**Supplementary Information**

**Table of Contents**

**Supplementary Texts**

* **Text S1.** Initial full-length model generation by sliding-window based alignment
* **Text S2.** Hybrid energy function for domain assembly
* **Text S3.** Full-length decoys generation using rotation angles and translation vectors

**Supplementary Figures**

* **Figure S1.** Global and local templates identification
* **Figure S2.**Sliding-window procedure for domain-template alignment search and initial model construction
* **Figure S3**. Relationship between the eTM-score/eRMSD and the actual TM-score/RMSD to the native
* **Figure S4.** Example of continuous and discontinuous domain

**Supplementary Tables**

**Table S1.** Results of full-length models generated by different methods for different categories

**Supplementary References**

**Supplementary Texts**

**Text S1. Initial full-length model generation by sliding-window based alignment**

The initial full-length conformations are built based on the top 10 global templates and local templates selected according to TM-scoreh. Since the domain alignments are performed separately, the aligned regions of domains may be far away from each other. In this case, a sliding-window based procedure is employed to recreate domain alignments so that neighboring domains have the initial structure constructed from the neighboring regions of the template. Take a protein with 2 domains shown in **Figure S2** as the example, 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 (1) 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 (2). 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.

**Text S2. Hybrid energy function for DEMO2 domain structure assembly**

The energy function for domain assembly of DEMO2 is a sum of the ten terms:

where *m* and *n* are domain index, and is the total number of domains.

The first term is the *inter-domain distance map*:

where and represent the sequence length of the *m-*th and *n-*th domain, respectively. is the distance between the *i-*th ( for Glycine) atom in the *m-*th domain and *j-*th atom in the *n-*th domain, is the predicted probability of the distance located in the *k-*th distance bin, and is the pseudo count to offset low-probability bins. In the calculation, we only consider atom pairs with probability peak located in [2Å, 20Å], and these atom pairs with predicted probabilities >0.5 in the last bin [>20 Å], which represents a low prediction confidence in [2Å, 20Å], are excluded.

The second term is the *inter-domain orientations*:

where represents the inter-residue θ, , or angles defined in Ref. (3), is the predicted probability of the angle located in the *k-*th angle bin.

The third term is the *domain-domain interface contact energy*:

where is the confidence score of the *i*-th residue and *j*-th residue with the C distance <18. A similar potential is also used to count for cross-link restraints when they are available, where is set to 1 with the distance cutoffs taken directly from the user-input cross-link data.

The fourth term is the *hydrogen bond restraints*. The predicted probability distribution of angles is converted into an energy potential with a similar from as the distance energy, where the potential is described as follows:

where is the hydrogen angle between residue pair *i* and *j*, i.e. the angle between vector and , which follows a probability distribution predicted by DeepPotential, is the probability that the angle is located at . The illustration of the hydrogen bond restraints is shown in(4).

The fifth term is designed to eliminate *steric clashes* between domains, i.e.,

where is set as the clash distance cutoff.

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

where the scale parameter depends on the hydrophobic and hydrophilic features of the residue pairs. , if both of the residues are hydrophobic (ALA, CYS, VAL, ILE, PRO, MET, LEU, PHE, TYR, TRP); if the two residues are hydrophilic (SER, THR, ASP, ASN, LYS, GLU, GLN, ARG, HIS); or , otherwise. This energy item is used to control the inter-domain distance, which will push the two domains together if they are two far away each other.

The seventh term is the *domain-domain distance profile* deducedfrom the templates identified by TM-align, which is calculated by:

For a residue pair (*i* and *j*, with *i* from N-terminal domain and *j* from C-terminal domain), 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) , where and are the indexes of the aligned residues of *i* and *j* on the template. is the distance between the residue and in the *t-*th template.

The eighth term is the *domain boundary energy* is defined as

where is the C distance between two consecutive domains, and is the standard length of C-C bond.

The nineth term is the *local domain distance restraint*:

where represents the distance between the *i-*th atom () and its corresponding atom in the initial structure generated in the template superposition process, and is the length of the protein. This term is to prevent the assembly deviating too much from the orientation obtained from the template.

The last term is *radius of gyration restraint*, defined as

where is the radius of gyration of the decoy structure, and are the maximum and minimum estimated radius of gyration, respectively. (*L* is the query sequence length) is the statistical minimum radius of gyration based on the known multi-domain protein models in the PDB. is the statistical maximum radius of gyration based on the known multi-domain protein models in the PDB, where is the number of residues of the longest helix.

The weighting parameters in Eq. (S1) are determined by maximizing the correlation between total energy and RMSD to the native on the structure decoys over a training set of 425 non-redundant proteins through a improved differential evolution algorithm (5,6). This resulted in , , , , , , , , , and for proteins with the template score (TplScore) <0.85, and , , , , , , , , , and for other proteins.

**Text S3. Full-length structure decoy generation using rotation angles and translation vectors**

According to inter-domain rotation angles ∅, 𝛉, and 𝝍, the rotation matrix can be calculated by

where , *i*=1,2,3, *j*=1,2,3 indicates the element of the matrix. Based on the inter-domain rotation matrix and translation vector, the position of each atom in the domain can be calculated by

where (, , ) is the translation vector of the domain, (, , ) is the initial position of the *m*-th atom, (, , ) is the new position of the *m*-th atom after the transition, (, , ) is the center point of the domain model. The new full-length structural decoy is generated by calculating the position of each atom in each domain according to the corresponding rotation angles and translation vector.

**Supplementary Figures**

**D:\论文\所写论文\在审—DEMO2_Multi-domain protein structures assembly by coupling quaternary structural alignment with deep-learning inter-domain restraint prediction\Figure\SI_Figure\FigureS_tmp_detect\template\fig_template_search.tif**

**Figure S1.**Global and local templates identification. (**A**) Flowchart of the template identification. (**B**) Template local evaluation, where the overlap between the alignments of different domains is allowed. (**C**) Template global evaluation with no overlap allowed in the alignments of different domains. The local template is evaluated by the global evaluation for every two consecutive domains. (**D**) Global template identification, where the fourth domain cannot be covered by the template. Therefore, the templates that can cover domains 1-3 and the templates that can cover domains 3-4 are independently detected from the library. Finally, the initial full-length model is generated by connecting the two templates according to the alignment of domain 3.

**D:\论文\所写论文\在审—DEMO2_Multi-domain protein structures assembly by coupling quaternary structural alignment with deep-learning inter-domain restraint prediction\Figure\SI_Figure\FigureS_superpose\Figure.tif**

**Figure S2.**Sliding-window procedure for domain-template alignment search and initial model construction. In this procedure, the N domain is superposed with every position along the template, where at each position, the C domain is allowed to superpose in the remaining regions of the template at a maximum of 10 residues away from the N domain. The alignment with the highest average TM-score is finally selected to construct the initial full-length model for the query sequence.

**D:\论文\所写论文\在审—DEMO2_Multi-domain protein structures assembly by coupling quaternary structural alignment with deep-learning inter-domain restraint prediction\Figure\SI_Figure\FigureS_eTM\Cscore.tif**

**Figure S3.** Relationship between the eTM-score/eRMSD and the actual TM-score/RMSD to the native. (**A**) The relationship between the eTM-score and the actual TM-score of the first model assembled by DEMO2, where TP, FP, TN, and FN represent the number of true positive, false positive, true negative, and false negative cases with correct global fold (TM-score > 0.5). (**B**) The relationship between the eRMSD and the actual RMSD of the first model assembled by DEMO2.

**D:\论文\所写论文\在审—DEMO2_Multi-domain protein structures assembly by coupling quaternary structural alignment with deep-learning inter-domain restraint prediction\Figure\SI_Figure\FigureS_example_discon\domainexample.tif**

**Figure S4.** Example of continuous and discontinuous domain. (**A**) A protein (PDBID: 4gslA) contains two continuous domains, where the first domain (blue) ranges from residue 1 to residue 287 and the second domain (red) covers residues from 288 to 598. (**B**) A protein (PDBID: 1itwA) consists of a discontinuous domain and a continuous domain. The first domain is a discontinuous domain which contains two separate segments at the sequence level, where the first segment (blue) ranges from residue 1 to residue 139, and the second segment (yellow) ranges from residue 572 to residue 740. The second domain (red) is a continuous domain inserted between the two segments of the discontinuous domain, and it covers the residues from 140 to 571.

**Supplementary Tables**

**Table S1.** Results of full-length models generated by different methods for different categories. Bold font highlights the best results from each category.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Continuous domain | Category | Method | TM-score | rTM-score | RMSD(Å) |
| 2dom  (*N* = 155) | AIDA | 0.57 | 0.27 | 13.6 |
| DEMO | 0.64 | 0.39 | 11.0 |
| DMPfold | 0.56 | 0.33 | 12.8 |
| trRosetta | 0.63 | 0.42 | 10.3 |
| DEMO2 | **0.70** | **0.48** | **8.9** |
| 3dom  (*N* = 65) | AIDA | 0.47 | 0.14 | 19.0 |
| DEMO | 0.57 | 0.28 | 14.1 |
| DMPfold | 0.51 | 0.23 | 16.4 |
| trRosetta | 0.57 | 0.31 | 13.1 |
| DEMO2 | **0.64** | **0.36** | **11.0** |
| m4dom  (*N* = 40) | AIDA | 0.37 | 0.08 | 25.1 |
| DEMO | 0.44 | 0.15 | 21.0 |
| DMPfold | 0.44 | 0.15 | 23.4 |
| trRosetta | 0.54 | 0.23 | 16.5 |
| DEMO2 | **0.60** | **0.27** | **15.3** |
| All  (*N* = 260) | AIDA | 0.52 | 0.20 | 16.7 |
| DEMO | 0.59 | 0.32 | 13.4 |
| DMPfold | 0.53 | 0.28 | 15.4 |
| trRosetta | 0.60 | 0.36 | 12.0 |
| DEMO2 | **0.67** | **0.42** | **10.4** |
| Discontinuous domain | 2dom  (*N* = 149) | AIDA | 0.58 | 0.28 | 14.1 |
| DEMO | 0.69 | 0.50 | 10.0 |
| DMPfold | 0.63 | 0.45 | 11.0 |
| trRosetta | 0.69 | 0.51 | 9.7 |
| DEMO2 | **0.75** | **0.60** | **7.4** |
| 3dom  (*N* = 33) | AIDA | 0.49 | 0.28 | 15.4 |
| DEMO | 0.69 | 0.30 | 12.1 |
| DMPfold | 0.63 | 0.36 | 12.6 |
| trRosetta | 0.73 | 0.46 | 9.6 |
| DEMO2 | **0.78** | **0.52** | **8.5** |
| m4dom  (*N* = 19) | AIDA | 0.31 | 0.17 | 27.4 |
| DEMO | 0.54 | 0.27 | 23.0 |
| DMPfold | 0.58 | 0.25 | 20.0 |
| trRosetta | 0.66 | 0.30 | 14.6 |
| DEMO2 | **0.70** | **0.34** | **13.0** |
| All  (*N* = 201) | AIDA | 0.54 | 0.27 | 15.5 |
| DEMO | 0.68 | 0.46 | 11.6 |
| DMPfold | 0.63 | 0.42 | 12.1 |
| trRosetta | 0.69 | 0.48 | 10.1 |
| DEMO2 | **0.75** | **0.56** | **8.1** |

2dom: protein with 2 domains.

3dom: protein with 3 domains.

m4dom: protein with 4 or more domains.

Discontinuous domain: protein contains domains which consist of segments from separate regions of the query sequence.

**Supplementary References**

1. 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**, 827-828.

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

3. Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S. and Baker, D. (2020) Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, **117**, 1496-1503.

4. Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J. and Zhang, Y. (2015) The I-TASSER Suite: protein structure and function prediction. *Nature methods*, **12**, 7-8.

5. Zhou, X.G. and Zhang, G.J. (2017) Abstract convex underestimation assisted multistage differential evolution. *IEEE transactions on cybernetics*, **47**, 2730-2741.

6. Zhou, X.G., Peng, C.X., Liu, J., Zhang, Y. and Zhang, G.J. (2020) Underestimation-assisted global-local cooperative differential evolution and the application to protein structure prediction. *IEEE Transactions on Evolutionary Computation*, **24**, 536-550.