Bioinformatics, 38(19), 2022, 4513–4521
https://doi.org/10.1093/bioinformatics/btac553
Advance Access Publication Date: 13 August 2022
Original Paper

Structural bioinformatics

Structural analogue-based protein structure domain assembly assisted by deep learning

Chun-Xiang Peng†, Xiao-Gen Zhou †, Yu-Hao Xia, Jun Liu, Ming-Hua Hou and Gui-Jun Zhang *

College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Lenore Cowen

Received on March 6, 2022; revised on July 27, 2022; editorial decision on August 6, 2022; accepted on August 8, 2022

Abstract

Motivation: With the breakthrough of AlphaFold2, the protein structure prediction problem has made remarkable progress through deep learning end-to-end techniques, in which correct folds could be built for nearly all single-domain proteins. However, the full-chain modelling appears to be lower on average accuracy than that for the constituent domains and requires higher demand on computing hardware, indicating the performance of full-chain modelling still needs to be improved. In this study, we investigate whether the predicted accuracy of the full-chain model can be further improved by domain assembly assisted by deep learning.

Results: In this article, we developed a structural analogue-based protein structure domain assembly method assisted by deep learning, named SADA. In SADA, a multi-domain protein structure database was constructed for the full-chain analogue detection using individual domain models. Starting from the initial model constructed from the analogue, the domain assembly simulation was performed to generate the full-chain model through a two-stage differential evolution algorithm guided by the energy function with an inter-residue distance potential predicted by deep learning. SADA was compared with the state-of-the-art domain assembly methods on 356 benchmark proteins, and the average TM-score of SADA models is 8.1% and 27.0% higher than that of DEMO and AIDA, respectively. We also assembled 293 human multi-domain proteins, where the average TM-score of the full-chain model after the assembly by SADA is 1.1% higher than that of the model by AlphaFold2. To conclude, we find that the domains often interact in the similar way in the quaternary orientations if the domains have similar tertiary structures. Furthermore, homologous templates and structural analogues are complementary for multi-domain protein full-chain modelling.

Availability and implementation: http://zhanglab-bioinf.com/SADA

Contact: zgj@zjut.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Multi-domain proteins are an important class of proteins and consist of more than one unique folding unit that is connected into a single chain. Many biological functions rely on the interaction of different domains (Zhou *et al.*, 2022). For example, in ribose-binding protein in bacterial transport and chemotaxis, the ligand-binding site is located in a pocket formed by two domains. More than 80% of eukaryotic proteins and 67% of prokaryotic proteins contain multiple domains. However, only one-third of structures in the Protein Data Bank (PDB) contains multiple domains, which indicates that it may be more difficult to experimentally determine the structure of multi-domain protein structure.

With the development of deep learning, protein structure prediction methods have been developed successively, such as trRosetta (Du *et al.*, 2021; Su *et al.*, 2021; Yang *et al.*, 2020), RaptorX (Xu, 2019; Xu and Wang, 2019), RocketX (Liu *et al.*, 2022) and D-I-TASSER (Zheng *et al.*, 2021). Recently, end-to-end methods, such as AlphaFold2 (Jumper *et al.*, 2021) and RoseTTAFold (Baek *et al.*, 2021), have been proposed to accurately predict more complex proteins including some multi-domain proteins. The performance of these methods relies to some extent on the quality of the MSA or the

homologous template (Pearce and Zhang, 2021). For example, the distributions of average confidence scores for AlphaFold2 models of human proteins with and without homologues available in the PDB were different (Jones and Thornton, 2022). However, homologues available in the PDB may be fewer for multi-domain proteins, which may further affect the performance of multi-domain protein structure prediction. Meanwhile, the end-to-end approach entails high demand on computing hardware. For relatively large proteins (i.e. >800 amino acids), most computer memory hardly satisfies its training and running requirements, which may not be helpful to the study of their intrinsic mechanisms. Therefore, based on the divide-and-conquer strategy, modelling the full-chain structure of multi-domain protein by domain assembly may be a lightweight alternative way. It may be also helpful for further improving the accuracy of end-to-end methods and revealing the folding mechanism but has been largely ignored by the community (Zhou *et al.*, 2019).

In general, domain assembly methods are mainly divided into two categories: *de novo*-based methods and template-based approaches. *De novo*-based methods focus mainly on construction of the linker models by some *de novo* or *ab initio* folding potentials. In Rosetta (Wollacott *et al.*, 2007), the method of assembling structures of multi-domain proteins consists of an initial low-resolution search, in which the conformational space of the domain linker is explored using the Rosetta *de novo* structure prediction method, followed by a high-resolution search, in which all atoms are treated explicitly, and the backbone and side chain degrees of freedom are simultaneously optimized. In AIDA (Xu *et al.*, 2015), domain assembly for assembling multi-domain protein structures is simulated by a fast-docking algorithm with an *ab initio* folding potential. The *de novo*- or *ab initio*-based methods may leave the domain structures largely randomly oriented in the final model (Zhou *et al.*, 2019). Template-based approaches often detect available templates to guide domain assembly. DEMO (Zhou *et al.*, 2019) was designed for constructing multi-domain protein structures by docking-based domain assembly simulations, in which inter-domain orientations are determined by the distance profiles from analogous templates as detected through domain-level structure alignments. However, the template-based approaches also have limitations because of the limited number of multi-domain proteins in PDB, and the difficulty of capturing the orientation between domains from the template may increase as the number of domains increases. Using deep learning technology may be helpful to capture the orientation information between domains that cannot be obtained from templates.

In this work, we proposed a new domain assembly method, SADA, in which the domain assembly simulation was performed to generate the full-chain model through a two-stage differential evolution algorithm guided by the energy function with an inter-residue distance potential predicted by deep learning. SADA was tested on a benchmark set of 356 proteins, for which the performance of SADA significantly outperformed most state-of-the-art domain assembly methods. Especially, on 40 benchmark proteins with the number of domains ≥4, the average TM-score of SADA is 0.60, which is 13.2% higher than that of the second-best method. Further, SADA was used to assemble 293 human multi-domain proteins from AlphaFold Protein Structure Database. On the 96 human proteins with the TM-score of AlphaFold2 model <0.9, the average TM-score of models assembled by SADA is 0.70, which is 6.1% higher than that of AlphaFold2 (0.66). Especially for these models of AlphaFold2 with average pLDDT ≤75, ~89% protein models were improved after SADA assembly. In addition, we also provided two additional function modules in the webserver, including a culling module to filter the whole multi-domain protein structure database (MPDB) according to input criteria and a detection module to identify the structural analogues of full-chain according to input domain models.

2 Materials and methods

SADA is a structural analogue-based protein structure domain assembly method assisted by deep learning, which involves five steps as follows: (i) detects structural analogues of the full-chain from the constructed MPDB according to the input protein domain models; (ii) constructs an initial model based on the detected first-ranked analogue; (iii) utilizes a deep learning network to predict the inter-residue distance distribution; (iv) builds a multi-domain protein specificity energy function for guiding domain assembly based on the predicted residue distance distribution and the property of multi-domain protein; and (v) assembles the domain models to generate final full-chain model by the proposed two-stage differential evolution algorithm from the initial model. The pipeline is displayed in Figure 1a.

2.1 MPDB construction

The flowchart of MPDB construction is shown in Figure 1b. The collection and processing steps of multi-domain protein data in MPDB are as follows: (i) CD-HIT (Fu *et al.*, 2012) was used to remove the redundancy of protein structures with a sequence identity cut-off of 100% in PDB, and then protein structures with sequence identity of <100% were obtained from PDB; (ii) DomainParser (Xu *et al.*, 2000) was next used to determine whether these proteins are multi-domain proteins or not; and (iii) the single-domain proteins determined by DomainParser were further confirmed by the definition of CATH (Lam *et al.*, 2016; Orengo *et al.*, 1997) and SCOPe (Chandonia *et al.*, 2017) on whether they were multi-domain proteins. All the multi-domain proteins selected in the three steps above were finally collected to construct the MPDB. The parameters of CD-HIT are listed in Supplementary Test S1.

As of September 2021, MPDB contains 48 225 multi-domain proteins, in which 37 495 proteins have 2 domains, 7539 proteins have 3 domains, 2182 proteins have 4 domains and 1009 proteins have more than 4 domains. The statistics of the number of domains in the MPDB are shown in Supplementary Figure S1.

For many purposes, it is useful to obtain a subset from MPDB. It is often the case that additional criteria are desirable, such as resolution, length, sequence identity or the number of domains. For example, in protein structure prediction methods based on machine learning, the proteins that satisfy specific criteria are used to train or test. Inspired by PISCES (Wang and Dunbrack, 2003), we developed a module for retrieving multi-domain proteins in MPDB, which filters the whole MPDB according to user's input criteria including protein chain length, resolution, number of domains, R-factor and sequence identity of multi-domain protein, and then provides the protein structures and related information that satisfy the criteria to users. The module is described in Supplementary Text S2.

2.2 Structural analogue detection

Homologues inherit similarities from their common ancestor, while analogues converge to similar structures due to a limited number of energetically favourable ways to pack secondary structural elements. In some cases, inferring structural analogues of the full-chain based on the domain structures may be more critical for research and modelling of multi-domain proteins.

Based on MPDB, we designed an algorithm to detect the multi-domain protein structural analogues. The structural analogue detection is illustrated in Figure 1c. To prevent structurally similar domains from being matched to the same part of the protein, the input individual domains were aligned on each protein in MPDB by using TM-align (Zhang and Skolnick, 2005, 2004), with no overlap allowed in the alignments of different domains. According to the structural similarity of individual domain models, a local similarity score LS_{score} was designed to evaluate the quality of structural analogues. LS_{score} is defined as follows:

$$LS_{\text{score}} = N \Big/ \sum_{n=1}^{N} \frac{1}{\text{TM} - \text{score}(n)} \tag{1}$$

where TM-score(n) is the TM-score between the n-th domain of the target and the template protein in MPDB after aligning the domain on the template protein by TM-align. N is the number of query individual domains. The LS_{score} has a value range of (0, 1], where 1

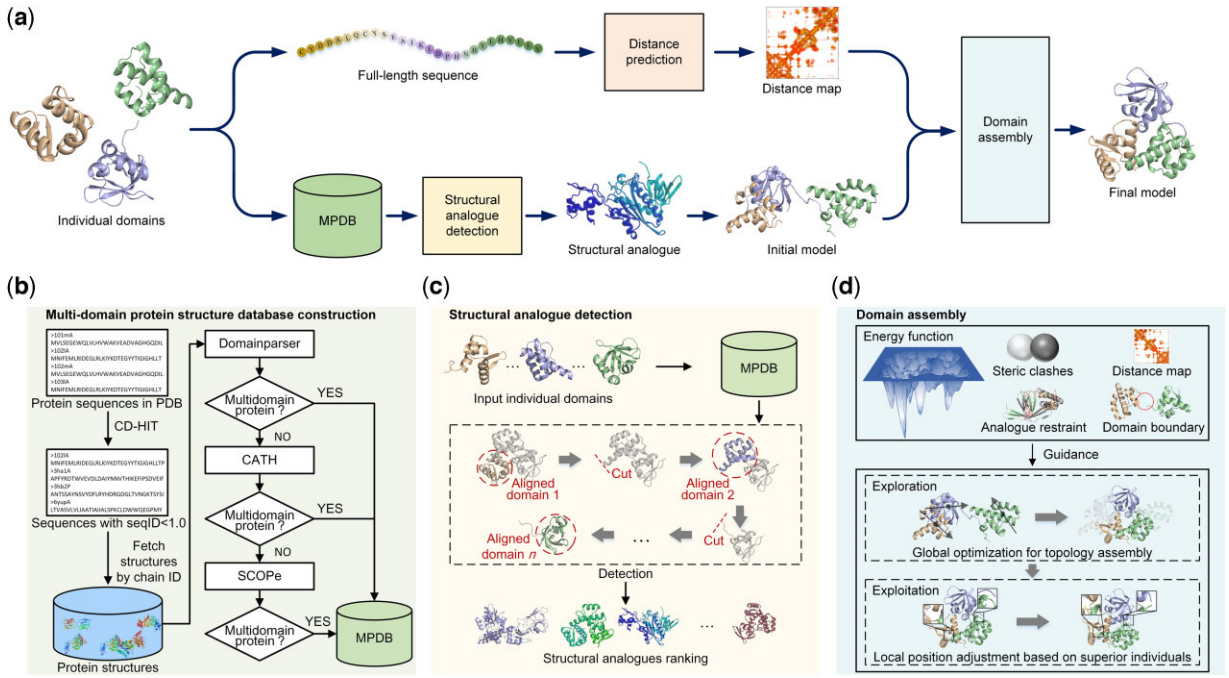

Fig. 1. (**a**) Pipeline of SADA for domain assembly. (**b**) Flowchart of multi-domain protein structure database construction. (**c**) Illustration of structural analogue detection. (**d**) Illustration of two-stage differential evolution algorithm for domain assembly

indicates a perfect match between domain models and the structural analogues of the full-chain.

In addition, although structural analogues serve as important parts of protein structural modelling and research, limited studies have attempted to recognize structural analogues (Javier *et al.*, 2020). Thus, we provide the function module to detect structural analogues from MPDB, and the module is described in Supplementary Text S3.

2.3 Energy function for domain assembly

The energy function was designed based on Anfinsen's hypothesis, in which native-like protein conformations represent unique, low-energy, thermodynamically stable conformations. Domain assembly was guided by the energy function, which was computed as a linear combination of four energy terms below,

$$E_{\text{total}} = w_1 E_{\text{clash}} + w_2 E_{\text{bd}} + w_3 E_{\text{ana}} + w_4 E_{\text{dist}} \tag{2}$$

in which each energy function term is explained.

E_{clash} represents the C_α clashes between domains, which is used to prevent the domain model from getting too close during assembly to satisfy the physical concept. This term is computed using Equation (3):

$$E_{\text{clash}} = \sum_{n=1}^{N-1}\sum_{m=n+1}^{N}\sum_{i=1}^{Rdom_n}\sum_{j=1}^{Rdom_m} E_{\text{clash}}^{i,j}$$
$$E_{\text{clash}}^{i,j} = \begin{cases} \dfrac{1}{d_{C_\alpha}^{i,j}}, & d_{C_\alpha}^{i,j} < d_{C_\alpha}^{\text{cut}} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $d_{C_\alpha}^{\text{cut}} = 3.75$ Å, N is the number of domain models, $Rdom_n$ and $Rdom_m$ represent the number of residues in the n-th and m-th domain, respectively and $d_{C_\alpha}^{i,j}$ is the distance between C_α of the residue i and j in the evaluated decoy.

E_{bd} is the boundary distance energy, which is defined as follows:

$$E_{\text{bd}} = \sum_{n=1}^{N-1}(bd_{C_\alpha}^{n,n+1} - bd_{C_\alpha}^{\text{cut}})^2 \tag{4}$$

where the $bd_{C_\alpha}^{\text{cut}} = 3.8$ Å, and $bd_{C_\alpha}^{n,n+1}$ is the C_α atom distance between the C-terminal residue of n-th domain and N-terminal residue of the $(n+1)$-th domain in the evaluated decoy. In the discontinuous domain proteins, the discontinuous domain is split into multiple segments because of the insertion of the continuous domains. Therefore, for discontinuous domain proteins, these discontinuous segments were treated as domains to calculate the energy.

E_{ana} is the structural analogue restraint energy. This term aims to prevent the assembly from deviating too much from the orientation obtained from the initial model. E_{ana} is calculated as follows:

$$E_{\text{ana}} = \frac{1}{L}\sum_{i=1}^{L} f(decoy_i, initModel_i) \tag{5}$$

where $f(decoy_i, initModel_i)$ represents the distance between the i-th C_α atom of the evaluated decoy and the i-th C_α atom of the initial model, and L is the length of the full-chain model.

E_{dist} is the inter-domain distance potential. The predicted inter-residue distance provides abundant spatial constraint information for domain assembly, which can compensate for the effects of low-quality structural analogues on domain assembly.

In this study, we used GeomNet, a geometric constraints prediction network in our recently developed structure prediction server (Liu *et al.*, 2022), to predict the inter-residue distance of full-chain. For the query sequence, MSA was generated by iterative search against UniRef30 (Mirdita *et al.*, 2017) and BFD (Steinegger *et al.*, 2019) databases by using HHblits (Steinegger *et al.*, 2019) with gradually relaxed e-values of $1e^{-30}$, $1e^{-10}$, $1e^{-6}$ and $1e^{-3}$. The input features of GeomNet were extracted from MSA, including the inverse of the covariance matrix, the position specific scoring matrix, the residue position entropy, and the one-hot encoding of the query sequence. GeomNet contains 66 residual blocks, each of which consists of two 2D convolutional layers, two instance normalization layers, two ELU activation layers and a dropout layer.

The potential is defined as follows:

$$E_{\text{dist}} = \sum_{n=1}^{N-1}\sum_{m=n+1}^{N}\sum_{i=1}^{Rdom_n}\sum_{j=1}^{Rdom_m} \frac{\log((d_{C_\beta}^{i,j} - u_{i,j})^2 + 1)}{\sigma_{i,j}} \tag{6}$$

where $d_{C_\beta}^{i,j}$ is the real distance between C_β atom (C_α for glycine) of the residue pair (i, j) in the evaluated decoy, and $u_{i,j}$ and $\sigma_{i,j}$ are the

mean and standard deviation obtained by Gaussian fitting of the residue pair (i, j) distance distribution, respectively.

The weighting parameters in Equation (2) were set as $w_1 = 0.52, w_2 = 0.20, w_3 = 0.50$ and $w_4 = (1 - LS_{\text{score}})$. If the local similarity score LS_{score} is greater than or equal to 0.82, w_3 is set to 6.0. A high LS_{score} indicates that the interaction direction inferred from structural analogue is reliable. The weighting parameters are determined by maximizing the TM-score between the SADA models and the native structures, which are optimized through an improved differential evolution algorithm (Zhou *et al.*, 2020). Details on the determination of the weighting parameters can be found in Supplementary Test S4.

2.4 Two-stage differential evolution algorithm for domain assembly

The assembly engine for domain assembly was carried out through simultaneous rotation and translation of each domain. For each domain, its movement can be represented by a translation vector and three rotation angles. Therefore, for a multi-domain protein with N domains, the solution of domain assembly can be represented as a $(6 \times N)$-dimensional target vector, and the solution S can be represented as follows:

$$S = (x_1, y_1, z_1, \phi_1, \psi_1, \omega_1, \ldots, x_N, y_N, z_N, \phi_N, \psi_N, \omega_N) \quad (7)$$

where x_N, y_N and z_N represent the translation vector of the N-domain, and ϕ_N, ψ_N and ω_N represent the three rotation angles of the N-domain.

Under the guidance of the energy function, a two-stage differential evolution algorithm was proposed to determine the optimal solution. The exploration stage aims to prevent the algorithm from getting stuck in local optima and generate multiple superior topology structures for the exploitation stage. Thus, the mutation strategy of slow convergence speed and strong exploration capability was used in this stage.

In the exploitation stage, the explored superior solutions rapidly converged to the minimum. Therefore, on the basis of the superior solutions generated in the previous stage, we further adjusted the local position of these solutions to generate the optimal solution, and the mutation strategy with fast convergence speed was used at this stage. The algorithm description and parameters setting of the two-stage differential evolution algorithm is described in Supplementary Text S5.

3 Results and discussion

3.1 Dataset

To fairly compare the performance of SADA with other methods, we used all the 356 proteins from the DEMO benchmark dataset as the test targets (Zhou *et al.*, 2019). The 356 proteins were generated by separately clustering the proteins with different domain types and structures from the template library of DEMO with a 30% sequence identity cut-off. This benchmark included 166 2-domain (2dom), 69 3-domain (3dom), 40 ≥4-domain (m4dom) and 81 discontinuous-domain (2dis) proteins. The maximum number of domains in m4dom is 7. Supplementary Table S1 summarizes the details of the 356 test proteins, including PDB ID, protein length and the number of domains.

3.2 Results of benchmark set

In this section, we reassembled the individual domain structures excised from the experimental structure. The initial domain structure was randomly rotated and translated before assembly, and structural analogues with a sequence identity >30% to the query were excluded. SADA was compared with two well-known domain assembly methods, namely, DEMO (Zhou *et al.*, 2019) and AIDA (Xu *et al.*, 2015). The average results of the final models of SADA, DEMO and AIDA on the benchmark set are shown in Table 1. The detailed results of each protein are shown in Supplementary Tables S2 and S3.

Table 1. Summary of domain structure assembly by using experimentally solved domains on 356 test proteins

Method	TM-score	#TM-score ≥0.5	*P*-value
SADA	0.80	328	–
DEMO	0.74	318	1.28E−21
AIDA	0.63	288	8.91E−51

Note: TM-score represents the average TM-score of final full-chain models. #TM-score ≥0.5 represents the number of models with TM-score ≥0.5. The values in the last column are the results of the Wilcoxon signed-rank test based on the comparison with the TM-score of SADA.

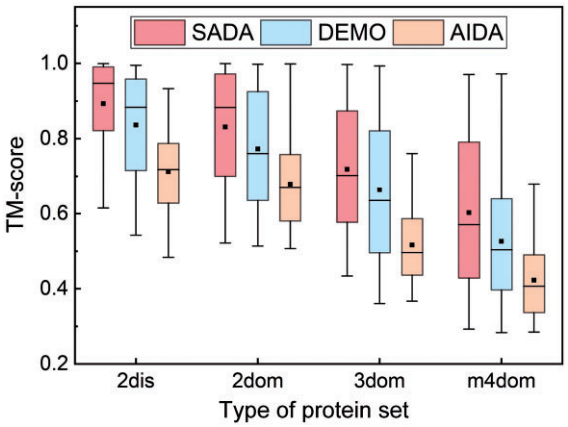

Fig. 2. Boxplot for the TM-score of the assembly models by SADA, DEMO and AIDA. The square and horizontal lines in the box represent the mean and median TM-scores, and the horizontal lines on the top and bottom are the maximum and minimum TM-scores, respectively

Overall, the average TM-score of SADA models is 0.80, which is 8.1% higher than that (0.74) of DEMO (with *P*-value = 1.28E−21) and 27.0% higher than that (0.63) of AIDA (with *P*-value = 8.91E−51), respectively. SADA correctly assembles (i.e. TM-score ≥ 0.5) 328 out of 356 targets, accounting for 92.1% of the total, which is 3.1% and 13.9% more than that of DEMO and AIDA, respectively. The respective *P*-values of DEMO and AIDA indicate statistically significant differences between the methods. An intuitive comparison of the TM-score between different methods is shown in Figure 2.

Figure 2 shows the TM-score histograms in separate categories, indicating that SADA assembles more accurate full-chain models for the proteins of different types of domains. As the number of domains increased, the performance of these methods decreased, possibly because the search space of domain assembly simulations increased as the degrees of freedom increased for proteins of more domains. In addition, for SADA, 2dom and 3dom proteins account for the majority in MPDB, and with the increase of the number of domains, high-quality full-chain structural analogues were hard to identify, thus affecting the performance. However, on the 40 m4dom proteins, the average TM-score is 0.60, which is 13.2% higher than that of the second-best method (DEMO). It is because we used GeomNet to capture the distance information between domains, thus compensating for the quality decline of structural analogues.

Generally, models of SADA have a higher TM-score than that of other methods, and the overall quality of the multi-domain models is acceptable, with an average TM-score of 0.68 for proteins with 3 or more domains, 74.3% of which had a TM-score ≥0.5.

To more rigorously compare the performance of SADA with the other two methods, we examined the overlap proteins between

DEMO benchmark set and GeomNet training set with CD-HIT. After checking by CD-HIT, we found that 58 proteins of the DEMO benchmark set have the sequence identity >40% to proteins in the GeomNet training set. Therefore, we removed these 58 proteins from the DEMO test set, and re-analysed the performance of SADA on the remaining 298 test proteins. On the remaining 298 test proteins, the average results of the final models of SADA, DEMO and AIDA were summarized in Supplementary Table S4. It was found that the conclusion for the performance of SADA and its advantage over the control methods were not changed.

In addition, to test the robustness of SADA, we run SADA 10 times on the benchmark set. The average results and the *P*-value between results of each run are shown in Supplementary Tables S5 and S6, respectively. The results indicate that there was no significant difference between the results of 10 independent runs of SADA.

We also recorded the algorithm's running time for each test protein and summarized in Supplementary Figure S2 and Table S7. The runtime of each protein is shown in Supplementary Table S8. Here, the final model of SADA is generated by running the whole pipeline of SADA once, and the structural analogous templates with a sequence identity >30% to the target were excluded. The hardware used is a computer cluster with CentOS 7.9 system, 10 × 80-core DELL PowerEdge R940XA 4U nodes running at 2.5 GHz.

3.3 Coverage of multi-domain proteins in the MPDB

On the benchmark set, we compared the coverage of MPDB and that of template library used in DEMO (Zhou *et al.*, 2019). Here, the template library used in DEMO is denoted as DEMO-lib. In the MPDB and DEMO-lib, TM-align (Zhang and Skolnick, 2005) was used to calculate the TM-score (Zhang and Skolnick, 2004) between each test protein and the proteins in the two databases, respectively. For each test protein, the average TM-score of the top 10 templates with the highest TM-score was considered as the coverage score of the test protein in the corresponding database.

At a sequence identity cut-off of 30%, the average coverage score of MPDB is 0.57, which is 7.5% higher than that of DEMO-lib (0.53) on the 166 2dom proteins. The average coverage score of MPDB is 0.56, which is 12.0% higher than that of DEMO-lib (0.50) on the 69 3dom proteins. On the m4dom proteins, the average coverage score of MPDB is 0.48, which is 6.7% higher than that of DEMO-lib (0.45). The average coverage score of MPDB is 0.56 on the 81 2dis proteins, which is 5.7% higher than that of DEMO-lib (0.53). The *P*-value between the coverage scores of MPDB and that of DEMO-lib is 5.54E−53, indicating that there is a significant difference between the coverage score of MPDB and DEMO-lib. The head-to-head comparisons between the coverage score of each test protein in DEMO-lib and MPDB are shown in Figure 3a. At a sequence identity cut-off of 50%, the average coverage score of MPDB is 0.60, which is 9.1% higher than that of DEMO-lib (0.55) for the 356 test proteins. At a sequence identity cut-off of 70%, the average coverage score of MPDB is 0.61, which is 10.9% higher than that of DEMO-lib (0.55). The detailed coverage score for each test protein and the summaries of different databases are listed in Supplementary Tables S9–S14. At sequence identity cut-offs of 50% and 70%, the head-to-head comparisons between the coverage score of each test protein in DEMO-lib and MPDB are shown in Supplementary Figure S3.

These results show that the constructed MPDB covers remarkably more multi-domain protein structures than DEMO-lib, and there is the significant difference between the coverage score of MPDB and that of DEMO-lib.

3.4 Performance of the structural analogue detection

There are some proteins that the sequence identity between them is low, but they have similar topologies. These proteins may be important for structural modelling of multi-domain proteins.

Structural analogue serves as the basis for SADA, and in this study, we tested the performance of the structural analogue detection algorithm. On the 356 test proteins, the proposed structural analogue detection algorithm was used to detect structural analogues of the full-chain according to the individual domain models, and the structural analogue with the highest LS_{score} was used to calculate TM-score between the analogue and native structure of target protein. When the sequence identity was <30%, the average TM-score of the 356 structural analogues with the highest LS_{score} was 0.56, where 192 cases have the highly similar topologies to the native structures with TM-score ≥ 0.5. When the sequence identities are <50% and 70%, the average TM-score of the structural analogues increased to 0.64 and 0.65, respectively. The number of structural analogues with TM-score ≥0.5 to the full-chain native structures were 239 and 247, respectively. The detailed TM-score of structural analogues for each test protein at different sequence identity cut-offs are shown in Supplementary Tables S15 and S16. At different sequence identity cut-offs, the average TM-scores of structural analogues are always more than 0.5, indicating that the detected structural analogues can describe the global topological structure of the full-chain for the majority of multi-domain proteins in the test set. For the 356 test proteins, the LS_{score} of the top 1 structural analogue is shown in Supplementary Table S17. Figure 3b shows the distributions between LS_{score} and TM-score at a sequence identity cut-off of 30%. At sequence identity cut-offs of 50% and

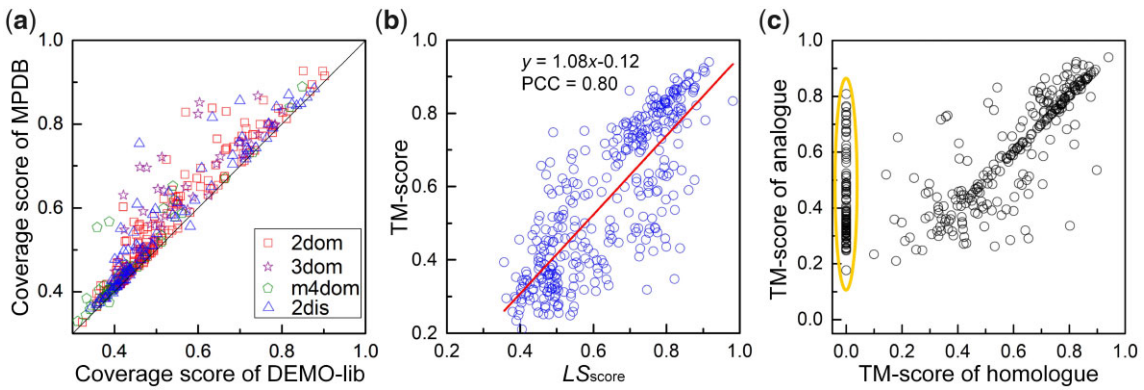

Fig. 3. (**a**) Head-to-head comparison between the coverage score of the test proteins in DEMO-lib and MPDB when the sequence identity between the test protein and the proteins in MPDB and DEMO-lib is <30%. The *x*-axis represents the coverage score of the test proteins in the DEMO-lib. The *y*-axis represents the coverage score of the test proteins in the MPDB. (**b**) Distribution of LS_{score} and TM-score under sequence identity of <30%. For each blue circle, the *x*-axis represents the LS_{score} of the structural analogue with the highest LS_{score} in the corresponding test protein, and the *y*-axis represents the TM-score between the native structure of the test protein and the detected structural analogue. The least-squares linear fit and Pearson correlation coefficient (PCC) are listed in (b). (**c**) Head-to-head analysis between the TM-score of analogue and that of homologue under the sequence identity cut-off of 30%. For each black circle, the *x*-axis represents the TM-score between the native structure of the test protein and the top 1 homologous template, and the *y*-axis represents the TM-score between the native structure of the test protein and the structural analogue with the highest LS_{score}. The black circles in the yellow area represent that the homologous templates of these test proteins cannot be detected by JackHMMER

70%, the distributions between LS_{score} and TM-score are shown in Supplementary Figure S4.

To further prove the effectiveness of the proposed structural analogue detection algorithm, the Pearson correlation coefficient between LS_{score} of the top 1 structural analogue and TM-score of the structural analogue to the native structure was calculated. The Pearson correlation coefficients between the LS_{score} and TM-score are 0.80, 0.80 and 0.81 at sequence identity cut-off values of 30%, 50% and 70%, respectively.

The results show that the proposed structural analogue detection algorithm can effectively detect the structural analogues of full-chain and verify that domains often interact in the similar way in the quaternary orientations if the domains have similar tertiary structures.

3.5 Impact of distance and structural analogue on SADA

In SADA, an energy function with the inter-residue distance potential predicted by GeomNet was used to guide domain assembly. The inter-residue distance information was not used in SADA to discuss the effect of the predicted distance information, and experiments were conducted on 356 test proteins. The SADA without using predicted distance information, namely, SADA-w/o-D, was used for comparative experiments to explore the contribution of the distance information predicted by GeomNet. In order to study the contribution of templates to the performance of SADA, we assembled the 356 DEMO test proteins without using any structural analogous templates (named SADA-w/o-T). The domain assembly results of SADA, SADA-w/o-D and SADA-w/o-T on the benchmark set are shown in Supplementary Table S18 and summarized in Supplementary Table S19. For an intuitive comparison of the effect of the structural analogous templates and predicted distance, the average TM-score histogram in separate categories is depicted in Figure 4.

Based on the results, SADA achieves a better TM-score than SADA-w/o-D in 275 out of 356 proteins. The average TM-score of SADA is 0.80 on 356 test proteins, which is 8.1% higher than that of SADA-w/o-D (0.74). On 166 2dom proteins and 81 2dis proteins, the average TM-scores of SADA (0.83 for 2dom and 0.89 for 2dis) are 5.1% and 4.7% higher than that of SADA-w/o-D (0.79 for 2dom and 0.85 for 2dis), respectively. Especially, on 69 3dom proteins and 40 m4dom proteins, the average TM-scores of SADA (0.72 for 3dom and 0.60 for m4dom) are 10.8% and 25.0% higher than that of SADA-w/o-D (0.65 for 3dom and 0.48 for m4dom), respectively. The performance of SADA degrades when the template information is not used. On the 356 test proteins, the average TM-score of SADA-w/o-T is 0.75, which is 6.3% lower than that of SADA.

The performance of SADA is significantly improved when the inter-residue distance potential and structural analogous template information are combined simultaneously to guide the domain assembly. These results indicate that the predicted distance and analogous templates are complementary to the improvement in SADA performance.

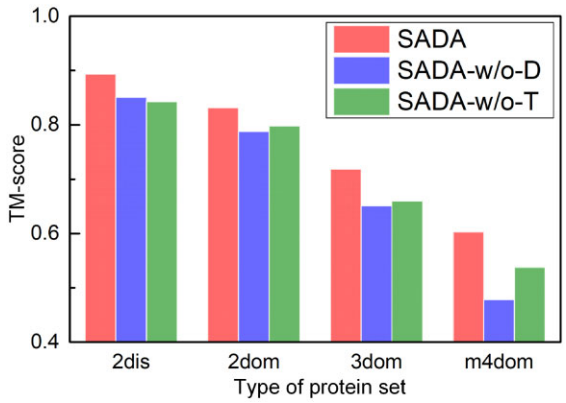

Fig. 4. Average TM-score of the final model assembled by SADA, SADA-w/o-D and SADA-w/o-T on 356 test proteins

3.6 Complementarity analysis between structural analogue and homologue

Template, as an important part of protein structure modelling, is mainly derived from homologous proteins. However, homologous templates for individual domains can often be detected, but templates that can be used to model the entire query protein are often unavailable. Therefore, structural analogues are suitable for multi-domain protein modelling in some cases, because some proteins are formed by duplication, divergence and recombination of domains.

Among the 356 test proteins, we analysed the quality of the detected structural analogues and the homologues searched by JackHMMER (Eddy, 1998). Here, the parameters of JackHMMER were set as default, and each domain model of test proteins was a native model. When the sequence identity cut-offs were 30%, 50% and 70%, the structural analogue with the highest LS_{score} and the top 1 homologue were used to calculate the TM-score between the analogue (homologue) and native structure of target protein, respectively. The detailed results of each protein are shown in Supplementary Tables S15 and S20, respectively. At a sequence identity cut-off of 30%, the head-to-head analysis between the TM-score of the analogue with the highest LS_{score} and that of the top 1 homologue are shown in Figure 3c. At a sequence identity cut-off of 30%, homologous templates of 100 test proteins cannot be searched in MPDB, and the TM-score of the structural analogues is greater than or equal to that of homologues in 169 out of the 256 test proteins. For the 100 proteins without homologous templates, there are 24 structural analogues with TM-score $\geq$0.5. The results are summarized in Supplementary Table S21.

At a sequence identity cut-off of 50%, homologous templates of 71 test proteins cannot be searched. Among the 285 test proteins, the average TM-score of the top 1 homologue is 0.66 and the average TM-score of the analogues with the highest LS_{score} is 0.68. The TM-score of the analogues is greater than or equal to that of homologues in 213 out of the 285 test proteins. Among the 71 test proteins, 22 structural analogues have TM-score $\geq$0.5. The results are summarized in Supplementary Table S22. At a sequence identity cut-off of 70%, homologues of 65 test proteins cannot be searched. Among the 65 test proteins, 20 structural analogues have TM-score $\geq$0.5. The TM-score of the analogues is greater than or equal to that of homologues in 218 out of the 291 test proteins. The results are summarized in Supplementary Table S23. The head-to-head analysis at identity cut-offs of 50% and 70% are shown in Supplementary Figure S5.

These results show that the structural analogues and homologues are complementary. For the test proteins in which the JackHMMER cannot search for homologues, the proposed structural analogues detection algorithm can detect some structural analogues with a TM-score $\geq$0.5.

3.7 Assembly multi-domain proteins using analogue and homologue

To further study the difference of analogues and homologues on the accuracy of domain assembly, we used SADA-w/o-D to assemble the full-chain structure of multi-domain proteins according to the homologous templates and analogous templates, respectively. In this study, SADA-w/o-D was used to assemble domains, because the influence of distance information predicted by deep learning needs to be excluded.

At a sequence identity cut-off of 30%, the initial models were generated based on the analogue with the highest LS_{score} and top 1 homologue, respectively. The final results are shown in Supplementary Table S24. Among the 256 test proteins with homologues detected by JackHMMER, the average TM-scores were 0.75 and 0.76 for the full-chain models generated by SADA-w/o-D using homologues and analogues, respectively. SADA-w/o-D using analogues obtained full-chain models with TM-score $\geq$0.5 in 218 out of 256. SADA-w/o-D using homologues obtained models with

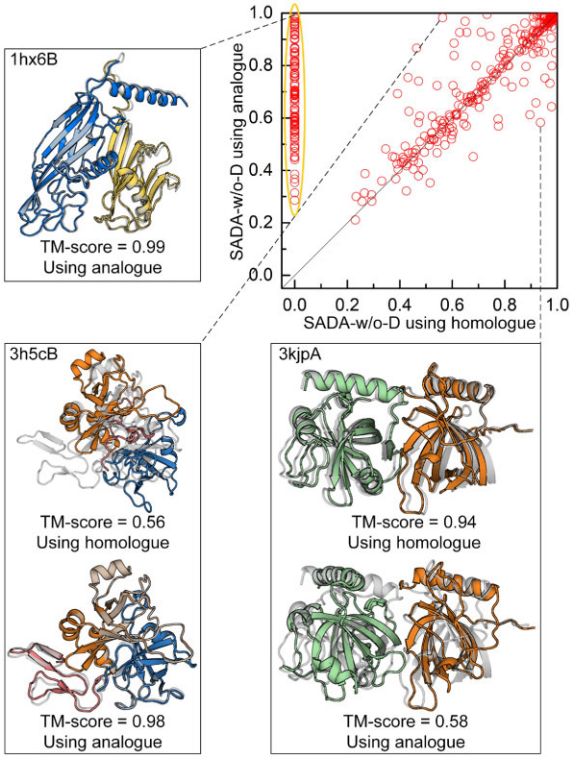

Fig. 5. Head-to-head comparison between TM-scores of the models assembled by SADA-w/o-D using homologue and analogue. The circles in the left marked region represent that the homologous templates of these test proteins cannot be detected by JackHMMER. The different colours in protein structures represent different domains, and transparent grey is native structure

TM-score ≥0.5 in 214 out of 256. The full-chain models generated by SADA-w/o-D using analogue are better than those using homologue in 132 out of the 256 test proteins. Especially, on the 100 test proteins without homologous templates, 88 full-chain models have a TM-score ≥0.5. For an intuitive analysis of the quality of full-chain models assembled by using different templates, the head-to-head comparison is shown in Figure 5.

Three illustrative examples are shown from Figure 5. For the 2dom protein 1hx6B, JackHMMER cannot detect homologous template at a sequence identity cut-off of 30%, while SADA-w/o-D assembled the full-chain model with TM-score of 0.99 by using analogous template. For the 3dom protein 3h5cB, the full-chain model with TM-score of 0.56 was generated using homologous template, while SADA-w/o-D using analogous template assembled the full-chain model with TM-score of 0.98. For the 2dom protein 3kjpA, SADA-w/o-D using homologous and analogous template assembled the full-chain models with TM-scores of 0.94 and 0.58, respectively.

These results further demonstrate that analogues and homologues are complementary, and the combination of analogues and homologues may improve the modelling accuracy of multi-domain proteins.

3.8 Assembly human multi-domain proteins

Although AlphaFold2 predicted the structures of many protein targets at or near experimental resolution, for some multi-domain proteins, the quality of its predicted structures may be further improved by SADA. In this section, we reassembled 293 human multi-domain proteins randomly selected from AlphaFold Protein Structure Database (AFDB) according to three criteria: ≥90% residues were solved in the native structure, the sequence identity to the training set of GeomNet is <40%, and the <30% sequence identity to each other (Tunyasuvunakool *et al.*, 2021). Here, the individual domain models for SADA assembly were decomposed from the full-chain structures of AlphaFold2. Under 30%, 70%, and 100% sequence

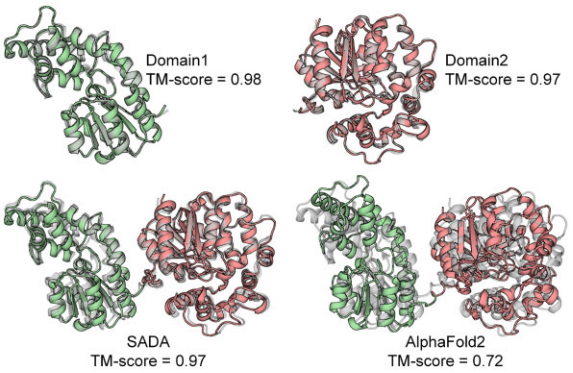

Fig. 6. The representative example (1s8oA) is showing SADA generates more accurate full-chain model than AlphaFold2. Grey is native structure, and different domain models are represented by different colours

identity cut-offs used to exclude structural analogous templates for SADA assembly, the results are summarized in Supplementary Tables S25–S27, and the head-to-head comparison between the TM-score of full-chain models generated by SADA and AlphaFold2 are shown in Supplementary Figure S6.

On the 293 human multi-domain proteins, the average TM-score of the models deposited in AFDB is 0.87, where the AlphaFold2 did not remove any homologous proteins, including the native for modelling, and it is trained on a copy of the PDB (Jumper *et al.*, 2021). Therefore, AlphaFold2 can achieve high-accuracy structure prediction on most proteins. When structural analogous templates with a sequence identity >30% to the target were excluded for SADA assembly, the average TM-score of the models assembled by SADA is 0.80. For the 96 proteins with the TM-score of AlphaFold2 model <0.90, the average TM-score of models of AlphaFold2 is 0.66, and the average TM-score of models assembled by SADA is 0.59, where the models assembled by SADA achieves better TM-score in 36 out of 96 proteins. When structural analogous templates with a sequence identity >70% to the target were excluded for SADA assembly, the average TM-score of the models assembled by SADA is 0.81. For the 96 proteins with the TM-score of AlphaFold2 model <0.90, the average TM-score of models assembled by SADA is 0.61, and some multi-domain protein models of AlphaFold2 are significantly improved after SADA assembly. As the protein 1s8oA is shown in Figure 6, it was composed of two domains, where the TM-scores for the models decomposed from the full-chain structures of AlphaFold2 for domains 1 and 2 were 0.98 and 0.97, respectively, but the full-length structure achieved a TM score of 0.72. However, the TM-score of the full-chain model generated by SADA achieved 0.97. When the sequence identity cut-off for SADA assembly is also set to 100%, which is same with that used by AlphaFold2, the average TM-score of SADA is 0.88, which is higher that of AlphaFold2 (0.87). On the 22 out of 293 proteins, the improvement of TM-score after the reassembly is ≥0.10. Especially for the 96 proteins of with the TM-score of AlphaFold2 model <0.9, the average TM-score of models assembled by SADA is 0.70, which is 6.1% higher than that of AlphaFold2 (0.66), and SADA achieves better TM-score than AlphaFold2 in 57 out of 96 proteins.

These results also show that SADA probably can improve the quality of AlphaFold2 models to a certain extent. To help determine in which scenarios SADA may be used to be complementary to the AlphaFold2, the relationship between the average pLDDT cut-off of the AlphaFold2 full-chain model and the proportion of the cases improved after SADA reassembly is shown in Supplementary Figure S7. On the 293 human multi-domain proteins, these models of AlphaFold2 with average pLDDT ≤75, ∼89% models were improved by SADA.

3.9 Results of CASP14 multi-domain targets

SADA is also compared with the latest version of AlphaFold2 and RoseTTAFold on all CASP14 multi-domain targets. It should be

noted that the AlphaFold2 models were regenerated using its latest standalone package rather than from the results reported in CASP14. This is because the models of AlphaFold2 belong to the type of human group in CASP14. RoseTTAFold models were generated by the RoseTTAFold standalone package. The head-to-head comparison between the TM-score of full-chain models assembled by SADA and that of AlphaFold2 and RoseTTAFold is shown in Supplementary Figure S8, and the results are summarized in Supplementary Table S28. Here, the individual domain models for SADA assembly were independently predicted by AlphaFold2.

On average, the models assembled by SADA achieve an average TM-score of 0.82, which is slightly lower than that of AlphaFold2 (0.84), but obviously higher than that of RoseTTAFold (0.52). It is probably mainly due to some incorrect domain models predicted by AlphaFold2, which affects the performance of SADA. When removing the targets with TM-score of domains ≤ 0.60 (T1058, T1061 and T1070), however, SADA obtains a comparable (or slightly higher) TM-score (0.871) to AlphaFold2 (0.867). These results show that the quality of full-chain structures assembled by SADA using the independently predicted domains is comparable to that of AlphaFold2, but higher than that of RoseTTAFold. If we further input the experimental structure of individual domains, SADA achieves an even higher TM-score (0.873) on all CASP14 multi-domain targets, where the native structures of targets are excluded.

These results demonstrated that although the overall model quality of SADA relies on the quality of individual domains, SADA has the ability to assemble the domain orientation better than the state-of-the-art neural-network models when starting from correct domain models. In addition, SADA allows the algorithm to split proteins into domains for independent modelling followed by domain assembly and is helpful for saving the computational resources for modelling large-size proteins. This is important for the large-size proteins that usually cannot be handled by the end-to-end method. For AlphaFold2, predicting large proteins can easily exceed the memory of a single GPU. For example, for the target T1050 from CASP14 (779 residues), AlphaFold2 exceeds the memory of the GPU (NVIDIA TITAN RTX, 24GB), but the full-chain model can be generated by SADA based on the domain structures independently modelled by AlphaFold2. The accuracy of the assembled full-chain model (TM-score = 0.97) is consistent with the full-chain model directly predicted by AlphaFold2 (TM-score = 0.97) using a more advanced GPU (TESLA V100, 32GB).

3.10 The potential of SADA in predicting protein complexes

We randomly selected eight proteins from the (Vreven *et al.*, 2015) benchmark set to test the potential of SADA for assembling the domain-domain interactions as being in different chains, and compared it with AlphaFold-Multimer (Evans *et al.*, 2021) and RoseTTAFold (Baek *et al.*, 2021). Here, the structures of the eight proteins were constructed using a simply extended version of SADA in which each chain was treated as a virtual 'domain' but the connectivity requirement between the virtual 'domains' was ignored, and the structures of single chain were predicted by the AlphaFold2's latest standalone package.

Compared with the models predicted by RoseTTAFold, the models assembled by SADA achieve an average TM-score of 0.74, which is higher than that of RoseTTAFold (0.71). The comparison results of these 8 proteins are shown in Supplementary Figure S9, which indicates that SADA can generate comparable or better domain-domain interactions as being in different chains than RoseTTAFold in seven out of eight cases. For 1grn, the structures of the single chain were correctly modelled with TM-scores >0.90. However, the domain orientations were not correctly predicted by RoseTTAFold, which resulted in the protein model with a relative low TM-score = 0.78 for RoseTTAFold. SADA correctly modelled the domain orientations as being in different chains and obtained a model with TM-score = 0.97. For 1rke, the domain-domain interactions as being in different chains were predicted by SADA (0.92) and RoseTTAFold (0.89), achieving a comparable TM-score.

Further, we compared SADA and AlphaFold-Multimer on the eight proteins. Overall, the average TM-score of models generated by AlphaFold-Multimer is 0.87, which is higher than that of SADA. This may be because AlphaFold-Multimer is specifically optimized for complex modelling and trained on a dataset that includes multimers. However, all templates in MPDB are single chain proteins, and the GeomNet is also trained on single chain proteins. Therefore, compared with AlphaFold-Multimer, the interaction between different chains may not be easily captured by SADA. It is worth mentioning that SADA can produce comparable results than AlphaFold-Multimer in proteins 1ab9, 1grn, 1lya and 1rke, as shown in Supplementary Figure S10.

These results show that although SADA is proposed for assembling domain–domain interactions as being in one chain, it also has the potential to be extended to assemble the domain–domain interactions as being in different chains.

4 Conclusion

We developed a structural analogue-based protein structure domain assembly method assisted by deep learning, named SADA. In SADA, an MPDB was constructed using DomainParser and the domain knowledge defined in CATH and SCOPe databases, named MPDB. Based on MPDB, a structural analogue detection algorithm was proposed, to identify the structural analogues of full-chain from MPDB through structural alignment of individual domain models. Based on the detected analogue, the initial full-chain model was generated. Under the guidance of the energy function with an inter-residue distance potential predicted by GeomNet, domain assembly was simulated by a two-stage differential evolution algorithm to generate the final full-chain model. In the exploration stage, some superior topology structures were generated. In the exploitation stage, the local position was optimized based on the superior topology structures. Results on 356 tested proteins show that the proposed SADA significantly outperforms most state-of-the-art domain assembly methods. In addition, SADA was also applied to assemble 293 human multi-domain proteins from AlphaFold Protein Structure Database, and some cases achieved better models. Especially for the 96 proteins with the TM-score of AlphaFold2 model <0.9, the average TM-score of models assembled by SADA is 0.70, which is 6.1% higher than that of AlphaFold2 (0.66). The results show that SADA can be complementary to the end-to-end protein structure prediction methods (e.g. AlphaFold2) to generate alternative or better models.

In this study, the Pearson correlation coefficient between the local similarity score LS_{score} and TM-score is 0.80 at sequence identity cut-off of 30%. The results indicate that domains often interact in the similar way in the quaternary orientations if they have similar tertiary structures. Moreover, the analysis of homologues searched by JackHMMER and analogues detected by the proposed algorithm shows that they are complementary. If more suitable templates can be selected from homologues and analogues, the accuracy of multi-domain protein structure modelling can be further improved.

Furthermore, we also assembled the domain-domain interactions as being in different chains using a simply extended version of SADA. These results indicate that SADA has the potential to be extended to assemble the domain–domain interactions as being in different chains. In addition, compared with the AlphaFold-Multimer directly generating final models, SADA reports structural analogues and structural alignments, which helps to further explain the domain–domain interactions and provides functional insights for further studies on the protein. A study of more advanced SADA for the domain–domain interactions as being in different chain assembly strategies will be our next research direction.

Funding

This work was supported by the 'New Generation Artificial Intelligence' major project of Science and Technology Innovation 2030 of the Ministry of Science and Technology of the People's Republic of China [2021ZD0150100], the National Nature Science Foundation of China

[62173304 and 61773346] and the Key Project of Zhejiang Provincial Natural Science Foundation of China [LZ20F030002].

Conflict of Interest: none declared.

References

Baek,M. *et al.* (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, **373**, 871–876.

Chandonia,J.M. *et al.* (2017) SCOPe: manual curation and artifact removal in the structural classification of proteins-extended database. *J. Mol. Biol.*, **429**, 348–355.

Du,Z.Y. *et al.* (2021) The trRosetta server for fast and accurate protein structure prediction. *Nat. Protoc.*, **16**, 5634–5651.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Evans,R. *et al.* (2021) Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, doi:10.1101/2021.10.04.463034.

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

Javier,C.D. *et al.* (2020) Deep learning enables the design of functional de novo antimicrobial proteins. *bioRxiv*, doi:10.1101/2020.08.26.266940.

Jones,D.T. and Thornton,J.M. (2022) The impact of AlphaFold2 one year on. *Nat. Methods*, **19**, 15–20.

Jumper,J. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.

Lam,S.D. *et al.* (2016) Gene3D: expanding the utility of domain assignments. *Nucleic Acids Res.*, **44**, D404–D409.

Liu,J. *et al.* (2022) *De novo* protein structure prediction by incremental inter-residue geometries prediction and model quality assessment using deep learning. *bioRxiv*, doi:10.1101/2022.01.11.475831.

Mirdita,M. *et al.* (2017) Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.*, **45**, D170–D176.

Orengo,C.A. *et al.* (1997) CATH–a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1109.

Pearce,R. and Zhang,Y. (2021) Toward the solution of the protein structure prediction problem. *J. Biol. Chem.*, **297**, 100870.

Steinegger,M. *et al.* (2019) Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat. Methods*, **16**, 603–606.

Steinegger,M. *et al.* (2019) HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*, **20**, 15.

Su,H. *et al.* (2021) Improved protein structure prediction using a new multi-scale network and homologous templates. *Adv. Sci.*, **8**, 2102592–2102602.

Tunyasuvunakool,K. *et al.* (2021) Highly accurate protein structure prediction for the human proteome. *Nature*, **596**, 590–596.

Vreven,T. *et al.* (2015) Updates to the integrated protein-protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. *J. Mol. Biol.*, **427**, 3031–3041.

Wang,G.L. and Dunbrack,R.L. (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.

Wollacott,A.M. *et al.* (2007) Prediction of structures of multidomain proteins from structures of the individual domains. *Protein Sci.*, **16**, 165–175.

Xu,D. *et al.* (2015) AIDA: ab initio domain assembly for automated multi-domain protein structure prediction and domain-domain interaction prediction. *Bioinformatics*, **31**, 2098–2105.

Xu,J.B. (2019) Distance-based protein folding powered by deep learning. *Proc. Natl. Acad. Sci. USA*, **116**, 16856–16865.

Xu,J.B. and Wang,S. (2019) Analysis of distance-based protein structure prediction by deep learning in CASP13. *Proteins*, **87**, 1069–1081.

Xu,Y. *et al.* (2000) Protein domain decomposition using a graph-theoretic approach. *Bioinformatics*, **16**, 1091–1104.

Yang,J.Y. *et al.* (2020) Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. USA*, **117**, 1496–1503.

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

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

Zheng,W. *et al.* (2021) Protein structure prediction using deep learning distance and hydrogen-bonding restraints in CASP14. *Proteins*, **89**, 1734–1751.

Zhou,X.G. *et al.* (2019) Assembling multidomain protein structures through analogous global structural alignments. *Proc. Natl. Acad. Sci. USA*, **116**, 15930–15938.

Zhou,X.G. *et al.* (2020) Underestimation-assisted global-local cooperative differential evolution and the application to protein structure prediction. *IEEE Trans. Evol. Comput.*, **24**, 536–550.

Zhou,X.G. *et al.* (2022) Progressive and accurate assembly of multi-domain protein structures from cryo-EM density maps. *Nat. Comput. Sci.*, **2**, 265–275.