DINE: A Novel Score Function for Modeling Multidomain Protein Structures with Domain Linker and Interface Restraints

SATORU HIRAKO¹ MASAFUMI SHIONYU^{2,a)}

Received: November 26, 2011, Accepted: January 5, 2012, Released: April 19, 2012

Abstract: The functional sites of multidomain proteins are often found at the interfaces of two or more domains. Therefore, the spatial arrangement of the domains is essential in understanding the functional mechanisms of multidomain proteins. However, an experimental determination of the whole structure of a multidomain protein is often difficult due to flexibility in inter-domain arrangement. We have developed a score function, named DINE, to detect probable docking poses generated in a rigid-body docking simulation. This score function takes into account the binding energy, information about the domain interfaces of homologous proteins, and the end-to-end distance spanned by the domain linker. We have examined the performance of DINE on 55 non-redundant known structures of two-domain proteins. In the results, the near-native docking poses were scored within the top 10 in 65.5% of the test cases. DINE scored the near-native poses higher in comparison with an existing domain assembly method, which also used binding energy and linker distance restraints. The results demonstrate that the domain-interface restraints of DINE are quite efficient in selecting near-native domain assemblies.

Keywords: protein structure prediction, protein-protein docking, domain interface, domain linker, reranking

1. Introduction

More than half of the genes in eukaryotic and prokaryotic genomes encode multidomain proteins that are composed of multiple well-folded regions (domains) connected by linkers [1]. In many cases, domains serve as functional units, and cooperation between domains is important for the functions of multidomain proteins [2]. Therefore, the spatial arrangement of domains is essential in understanding the functional mechanisms of multidomain proteins. When a multidomain protein is composed of domains that stably interact with each other, the experimental determination of the whole structure would be expected. However, because the domains in a multidomain protein frequently interact only weakly, the experimental determination of the whole structure of a multidomain protein is often difficult. Consequently, multidomain proteins have been dissected into individual domains for experimental structure determination. Because such dissected domain structures represent a considerable part of the current structure database, efficient computational methods to assemble domain structures into whole multidomain protein structures are required.

Wollacott et al. developed a domain assembly method that was based on the structure predictions of domain linkers. In this method, domain linker structures were predicted by an ab initio

method, and the domains were placed at the N- and C-terminals of the linkers [3]. Near-native assemblies were constructed in 50% of 76 two-domain proteins with this method. The most popular method of the domain assembly approaches is rigid-body docking of domain structures. Inbar et al. generated domain assemblies of three multidomain proteins with a combinatorial docking algorithm, CombDock, and found near-native assemblies within the top 10 docking poses in all three cases [4]. Lise et al. developed the domain docking method using a pair-wise residue contact function based on structural, physicochemical, and evolutionary information [5]. It was shown that the predicted inter-domain contacts could detect the best assembly model among a set of optimal solutions generated by a standard docking procedure in 13 out of 20 cases. Cheng et al. developed pyDockTET for modeling two-domain proteins [6]. This program executed rigid-body docking to generate domain-domain poses, which were further scored by the binding energy and the pseudo-energy derived from the linker end-to-end distances. The near-native assemblies were detected within the top 10 poses for more than 60% of 51 two-domain proteins. They reported that the combination of energy scoring and linker-based restraints was useful for modeling domain assemblies.

It was shown that the spatial arrangements of domains were conserved among homologous proteins (> 30% sequence identity) [7]. Korkin et al. also revealed that the locations of the binding sites for protein-ligands were often conserved, irrespective of the folds of their binding partners [8]. Based on these observations, they obtained eight near-native domain assemblies for nine

¹ Graduate School of Bioscience, Nagahama Institute of Bio-Science and Technology, Nagahama, Shiga 526-0829, Japan

² Faculty of Bioscience, Nagahama Institute of Bio-Science and Technology, Nagahama, Shiga 526-0829, Japan

^{a)} m.shionyu@nagahama-i-bio.ac.jp

multidomain proteins through comparative patch analysis, which was a combined method involving protein docking and comparative modeling on complex templates [9]. Therefore, in addition to the linker-based restraints, homology-based interface predictions might be effective in selecting near-native domain assemblies from rigid-body docking poses.

We have developed a new score function, DINE, for modeling domain assemblies. DINE is the linear sum of a docking energy score, a homology-based domain interface score, and a linker end-to-end distance score. The scores were calculated for domain docking poses obtained from a conventional rigid-body docking program, ZDOCK [10]. The docking energy score was based on that obtained from the ZRANK program [11], which re-ranked docking poses with the detailed binding energy. The domain interface score was proportional to the number of predicted domain interface residues that were actually involved in the interface of each docking pose. The end-to-end distance score was the deviation of the end-to-end distance of the domain terminals in each docking pose from the average distance of linkers of equal length observed in the known multidomain protein structures. We evaluated the score function against the same set of test proteins used in the previous pyDockTET study. The new function detected the correct domain assemblies within the top 10 poses in 65.5% of the test cases and tended to score near-native structures higher in comparison with pyDockTET.

2. Materials & Methods

2.1 Construction of a Domain Interface Database (DiD) for Interface Residue Prediction

For the prediction of domain interface residues, we constructed a non-redundant domain interface database, called the DiD. In this study, the definition of domains in the SCOP database (release 1.75) was adopted [12]. First, the protein structure data that had SCOP definitions were retrieved from the PDB [13]. Single-domain and small (< 50 residues) proteins were removed. The proteins were dissected into domains according to the SCOP definition. All of the non-sequential domains were removed. Then, with the CD-HIT program [14], the amino acid sequences of the domains were clustered with the threshold of 80% sequence identity. The cluster-representative domains were selected from each cluster, primarily based on the resolution and the fraction of disordered residues. The cluster-representative domains had been determined to higher than 3.0-Å resolution and contained the least disordered residues in the cluster. A total of 5,654 domains were selected. Finally, the domain interface residues of the selected domains were determined. If the accessible surface area (ASA) of residues in a domain was reduced by more than 10 Å² in the complete structures with the other domain, they were defined as interface residues. The ASA calculations were executed with an in-house implementation of the algorithm of Lee and Richards [15]. For the proteins with more than two domains, the domain interface residues were defined for each binary domain combination. If no interface residue was detected, the domain was discarded. The current DiD contains 7,040 domain interface data derived from 5,630 domains.

2.2 Prediction of Domain Interface Residues by the KIP Method

To predict domain interface residues using the DiD, we have developed the KIP (Known Interface residue's Positions) method. Using an amino acid sequence of a query domain, a BLAST [16] search was performed against the DiD with an E-value smaller than 0.01. The most similar one among the detected sequences with $\leq T_{iden}\%$ pair-wise sequence identity and with $\leq 90\%$ sequence coverage was aligned with the query domain sequence. The amino acid residues of the query domain that aligned with the domain interface residues of the DiD entry were assigned to be domain interface residues.

To determine the threshold value of pair-wise sequence identity (T_{iden}), we prepared a dataset of 2-domain structures. The current DiD contains 4,190 domains from 3,066 two-domain structures. If both of the two domains were allocated on a single peptide chain in the DiD (i.e., neither of the two domains was discarded in the selection process described above), the amino acid sequence of the chain was used for the dataset construction. As a result, 1,662 two-domain structures were retained in the dataset. The chain sequences were clustered with the threshold of 30% sequence identity with the BLASTCLUST program [16]. From each cluster, we selected the cluster-representative structure that contained the least disordered residues in the cluster, and a total of 2,232 domains were obtained from the two-domain structures. The interface residues of these domains were predicted by the KIP method with a leave-one-out cross validation, that is, when a domain was applied to the KIP method, the query domain was excluded from the DiD. The prediction performance of the KIP method was evaluated with precision, recall, and F-measure defined as

$$precision = \frac{TP}{TP + FP}, \quad (1)$$

$$recall = \frac{TP}{TP + FN}, \quad (2)$$

$$F - measure = \frac{2 \cdot precision \cdot recall}{precision + recall}, \quad (3)$$

where TP (true positive) is the number of the correctly predicted true interface residues, and FP (false positive) and FN (false negative) are the numbers of false pickups and overlooks of interface residues, respectively.

2.3 Analysis of End-to-end Distance between the Domains

Because the domain linker was not always clearly defined in the SCOP database, the domain linker was defined, in this study, as the region between the last residue of the last secondary structure of the N-terminal domain and the first residue of the first secondary structure of the C-terminal domain. The end-to-end distance between the domains was the distance between the preceding and the following $C\alpha$ atoms of a defined linker. The secondary structures were identified using the DSSP program [17].

The domain linkers and end-to-end distances were defined for 1,662 two-domain structures in the DiD, which were described in the previous section. Because the amino acid sequences and/or conformations of the domain linkers could be variable among

highly similar proteins, a higher clustering threshold of 80% was adopted for the statistics. The average and the standard deviation of the end-to-end distance were calculated for each linker length (residue size).

2.4 Score Function for Modeling

We generated 2,000 rigid-body docking poses with ZDOCK version 3.0.1 for each domain assembly [18]. For 1,000 poses, the N- and C-terminal domains were used as the receptor and the ligand, respectively. For the remaining 1,000 poses, the receptor and the ligand were swapped. All of the docking poses were scored by the DINE score function, which was a weighted linear sum of the docking score (S_{dock}), the domain interface score (S_{int}), and the end-to-end distance score (S_{ete}) as

$$Score = w_{dock}S_{dock} + w_{int}S_{int} + w_{ete}S_{ete}, \quad (4)$$

where w_{dock} , w_{int} , and w_{ete} were the weights of the docking score, the domain interface score, and the end-to-end distance score, respectively.

Docking score: the docking score was derived with ZRANK [11]. ZRANK is a re-ranking program for docking poses using a scoring function that is a weighted linear sum of the van der Waals attractive and repulsive energies, the electrostatic short- and long-range attractive and repulsive energies, and the desolvation energy. The polar hydrogen atoms were added to domains using the HBPlus program [19]. The docking score (S_{dock}) was defined as

$$S_{dock} = \frac{-zr + max_{zr}}{max_{zr} - min_{zr}}, \quad (5)$$

where zr was the ZRANK score of a docking pose and max_{zr} and min_{zr} were the maximum and the minimum of the ZRANK scores of 2,000 docking poses, respectively.

Domain interface score: The domain interface residues of two query domains were predicted using the KIP method described above. When a homologous protein with more than one domain interface (having more than two domains) was detected in the database, one of the interfaces was selected so that the sequential order of the domains would be consistent with the query domain, i.e., if the query was the N-terminal domain, the interface of the homologue toward the C-terminal domain was selected.

Next, the contacting residues were identified for all docking poses. The contacting residues were those with a $C\alpha$ atom at a distance $\leq 8.0 \text{ \AA}$ from the $C\alpha$ atoms of any residue of the other domain. For each domain, the ratio of the number of predicted interface residues, which were identified as the contact residues in the docking pose, to the total number of predicted interface residues was calculated. The mean of the ratios for two query domains was defined as the domain interface score (S_{int}).

End-to-end distance score: To identify the linker length (in number of residues) between query domains, we searched the UniProtKB/Swiss-Prot [20] or NCBI (the National Center for Biotechnology Information) nr database for the full-length sequence using the BLAST program. The query domains were mapped to the corresponding full-length sequence. Next, the end-to-end distances of the query domains were calculated for all of

the docking poses. If the number of residues in a linker (L) was ≤ 26 residues, the end-to-end distance score (S_{end}) was defined as:

$$S_{end} = \begin{cases} 1, & \text{if } |d_e - m_e(L)| \leq S D_e(L), \\ 2 - \frac{|d_e - m_e(L)|}{S D_e(L)}, & \text{if } S D_e(L) < |d_e - m_e(L)| \leq 2S D_e(L), \\ 0, & \text{if } |d_e - m_e(L)| > 2S D_e(L), \end{cases} \quad (6)$$

where d_e was the end-to-end distance in the docking pose, and $m_e(L)$ and $S D_e(L)$ were the average and standard deviation of the end-to-end distances of linkers of length L . If $L > 26$ residues, the score was set to 0 for any docking pose.

2.5 Evaluation of the Scoring Function

To estimate the performance of DINE, the success rate of ranking near-native domain assemblies (poses) within the top N ranks ($N = 10, 20, 30, 40, 50, 100, 200, 300, 400$, and 500) was examined. According to the criteria adopted in the pyDockTET evaluation [6], a domain assembly was considered to be good if the root mean square deviation (RMSD) of the smaller domain was $\leq 5.0 \text{ \AA}$ from the corresponding domain of the experimentally determined structure when the larger domains were superposed. Similarly, a domain assembly was considered to be acceptable if the RMSD of the smaller domain was $\leq 10.0 \text{ \AA}$ in the same superposition. To evaluate a random success rate, we adopted the definition introduced by Cheng et al. [21].

3. Results & Discussion

3.1 The Prediction of Domain Interface Residues by the KIP Method

The performance of the KIP method, which predicts domain interface residues based on homology, was tested through a leave-one-out validation against 2,232 domain interface data in the DiD. When the threshold of sequence identity T_{iden} was changed from 80 to 20% with an interval of 10%, the recall rate was markedly increased, and the precision rate was slightly decreased (Fig. 1 A). The high precision rate ($> 78\%$) of the KIP method suggested that the locations of domain interface sites were highly conserved among homologous domains ($\leq 20\%$ sequence identity). However, the low recall rate ($< 35\%$) implied that no homologous interface was detected in the DiD for many of the test proteins.

The F-measure was maximized at $T_{iden} = 20\%$ (Fig. 1 B). Therefore, this parameter was employed for the calculation of the domain interface score. With this parameter, the recall and precision rates were 34.8% and 78.4%, respectively, and we could detect true domain interface residues for 44.4% of the test domains.

3.2 The Correlation of the Average End-to-end Distance to the Domain Linker Length

For a total of 1,662 two-domain structures in the DiD, 1,657 full-length amino acid sequences were found in the UniProtKB/Swiss-Prot or NCBI nr database for the identification of domain linkers. The lengths of the domain linkers varied from 2 to 282 amino acids. The frequency of linkers longer than 26

Fig. 1 The accuracy of domain interface residue prediction with the KIP method. (A) A recall-precision plot for the prediction of domain interface residues by the KIP method with the sequence identity threshold T_{iden} ranging from 20% (cross) to 80% (circle). (B) A plot of the F-measure values against the sequence identity threshold T_{iden} .

Fig. 2 The relationships between linker length, frequency and end-to-end distance. (A) The length distribution of 1,657 domain linkers. (B) The relationship between the length and the end-to-end distance of the domain linkers. The circle indicates the average distance. The upward and downward triangles show the maximum and the minimum distances at each linker length, respectively. The error bars represent two standard deviations of the end-to-end distance at each linker length.

amino acids was less than 10 for each length (Fig. 2 A). Therefore, the correlation of the end-to-end distance and the linker length was not evaluated for the long linkers.

For linkers containing 26 or fewer residues, the average end-to-end distances and their standard deviations (SDs) increased as the linker length increased (Fig. 2 B). This observation was consistent with the previous report by Cheng et al. for 542 domain linkers [6]. The end-to-end distances appeared to adequately obey a normal distribution. On average, 64.5% and 98.5% of the distances fell within one and two standard deviations from the average over linker lengths of 2 to 26 residues, respectively. Based on these observations, we defined the end-to-end distance score as described in the Materials and Methods.

3.3 The Optimization of the Weights in the DINE Score Function

Wollacott et al. selected 76 benchmark proteins that contained no cofactors or ligands proximal to their domain interfaces from a database of non-redundant protein structures [3]. Because 14 out of the 76 proteins in the benchmark set were defined as single-domain proteins in SCOP, we used the remaining 62 proteins for an optimization of DINE (Wollacott dataset in Table 1). The protein structures were dissected into domains according to the SCOP definition. Then, all of the side-chains of the isolated domains were remodeled with SCWRL 4.0 [22], and each domain was randomly rotated to avoid bias from the native structures. To calculate the interface score (S_{int}), we applied the KIP method using a customized DiD that did not contain the query domain itself or homologous domains of the query domain with more than

90% sequence identity.

When we scored the Wollacott dataset using DINE with the initial weights, namely, $w_{dock} = w_{int} = w_{ete} = 1$, at least one good domain assembly (the RMSD of the smaller domains was ≤ 5.0 Å when the larger domains were superposed) was found within the top 10 poses for 40 out of 62 proteins (a success rate of 64.5%; Table 1). Then, the weights were optimized against the same dataset by changing each weight from 1 to 10 by intervals of 1 and maximizing the number of good domain assemblies within the top 10 docking poses. As a result, at least one good assembly was found in 49 out of 62 cases (a success rate of 79.0%) for five weight sets. In these weight sets, the values of w_{dock} and w_{int} were higher than 4, and w_{ete} was always 1. This result suggested that the interface score was more effective for prediction than the linker score.

The five weight sets were further examined in maximizing the total number of good assemblies and acceptable assemblies (the RMSD of the smaller domain was ≤ 10.0 Å) within the top N poses ($N = 10, 20, 30, 40, 50, 100, 200, 300, 400$, and 500). The optimal values of the w_{dock} , w_{int} , and w_{ete} were 7, 8, and 1, respectively. The improvement of the success rates by the weight optimization was remarkable for the higher range of ranking from the top 10 to the top 100 (Fig. 3).

3.4 The Evaluation of the Domain Assemblies Selected by DINE

Cheng et al. developed pyDockTET for building multidomain protein structures from individual domains, using a combination of rigid-body docking, binding energy scoring, and linker-length

Table 1 The prediction results on the Wollacott dataset.

PDB	Domain 1	Domain 2	Linker length ^b	Best rank of docking poses ^c	
				Initial weight	Optimized weight
1a62A	1–47	48–125	4 (45–48)	3 (2.4)	3 (2.4)
1a6qA	2–296	297–368	8 (291–298)	244 (8.8)	212 (8.8)
1a79C	9–82	83–179	3 (82–84)	6 (8.1)	16 (8.1)
1a8dA	1–247	248–452	22 (244–265)	1 (1.3)	1 (1.3)
1a8lA	1–119	120–226	6 (117–122)	1 (1.3)	1 (1.3)
1ammA ^a	1–85	86–174	8 (81–88)	5 (6.9)	3 (6.9)
1aoaA	121–251	260–375	16 (245–260)	2 (4.7)	2 (4.7)
1avaA	1–346	347–403	8 (344–351)	1 (1.3)	1 (1.3)
1b63A	–2–216	217–331	2 (216–217)	7 (8.9)	9 (1.0)
1bagA ^a	1–347	348–425	7 (345–351)	2 (3.3)	2 (3.3)
1bg6A	4–187	188–359	3 (186–188)	2 (4.9)	2 (1.1)
1bi5A	1–235	236–389	4 (233–236)	1 (1.1)	1 (1.1)
1bkbA	4–74	75–139	3 (74–76)	11 (9.8)	12 (9.8)
1bu6O	3–253	254–499	7 (253–259)	1 (1.1)	1 (1.1)
1c2aA	4–64	65–123	15 (57–71)	427 (13.6)	510 (13.6)
1ca1A ^a	1–249	250–370	11 (246–256)	1 (1.8)	1 (1.8)
1cdyA	1–97	98–178	2 (97–98)	110 (3.7)	2 (3.7)
1cjbA	4–153	154–356	12 (144–155)	1 (2.3)	1 (2.3)
1clcA ^a	35–134	135–575	2 (134–135)	2 (7.5)	9 (1.6)
1cliB	1,021–1,170	1,171–1,345	11 (1,169–1,179)	1 (2.4)	1 (2.4)
1crzA	7–140	141–409	8 (137–144)	1 (4.7)	1 (4.7)
1ctuA	1–150	151–294	8 (149–156)	1 (1.7)	1 (1.7)
1cvrA	1–350	351–432	16 (340–355)	1 (1.1)	1 (1.1)
1cx4A	130–265	266–412	2 (265–266)	2 (4.3)	6 (4.3)
1d09B	1–100	101–153	6 (96–101)	2 (7.8)	3 (7.8)
1d5rA	14–187	188–351	7 (185–191)	1 (4.1)	1 (4.1)
1dzfA	5–143	144–215	11 (141–151)	26 (9.4)	30 (9.4)
1egaB	4–182	183–296	18 (172–189)	35 (7.8)	54 (7.8)
1eovA	71–204	205–557	9 (199–207)	6 (21.1)	6 (21.1)
1eudA	1–130	131–306	5 (129–133)	5 (1.5)	1 (1.5)
1f1zA	8–168	169–267	2 (168–169)	12 (7.3)	1 (4.1)
1f3aA	1–79	80–221	7 (78–84)	5 (4.8)	5 (4.8)
1f5nA	7–283	284–583	13 (277–289)	1 (4.8)	1 (1.3)
1fmtA	1–206	207–314	8 (201–208)	999 (8.8)	814 (9.0)
1ftsA	201–284	285–495	13 (281–293)	1 (1.4)	1 (1.4)
1gcyA	1–357	358–418	6 (356–361)	15 (6.3)	1 (4.4)
1gv1B	1–142	143–299	2 (142–143)	1 (4.5)	1 (1.6)
1hanA	2–132	133–289	21 (119–139)	1 (2.1)	1 (2.1)
1hyeA	1–145	146–313	4 (145–148)	1 (2.3)	1 (1.2)
1i8dB ^a	1–93	94–206	18 (87–104)	2 (1.1)	2 (1.1)
1j8mF	3–86	87–297	15 (84–98)	1 (1.4)	4 (1.4)
1jakA	8–150	151–506	5 (148–152)	1 (2.4)	1 (2.4)
1jgtB	2–209	210–508	12 (207–218)	1 (4.0)	1 (4.0)
1jpmA	1–125	126–359	6 (123–128)	1 (1.2)	1 (1.2)
1jpnA	1–88	89–296	13 (86–98)	1 (3.2)	1 (3.2)
1k0mA	6–91	92–240	13 (88–100)	1 (1.8)	1 (1.8)
1kbwC	13–163	164–314	9 (156–164)	1 (1.5)	1 (1.5)
1knyA	1–125	126–253	3 (125–127)	3 (5.6)	1 (4.3)
1ks9A	1–167	168–291	3 (166–168)	1 (9.2)	6 (1.3)
1lamA	1–159	160–484	2 (159–160)	47 (9.2)	158 (9.2)
1lbuA	1–83	84–213	11 (77–87)	31 (4.2)	1 (4.2)
1mgtA	1–88	89–169	7 (84–90)	1 (1.3)	1 (1.3)
1nkrA	6–101	102–200	9 (100–108)	2 (2.0)	2 (2.0)
1pgsA ^a	4–140	141–314	6 (136–141)	5 (9.2)	5 (9.2)
1piiA	1–254	255–452	4 (254–257)	17 (6.8)	56 (6.8)
1qcsA	0–85	86–201	3 (84–86)	1 (7.1)	3 (1.0)
1qfjC	1–97	98–232	12 (93–104)	1 (1.5)	1 (1.5)
1qh4B	2–102	103–381	12 (100–111)	1 (2.4)	1 (2.4)
1rhsA	1–149	150–293	22 (136–157)	1 (1.3)	1 (1.3)
1smdA	1–403	404–496	4 (402–405)	1 (2.3)	1 (2.3)
1tf4B	1–460	461–605	18 (445–462)	1 (1.8)	1 (4.2)
2rebA	3–268	269–328	3 (267–269)	1 (4.4)	1 (1.1)
No. of good cases				40 (64.5%)	49 (79.0%)

^a Proteins also included in the pyDockTET dataset.^b The residue ranges of each linker are in parentheses.^c The best ranking and RMSD of the docking poses (in parentheses) in the 2,000 poses.

based distance restraints [6]. To evaluate the prediction accuracy of DINE in comparison with pyDockTET, the 77 non-redundant protein structures that were used for the evaluation of pyDock-

TET were also considered in this study. The structures were dissected into domains according to the domain definition in the literature [6]. To coordinate the prediction conditions, all of the side

Fig. 3 The success rates of DINE with the initial and optimized weight sets. The success rate is defined as the percentage of the cases in which at least one good docking pose has been detected with a rank higher than the number shown in the horizontal axis. The success rates of DINE with the initial weights, $w_{dock} = 1$, $w_{int} = 1$, and $w_{ete} = 1$, are shown with triangles, and those with the optimized weights, $w_{dock} = 7$, $w_{int} = 8$, and $w_{ete} = 1$, are shown with circles. The success rates expected from a random selection of the docking poses are indicated with squares.

chains of the individual domains were re-modeled by SCWRL 3.0 [23] instead of SCWRL 4.0, which was used for the weight optimization mentioned above, and 2,000 docking poses were generated using ZDOCK 2.1 instead of ZDOCK 3.0.1 from randomly rotated domain structures. The proteins that showed no acceptable docking poses within the top 2,000 were removed from the benchmark set. As a result, 55 out of the 77 proteins remained as the common benchmark set for the pyDockTET-DINE comparison, and they were called the pyDockTET dataset (Table 2 and Table 3). To avoid bias from the known 3D structures, we used a DiD that did not contain the query domains or the homologous domains of the query domain with more than 90% sequence identity when applying the KIP method to calculate the interface scores (S_{int}).

When the pyDockTET dataset was scored with DINE, the success rate in identifying good domain assemblies within the top 10 docking poses was 70.9% (Table 2). Moreover, the success rates in detecting acceptable domain assemblies in the top 10 and top 50 docking poses were 83.6% and 96.4%, respectively (Table 2 and Fig. 4). In contrast, pyDockTET provided an acceptable assembly in the top 10 and top 50 docking poses in 61% and 78% of the same dataset, respectively [6].

We compared the results of DINE for the 20 cases that were detailed in the previous study [6] with those of pyDockTET (Table 4). In seven cases (1nezA, 1s9vB, 1onqA, 1edhB, 1hnfA, 1jk8B, 1k2dB), no acceptable structure was found in the 2,000 docking poses generated by ZDOCK with both of the methods. The areas of domain interface in the experimental structures were found to be less than $322 \text{ \AA}^2$ for these false cases (Table 4). This result might be consistent with the observation that protein docking with interfaces smaller than $1,400 \text{ \AA}^2$ generally yielded poor results due to the weak interactions at the interface [24].

In two cases (1gk8C and 1mb8A), acceptable structures were detected with DINE but not with pyDockTET (Table 4). We re-generated the same number of docking poses for the two cases without swapping the receptor and ligand domains and found that

at least one acceptable pose was now detected for both cases. DINE scored near-native structures ($\text{RMSD} \leq 2.0 \text{ \AA}$) within the top 10 for eight cases (1a8pA, 1aw7A, 1b06A, 1dluB, 1e5mA, 1j3nA, 1ee0A, 1etpB). In contrast, pyDockTET succeeded in only three cases (1ffuF, 1dluB, 1e5mA) under the same conditions.

Thus, DINE demonstrated an improvement in the prediction result for 6 cases, whereas pyDockTET was better than DINE in only one case. The results also showed that DINE was able to score near-native structures higher than pyDockTET. The major difference between DINE and pyDockTET is the restraint of interface residues through the DiD. These results prove that the combination of the linker distance restraint and the interface restraint of DINE is efficient in scoring rigid-body docking poses.

3.5 The Contribution of the Domain Interface Score to Success Rates of Prediction

The DINE score is composed of the terms of docking score, domain interface score, and end-to-end distance score. The combination of the binding energy and the distance restraints of the domain linkers were also employed in the pyDockTET scoring function [6]. It was found that the inclusion of the linker distance restraints markedly improved the prediction accuracy, especially for the domains that had multiple interfaces of comparable binding energy. When the pyDockTET dataset was scored using the docking score (S_{dock}) alone, the success rate for identifying acceptable domain assemblies in the top 10 poses was only 50.9% (Table 2 and Fig. 4). This value was close to the success rate that was obtained with pyDock, which only evaluated the binding energy [6]. When the pyDockTET dataset was scored by combining S_{dock} and the end-to-end distance score (S_{ete}), the success rate slightly increased to 58.2% (Table 2 and Fig. 4). This value was also close to the success rate with pyDockTET, which evaluated the binding energy and the distance restraints based on the linker length [6]. Thus, by detecting acceptable domain assemblies in 83.6% of the cases, DINE outperformed pyDockTET. This result might be partially because the contribution from the end-to-end distance is not sufficient, especially for long linkers, to discriminate near-native structures. The inefficiency of the docking energy toward a small interface area may also be responsible.

When the DiD contains the 3D structures of homologous proteins with the same domain combination as the query protein, the contribution of the domain interface score (S_{int}) of DINE should be critical. It was demonstrated that the domain interface residues and the spatial arrangements of domains were well-conserved among proteins with more than 30% sequence identity [7]. To further evaluate the DINE method under the condition in which no apparent homologous protein was in the DiD, the pyDockTET dataset was scored with DINE using a subset of the DiD that did not contain the query proteins or any apparent homologous proteins of the query protein with more than 30% sequence identity and an E-value less than 0.01. As a result, the success rate in detecting acceptable domain assemblies in the top 10 docking poses was reduced to 65.5% (Fig. 4), although this rate was still $\sim 7\%$ larger than that obtained using the combination of S_{dock} and S_{ete} . This result suggested that S_{int} would contribute in selecting accu-

Table 2 The prediction results on the pyDockTET dataset.

PDB	Domain 1	Domain 2	Linker length ^b	Details of each score in the DINE score			Best rank of any acceptable docking poses ^c			
				$w_{dock} S_{dock} (S_{dock})$	$w_{int} S_{int} (S_{int})$	$w_{ete} S_{ete} (S_{ete})$	DINE	Docking	docking + end-to-end	Interface area (Å ²)
1a8pA	2–101	107–25	16 (93–108)	6.937 (0.991)	7.344 (0.918)	1.000 (1.000)	1 (1.6)	6 (1.6)	5 (1.6)	1,028
1ah5A	3–217	225–313	9 (217–225)	6.517 (0.931)	0.000 (0.000)	0.927 (0.927)	8 (8.9)	9 (8.9)	8 (8.9)	1,327
1ammA ^a	1–81	89–174	9 (81–89)	5.383 (0.769)	5.184 (0.648)	1.000 (1.000)	3 (9.2)	713 (9.2)	340 (9.2)	816
1aorB	1–207	217–605	13 (205–217)	7.000 (1.000)	6.232 (0.779)	1.000 (1.000)	1 (1.7)	1 (1.7)	1 (1.7)	3,201
1aqhA	1–349	359–448	13 (349–361)	7.000 (1.000)	6.704 (0.838)	0.460 (0.460)	1 (2.9)	1 (2.9)	4 (2.9)	1,270
1ar4A	1–83	88–201	13 (81–93)	5.642 (0.806)	4.824 (0.603)	1.000 (1.000)	16 (6.3)	33 (9.3)	46 (7.9)	1,230
1aw7A	1–92	97–194	8 (90–97)	7.000 (1.000)	3.200 (0.400)	1.000 (1.000)	1 (1.0)	1 (1.0)	1 (1.0)	1,225
1aw9A	2–77	94–217	18 (77–94)	5.691 (0.813)	5.568 (0.696)	0.736 (0.736)	8 (3.4) ^d	93 (3.1)	45 (3.1)	1,035
1b06A	3–92	98–210	13 (90–102)	4.802 (0.686)	6.568 (0.821)	0.428 (0.428)	8 (1.2) ^d	138 (8.0)	294 (8.0)	1,231
1b25A	1–205	215–619	13 (203–215)	7.000 (1.000)	6.936 (0.867)	0.851 (0.851)	1 (1.6)	1 (1.6)	2 (1.6)	3,160
1b8pA	3–157	160–329	4 (157–160)	7.000 (1.000)	6.248 (0.781)	0.975 (0.975)	1 (1.5)	1 (1.5)	1 (1.5)	1,589
1bagA ^a	1–342	359–425	19 (342–360)	6.608 (0.944)	2.000 (0.250)	1.000 (1.000)	2 (1.4)	3 (3.3)	3 (3.3)	1,165
1bayA	1–74	84–209	11 (74–84)	5.908 (0.844)	5.952 (0.744)	1.000 (1.000)	1 (4.4) ^d	35 (8.7)	8 (7.6)	1,149
1bikA	25–77	81–134	18 (77–94)	6.279 (0.897)	0.000 (0.000)	0.548 (0.548)	128 (8.4)	29 (8.4)	128 (8.4)	488
1calA ^a	1–250	258–370	13 (246–258)	4.620 (0.660)	5.792 (0.724)	1.000 (1.000)	3 (5.7)	85 (9.7)	173 (7.4)	1,015
1chmB	2–156	165–402	10 (156–165)	4.298 (0.614)	3.368 (0.421)	1.000 (1.000)	136 (9.4)	1,174 (9.4)	515 (9.4)	884
1clcA ^a	35–130	138–575	11 (129–139)	4.998 (0.714)	3.232 (0.404)	1.000 (1.000)	9 (1.4) ^d	927 (1.4)	272 (1.4)	1,262
1clvA	1–374	383–471	12 (374–385)	6.615 (0.945)	5.520 (0.690)	1.000 (1.000)	3 (3.8)	7 (3.8)	4 (3.8)	1,211
1cr5B	23–105	120–207	19 (102–120)	5.964 (0.852)	5.648 (0.706)	1.000 (1.000)	8 (4.5) ^d	99 (4.3)	67 (4.3)	957
1dluB	4–262	269–392	13 (261–273)	7.000 (1.000)	6.328 (0.791)	0.000 (0.000)	1 (1.3)	1 (1.3)	1 (1.3)	2,886
1e5mA	6–250	258–416	10 (250–259)	7.000 (1.000)	7.024 (0.878)	0.492 (0.492)	1 (1.8)	1 (1.8)	1 (4.1)	2,653
1e9IB	1–133	142–430	14 (133–146)	6.328 (0.904)	6.216 (0.777)	0.964 (0.964)	1 (3.1) ^d	41 (3.1)	20 (3.1)	1,532
1ebgA	1–133	143–436	11 (133–143)	5.873 (0.839)	5.952 (0.744)	1.000 (1.000)	1 (3.9) ^d	123 (3.9)	24 (3.9)	1,821
1ee0A	20–233	243–395	13 (231–243)	7.000 (1.000)	6.832 (0.854)	0.000 (0.000)	1 (1.1)	1 (1.1)	1 (1.1)	2,471
1et6B	4–95	101–209	15 (94–104)	6.839 (0.977)	5.968 (0.746)	1.000 (1.000)	1 (3.1)	4 (3.1)	3 (3.1)	1,456
1et9A	1–94	100–204	8 (93–100)	7.000 (1.000)	3.384 (0.423)	1.000 (1.000)	1 (1.4)	1 (1.4)	1 (1.4)	1,704
1etpB	1–89	101–190	15 (87–101)	5.103 (0.729)	5.776 (0.722)	1.000 (1.000)	1 (2.0)	203 (9.4)	154 (2.0)	783
1f2eA	1–75	89–201	15 (75–89)	6.069 (0.867)	3.968 (0.496)	1.000 (1.000)	15 (8.8)	37 (8.8)	22 (8.8)	997
1fdrA	2–92	106–248	16 (92–107)	5.320 (0.760)	5.824 (0.728)	1.000 (1.000)	4 (3.2) ^d	2 (5.5)	2 (5.5)	1,043
1fthA	2–87	97–295	13 (86–98)	6.510 (0.930)	3.552 (0.444)	1.000 (1.000)	1 (4.1)	3 (1.5)	3 (1.5)	1,203
1fufA	1–175	180–287	7 (174–180)	7.000 (1.000)	6.728 (0.841)	0.783 (0.783)	1 (2.3)	1 (2.3)	1 (3.9)	1,062
1fiqB	224–412	418–528	8 (411–418)	7.000 (1.000)	3.000 (0.375)	0.443 (0.443)	1 (3.4)	1 (3.4)	2 (3.4)	1,197
1gk8C	7–146	154–475	9 (146–154)	4.214 (0.602)	6.896 (0.862)	1.000 (1.000)	1 (4.0) ^d	126 (8.9)	10 (8.9)	1,384
1gsaA	1–72	82–202	11 (72–82)	6.188 (0.884)	5.376 (0.672)	1.000 (1.000)	4 (1.4) ^d	7 (7.9)	4 (7.9)	1,089
1h1oB	213–287	301–383	16 (286–301)	4.865 (0.695)	6.384 (0.798)	1.000 (1.000)	9 (4.4) ^d	188 (8.3)	132 (8.3)	927
1i8dB ^a	1–87	100–206	18 (87–104)	6.111 (0.873)	3.296 (0.412)	0.946 (0.946)	4 (4.4) ^d	26 (4.4)	26 (4.4)	821
1ik6A	1–184	198–326	20 (181–200)	6.111 (0.873)	5.128 (0.641)	1.000 (1.000)	12 (3.8)	104 (6.4)	47 (6.4)	1,090
1j3nA	1–243	252–408	13 (241–253)	7.000 (1.000)	7.024 (0.878)	0.000 (0.000)	1 (1.2)	1 (1.2)	4 (1.2)	2,709
1jlvA	1–74	88–207	15 (74–88)	5.019 (0.717)	5.512 (0.689)	1.000 (1.000)	11 (8.6)	117 (9.6)	57 (9.6)	1,132
1kgzA	12–77	83–344	11 (76–86)	6.881 (0.983)	7.480 (0.935)	1.000 (1.000)	1 (1.3)	1 (1.5)	1 (1.5)	1,325
1kzlA	1–87	100–202	18 (87–104)	6.265 (0.895)	7.064 (0.883)	0.964 (0.964)	1 (3.2) ^d	12 (4.3)	12 (4.3)	979
1mb8A	56–172	187–293	16 (172–187)	5.033 (0.719)	5.176 (0.647)	0.841 (0.841)	45 (9.7)	330 (9.7)	324 (9.7)	1,159
1n5wF	1–175	182–286	9 (174–182)	4.690 (0.670)	4.952 (0.619)	1.000 (1.000)	22 (5.3)	398 (9.3)	118 (9.3)	1,121
1oe7A	4–80	97–207	18 (80–97)	6.069 (0.867)	5.488 (0.686)	0.368 (0.368)	7 (9.0)	22 (6.7)	30 (8.0)	924
1oi7A	1–114	132–288	25 (114–138)	6.510 (0.930)	5.720 (0.715)	1.000 (1.000)	2 (1.4)	9 (1.4)	5 (1.4)	1,284
1p5uA	9–148	152–234	11 (144–154)	5.516 (0.788)	4.944 (0.618)	1.000 (1.000)	8 (8.4)	264 (8.4)	56 (8.4)	585
1pbjA	2–57	69–121	17 (57–73)	6.384 (0.912)	2.512 (0.314)	1.000 (1.000)	39 (7.8)	2 (8.6)	2 (8.6)	961
1pdyA	1–133	141–433	10 (133–142)	5.803 (0.829)	4.936 (0.617)	0.035 (0.035)	2 (4.6) ^d	177 (4.6)	189 (3.6)	1,593
1pgsA ^a	4–140	149–314	14 (136–149)	6.230 (0.890)	0.000 (0.000)	0.990 (0.990)	9 (4.1)	31 (4.1)	9 (4.1)	1,585
1pmtA	1–75	87–201	14 (75–88)	6.034 (0.862)	5.616 (0.702)	0.690 (0.690)	1 (3.2) ^d	11 (8.7)	4 (8.7)	1,102
1qnnA	1–82	86–191	11 (80–90)	6.279 (0.897)	6.280 (0.785)	1.000 (1.000)	1 (5.0)	4 (3.9)	2 (3.9)	1,253
1r5aA	2–76	90–215	15 (76–90)	6.888 (0.984)	4.928 (0.616)	1.000 (1.000)	2 (8.6)	8 (8.6)	6 (8.6)	1,121
1r6IA	1–144	158–239	16 (144–159)	7.000 (1.000)	6.184 (0.773)	1.000 (1.000)	1 (7.8)	1 (7.8)	1 (7.8)	1,772
1rwzA	1–114	121–244	19 (110–128)	6.566 (0.938)	6.168 (0.771)	1.000 (1.000)	1 (4.9)	5 (4.9)	5 (4.9)	837
1v8gA	1–64	70–329	11 (62–72)	5.985 (0.855)	7.248 (0.906)	1.000 (1.000)	1 (1.6)	3 (5.7)	6 (1.6)	1,223
No. of good cases							39 (70.9%)	21 (38.2%)	23 (41.8%)	
No. of acceptable cases							46 (83.6%)	28 (50.9%)	32 (58.2%)	

^a Proteins are also included in the Wollacott dataset.

^b The residue ranges of each linker are in parentheses.

^c The best ranking and RMSD of the acceptable domain assembly in the 2,000 poses.

^d Good domain assembly was detected within the top 10 poses with DINE.

rate docking assemblies even if only proteins of marginal similarity to a query were known.

In the 16 cases in which DINE but not $S_{dock} + S_{ete}$ detected good domain assemblies within the top 10 (Table 2), the average of the inter-domain interface areas was small (1,169 Å²) compared with that of the remaining proteins (1,328 Å²). Therefore, the additional term of the domain interface score of DINE appeared to contribute largely by detecting smaller domain interfaces. For example, in the case of 1etpB, the docking score could detect an acceptable (9.4 Å RMSD) domain assembly model (Table 2 and Fig. 5 A). However, the rank of the domain assembly was as low as 203 (Table 2). In this case, because the inter-domain interface area was very small (783 Å²), the binding energy alone was not sufficient to highly rank the acceptable pose. When the same

protein was scored using $S_{dock} + S_{ete}$, a near-native model (2.0-Å RMSD) was obtained (Table 2 and Fig. 5 B), but its rank was still as low as 153. However, the scoring with DINE ranked the near-native model at the top. This example well indicated the important contribution of the domain interface score in the DINE method.

4. Conclusion

We presented the novel score function, DINE, for selecting plausible domain assemblies from docking poses generated in a rigid-body docking simulation. DINE is based on a combination of the binding energy, domain interface restraints based on homologous domains, and end-to-end distance restraints based on the domain linker length. A comparison with the pyDockTET

Table 3 The pyDockTET benchmark set with no acceptable pose detected.

PDB	Domain 1	Domain 2	Linker	RMSD ^a
1b4aA	4–70	79–149	16 (64–79)	14.7
1b9kA	702–819	825–938	8 (818–825)	12.9
1bhtB	336–424	428–509	42 (422–463)	11.1
1cfbA	610–700	713–814	24 (697–720)	14.5
1edhB	3–99	113–213	15 (99–113)	20.2
1epfA	1–96	113–189	20 (96–115)	17.4
1fpoA	1–74	86–171	15 (74–88)	15.0
1ghqB	1–63	72–129	23 (57–79)	11.5
1grjA	2–74	80–158	17 (74–90)	15.5
1h03Q	5–64	69–129	20 (58–77)	14.7
1hdmB	3–82	89–185	11 (81–91)	18.4
1hnfA	4–96	101–182	14 (96–109)	12.5
1j5yA	3–67	75–174	12 (64–75)	12.0
1jk8B	3–87	103–192	27 (87–113)	22.7
1k2dB	5–86	103–190	30 (83–112)	20.4
1lslA	416–471	478–528	25 (468–492)	26.2
1nezA	1–179	189–274	11 (179–189)	17.4
1onqA	7–181	194–280	20 (181–200)	18.4
1piiA	1–252	257–452	6 (252–257)	20.0
1qg3A	1,126–1,208	1,220–1,320	27 (1,200–1,226)	18.7
1s9vB	3–87	97–190	11 (87–97)	18.0
1vk6A	0–124	127–256	4 (124–127)	14.3

^a The minimum RMSD (Å) of the docking poses in the 2,000 solutions.

Fig. 4 The success rates of DINE in detecting acceptable domain assemblies among the pyDockTET dataset. The success rates of DINE, the same rates using the DiD that did not contain apparent homologous proteins ($\leq 30\%$ identity and < 0.01 E-value) of the query protein, the docking score (S_{dock}) alone, and the combination of the docking score with the end-to-end distance score ($S_{dock} + S_{ete}$) are shown with circles, downward triangles, diamonds, and upward triangles, respectively. The success rates expected from a random selection of the docking poses are shown with squares.

method demonstrated the superior ability of DINE, which might be attributed to the domain interface restraint term. DINE detected the correct domain assembly within the top 10 docking poses in 65.5% of the test proteins and, in most cases, ranked near-native structures higher in comparison with the previous method.

Acknowledgments We thank Dr. Asako Tomiyama and Mr. Hideaki Ishikawa for their assistance in preparing DINE data and programs and Dr. Tsuyoshi Shirai for helpful discussion, critical reading, and kind revision of this manuscript. This work was partially supported by the Targeted Proteins Research Program (TPRP) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

Fig. 5 An example of the best model for 1etpB prediction. (A) The best domain assembly detected by the docking score (S_{dock}) alone. (B) The best domain assembly detected by the combination of the docking score with the end-to-end distance score ($S_{dock} + S_{ete}$) and DINE. The larger domains of the experimental and model structures are superposed and colored in orange. The smaller domains of the experimental and model structures are colored in red and blue, respectively. The Ca atoms of the interface residues predicted by the KIP method are shown as spheres.

Table 4 Comparison of the best poses between DINE and pyDockTET.

PDB	Best rank of any acceptable docking poses ^a		
	DINE	pyDockTET	Interface area (Å) ^b
1a8pA	1 (1.6)	1 (6.8)	1,028
1ar4A	16 (6.3)	39 (2.2)	1,229
1aw7A	1 (1.0)	1 (2.2)	1,225
1ffuF	1 (2.3)	1 (1.9)	1,062
1b06A	8 (1.2)	1 (3.9)	1,231
1dluB	1 (1.3)	1 (1.8)	2,886
1ca1A	3 (5.7)	10 (4.8)	1,015
1e5mA	1 (1.8)	1 (1.3)	2,653
1gk8C	1 (4.0)	-	1,384
1j3nA	1 (1.2)	1 (5.6)	2,709
1ee0A	1 (1.1)	2 (8.1)	2,471
1nezA	-	-	250
1s9vB	-	-	249
1etpB	1 (2.0)	204 (7.4)	783
1onqA	-	-	231
1edhB	-	-	227
1hnfA	-	-	322
1mb8A	45 (9.7)	-	1,159
1jk8B	-	-	253
1k2dB	-	-	121

^a The RMSDs of the smaller domains are in parentheses. A dash indicates that no acceptable structure was found in 2,000 docking poses generated by ZDOCK.

^b The mean value of the inter-domain interface area of each domain in the experimental structure.

Reference

- [1] Apic, G., Gough, J. and Teichmann, S.A.: Domain combinations in archaeal, eubacterial and eukaryotic proteomes, *J. Mol. Biol.*, Vol.310, No.2, pp.311–325 (2001).
- [2] Vogel, C., Bashton, M., Kerrison, N.D., Chothia, C. and Teichmann, S.A.: Structure, function and evolution of multidomain proteins, *Curr. Opin. Struct. Biol.*, Vol.14, No.2, pp.208–216 (2004).
- [3] Wollacott, A.M., Zanghellini, A., Murphy, P. and Baker, D.: Prediction of structures of multidomain proteins from structures of the individual domains, *Protein Sci.*, Vol.16, No.2, pp.165–175 (2007).
- [4] Inbar, Y., Benyamini, H., Nussinov, R. and Wolfson, H.J.: Combinatorial docking approach for structure prediction of large proteins and multi-molecular assemblies, *Phys. Biol.*, Vol.2, No.4, pp.S156–65 (2005).
- [5] Lise, S., Walker-Taylor, A. and Jones, D.T.: Docking protein domains in contact space, *BMC Bioinformatics*, Vol.7, p. 310 (2006).

- [6] Cheng, T.M.K., Blundell, T.L. and Fernandez-Recio, J.: Structural assembly of two-domain proteins by rigid-body docking, *BMC Bioinformatics*, Vol.9, p.441 (2008).
- [7] Aloy, P., Ceulemans, H., Stark, A. and Russell, R.B.: The relationship between sequence and interaction divergence in proteins, *J. Mol. Biol.*, Vol.332, No.5, pp.989–998 (2003).
- [8] Korkin, D., Davis, F.P. and Sali, A.: Localization of protein-binding sites within families of proteins, *Protein Sci.*, Vol.14, No.9, pp.2350–2360 (2005).
- [9] Korkin, D., Davis, F.P., Alber, F., Luong, T., Shen, M.-Y., Lucic, V., Kennedy, M.B. and Sali, A.: Structural modeling of protein interactions by analogy: application to PSD-95, *PLoS Comput. Biol.*, Vol.2, No.11, p.e153 (2006).
- [10] Chen, R., Li, L. and Weng, Z.: ZDOCK: An initial-stage protein-docking algorithm, *Proteins*, Vol.52, No.1, pp.80–87 (2003).
- [11] Pierce, B. and Weng, Z.: ZRANK: Reranking protein docking predictions with an optimized energy function, *Proteins*, Vol.67, No.4, pp.1078–1086 (2007).
- [12] Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C.: SCOP: A structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.*, Vol.247, No.4, pp.536–540 (1995).
- [13] Berman, H., Henrick, K. and Nakamura, H.: Announcing the worldwide Protein Data Bank, *Nat. Struct. Biol.*, Vol.10, No.12, p.980 (2003).
- [14] Li, W. and Godzik, A.: Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics*, Vol.22, No.13, pp.1658–1659 (2006).
- [15] Lee, B. and Richards, F.M.: The interpretation of protein structures: estimation of static accessibility, *J. Mol. Biol.*, Vol.55, No.3, pp.379–400 (1971).
- [16] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J.: Basic local alignment search tool, *J. Mol. Biol.*, Vol.215, No.3, pp.403–410 (1990).
- [17] Kabsch, W. and Sander, C.: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, Vol.22, No.12, pp.2577–2637 (1983).
- [18] Mintseris, J., Pierce, B., Wiehe, K., Anderson, R., Chen, R. and Weng, Z.: Integrating statistical pair potentials into protein complex prediction, *Proteins*, Vol.69, No.3, pp.511–520 (2007).
- [19] McDonald, I.K. and Thornton, J.M.: Satisfying hydrogen bonding potential in proteins, *J. Mol. Biol.*, Vol.238, No.5, pp.777–793 (1994).
- [20] UniProt Consortium: Ongoing and future developments at the Universal Protein Resource, *Nucleic Acids Res.*, Vol.39, No.Database issue, pp.D214–9 (2011).
- [21] Cheng, T.M.-K., Blundell, T.L. and Fernandez-Recio, J.: pyDock: Electrostatics and desolvation for effective scoring of rigid-body protein-protein docking, *Proteins*, Vol.68, No.2, pp.503–515 (2007).
- [22] Krivov, G.G., Shapovalov, M.V. and Dunbrack, R.L.: Improved prediction of protein side-chain conformations with SCWRL4, *Proteins*, Vol.77, No.4, pp.778–795 (2009).
- [23] Canutescu, A.A., Shelenkov, A.A. and Dunbrack, R.L.: A graph-theory algorithm for rapid protein side-chain prediction, *Protein Sci.*, Vol.12, No.9, pp.2001–2014 (2003).
- [24] Vajda, S.: Classification of protein complexes based on docking difficulty, *Proteins*, Vol.60, No.2, pp.176–180 (2005).

scriptome/proteome. He is a member of IPSJ, BSJ, PSSJ, MBSJ and GSJ.

(Communicated by Toshinori Endo)

Masafumi Shionyu is a Lecturer of Faculty of Bio-Science, Nagahama Institute of Bio-Science and Technology. He received his Ph.D. degree from Nagoya University in 2002. His current research interests are protein-ligand interaction prediction, protein structure prediction, and computational analysis of trans-

Satoru Hirako received his B.Bio. and M.Bio. degree from Nagahama Institute of Bio-Science and Technology in 2005 and 2008, respectively.