Supporting Information

Table of Contents

Supporting Tables

Supporting Texts

Supporting Figures

References

Supporting Tables

Table S1. List of multi-domain proteins in CASP12 and CASP13 experiments as defined by the CASP assessors.

Domain type[a]	Number	Target ID
2dom	20	T0863, T0880, T0890, T0892, T0893, T0894, T0897, T0898, T0914, T0920, T0942, T0976, T0977, T0982, T0987, T0989, T1000, T1014, T1021s3, T1022s1
3dom	4	T0896, T0918, T1002, T1004
m4dom	4	T0960, T0963, T0996, T0999
2dis	9	T0886, T0899, T0901, T0905, T0943, T0946, T0957s1,T0984, T1011
3dom_dis	2	T0953s2, T0990
3dom_2dis	1	T0912
5dom_dis	1	T0981

[a]2dom, 3dom, m4dom, 2dis, 3dom_dis, 3dom_2dis and 5dom_dis represent the proteins with 2 continuous domains, 3 continuous domains, 4 or more continuous domains, 2 domains with one discontinuous domain, 3 domains with one discontinuous domain, 3 domains with 2 discontinuous domains and 5 domains with one discontinuous domain, respectively.

Table S2. Summary of DEMO modeling results on 166 2dom test proteins when using different template structural libraries. The domain assemblies start from the experimentally solved domain structures. Library-1 and Library-2 are the template libraries with and without including the 1,459 entries from the structure-based TM-score<0.5 cutoff.

	TM-score	RMSD(Å)	iRMSD(Å)	#clashes
Library-1	0.78	7.3	5.5	0.59
Library-2	0.75	8.1	5.9	0.62

Table S3. Summary of DEMO domain assemble results on 166 2dom test proteins starting from the target domain structures. 'DEMO(TM)' and 'DEMO(TM+LOMETS)' refer, respectively, to the pipelines with templates identified by TM-align only and a combination of TM-align and LOMETS programs.

	TM-score	RMSD(Å)	iRMSD(Å)	#clashes
DEMO(TM)	0.78	7.3	5.5	0.59
DEMO(TM+LOMETS)	0.79	7.2	5.4	0.68

Supporting Texts

Text S1. Multi-domain Template Structure Library Construction

The DEMO multi-domain template library is constructed in the following 4 steps:

1) Collect all non-redundant multi-domain protein structures from the PDB with a sequence identity cutoff 70%, where the domains of protein structures are defined by DomainParser (1).
2) Add all the non-redundant multi-domain protein structures from the CATH 4.1 library (2) if they have a sequence identity <70% to all entries in the existing library
3) Add all the non-redundant multi-domain protein structures from the SCOPe 2.06 library (3) if they have a sequence identity <70% to all entries in the existing library
4) Scan all other multi-domain structures from the PDB, CATH and SCOPe and add them to the library if they have a TM-score <0.5 to the existing templates regardless their sequence identity.

Here, the selection of sequence identity cutoff of 70% follows that used in the I-TASSER template library, which has been extensively tested in previous studies to optimize the trade-off of structural coverage and template search efficiency (4-6). Since some proteins may have different domain orientations even with a high sequence identity due to diverse evolution, the last step of structure-based template addition is designed to include those structures, which counts in total 1,459 entries. Table S2 lists a comparison of the DEMO models using the template library-1 (full DEMO library) and library-2 (without the addition of the 1,459 entries) based on 166 2-domain proteins. The average TM-score of the final models was increased by 4% (from 0.75 to 0.78) by the addition of the 1,459 entries, showing the importance of the last step of template library collection.

We note that the sequence identity of template collection (70%) is different from that used for benchmark test and training dataset collection (30%), since the former is designed to maximize the structural space coverage with a limited entry number for template search efficiency while the latter is to minimize the homologous contaminations between the test proteins and between training and testing datasets. The 30% sequence identity cutoff has been widely used in previous method development and benchmark studies (7-10) due to its sensitivity at sequence alignment and fold discriminations (11).

Text S2. Force Filed of DEMO for Domain Assembly Simulations

During the rigid-body domain assembly simulations, the DEMO force field is a sum of the five terms:

$$E = \sum_{\mathrm{i}=1}^{\mathrm{N}_1}\sum_{j=1}^{N_2}[w_1E_{sc}(i,j) + w_2E_{ct}(i,j) + w_3E_{dp}(i,j)] + w_4E_{bd} + w_5E_{dr}, \quad (S1)$$

where i and j are residue index running through the sequence of two domains separately, which have the size of N_1 and N_2, respectively.

The first term in Eq. (S1) is designed to eliminate ***steric clashes*** between domains, i.e.,

$$E_{sc}(i,j) = \begin{cases} 1/d_{ij}, & \text{if } d_{ij} < d_{\mathrm{cut}} \\ 0, & \text{otherwise} \end{cases} \quad (S2)$$

where d_{ij} is the distance between the ith C_α atom of the N-terminal domain and the jth C_α atom of the C-terminal domain in the decoy structure. $d_{\mathrm{cut}} = 3.75$ Å is set as the clash distance cutoff.

The second term is the ***generic domain-domain contact energy*** computed by:

$$E_{ct}(i,j) = \begin{cases} -u_{ij}, & \text{if } d_{ij} < 8\text{Å} \\ -\frac{1}{2}u_{ij}\left[1 - \sin\left(\frac{d_{ij}-9}{2}\pi\right)\right], & \text{if } 8\text{Å} \le d_{ij} \le 10\text{Å} \\ \frac{1}{2}u_{ij}\left[1 - \sin\left(\frac{d_{ij}-45}{70}\pi\right)\right], & \text{if } 10\text{Å} < d_{ij} \le 80\text{Å} \\ u_{ij}, & \text{otherwise} \end{cases} \quad (S3)$$

where the scale parameter u_{ij} depends on the hydrophobic and hydrophilic features of the residue pairs. $u_{ij} = 0.1$, if both of the residues are hydrophobic (ALA, CYS, VAL, ILE, PRO, MET, LEU, PHE, TYR, TRP); $u_{ij} = 0.01$, if the

two residues are hydrophilic (SER, THR, ASP, ASN, LYS, GLU, GLN, ARG, HIS); or $u_{ij} = 0.05$, otherwise. This energy item is used to control the inter-domain distance, which will push the two domains together if they are too far away from each other.

The third term is the ***domain-domain distance profile*** from the templates identified by TM-align (12), which is calculated by:

$$E_{dp}(i,j) = -\frac{1}{T_{ij}} \sum_{t=1}^{T_{ij}} \frac{1}{\left|d_{ij} - D_{ij}^{t}\right|} \quad (S4)$$

For a residue pair (i and j, with i from N-terminal domain and j from C-terminal domain, see Figure S9A), T_{ij} is the number of templates that satisfy the following two conditions: (1) the template has both residue i and j aligned by TM-align; (2) $0.6|i-j| < \left|a_i - a_j\right| < 1.5|i-j|$, where a_i and a_j are the indexes of the aligned residues of i and j on the template. D_{ij}^{t} is the C_α distance between the residue a_i and a_j in the t-th template (see Figure S9B).

Here, to ensure the quality of the distance profile, only the templates, which have the G-score (equal to the average TM-score of the two target domains to the template) higher than 0.6 and are aligned with more than 60% of residues in both N-terminal and C-terminal domains, are considered (see Figure S8). In case more than 200 templates satisfy the criterion, only top 200 templates with the highest G-scores are used for deriving the distance profile term.

The fourth term in Eq. (S1) is the ***boundary distance energy*** is defined as

$$E_{bd} = |d_1 - 3.8| + |d_2 - 3.8|. \quad (S5)$$

This term is only applied to the case of discontinuous domains (Figure S10), where the discontinuous domain is split into two segments due to the insertion of the continuous domain. d_1 is the Cα distance between the C-terminal residue of the first segment of the discontinuous domain (Seg-1) and the N-terminal residue of the continuous domain, and d_2 is that between the N-terminal residue of the second segment of the discontinuous domain (Seg-2) and the C-terminal residue of the continuous domain. This term is to constrain the connection of the two domains with neighboring Cα-Cα bond equal to the standard distance (3.8 Å).

The last term in Eq. (S1) is the ***local domain distance restraint***:

$$E_{dr} = \frac{1}{L_{\mathrm{M}}} \sum_{i=1}^{L_{\mathrm{M}}} d\,(S_i, S'_i) \quad (S6)$$

where $d(S_i, S'_i)$ represents the distance between the ith C_α atom (S_i) of the smaller domain that is at moving and its corresponding atom S'_i in the initial structure generated in the sliding-window based template superposition process, and L_{M} is the length of the domain. This term is to prevent the assembly deviating too much from the orientation obtained from the template.

The weighting parameters in Eq. (S1) are determined by maximizing the correlation between total energy and RMSD to the native on the structure decoys following the protocol in (8). This was performed on a training set of 425 proteins sharing <30% sequence identity to the test proteins reported in this study. This resulted in w_1=1.0, w_2=0.01, w_3=0.25, w_4=0, and w_5=0.30 for the continuous domain assembly, or w_1=0.2, w_2=0.01, w_3=0.08, w_4=0.88, and w_5=0.25 for the discontinuous domain assemble with its inserted domain.

Text S3. Replica Exchange Monte Carlo Simulation

The domain assembly conformational space in DEMO is searched through Monte Carlo (MC) simulations. In the classic Metropolis MC protocol (13), a Markov chain of conformations are created by randomly moving the relative domain orientations. At each step, the modified conformation is accepted by the probability of $p_{local} = min\{1, \exp(-\Delta E/kT)\}$, where ΔE is the energy difference between new and old conformations and kT is the temperature parameter. Since the acceptance rate is exponentially reduced with the energy difference at a given temperature, the simulation can be easily trapped at local minimum. To improve the sampling efficiency, DEMO implements the replica-exchange Monte Carlo (REMC) protocol (14), in which $N_{\mathrm{rep}} = 30$ replicas of the domain assembly system are sampled in parallel. The temperature of the ith replica is set by

$$T_i = T_{\min} \times \left(\frac{T_{\max}}{T_{\min}}\right)^{\frac{i-1}{N_{\text{rep}}-1}} \quad (S7)$$

where $T_{\min} = 10/15$ and $T_{\max} = 20$ represent the temperatures of the first and the last replicas, respectively. In every $N_{\text{step}} = 200$ MC movements (rigid-body rotation and translation of the smaller domain), a global swap movement between two contiguous replicas (*i* and *j*) is attempted with the acceptance probability of

$$P_{\text{swap}} = \min\left\{1, \exp\left(\left(E_j - E_i\right)\left(\frac{1}{KT_j} - \frac{1}{KT_i}\right)\right)\right\} \quad (S8)$$

where E_i and E_j are the energies of the *i*th and the *j*th replicas, and T_i and T_j are the corresponding temperatures, respectively. *K* is a constant which equals to 1 in DEMO. This global movement can help to drive the simulation of low-temperature replicas out of local energy basins by swapping conformations with high-temperature replicas.

The ranges of rotation and translation are $[-57.3°, 57.3°]$ and $[-0.5\text{Å}, 0.5\text{Å}]$ respectively. Each replica is terminated when the number of movements reaches to 10,000. The decoy conformation with the lowest energy in the entire simulation is selected as the final model for linker reconstruction and side-chain refinement.

Text S4. Energy Terms Counting for the Cross-Linking and Cryo-EM Density Map Restraints

When the experiment data is available, the corresponding energy terms are added to Eq. (S1) to guide the simulation. For the cross-linking data which specify the maximum distance ($d_{\max}$) for a given residue pair, a ***cross-linking restraint energy*** is calculated by

$$E_{\text{CL}}(i,j) = \begin{cases} -1, & \text{if } d_{ij} < d_{\max} \\ -\frac{1}{2}\left[1 - \sin\left(\frac{d_{ij} - d_a}{d_b}\pi\right)\right], & \text{if } d_{\max} \le d_{ij} \le d_0 \\ \frac{1}{2}\left[1 - \sin\left(\frac{d_{ij} - d_c}{d_d}\pi\right)\right], & \text{if } d_0 < d_{ij} \le 80\text{Å} \\ 1, & \text{otherwise} \end{cases} \quad (S9)$$

where d_{ij} is the distance between the *i*th C_α atom of the N-terminal domain and the *j*th C_α atom of the C-terminal domain in the decoy structure, and we only consider cross-links on residues with $|i - j| > 5$. This term only involves two free parameters, i.e., $d_{\text{well}} = 2.0$ and the weight w_{CL}=1.2, which determine the strength of the CL restraints and control the speed of the convergence of the simulations towards the target distance (d_{max}). Accordingly, other related parameters are determined by $d_0 = d_{\max} + d_{\text{well}}$, $d_a = (d_{\max} + d_0)/2$, $d_b = d_{\text{well}}$, $d_c = (d_0 + 80)/2$, and $d_d = 80 - d_0$.

For the cryo-EM density map, the ***cryo-EM density correlation restraint*** is calculated by

$$E_{\text{EM}} = 1 - \frac{\sum_{i=1}^{N} \varepsilon(\boldsymbol{y_i})(\rho_0(\boldsymbol{y_i}) - \bar{\rho}_0)(\rho_c(\boldsymbol{y_i}) - \bar{\rho}_c)}{\sqrt{\sum_{i=1}^{N}(\rho_0(\boldsymbol{y_i}) - \bar{\rho}_0)^2}\sqrt{\sum_{i=1}^{N}(\rho_c(\boldsymbol{y_i}) - \bar{\rho}_c)^2}} \quad (S10)$$

where *N* is the total number of grid points in the density map and $\rho_0(\boldsymbol{y_i})$ is the observed density of the *i*th grid point $\boldsymbol{y_i}$. $\rho_c(\boldsymbol{y_i}) = \sum_{j=1}^{L} C \cdot m \cdot e^{(-k|y_i - x_j|^2)}$) is the expected density of $\boldsymbol{y_i}$ calculated from the decoy conformation, where *L* is the length of the sequence, $\boldsymbol{x_j}$ is the position of the *j*th C_α atom in the decoy, *m* is its mass, and $k = [\pi/(2.4 + 2.45R)]^2$ and $C = (k/\pi)^{3/2}$ are parameters controlling the Gaussian damping rate of the electric density with *R* being the resolution of the density map data. $\bar{\rho}_0$ and $\bar{\rho}_c$ are the average values of observed and calculated densities, respectively. $\varepsilon(\boldsymbol{y_i}) = 1 - \prod_{j=1}^{L}[1 - 1/(1 + e^{-(M - |x_j - y_i|)})]$ is the masking function introduced following (15), with *M* = 8Å being the masking distance. The calculation of the density-map correlations is under the Cartesian coordinates system, where the grip point of the density map is transformed by the method in Situs (16). The weight of E_{EM} is set to 120 when it is added to Eq. (S1), and $w_4 = 0$ for the discontinuous domain.

Text S5. Force Field of DEMO for Linker Reconstruction Simulations

The total energy of DEMO for the linker modeling is a sum of the following 4 terms:

$$E_{\text{link}} = w_1 E_{\text{ta}} + w_2 E_{\text{lcl}} + w_3 E_{\text{ba}} + w_4 E_{\text{sct}} \quad (S11)$$

The first term describes the ***torsion angle energy*** by:

$$E_{ta} = -\sum_{i=1}^{l} \log\big(P(\phi_i, \psi_i | R_i, S_i)\big) \quad (S12)$$

where l is the length of the linker; ϕ_i and ψ_i represent the torsion angle pair of the ith residue; R_i and S_i are the residue type and secondary structure type of the ith residue, respectively; $P(\phi_i, \psi_i | R_i, S_i)$ is the conditional probability calculated based on the Ramachandran plot of 6,023 high-resolution protein structures from the PDB (17, 18).

The second term is for reducing the ***linker-domain clash***, which is in the same form as Eq. (S2) but with the distance index running for all residue pairs between linker and domain structures.

The third term is for the ***N-C_α-C bond angle potential*** calculated by

$$E_{\text{ba}} = \sum_{i=1}^{l} 0.5(\alpha_i - \bar{\alpha})^2 \quad (S13)$$

where α_i is the bond angle formed by the N, C_α, and C atoms of the ith residue; $\bar{\alpha} = 110.86°$ is the average value of the bond angle in the PDB structures.

The last term in Eq. (S11) is the generic ***orientation-dependent side-chain contact potential*** extended from I-TASSER (19), i.e., $E_{sct} = \sum_{i,j} E_{pair}\big(g_{ij}, c_{ij}, A_i, A_j\big)$, where

$$E_{pair}(g_{ij}, c_{ij}, A_i, A_j) = \begin{cases} E_p\big(g_{ij}, c_{ij}, A_i, A_j\big), & if\ g_p^1(A_i, A_j) < g_{ij} < g_p^2(A_i, A_j)\ \&\ c_{ij} > 0.5 \\ E_a\big(g_{ij}, c_{ij}, A_i, A_j\big), & if\ g_a^1(A_i, A_j) < g_{ij} < g_a^2(A_i, A_j)\ \&\ c_{ij} < -0.5 \\ E_t\big(g_{ij}, c_{ij}, A_i, A_j\big), & if\ g_t^1(A_i, A_j) < g_{ij} < g_t^2(A_i, A_j)\ \&\ -0.5 \le c_{ij} \le 0.5 \\ 0, \quad \text{otherwise} \end{cases} \quad (S14)$$

Here, g_{ij} is the distance between the side-chain centers of ith and jth residues. c_{ij} measures the relative orientation of the side-chain vectors of the two residues, i.e., $c_{ij} = \vec{c_i} \cdot \vec{c_j}$, where $\vec{c} = (\vec{p} - \vec{q})/|\vec{p} - \vec{q}|$, and $\vec{p}$ and $\vec{q}$ are the Cα vectors, as defined in Figure S11. A_i and A_j are the amino acid type of the residues. $E_{p,a,t}(g_{ij}, c_{ij}, A_i, A_j)$ is the orientation and amino acid specific contact potential derived from 6,500 non-redundant high-resolution PDB structures, where p, a and t refer to the side-chain vectors being in parallel, antiparallel and perpendicular, respectively; $g_{p,a,t}^{1,2}(A_i, A_j)$ denotes the corresponding distance cutoffs used for defining the contacts between the two residues (19).

The weighting parameters in Eq. (S11) are decided by maximizing the correlation between the total energy and RMSD based on the structure decoys of the same 425 training proteins as used above. This results in w_1=0.04, w_2=9.0, w_3=1.35, and w_4=0.25.

Text S6. Force Field for Assembly Refinement Simulations on 3 or More Domains

The energy of the global structural refinement for proteins of 3 or more domains is the sum of all pair-wise domain interactions in every two consecutive domains assembly, i.e.,

$$E_{\text{total}} = \sum_{k=1}^{N_D - 1} E(k, k+1) \quad (S15)$$

where N_D is the total number of domains; $E(k, k+1)$ is the energy calculated by Eq. (S1) for the k-th domain and the (k+1)-th domain assembly.

Text S7. Comparison of Modeling Results Based on Different Template Recognition Programs

To examine the impact of homologous template identification on the domain assembly process, we have tested an alternative pipeline that combines templates from both TM-align and LOMETS (20), the latter of which is a meta-server threading program to search for homologous templates based on advanced sequence and sequence profile alignments. Table S3 lists the test results on a set of 166 2-domain proteins, where DEMO(TM+LOMETS) only achieves slightly better modeling results with TM-score increasing by 1% (0.78 vs. 0.79). Since the difference is

insignificant (with a p-value=0.07 in the Student's t-test) and the inclusion of LOMETS can increase the complexity and implementation time of the pipeline, we stick to the DEMO program based only on TM-align template search.

Supporting Figures

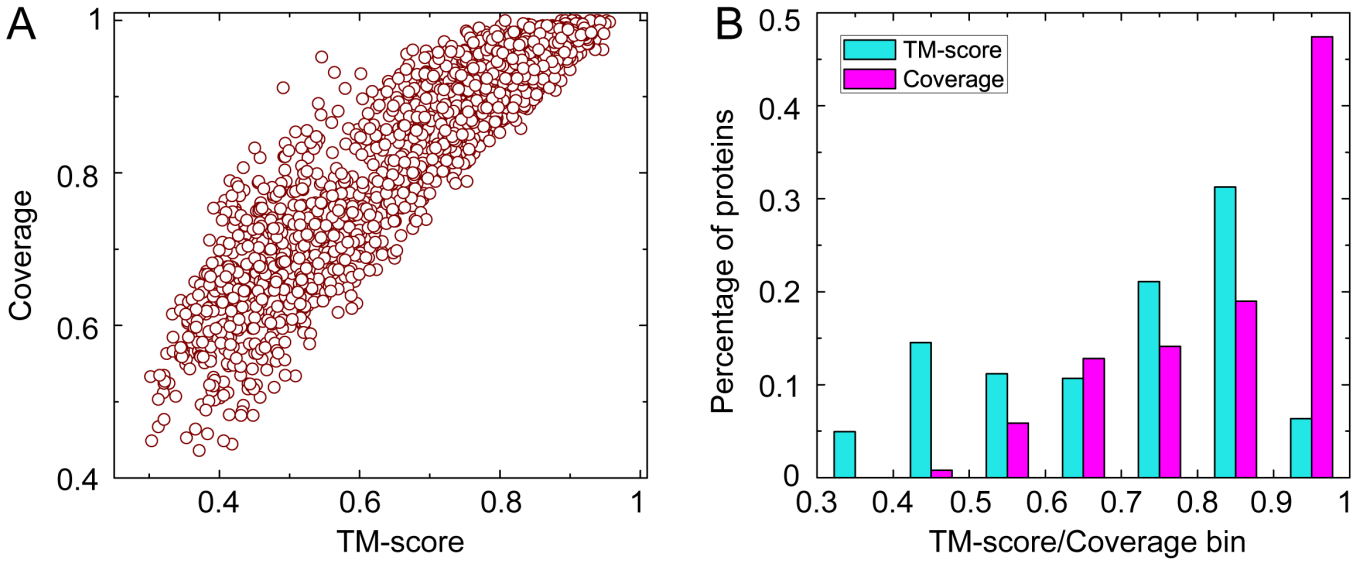

Figure S1. Completeness of multi-domain structure space is examined by structurally comparing 2,269 non-redundant multi-domain target proteins with other proteins in the DEMO template library using TM-align (12), where all homologous templates with a sequence identity >30% to the targets are excluded. (A) alignment coverage (number of aligned residues divided by the total number of residues on the target) versus TM-score. (B) Histogram of TM-score and alignment coverage.

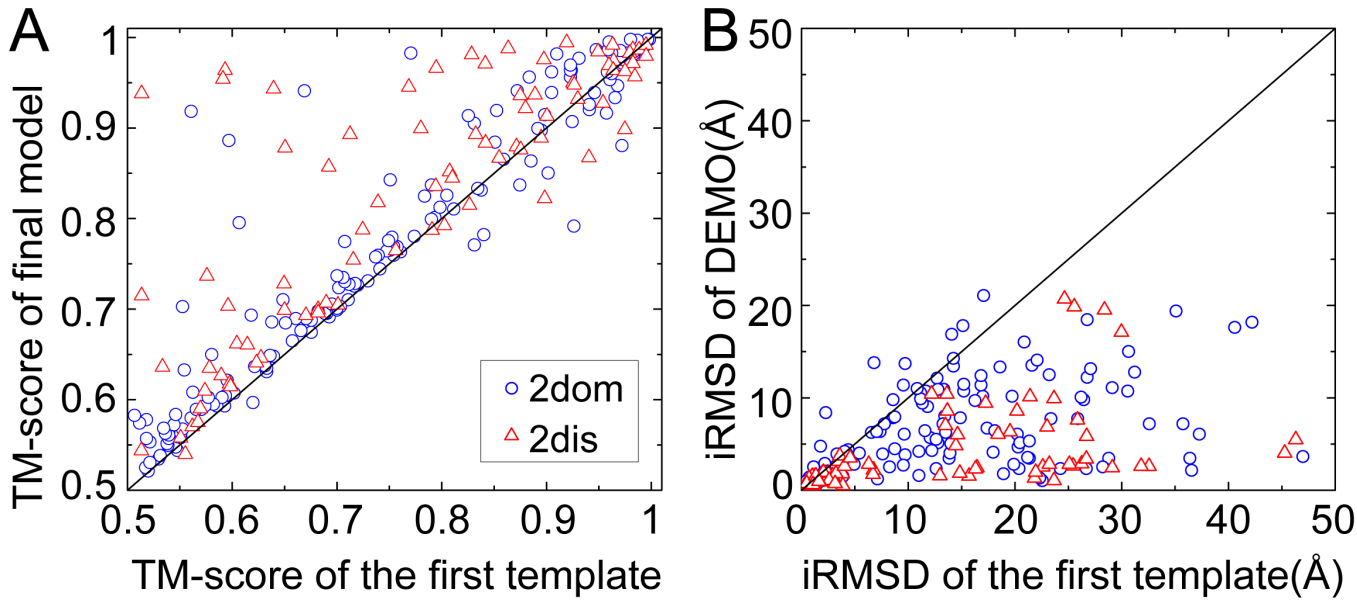

Figure S2. Comparison the final models generated by DEMO with the hybrid models generated by superimposing the experimentally solved domain structures onto the best-scoring structural template from TM-align. (A) TM-score of the first DEMO models versus that of the hybrid models. The figure shows that the DEMO models have the TM-score improved for 78% 2dom cases and 82% 2dis cases compared to the hybrid models. On average, the TM-score was improved from 0.75 to 0.78 for 2dom proteins and 0.77 to 0.84 2dis proteins. (B) iRMSD to native of the first DEMO model versus that of the hybrid models. The DEMO models have the iRMSD decreased for about 78% 2dom cases and 86% 2dis cases. On average, the RMSD was decreased from 11.8 Å to 5.5 Å for 2dom proteins and 11.9 Å to 3.8 Å for 2dis proteins.

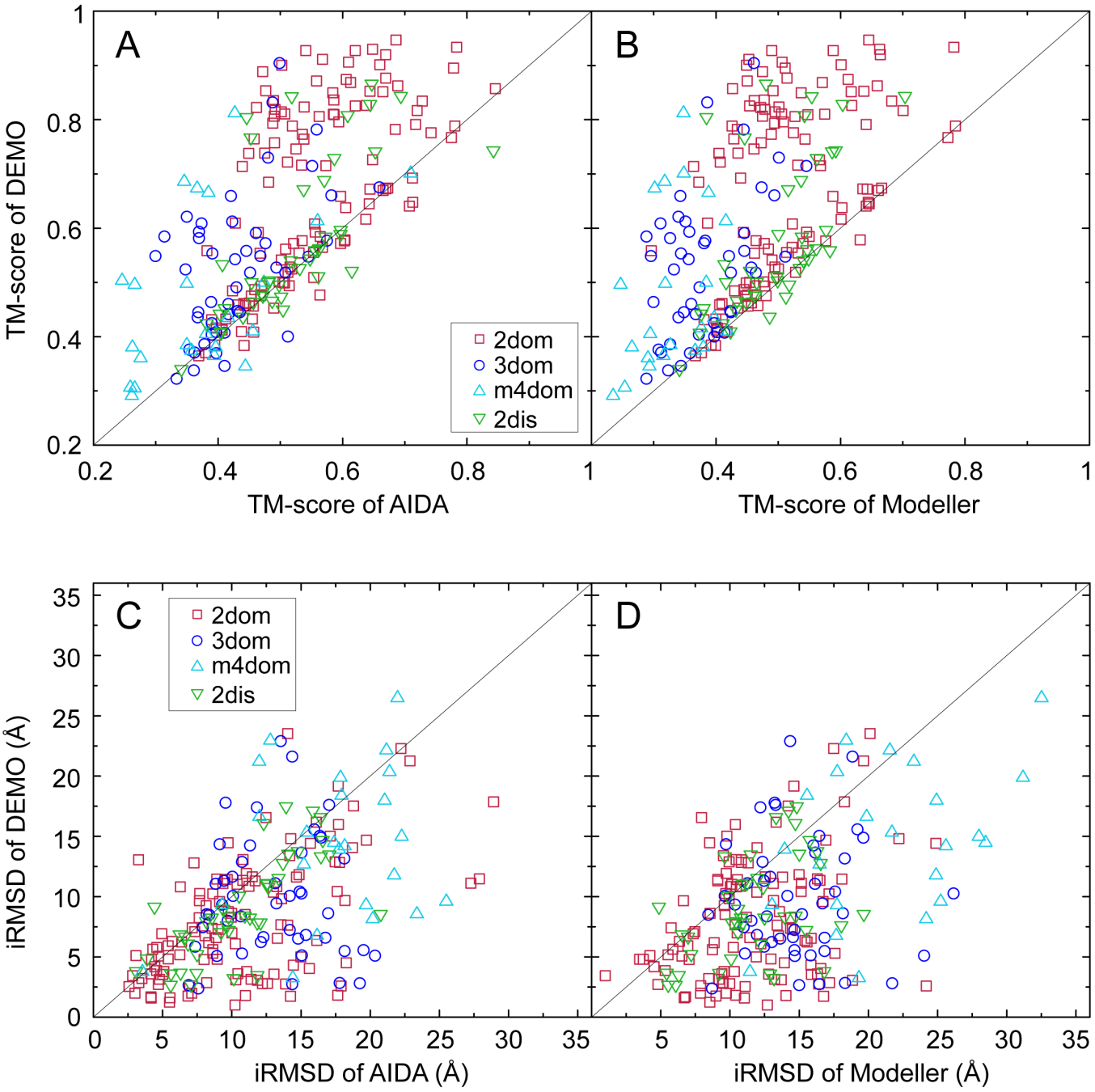

Figure S3. Results of full-length models assembled from domain models predicted by I-TASSER. (A) TM-score of models by DEMO versus that by AIDA. (B) TM-score of models by DEMO versus that by Modeller. (C) iRMSD of models by DEMO versus that by AIDA. (D) iRMSD of models by DEMO versus that by Modeller. Different points represent proteins of different categories of domain structures.

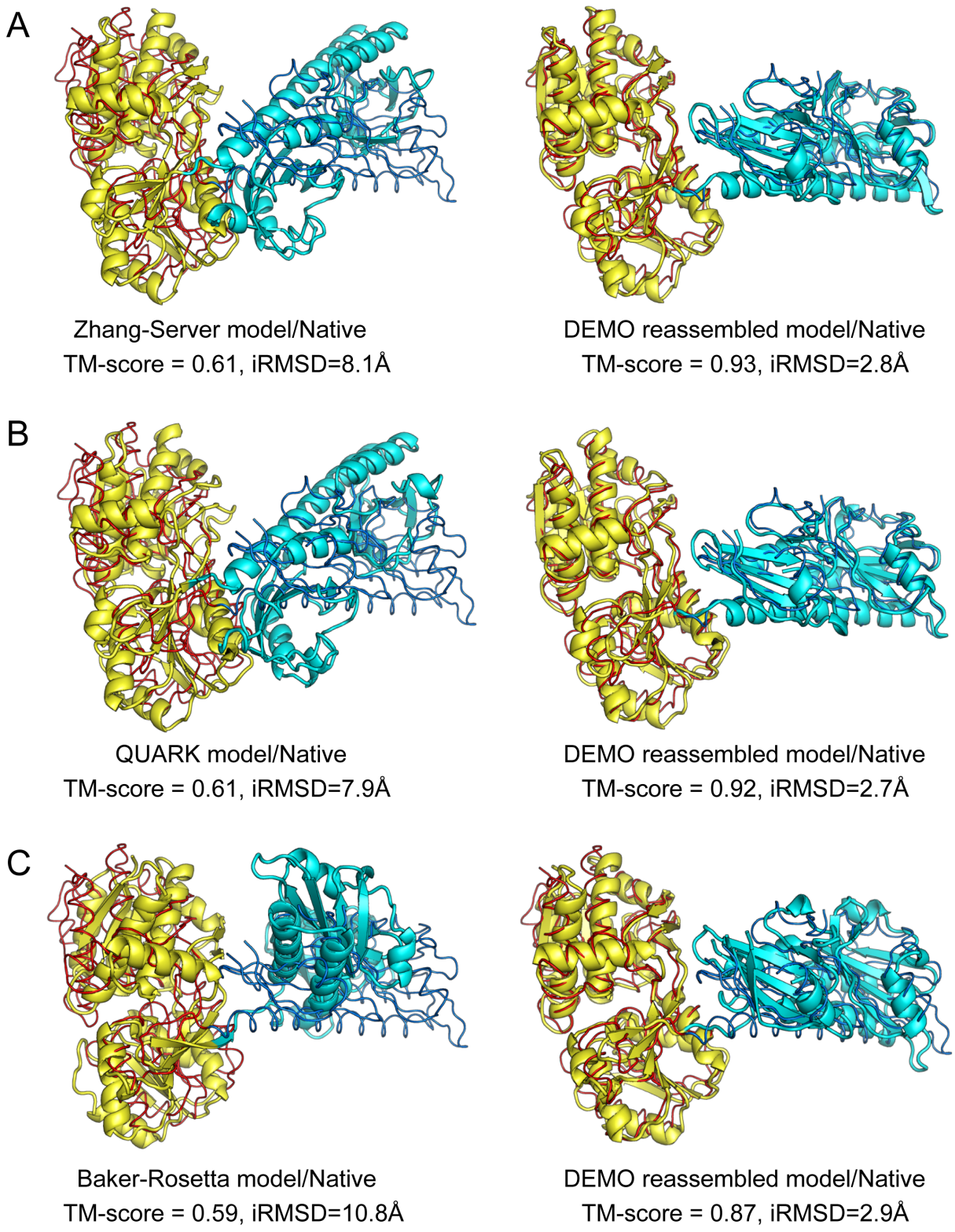

Figure S4. Domain assembly results on a representative example from T0920 in CASP12 which have the domain structure excised from the original full-length models predicted by three top servers. The thin lines represent the experimental structures and cartoons are final models by the server and DEMO (from left to right panel), with different colors indicating different domains. (A) Model from Zhang-Server; (B) Model from QUARK; (C) Model from Baker-Rosetta.

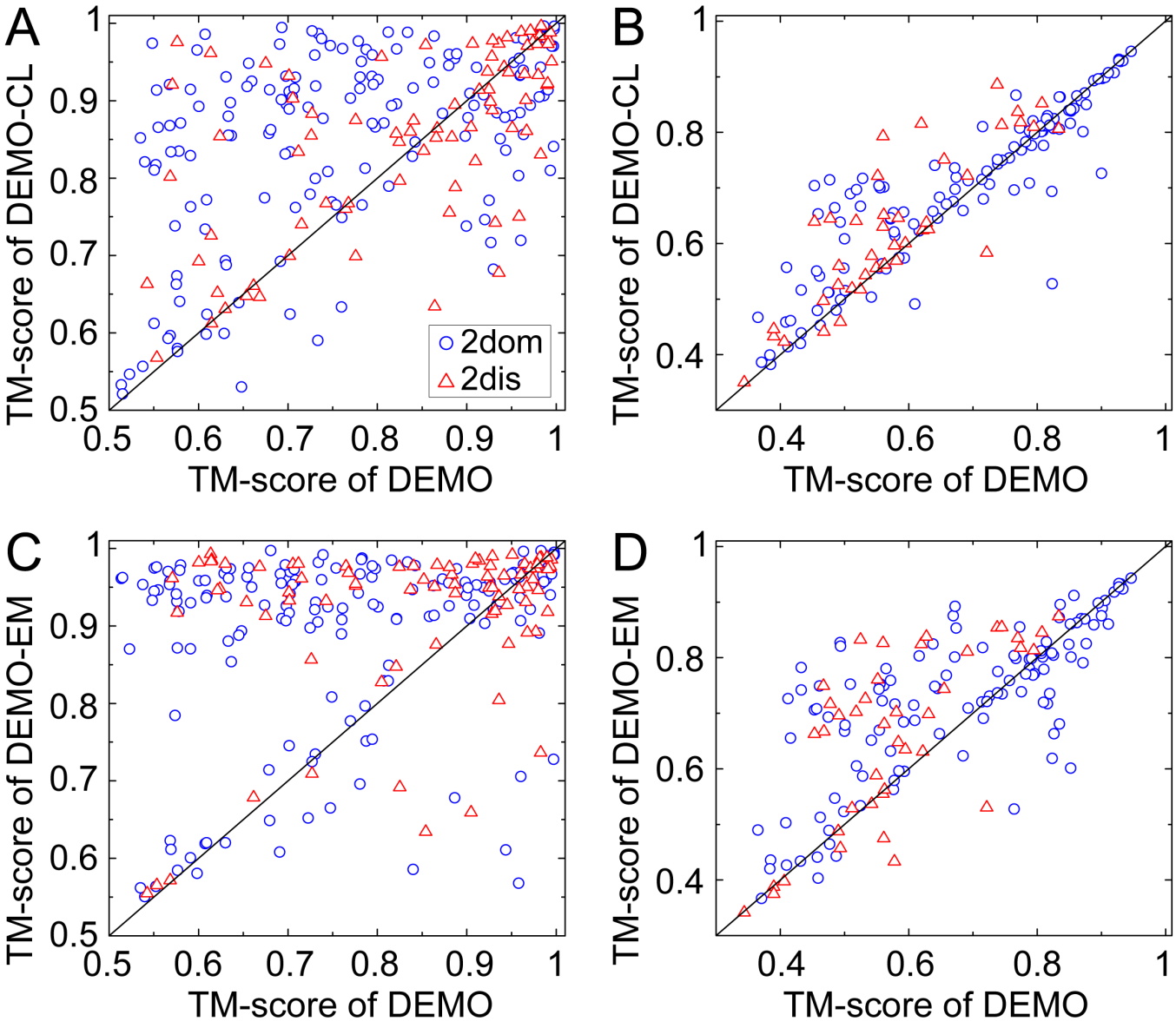

Figure S5. Comparison the final models generated by DEMO with or without using experimental data restraints. (A) TM-score of models by DEMO versus that by cross-link assisted DEMO (DEMO-CL) using the experimentally solved domain models. (B) TM-score of models by DEMO versus that by DEMO-CL using the I-TASSER predicted domain models. (C) TM-score of models by DEMO versus that by cryo-EM assisted DEMO (DEMO-EM) using the experimentally solved domain models. (D) TM-score of models by DEMO versus that by DEMO-EM using the I-TASSER predicted domain models.

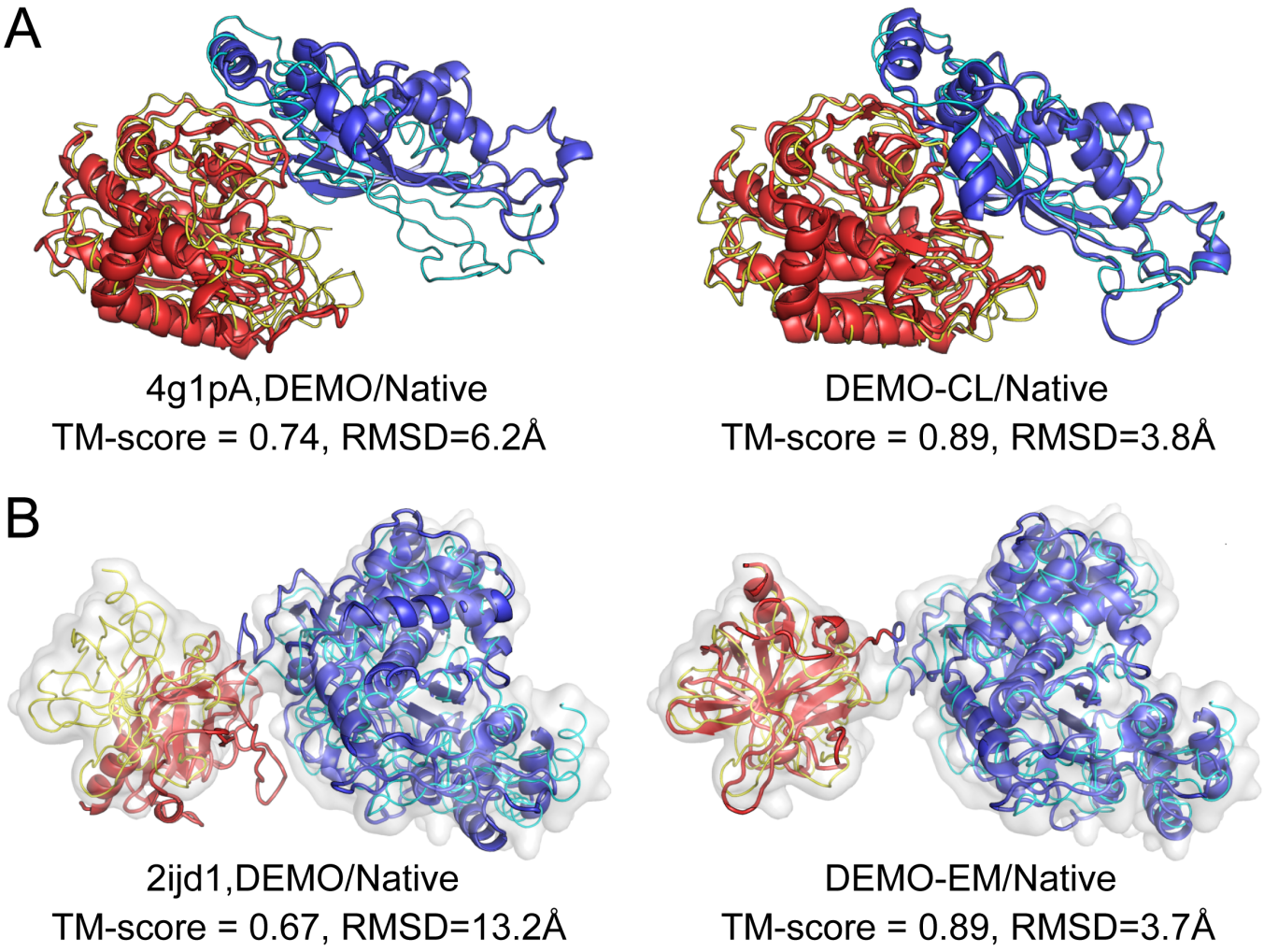

Figure S6. Illustrative examples of domain assembly based on I-TASSER predicted domain structures assisted with experimental restraints. The thin lines are experimental models and cartoons represent final model assembled by DEMO pipelines, with different colors indicating different domains. (A) 4g1pA assisted with cross-linking data; (B) 2ijd1 assisted with cryo-EM data.

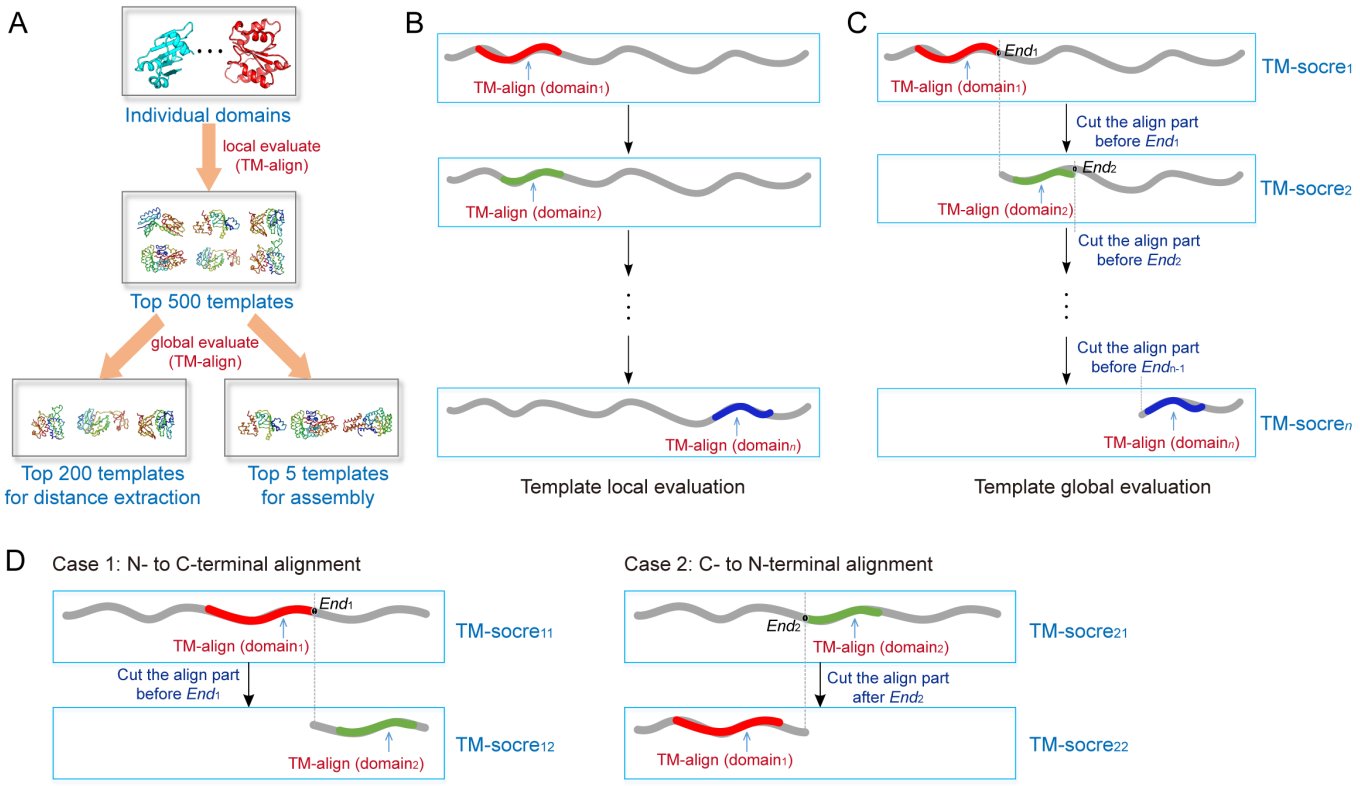

Figure S7. Domain-structure based template identification. (A) The overall process of template identification, which consists of two steps of local and global template searches. The top templates are selected, as ranked by G-score, for inter-domain distance profile derivation and initial full-length model construction respectively. (B) Procedure of local structure search, where the gray line represents the template chain and other color lines are for different domains of the query. In this step, individual domains of the query are matched to the complex templates by TM-align, regardless of the domain overlap, where the average TM-score of all domains is defined as the local score (L-score) for the template. (C) Procedure of global template search which is performed on 500 templates with the highest L-scores. The individual query domains are aligned, in a consecutive order from N- to C-terminal, to the template structure by TM-align (12), where no overlap is allowed between domains. The average TM-score of the domains is defined as the global score (G-score). (D) Illustration of the two cases in the template global search for a 2-domain protein. In Case 1, we first match the N-terminal domain (domain-1) to the template by TM-align and denote the C-terminal ending residue of the alignment as End_1 on the template sequence. Next, we match the C-terminal domain (domain2) to the remaining region of the template ranging from End_1 to the C-terminal of the template. In Case 2, we first match domain2 to the template by TM-align and mark the N-terminal ending residue as End_2 on the template, and then match the domain1 to the rest of the template region ranging from the N-terminal to End_2 along the template sequence. Since the structural result is asymmetric, the two cases have different alignment results. The G-score of Case1 and Case2 are G-score$_1$=(TM-score$_{11}$+TM-score$_{12}$)/2 and G-score$_2$= (TM-score$_{21}$ +TM-score$_{22}$)/2, respectively. The alignment with higher G-score between G-score$_1$ and G-score$_2$ is selected for the initial model generation.

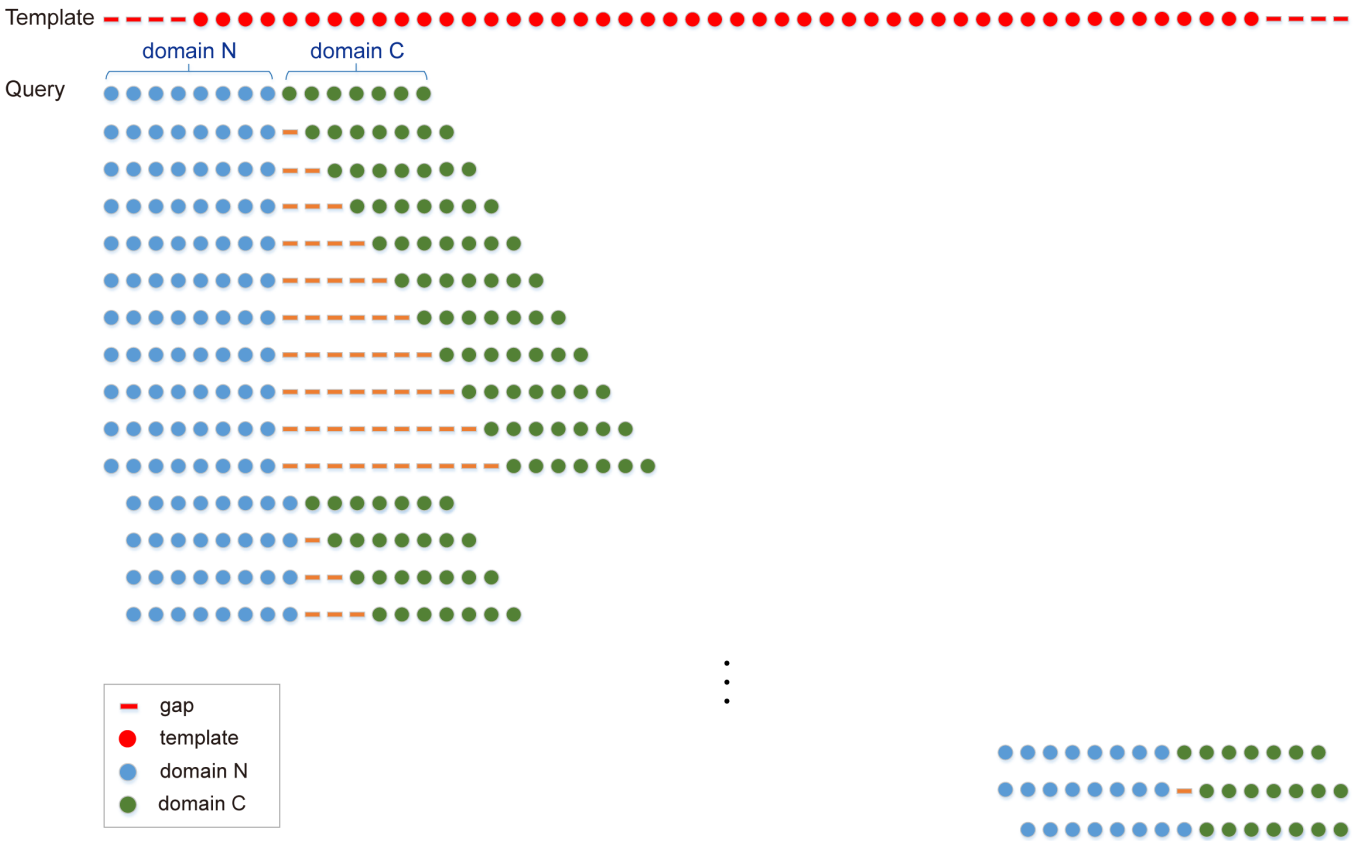

Figure S8. Sliding-window procedure for query-template alignment search and initial model construction. The N-terminal domain of the query is first superposed at the N-terminal of the template, where C-terminal domain is superposed at all the right-hand positions of the N-terminal domain along the template sequence, but with a maximum gap of 10 residues from N-terminal domain. Next, the superposition of N-terminal domain is shifted by one residue to the C-terminal of the template and redo the C-terminal superpositions. This procedure is repeated with the N-terminal domain sliding through all positions along the templates, where C-terminal domain is always on the right hand of the N-terminal domains. To save time, the superposition is initially performed by Kabsch RMSD rotation matrix (21) on all the positions. The top-10 alignment positions with the lowest average RMSD are selected, whose superpositions are then regenerated by the TM-score rotation matrix (22). The alignment with the highest average TM-score of the N/C-domains among all the positions is finally selected for initial model construction. Here, structural superposition without gap (instead of structural alignment with gap) is performed for each comparison of query domain and template structures. The two ending terminals of 20 residues were skipped during domain sliding to further save time.

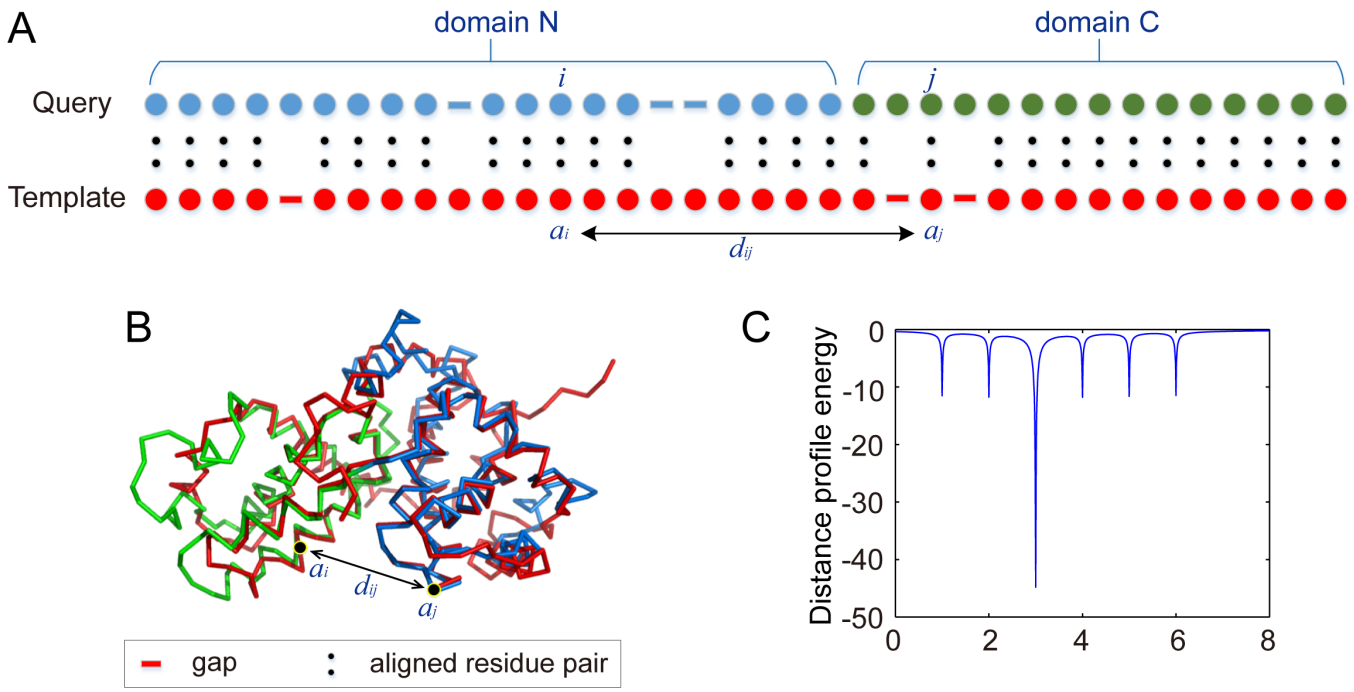

Figure S9. Illustration of the inter-domain distance profile. (A) In the top 200 templates, if the *i*th residue of domain-N and the *j*th residue of domain-C have the aligned residue a_i and a_j in the corresponding template, the C_α distance d_{ij} between a_i and a_j is calculated, with a distance profile constructed from the distances mapped from all the templates. (B) The distance d_{ij} between the aligned residue a_i and a_j in the template, where the red, green, and blue structure represents the template, domain N, and domain C of the query, respectively. (C) Example of the distance profile energy with the distance profile d_{ij}= [1, 2, 3, 3, 3, 3, 4, 5, 6 Å]. As shown in the figure, the residue pair will obtain the lowest energy if their distance is close to 3 Å which appears most often in the distance profile.

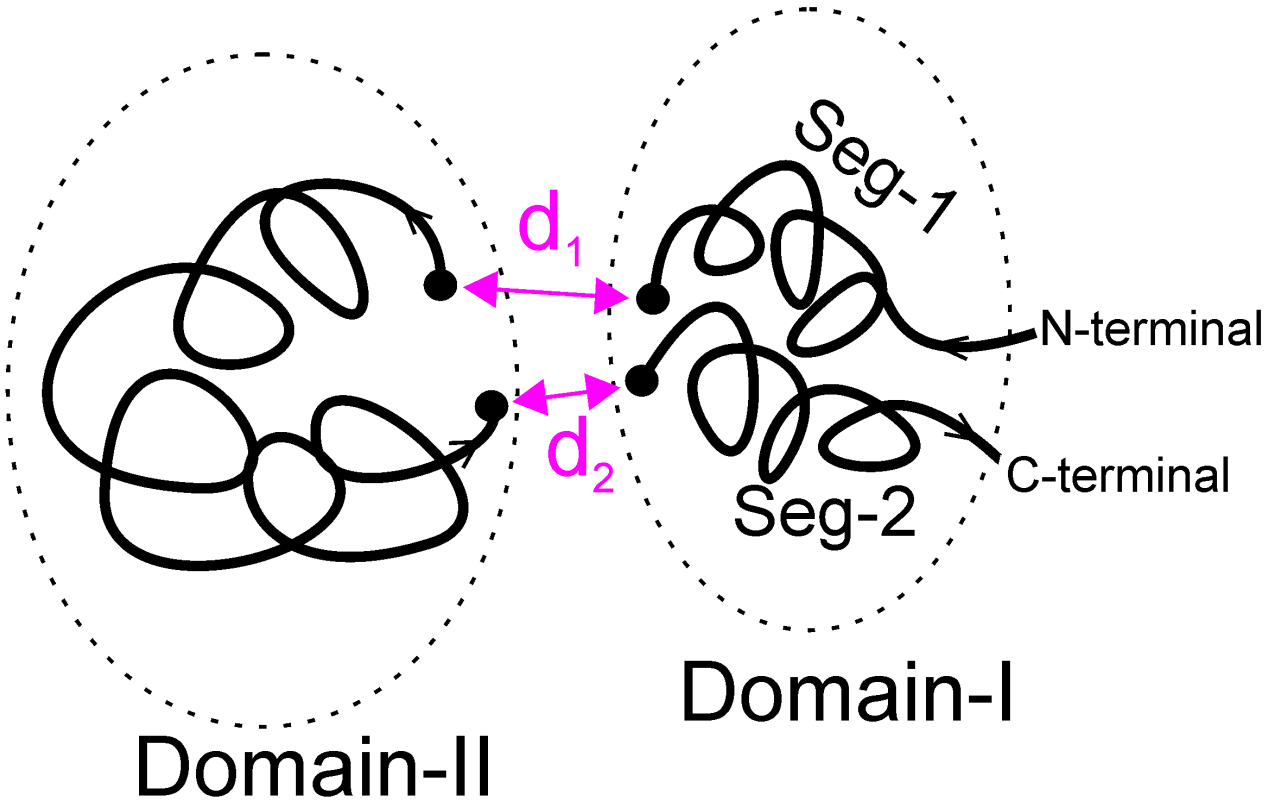

Figure S10. Illustration of domain boundary distance potential for a two-domain protein with discontinuous domains. The discontinuous domain (Domain-I) is split into two segments due to the insertion of the continuous domain (Domain-II). d_1 and d_2 are Cα-distances which are constrained to 3.8 Å by boundary distance energy Eq. (S5).

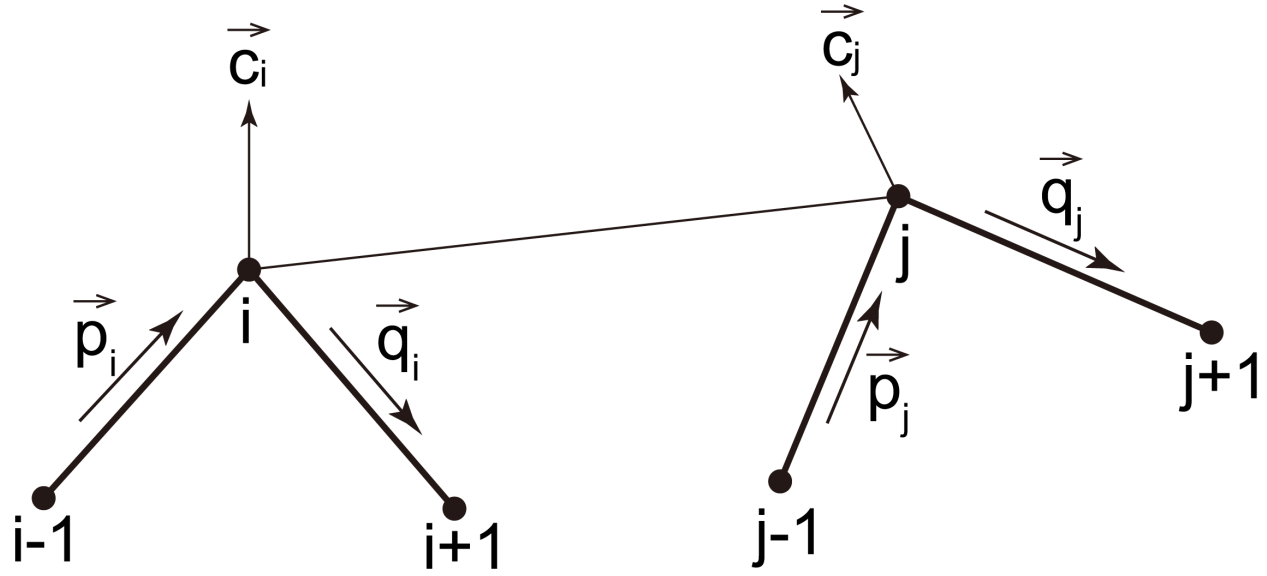

Figure S11. Definition of the relative orientation and orientation-dependent side-chain contact potential between Residues i and j. Here, $\vec{p}$ and $\vec{q}$ are C_α-C_α vectors, and $\vec{c} = (\vec{p} - \vec{q})/|\vec{p} - \vec{q}|$ is the unit vector defining the orientation of a local structure. The relative orientation between residues i and j is defined according to $cc_{ij} = \vec{c_i} \cdot \vec{c_j}$, i.e., parallel ($cc_{ij} > 0.5$), antiparallel ($cc_{ij} < -0.5$), and perpendicular ($-0.5 \leq cc_{ij} \leq 0.5$).

References

1. Xu Y, Xu D, Gabow HN (2000) Protein domain decomposition using a graph-theoretic approach. *Bioinformatics* 16(12):1091-1104.
2. Lam SD*, et al.* (2015) Gene3D: expanding the utility of domain assignments. *Nucleic acids research* 44(D1):D404-D409.
3. Chandonia J-M, Fox NK, Brenner SE (2017) SCOPe: manual curation and artifact removal in the structural classification of proteins–extended database. *Journal of molecular biology* 429(3):348-355.
4. Wu S, Zhang Y (2008) MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins* 72(2):547-556.
5. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* 5(4):725-738.
6. Zheng W*, et al.* (2019) LOMETS2: Improved meta-threading server for fold-recognition and structure-based function annotation for distant-homology proteins. *Nucleic acids research*:In press.
7. Simons KT, Strauss C, Baker D (2001) Prospects for *ab initio* protein structural genomics. *J. Mol. Biol.* 306:1191-1199.
8. Zhang Y, Kolinski A, Skolnick J (2003) TOUCHSTONE II: A new approach to ab initio protein structure prediction. *Biophys. J.* 85:1145-1164.
9. Xu D, Zhang Y (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* 80(7):1715-1735.
10. Park H, Ovchinnikov S, Kim DE, DiMaio F, Baker D (2018) Protein homology model refinement by large-scale energy optimization. *Proceedings of the National Academy of Sciences of the United States of America* 115(12):3054-3059.
11. Rost B (1999) Twilight zone of protein sequence alignments. *Protein Eng* 12(2):85-94.
12. Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic. Acids Res.* 33(7):2302-2309.
13. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. *The journal of chemical physics* 21(6):1087-1092.
14. Swendsen RH, Wang J-S (1986) Replica Monte Carlo simulation of spin-glasses. *Physical review letters* 57(21):2607-2609.
15. DiMaio F, Tyka MD, Baker ML, Chiu W, Baker D (2009) Refinement of protein structures into low-resolution density maps using rosetta. *Journal of molecular biology* 392(1):181-190.
16. Wriggers W, Milligan RA, McCammon JA (1999) Situs: a package for docking crystal structures into low-resolution maps from electron microscopy. *Journal of structural biology* 125(2-3):185-195.
17. Ramachandran GT, Sasisekharan V (1968) Conformation of polypeptides and proteins. *Advances in protein chemistry*, (Elsevier), Vol 23, pp 283-437.
18. Xu D, Zhang Y (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Structure, Function, and Bioinformatics* 80(7):1715-1735.
19. Yang J*, et al.* (2015) The I-TASSER Suite: protein structure and function prediction. *Nature methods* 12(1):7.
20. Wu S, Zhang Y (2007) LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic acids research* 35(10):3375-3382.
21. Kabsch W (1978) A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical General Crystallography* 34(5):827-828.
22. Zhang Y, Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. *Proteins* 57:702-710.