
1.0
0.8
0.6
0.4
0.2
0.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
1pgn.1
3k8v.0
3la9.1

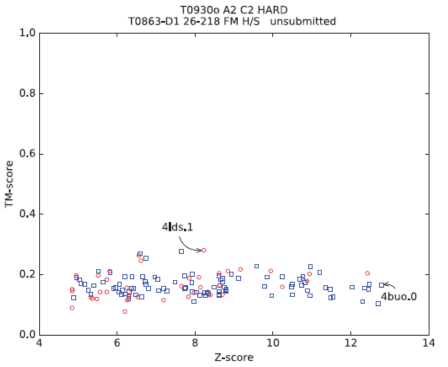

T0914-T0915 AB C1 HARD
T0915-D1 6-159 FM H/S   2PV4_A 5.9 0.14
TM-score
Z-score
1.0
0.8
0.6
0.4
0.2
0.0
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
4imi.1
2pv4.1
2csb.0

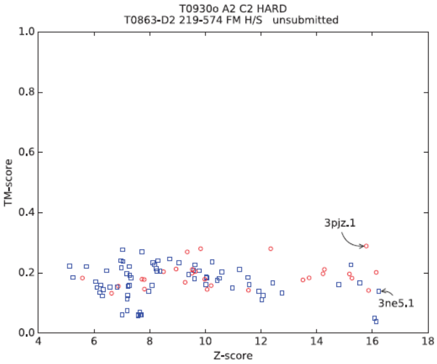

T0917o A2 C2 EASY
T0917-D1 19-409 TBM S   3OWO_D 49.7 0.86
TM-score
Z-score
1.0
0.8
0.6
0.4
0.2
0.0
20
25
30
35
40
45
50
55
60
65
1vlj.1
3owo.2
1vlj.1

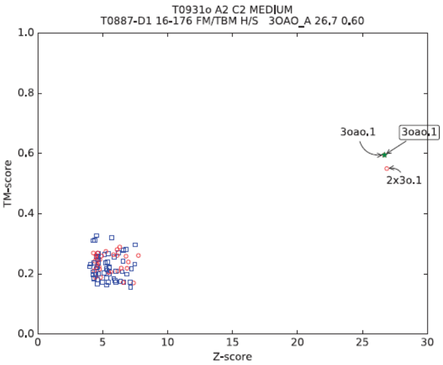

T0921-T0922 AB C1 EASY
T0921-D1 5-142 TBM S   4UYP 48.0 0.54
TM-score
Z-score
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
4iu2.1
4uyp.1
4uyp.1

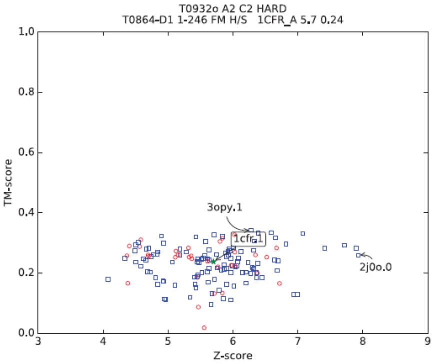

T0921-T0922 AB C1 EASY
T0922-D1 23-96 TBM S   4UYP 41.2 0.54
TM-score
Z-score
1.0
0.8
0.6
0.4
0.2
0.0
10
15
20
25
30
35
40
45
4dh2.1
4uyp.1
4uyp.1

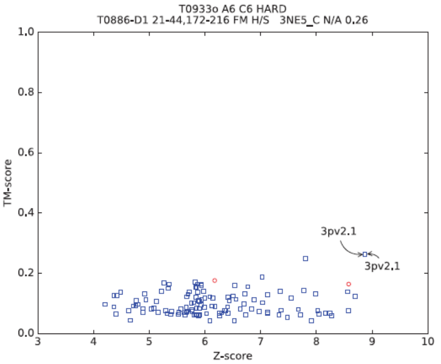

T0929o A2 C2 HARD
T0859-D1 4-132 FM H/S   N/A
TM-score
Z-score
1.0
0.8
0.6
0.4
0.2
0.0
3
4
5
6
7
8
9
3ouu.1
2cuy.1

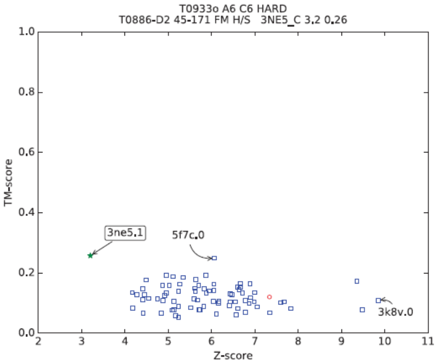

T0930o A2 C2 HARD
T0863-D1 26-218 FM H/S   unsubmitted
TM-score
Z-score
1.0
0.8
0.6
0.4
0.2
0.0
4
6
8
10
12
14
4lds.1
4buo.0

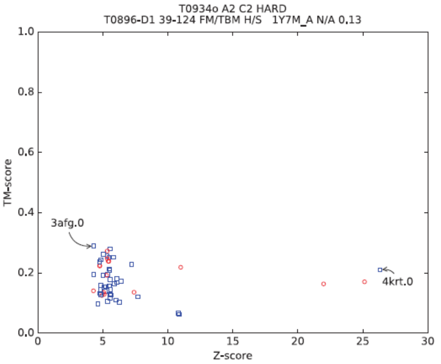

T0930o A2 C2 HARD
T0863-D2 219-574 FM H/S unsubmitted
TM-score
Z-score
3pjz.1
3ne5.1
0.0
0.2
0.4
0.6
0.8
1.0
4
6
8
10
12
14
16
18

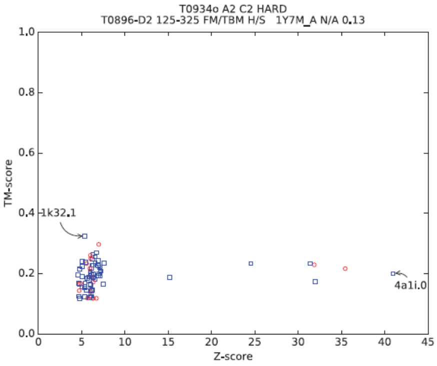

T0931o A2 C2 MEDIUM
T0887-D1 16-176 FM/TBM H/S 3OAO_A 26.7 0.60
TM-score
Z-score
3oao.1
3oao.1
2x3o.1
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30

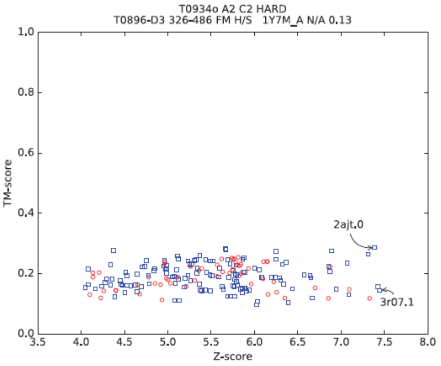

T0932o A2 C2 HARD
T0864-D1 1-246 FM H/S 1CFR_A 5.7 0.24
TM-score
Z-score
3opy.1
1cfr.1
2j0o.0
0.0
0.2
0.4
0.6
0.8
1.0
3
4
5
6
7
8
9


T0933o A6 C6 HARD
T0886-D1 21-44,172-216 FM H/S 3NE5_C N/A 0.26
TM-score
Z-score
3pv2.1
3pv2.1
0.0
0.2
0.4
0.6
0.8
1.0
3
4
5
6
7
8
9
10


T0933o A6 C6 HARD
T0886-D2 45-171 FM H/S 3NE5_C 3.2 0.26
TM-score
Z-score
3ne5.1
5f7c.0
3k8v.0
0.0
0.2
0.4
0.6
0.8
1.0
2
3
4
5
6
7
8
9
10
11


T0934o A2 C2 HARD
T0896-D1 39-124 FM/TBM H/S 1Y7M_A N/A 0.13
TM-score
Z-score
3afg.0
4krt.0
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30


T0934o A2 C2 HARD
T0896-D2 125-325 FM/TBM H/S 1Y7M_A N/A 0.13
TM-score
Z-score
1k32.1
4a1i.0
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
35
40
45


T0934o A2 C2 HARD
T0896-D3 326-486 FM H/S 1Y7M_A N/A 0.13
TM-score
Z-score
2ajt.0
3r07.1
0.0
0.2
0.4
0.6
0.8
1.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0

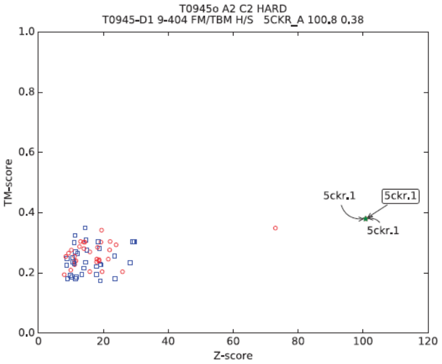


Figure A2 Plots of TM-scores vs. the highest Z-scores of templates. The horizontal axis shows Z-score of an alignment between a target domain sequence and a template sequence in PDB. We show the highest Z-score when the same template was identified within the top five hits using different profile-profile 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 serial number of biological assembly defined in the PDB. We gave 0 for an asymmetric unit.

Table A1: 138 binding pockets used as *Ito138* dataset.

chainID: Chain identifier.
resSeqNum: Residue sequence number.
altLoc: Alternate location indicator.

ligand	PDBID	chainID	resSeqNum	altLoc
FAD	3OF4	C	250	
FAD	3NVW	B	606	
FAD	2X8H	A	1594	
FAD	2GAG	B	501	
FAD	1F20	A	1501	
FAD	3GWN	A	334	
FAD	1ZR6	A	501	
FAD	3AN1	A	3006	
FAD	1OWL	A	485	
FAD	1GPE	A	600	
FAD	2YWL	A	1001	
FAD	1YOA	A	401	
FAD	1O26	A	615	
FAD	2DW4	A	1001	
NAD	1DQS	A	400	
NAD	1AD3	A	600	
NAD	3M6I	A	501	
NAD	2PH5	A	501	
NAD	1Z45	A	703	
NAD	3C7A	A	405	
NAI	3MW9	A	603	
NAD	1OBB	A	500	
NAD	1LW7	A	601	
NAP	3NRR	A	515	A
NAP	1PS9	A	703	
NAP	2O7S	A	1411	

NAP	1SUW	A	3075	
NAP	2D1C	A	1002	
NAP	2AZN	A	2001	
ATP	2Q0D	A	501	
ATP	1RDQ	E	600	B
ATP	3CIS	A	1101	
ATP	1A0I	A	1	
ATP	3C5E	A	801	
ATP	1KVK	A	535	
ATP	3FKQ	A	500	
ATP	2OLR	A	541	
ATP	1QHH	A	700	
ATP	3OPY	B	942	
ATP	2A5Y	B	551	
ATP	1B76	A	1552	
ATP	1SVM	A	800	
ATP	2NPI	A	600	
ATP	1HP1	A	606	
ATP	2P09	A	500	
ADP	2IUU	A	1723	
ADP	3A1D	A	997	
ADP	1F9V	A	998	
ADP	3KH5	A	281	
ADP	2WHX	A	1619	
ADP	2CVX	A	1002	
ADP	2BVC	A	501	
ADP	3VIU	A	800	
ADP	1X6V	B	800	
ADP	1M15	A	400	
ADP	2ZPA	A	800	
ADP	1HTW	A	560	
ADP	1IHU	A	590	
ADP	1IOW	A	310	

ADP	2C9O	A	1450
ADP	2R6F	A	1000
ADP	2HYD	A	700
ADP	1R6B	X	780
GTP	1A8R	A	401
GTP	2FH5	B	301
GTP	2QV6	A	300
GTP	1C4K	A	999
GTP	2DY1	A	700
GDP	3LVR	E	737
GDP	2PHN	A	2696
GDP	2HEK	A	401
GDP	4AC9	A	1469
GDP	3CB2	A	500
GDP	2ZEJ	A	1
GDP	2E87	A	400
GDP	3DM5	A	501
GDP	1VJJ	A	1004
GDP	1TQ4	A	500
GDP	2RCN	A	600
GDP	1MKY	A	500
GDP	3D45	A	652
GDP	2HCJ	A	999
GDP	2NZX	A	3001
GDP	3Q5D	A	3850
GDP	1VJ7	A	998
GLC	1Y4C	A	371
GLC	1V2B	A	1203
GLC	2MPR	A	429
GLC	2X42	A	1722
GLC	1AC0	A	617
GLC	2BVL	A	1546
GLC	1CZA	N	918

GLC	1EU1	A	2003	
GLC	2F2E	B	401	
GLC	2CN3	A	1769	
MAN	1OFL	A	507	
MAN	2GUD	A	122	A
MAN	1QMO	A	302	
MAN	3AIH	A	302	
GAL	3DH4	A	701	
GAL	2G7C	A	2	
GAL	2VU9	A	2301	
GAL	1C4Q	A	191	
GAL	1J8R	A	203	
GAL	1G1T	A	602	
GAL	2WNF	A	1346	
GAL	3EF2	A	295	
GAL	2WT0	A	1690	
HEM	2KIL	A	182	
HEM	1QHU	A	500	
HEM	3HX9	A	200	
HEM	3OV0	A	601	
HEM	1FGJ	A	547	
HEM	3LF5	A	1	A
HEM	2CZS	A	500	
HEM	1J0P	A	1001	
HEM	1DW0	A	113	
HEM	1PL3	A	401	
HEM	3NT1	A	619	
HEM	1FFT	A	1001	
HEM	3A15	A	354	
HEM	1V9Y	A	1140	
HEM	2FW5	A	803	
HEM	1ASH	A	301	
HEM	1IZO	A	501	

SF4	8ACN	A	999
SF4	1SU8	A	637
SF4	3LZD	A	343
SF4	1OLT	A	500
SF4	2JH3	A	650
SF4	1HUX	A	290
SF4	1U8V	A	491
SF4	3N5N	X	400
SF4	3A38	A	84
CA	1OAC	A	802
CA	3BWX	A	285
CA	2QM3	A	350
CA	1MIO	B	492