Bioinformatics, 36(8), 2020, 2443–2450
doi: 10.1093/bioinformatics/btz943
Advance Access Publication Date: 20 December 2019
Original Paper

Structural bioinformatics

CGLFold: a contact-assisted *de novo* protein structure prediction using global exploration and loop perturbation sampling algorithm

Jun Liu[1], Xiao-Gen Zhou[2], Yang Zhang[2,*] and Gui-Jun Zhang [1,*]

[1]College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China and [2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109-2218, USA

*To whom correspondence should be addressed.

Associate Editor: Yann Ponty

Received on July 3, 2019; revised on December 10, 2019; editorial decision on December 12, 2019; accepted on December 18, 2019

Abstract

Motivation: Regions that connect secondary structure elements in a protein are known as loops, whose slight change will produce dramatic effect on the entire topology. This study investigates whether the accuracy of protein structure prediction can be improved using a loop-specific sampling strategy.

Results: A novel *de novo* protein structure prediction method that combines global exploration and loop perturbation is proposed in this study. In the global exploration phase, the fragment recombination and assembly are used to explore the massive conformational space and generate native-like topology. In the loop perturbation phase, a loop-specific local perturbation model is designed to improve the accuracy of the conformation and is solved by differential evolution algorithm. These two phases enable a cooperation between global exploration and local exploitation. The filtered contact information is used to construct the conformation selection model for guiding the sampling. The proposed CGLFold is tested on 145 benchmark proteins, 14 free modeling (FM) targets of CASP13 and 29 FM targets of CASP12. The experimental results show that the loop-specific local perturbation can increase the structure diversity and success rate of conformational update and gradually improve conformation accuracy. CGLFold obtains template modeling score $\geq$ 0.5 models on 95 standard test proteins, 7 FM targets of CASP13 and 9 FM targets of CASP12.

Availability and implementation: The source code and executable versions are freely available at https://github.com/iobio-zjut/CGLFold.

Contact: zgj@zjut.edu.cn or zhng@umich.edu

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

1 Introduction

De novo protein structure prediction, directly predicting the tertiary structure from its amino acid sequence without relying on a template, remain an unsolved problem, one of the main reasons is that the conformational space to be searched is massive (Bradley, 2005). Various methods have been proposed to explore the conformational space (Dukka, 2017; Moult *et al.*, 2018), including Metropolis Monte Carlo (Rohl *et al.*, 2004), Replica Exchange Monte Carlo (Xu and Zhang, 2012), Evolutionary algorithm (Custodio *et al.*, 2014; Zhang *et al.*, 2017; Zhou *et al.*, 2019a,b), resampling (Shrestha and Zhang, 2014), multi-objective optimization (Olson and Shehu, 2014), sequential search (De Oliveira *et al.*, 2018) and so on. These methods have been proved effective and efficient for the conformational space sampling. However, the size of the conformational space increases with the target size (Kandathil *et al.*, 2018). The fragment assembly technique greatly reduces the conformational space that need to be searched during the process by using the fragments of known protein structures (Handl *et al.*, 2012), and is widely used in *de novo* protein structure prediction (Rohl *et al.*, 2004; Xu and Zhang, 2013). However, some potential conformation spaces cannot be sampled due to the limitations of fragment library. In particular, flexible loop regions often cannot be adequately sampled. In fully internal coordinate representations, small perturbations can result in large atomic movements due to lever arm effects (Ovchinnikov *et al.*, 2018). This effect is noticeable in flexible loop areas.

Many methods that focus on loop sampling, known as loop modeling, are dedicated to improving the prediction accuracy of the loop structure (Arnautova *et al.*, 2011; Liang *et al.*, 2014; Spassov *et al.*, 2008). In LoopBuilder (Soto *et al.*, 2010), a fast and accurate protocol for the prediction of loop conformations in proteins has

been described, including sampling of backbone conformations, side-chain addition, selection of conformation subset and all-atom energy minimization. (Heo *et al.*, 2017) proposes a protein loop structure prediction method that combines a new energy function designed for accurate protein loop structure determination with the conformational space annealing global optimization algorithm. In Sphinx (Marks *et al.*, 2017), a loop modeling approach has been proposed that can use the additional information present in different length loops by integrating aspects of both knowledge-based and *ab initio* methodologies in a novel way. Two strategies have been proposed in Marks *et al.* (2018) to improve the accuracy of protein loop structure prediction by using contact information. Some methods also sample the loop region by special operators to improve the prediction accuracy of the final model. Rosetta-based memetic algorithm (RMA) (Garza-Fabre *et al.*, 2016) proposed a loop-based recombination and mutation operators to increase the exploration of possible loop regions, and Rosetta is used as a local search routine. A local perturbation method for protein chain torsion angle space has been proposed (Favrin *et al.*, 2001). This method introduces deviation probability and Gaussian distribution to obtain a set of small torsion angle disturbances. In Rohl *et al.* (2004), two similarity estimation methods are used to select an insertion fragment and achieve a local move, avoiding large global perturbation.

In the recent years, residue–residue contact prediction has achieved remarkable results in recent CASP sessions (Abriata *et al.*, 2018; Schaarschmidt *et al.*, 2018), and increasingly used to *de novo* protein structure prediction. The predicted contact information is usually integrated into the energy function (Bhattacharya and Cheng, 2015; Kosciolek and Jones, 2014) or separately constructed as an evaluation model to guide the conformational update along with energy function (Evans, 2018; Zhang *et al.*, 2018). EVfold (Marks *et al.*, 2011) uses multiple sequence alignment and maximum entropy model to infer distance constraints from evolutionary sequence variations, and then utilizes the distance constraints to predict protein structure. CONFOLD (Adhikari *et al.*, 2015) predicts protein structure by converting contacts and secondary structures into distance, dihedral angle and hydrogen bond constraints. In PconsFold (Michel *et al.*, 2014), the contacts predicted from PconsC (Skwark *et al.*, 2013) are used within Rosetta to fold a given protein sequence from scratch. FRAGFOLD (Kosciolek and Jones, 2014) combines fragment assembly with statistical potentials and predicted contacts to construct the tertiary structure. RaptorX-Contact (Ma *et al.*, 2015; Wang *et al.*, 2016, 2017a,b, 2018; Xu, 2019) predicts the tertiary structure of the input sequence by feeding predicted distance and torsion angles into CNS as restraints.

In this study, we focus on whether the accuracy of protein structure can be improved using a loop-specific perturbation, and propose a contact-assisted global exploration and loop perturbation cooperative *de novo* protein structure prediction method (CGLFold). In the global exploration phase, conformational space is extensively searched using fragment recombination and assembly. In the loop perturbation phase, the loop-specific perturbation model is firstly constructed, then the differential evolution algorithm (DE) is used to solve the disturbance angles, and the conformational update is realized through the selection strategy finally. Loop perturbations can mitigate the constraints of the fragment library and explore more possible structures. DE is one of the powerful approaches for the global optimization problems (Storn and Price, 1997; Zhou *et al.*, 2016a,b,c; Zhou and Zhang, 2017). It can efficiently obtain multiple feasible solutions and improve the diversity of the loop structure. The contact-based selection model is designed to guide conformational search. The experimental results show that loop-specific local perturbation can significantly improve the accuracy of the prediction model.

2 Materials and methods

The pipeline of CGLFold is given in Figure 1. For the query sequence, the fragment library was built by the Robetta server (http://robetta.bakerlab.org/) with the Exclude Homologs option selected. The inter-residue contact map is predicted by ResTriplet (Li *et al.*, 2018) developed by Zhang group, and the contact map is filtered first before constructing the contact-assisted conformation selection model. The initial population is generated by the random fragment assembly. The initialization process will be terminated when all torsion angles in each conformation have been replaced or no movements are accepted in 4000 attempted insertions. For each individual in the population, the global exploration and loop perturbation are conducted. In the global exploration phase, the massive conformational space is explored using fragment recombination and assembly. A loop-specific perturbation model is constructed in the loop perturbation phase and solved using DE to increase the structural diversity of loop region and further improve the conformation obtained in the global phase. The two phases enable the cooperation in the global exploration and local exploitation. The iterative procedure of the global exploration and loop perturbation is conducted until the termination criterion is satisfied. The sampled conformations are clustered using SPICKER (Zhang and Skolnick, 2004a,b), and the center model of the first cluster is selected for the side-chain assembling by PULCHRA (Rotkiewicz and Skolnick, 2008). Detailed flowchart of CGLFold can be found in Supplementary Figure S1.

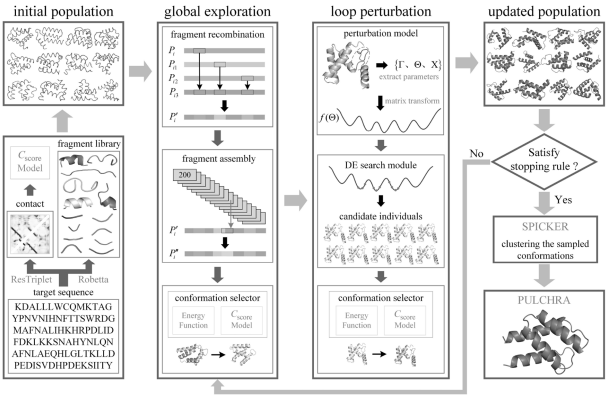

Fig. 1. The pipeline of CGLFold

2.1 Contact-based conformation selection model

The massive conformational space and inaccurate energy function usually result in low search efficiency and prediction accuracy. The residue–residue contact information, which combined with energy function, can be used for guiding the conformational space sampling effectively. Therefore, a residue–residue contact selection model, named C_{score}, is designed to guide conformational search. The Rosetta score3 model (Ovchinnikov *et al.*, 2018) is selected as the energy function in this study.

2.1.1 Contact map pre-processing

In order to include as much contact information as possible and avoid the prediction noise, the predicted contact map is filtered first before constructing the selection model. If the Euclidean distance of the residue index between any two residue pairs in the contact map is ≤ 2, only the contact with a high confidence is retained. Then, the top $L/2$ contacts with higher confidence are selected from the filtered contact map to construct the selection model, where L is the length of the query sequence. An example of this process is shown in Supplementary Figure S2. The experimental results in Section 3.5 show that the contact-based selection model constructed using the filtered contact can guide the conformational space search to obtain accurate prediction models.

2.1.2 Contact-based selection model

The contact-based conformation selection model is defined as:

$$C_{\text{score}} = \sum_{r=1}^{L/2} c_r \tag{1}$$

$$c_r = \begin{cases} 8^{p^r}(d_{\text{clash}} - d_{i,j}), & d_{i,j} \le d_{\text{clash}} \\ -8^{p^r}, & d_{\text{clash}} < d_{i,j} \le d_{\text{con}} \\ 8^{p^r}\ln(d_{i,j} - d_{\text{con}} + 1), & \text{otherwise} \end{cases} \tag{2}$$

where r is the index of contact in the selected contacts; i and j are the residue indexes corresponding to the rth contact; p^r is the confidence score that residue i and j are in contact; $d_{i,j}$ is the real distance between C_β atoms (C_α for glycine) of residue i and j; $d_{\text{con}} = 8$Å is the maximum distance that two residues are in contact (Di Lena *et al.*, 2012); and $d_{\text{clash}} = 3.8$Å is the minimum distance that two residues have not spatial clash. For each conformation, its C_{score} can be calculated using Equations (1) and (2). In this model, the smaller the C_{score} is, the more satisfying the contact constraint will be, and the conformation is also likely closer to the native structure. C_{score} reaches the desired minimum value C^*_{score} when conformation satisfies all contact constraints.

$$C^*_{\text{score}} = \sum_{r=1}^{L/2} -8^{p^r} \tag{3}$$

2.2 Global exploration

In the global exploration phase, the native-like topology is generated by using fragment recombination and assembly. Fragment recombination (Krasnogor *et al.*, 1999; Zhou *et al.*, 2016a,b,c) is designed on the basis of evolutionary algorithms to swap the fragment between different individuals. Fragment assembly is used to reduce the conformational search space by using the fragment information of known structures, which are widely used in the template-free protein structure prediction (Han and Baker, 1996; Simons *et al.*, 1997; Xu and Zhang, 2012).

For fragment recombination, three different individuals P_{i1}, P_{i2} and P_{i3}, which also differ from the target individual P_i, are randomly selected from the population. Then the new individual P'_i is generated by replacing the corresponding fragments in P_{i3} with three fragments from different positions in P_i, P_{i1} and P_{i2}. The schematic of fragment recombination is shown in Supplementary Figure S3. In fragment assembly, an insertion window is randomly selected from P'_i, and the fragment of the insertion window is replaced by a randomly selected fragment from the corresponding fragment library to generate a new individual P''_i. The energy function is used to evaluate P'_i and P''_i, and the Metropolis criterion (Metropolis *et al.*, 1953) is used to determine whether the fragment assembly is accepted in P'_i. If accepted, P''_i is considered as the trial individual P_i^{trial}; otherwise, the fragment assembly is performed again. If the 200 consecutive assemblies fail, P'_i is considered as the trial individual.

Having the trial individual, two conformational update strategies (i.e. G_E_C and G_C_E) are designed to determine whether the target should be replaced with the trial individual, as shown in Supplementary Figure S4. When the generation number g is less than or equal to quarter of the maximum generation G, G_E_C is used. Otherwise, G_C_E is used to accelerate C_{score} convergence.

2.3 Loop perturbation

A loop region connects α-helixes or β-sheets. It determines the positional relationship between them and influences overall topology. Therefore, in order to enhance the diversity of the loop region and search for possible structures without considerably changing the overall structure, a loop-based local perturbation strategy is designed in this study.

2.3.1 Loop perturbation model

To obtain the secondary structure of the target individual, Define Secondary Structure of Proteins (DSSP) (Kabsch and Sander, 1983) is employed in our method, and a loop region with both ends connected to α-helix region or β-sheet region and more than four residues is randomly selected (Favrin *et al.*, 2001). For the selected loop region, we can obtain the set of rotation axes Γ, the set of rotation angles Θ, and the set of rotation points X, which are defined as follows:

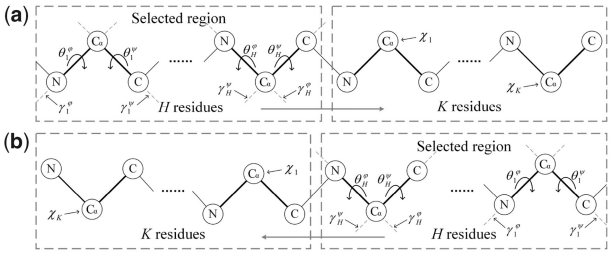

Fig. 2. The schematic of Equation (4). If the selected loop region is close to the C-terminal or in the middle of the protein, (**a**) is used to define Equation (4). Otherwise, (**b**) is used

$$\begin{cases} \Gamma = \{\gamma_1^\phi, \gamma_1^\psi, \ldots, \gamma_h^\phi, \gamma_h^\psi, \ldots, \gamma_H^\phi, \gamma_H^\psi\} \\ \Theta = \{\theta_1^\phi, \theta_1^\psi, \ldots, \theta_h^\phi, \theta_h^\psi, \ldots, \theta_H^\phi, \theta_H^\psi\}, \\ X = \{\chi_1, \ldots, \chi_k, \ldots, \chi_K\} \end{cases} \tag{4}$$

where γ_h^ϕ and θ_h^ϕ are the axis and disturbance angle on the atom bond of N-C_α for the hth residue in the selected loop region, respectively. γ_h^ψ and θ_h^ψ are the axis and disturbance angle on the atom bond of C_α-C for the hth residue in the selected loop region, respectively. H is the number of residues of the selected loop region, χ_k is the coordinate of backbone atomic C_α of the kth residue following the selected loop region, and K is the number of residues following the selected loop region. The schematic of Equation (4) is shown in Figure 2. If the selected loop region is close to the C-terminal or in the middle, Figure 2a is used to define Equation (4). Otherwise, Figure 2b is used.

In accordance with Equation (4), the rotation matrix of the coordinate transformation for the rotation points X can be calculated by

$$T(\Theta) = T_1(\Theta)\cdots T_l(\Theta)\cdots T_{2H}(\Theta) \tag{5}$$

$$T_l(\Theta) = \begin{bmatrix} C_l + (1-C_l)x_l^2 & (1-C_l)x_l y_l - z_l S_l & (1-C_l)x_l z_l + y_l S_l \\ (1-C_l)y_l x_l + z_l S_l & C_l + (1-C_l)y_l^2 & (1-C_l)y_l z_l - x_l S_l \\ (1-C_l)z_l x_l - y_l S_l & (1-C_l)z_l y_l + x_l S_l & C_l + (1-C_l)z_l^2 \end{bmatrix} \tag{6}$$

where $T_l(\Theta)$ is the rotation matrix corresponding to the lth rotation axis; C_l and S_l are $\cos\theta_l$ and $\sin\theta_l$, respectively; θ_l is the lth rotation angle; (x_l, y_l, z_l) is the unit vector of the lth rotation axis; and $l \in \{1, 2, \ldots, 2H\}$.

The rotated points $X' = \{\chi'_1, \ldots, \chi'_k, \ldots, \chi'_K\}$ are obtained by rotating points X around axes Γ with the angles of Θ, and each rotated point of χ'_k can be calculated as follows:

$$\chi'_k = T(\Theta)\cdot\chi_k. \tag{7}$$

The perturbation model is used to perturb the selected loop region without causing a large change in topology can be defined as:

$$\begin{aligned} \min \quad & f(\Theta) = \sum_{k=1}^{K} \sqrt{k}\cdot||\chi'_k - \chi_k||^2, \\ \text{s.t.} \quad & \Theta \neq 0 \end{aligned} \tag{8}$$

where $\sqrt{k}$ is designed to adjust the weight of the constraint based on the residue distance. The perturbation model will be solved in the following section.

2.3.2 Disturbance angles

In this section, the disturbance angles Θ is calculated by Equation (8) to achieve a satisfactory loop perturbation without causing a

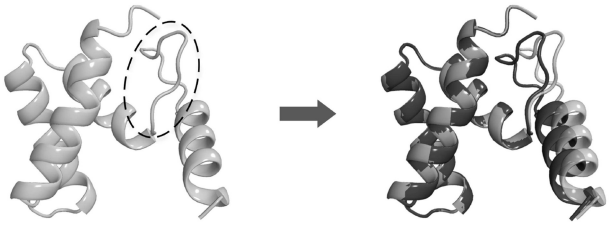

Fig. 3. An example of loop perturbation with 12 residues in the loop region selected for the perturbation model generation. The light gray structure is the target individual, and the dark gray structure is a candidate individual generated by loop perturbation. The RMSD between the two individuals is 1.09 Å. The average changes of the dihedral angles and C_α atomic coordinates of the perturbed loop region is 4.36 and 1.07 Å, respectively

large topological move. The analytic method is complicated and inefficient for this problem because of the degree of freedom of the perturbation model is dynamically changed, and the heuristic method is suitable for solving the model. Therefore, DE is used to solve the disturbance angles in this study, since it is one of the most effective approaches for global optimization problems (Hao, 2017; Zhou *et al.*, 2016a,b,c; Zhou and Zhang, 2019). The flowchart of the DE used to optimize disturbance angles Θ is shown in Supplementary Figure S5. Briefly, an initial disturbance angle population is randomly generated according to the input of the rotation axes Γ and rotation points X, and each component of the disturbance angle is randomly distributed between -5° and 5°. The mutation, crossover and selection operators are iterated to update the population. Last, 10 satisfactory disturbance angle solutions are selected.

2.3.3 Conformational update

According to the 10 selected disturbance angle sets, 10 candidate individuals are generated by disturbing the dihedral angles of the selected loop region. The C_{score} model and energy function are used to evaluate the quality of the candidate individuals and determine whether the target individual will be replaced by the candidate individual. Figure 3 shows an example to compare the structure of the generated candidate and target individual.

The energy of the candidate individuals is calculated and the candidate individuals are ranked according to the energy from low to high. The ordered individuals are used in turn as the trial individual to replace the target individual until the replacement succeeds or traverses all candidate individuals. Two conformation update strategies (i.e. L_E_C and L_C_E) are designed for different periods to select candidate individuals to replace the target individual, as shown in Supplementary Figure S6. When the generation number g is less than or equal to quarter of the maximum generation G, L_E_C is used for conformational update. Otherwise, L_C_E is used.

3 Result and discussion

In this study, the root mean square deviation (RMSD) and template modeling score (TM-score) (Xu and Zhang, 2010; Zhang and Skolnick, 2004a,b) are used to evaluate the quality of the predicted model. The parameters of CGLFold are described in Table 1. The proteins with sequence identify > 30% to the CGLFold test proteins are removed from the ResTriplet training set, and the ResTriplet is retrained to predict the contact map for CGLFold.

3.1 Dataset

In order to test the performance of CGLFold, the QUARK (Xu and Zhang, 2012) test set, 14 free modeling (FM) targets of CASP13 and 29 FM targets of CASP12 are employed in the following experiments. The QUARK test set contains 145 proteins with sequence lengths from 70 to 150, and the detailed information is shown in Supplementary Table S1. The 43 FM targets of CASP13 and CASP12 contain 9 proteins with sequence lengths from 155 to 375,

Table 1. The parameter descriptions in CGLFold

Parameters of conformational search		
Population size	NP	200
Maximum generation	G	800
Energy temperature scaling factor	KT_e	2
C_{score} temperature scaling factor	KT_c	4
Fragment length of fragment replace and assembly	f	3or9
Parameters of DE search		
Population size	NP'	50
Maximum generation	G'	50
Scaling factor	F	0.5
Crossover rate	CR	0.5
Temperature scaling factor	KT_l	1

Note: When the generation $g \leq 0.5*G$, nine-residue fragment ($f = 9$) is used for fragment recombination and fragment assembly. Otherwise, three-residue fragment ($f = 3$) is used.

and the detailed information are shown in Supplementary Tables S3 and S4. The sequence similarity of the test protein to the proteins corresponding to each fragment in the fragment library is calculated by using the NW-align (http://zhanglab.ccmb.med.umich.edu/NW-align). As shown in Supplementary Figure S7, the average sequence similarity of most test protein is <0.3, and the average value is 0.26.

3.2 Results of QUARK test set

CGLFold is compared with three state-of-the-art methods, i.e. Rosetta, QUARK and CONFOLD, on QUARK test set. Here, Rosetta and QUARK do not use contact information. CONFOLD uses predicted contacts and secondary structures to guide the prediction. For Rosetta, 1000 independent trajectories are run using the Rosetta's ClassicAbinitio protocol with the increase_cycles equals to 10. The center model of the first cluster determined by SPICKER (Zhang and Skolnick, 2004a,b) based on all decoys is considered as the final model. The result of QUARK and CONFOLD are predicted by their online server (https://zhanglab.ccmb.med.umich.edu/QUARK/) and (http://protein.rnet.missouri.edu/confold/), respectively. The secondary structure predicted by PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/) and the contact predicted by ResTriplet are used as inputs of CONFOLD.

The predicted results of CGLFold, Rosetta, QUARK and CONFOLD on the QUARK test set are summarized in Table 2, and the detailed results of each protein can be found in Supplementary Table S2. The average RMSD of CGLFold is 6.51 Å, which is 37.2, 37.3 and 15.9% lower than Rosetta, QUARK and CONFOLD, respectively. The average TM-score of CGLFold is 0.552, which is 52.5, 34.4 and 2.8% higher than Rosetta, QUARK and CONFOLD, respectively. CGLFold obtains TM-score $\geq$ 0.5 models in the 95 proteins, accounting for 65.5% of the total proteins, and higher than that of the comparison methods. The comparisons of CGLFold with these three methods on the QUARK test set are visually reflected in Figure 4. When compared with Rosetta, CGLFold achieves a lower RMSD on 124 proteins and a higher TM-score on 140 proteins. CGLFold achieves a lower RMSD on 125 proteins and a higher TM-score on 129 proteins than QUARK. When compared with CONFOLD, CGLFold achieves a lower RMSD on 98 proteins and a higher TM-score on 86 proteins. In order to analyze the performance of CGLFold more objectively, the Wilcoxon signed-rank test (Corder and Foreman, 2009) is used to analyze significant difference between CGLFold and the compared methods. The significance test results in columns five and six of Table 2 show that the performance of CGLFold is significantly better than those of Rosetta and QUARK. Although CGLFold is not significantly better than CONFOLD on TM-score, CGLFold obtains lower RMSD and higher TM-score models on more proteins. Figure 5 shows three illustrative examples of the comparison between the 3D structure

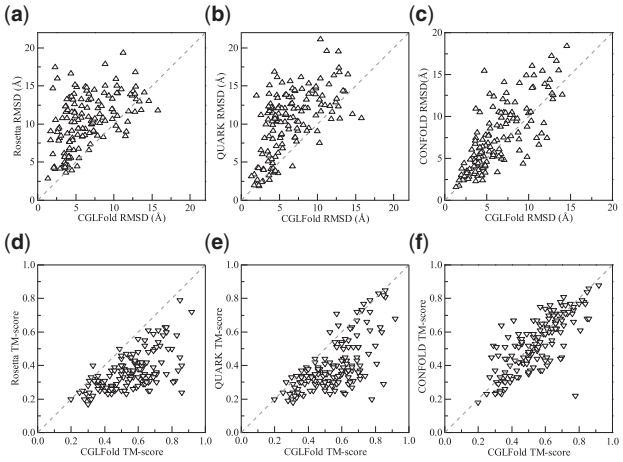

Fig. 4. The comparison of CGLFold with Rosetta, QUARK and CONFOLD. (**a–c**) are the RMSD of the predicted final models between CGLFold and Rosetta, QUARK and CONFOLD, respectively. (**d–f**) are the TM-score of the predicted final models of between CGLFold and Rosetta, QUARK and CONFOLD, respectively

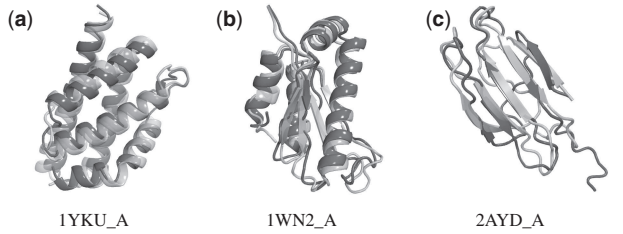

Fig. 5. The three illustrative examples of the comparison of the 3D structure predicted by CGLFold (light gray) with the corresponding native structure (dark gray)

predicted by CGLFold and the corresponding native structure. The folding types of these three proteins are α, α/β and β, respectively.

3.3 Result of CASP targets

The performance of CGLFold is also tested on 14 FM targets of CASP13 and 29 FM targets of CASP12, and is compared with C-QUARK (Mortuza *et al.*, 2018), BAKER-ROSETTASERVER (Anishchenko *et al.*, 2018), MULTICOM_cluster (Hou *et al.*, 2018) and RaptorX-Contact (Xu, 2018). The results of C-QUARK, BAKER-ROSETTASERVER, MULTICOM_cluster and RaptorX-Contact are from the CASP official website (http://predictioncenter.org) or updated results in the latest literature, and all of them use contact information. Figure 6 visually reflects the TM-score of each method, and the detailed results are listed in Supplementary Tables S3 and S4. CGLFold obtains the highest TM-score on 4 CASP13 targets and 13 CASP12 targets, and generates the model with TM-score $\geq$ 0.5 on 7 CASP13 targets and 9 CASP12 targets. CGLFold achieves an average TM-score of 0.49 and 0.40 on the CASP13 targets and CASP12 targets, respectively.

3.4 Effect of loop perturbation

In order to verify the effect of loop perturbation, CGLFold is compared with the method without loop perturbation (CGFold). The predicted results are summarized in Table 3, and the detailed results of each protein can be found in Supplementary Table S2. The comparison is visually reflected in Figure 7.

When compared with CGFold, CGLFold achieves a lower RMSD on 124 proteins and a higher TM-score on 130 proteins. The average RMSD and TM-score of CGFold are 8.44A and 0.464, respectively. In contrast, the average RMSD of CGLFold is decreased by 22.9%, and the average TM-score is increased by 19.0% when the loop perturbation is included. RMSD of 13 proteins in CGLFold

Table 2. The predicted results of CGLFold, Rosetta, QUARK and CONFOLD

Method	RMSD	TM-score	#TM $\geq$ 0.5	*P*-value	Significance
CGLFold	6.51	0.552	95	NA	NA
Rosetta	10.36	0.362	20	5.70E-25	+
QUARK	10.39	0.399	34	9.00E-23	+
CONFOLD	7.74	0.537	87	0.0842	≈

Note: #TM $\geq$ 0.5 is the number of models with TM-score $\geq$ 0.5. The last two columns are the results of Wilcoxon signed-rank test calculated in accordance with TM-score.

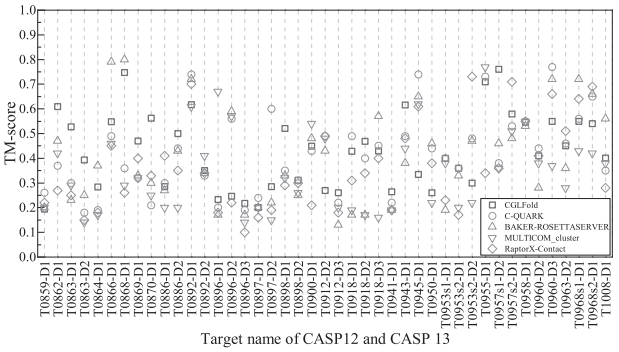

Fig. 6. The TM-score of the final models predicted by CGLFold, C-QUARK, BAKER-ROSETTASERVER, MULTICOM_cluster and RaptorX-Contact for the CASP12 and CASP13 targets

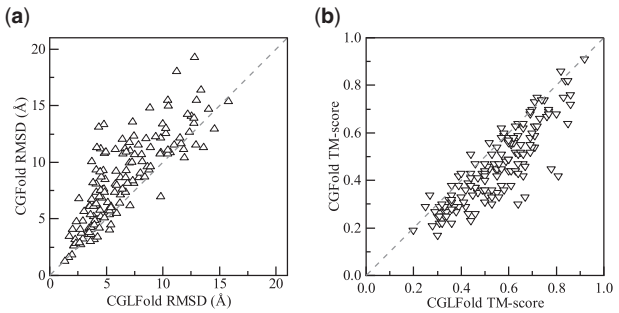

Fig. 7. The comparison of CGLFold and CGFold on QUARK test set. (**a**) and (**b**) are the comparison of RMSD and TM-score of the predicted final models of CGLFold and CGFold, respectively

is reduced by >50%, and the TM-score of 19 proteins is increased by >50%. CGLFold obtains more TM-score $\geq$ 0.5 models than CGFold. The Wilcoxon signed-rank test results in columns five and six of Table 3 show that the predicted results of CGLFold are significantly better than those of CGFold.

In order to analyze the effect of loop perturbation, we calculated the topological differences of the conformation before and after each perturbation and the changes of the dihedral angles (ϕ and ψ) and the C_α atomic coordinates in the loop region. Figure 8 shows the loop perturbation effect on 43 CASP targets. The statistical results of all the test proteins are listed in Supplementary Table S5. The torsion angles of the loop region are greatly changed, and the average variations of phi and psi torsion angles are 2.87° and 2.82°, respectively. The average change of C_α atomic coordinates in loop region is 0.45 Å, which indicates that the coordinates of the CA atoms are effectively limited to a certain range. The average RMSD between the conformations before and after loop perturbation is 0.27 Å, which indicates that each loop perturbation produces a small adjustment to the topology. In order to further analyze the cumulative effect of loop perturbation, we also calculated and

Table 3. The predicted results of CGLFold and CGFold

Method	RMSD	TM-score	#TM ≥ 0.5	*P*-value	Significance
CGLFold	6.51	0.552	95	NA	NA
CGFold	8.44	0.464	57	4.50E-21	+

Note: #TM ≥ 0.5 is the number of models with TM-score > 0.5. The last two columns are the results of Wilcoxon signed-rank test calculated in accordance with TM-score.

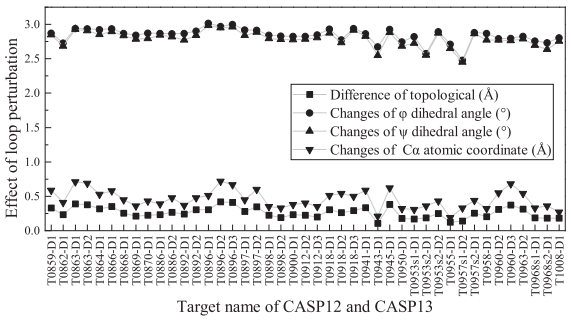

Fig. 8. The effect of loop perturbation on the targets of CASP12 and CASP13

accumulated the RMSD changes of the conformation before and after each perturbation. As shown in Supplementary Figure S8, the small adjustments of loop perturbation can be continuously accumulated and ultimately yield considerable benefits. The RMSD of 137 proteins are reduced. The RMSD of 1 and 7 proteins are unchanged and increased, respectively, because the accuracy of their residue–residue contacts is low.

Figure 9 shows the loop perturbation success rate of the QUARK test set proteins. A successful replacement of the target individual by the candidate individual is considered as a successful loop perturbation. It can be found that the loop perturbation success rate of all proteins is >60%, and the average success rate is 88.7%. The loop perturbation success rate is >80% in 88.5% test proteins.

Obviously, the loop perturbation enables small adjustments to the topology, and this small adjustment accumulates and ultimately yields considerable benefits. The mechanism of multi-perturbation solution can improve the diversity of the perturbation structure and improve the success rate of loop perturbation.

3.5 Effect of contact-based selection model

In CGLFold, the predicted contact map is filtered first before constructing the contact-based selection model. In order to verify the effectiveness of the filtering operation, the method involving the filtered contact (CGFold-filtered) is compared with the method using the unfiltered contact (CGFold-unfiltered). The prediction results are summarized in Table 4, and the detailed results can be found in Supplementary Table S6. The average RMSD of CGFold-filtered is decreased by 13.3% compared with that of CGFold-unfiltered, and the average TM-score increased by 6.9%. CGFold-filtered obtains more TM-score ≥ 0.5 models than CGFold-unfiltered. The Wilcoxon signed-rank test result shows that the predicted results of CGFold-filtered are significantly better than those of CGFold-unfiltered. The method using top $L/2$ filtered contacts (CGLFold-top_L/2) is compared with the method using top L filtered contacts (CGLFold-top_L). As shown in Supplementary Table S7, CGLFold-top_L/2 is superior to CGLFold-top_L in terms of the average RMSD, the average TM-score, and the number of models with TM-score ≥ 0.5. The Wilcoxon signed-rank test result shows that the predicted results of CGLFold-top_L/2 are significantly better than those of CGLFold-top_L.

The relationship between $\hat{C}_{\text{score}}$ and TM-score of the predicted models is analyzed to verify the rationality of the contact-based selection model, where $\hat{C}_{\text{score}}$ is the ratio between the average C_{score} of all conformations in the final population and C^*_{score}, indicating the

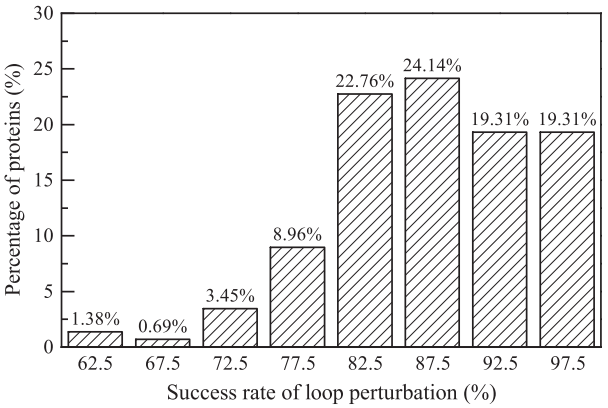

Fig. 9. The distribution of the success rate for all test proteins. The horizontal axis is the success rate of the loop perturbation, and the vertical axis is the percentage of proteins

Table 4. The predicted results of CGFold-filtered and CGFold-unfiltered

CGFold	RMSD	TM-score	#TM ≥ 0.5	*P*-value	Significance
Filtered	8.44	0.464	57	NA	NA
Unfiltered	9.73	0.434	41	2.20E-04	+

Note: #TM ≥ 0.5 is the number of models with TM-score ≥ 0.5. The last two columns are the results of Wilcoxon signed-rank test calculated in accordance with TM-score.

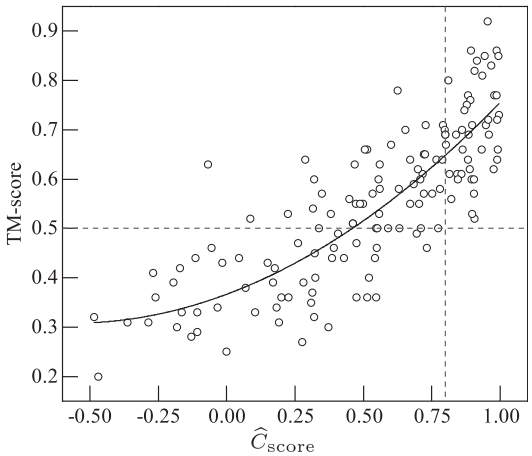

Fig. 10. The relationship between $\hat{C}_{\text{score}}$ and TM-score of the predicted models. The solid line is the result of a polynomial fitting

extent to which C_{score} is close to the desired value. In Figure 10, CGLFold obtains structures with a high TM-score for most of proteins when $\hat{C}_{\text{score}}$ is close to 1. The TM-score of all 43 proteins with $\hat{C}_{\text{score}} > 0.8$ is > 0.5. For these proteins, the average TM-score is 0.71, and the maximum TM-score is 0.92. The polynomial fitting curve indicates that the TM-score is positively correlated with $\hat{C}_{\text{score}}$, and the closer $\hat{C}_{\text{score}}$ is to 1, the stronger the correlation between them will be. Supplementary Figure S9 shows the relationship between the accuracy of the contacts used to construct C_{score} and the TM-score of the predicted models. The fitting results indicate that they are positively correlated, and most models with TM-score > 0.5 have a high contact accuracy. Supplementary Figure S10 reflects the effect of contact prediction accuracy and $\hat{C}_{\text{score}}$ on the accuracy of the prediction model. The fitted curve between them shows a strong correlation.

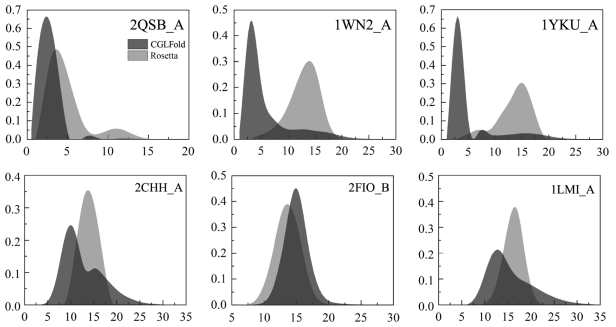

Fig. 11. RMSD distribution of the sampled conformations to native structures on six representative proteins. The horizontal axis is the RMSD(Å), and the vertical axis is the sampling probability. Dark gray and light gray represent CGLFold and Rosetta, respectively

To investigate the factors that led to a low $\hat{C}_{score}$, we analyzed the 18 proteins with $\hat{C}_{score}$ lower than 0. The statistical results are shown in Supplementary Table S8, $\hat{C}^{n}_{score}$ is the $\hat{C}_{score}$ of the native structure corresponding to the test protein. Figure S11 shows the relationship between the accuracy of contact used to construct C_{score} and $\hat{C}^{n}_{score}$. It can be found that their correlation is very strong, further verifying that the design of the C_{score} model is reasonable. For proteins with high $\hat{C}^{n}_{score}$ but low $\hat{C}_{score}$, indicate the sampling is not sufficient. C_{score} model can guide the sample to the near-native structural region, but the sampling algorithm cannot sample these potential structures. There are three major factors that lead to the insufficient sampling capability of the algorithm. (i) The poor fragment library leads to the lack of near-native structures in the conformational space that can be explored. (ii) The energy function is inaccurate. Since both the energy function and C_{score} model are used in CGLFold for the conformational sampling, the inaccurate energy function will limit the effect of C_{score}. (iii) The movement of the algorithm is not powerful enough to explore more potential structures. For proteins with low $\hat{C}^{n}_{score}$ and low $\hat{C}_{score}$, the accuracy of the contact used to construct the C_{score} model is low, which resulting in an inaccurate C_{score} model.

3.6 Near-native sampling ability

The RMSD distributions of the conformations sampled by CGLFold and Rosetta on six representative proteins are shown in Figure 11. For most proteins, the RMSD corresponding to the peak of the sampling probability of CGLFold is closer to 0 than Rosetta. In particular, the RMSD of the peak for 1WN2_A and 1YKU_A is 10 Å lower than that of Rosetta. For 2QSB_A, CGLFold and Rosetta have 97.9 and 80.1% sample distributions within 5 Å, and both of them have strong near-native sampling capabilities. CGLFold has 69.9 and 77.2% samples distributed within 5 Å for 1WN2_A and 1YKU_A, respectively, and sampling the conformations with RMSD below 5 Å is difficult for Rosetta. For 2CHH_A, although the near-native sampling ability of CGLFold and Rosetta are weak, the sampling distribution of CGLFold is closer to native states than that of Rosetta. For 2FIO_B, Rosetta can sample the conformations closer to the native structure than CGLFold. For 1LMI_A, Rosetta has a higher sampling probability peak than CGLFold, but its near-native sampling ability is inferior to CGLFold. The sampling distribution of all proteins can be found in Supplementary Figure S12.

4 Conclusion

A contact-assisted *de novo* protein structure prediction using global exploration and loop perturbation sampling algorithm, called CGLFold, is proposed in this article. The predicted contact map is filtered first before constructing the conformation selection model C_{score}. The C_{score} model and energy function are used to guide the conformational space sampling. In CGLFold, two phases, the global exploration phase and loop perturbation phase are performed for each generation. The global exploration phase is performed for each target individual of the current generation by using fragment recombination and assembly, while in the loop perturbation phase, the loop-specific perturbation model, which is solved by DE, is designed to refine all individuals obtained in the global exploration phase. The global exploration phase aims to explore the massive conformational space quickly and generate native-like topology, and the loop perturbation phase improves the accuracy of the topology by adjusting the dihedral angles of the selected loop region.

The performance of CGLFold is tested on the QUARK test proteins, 14 FM targets of CASP13, and 29 FM targets of CASP12, and is compared with state-of-the-art methods. The experimental results show that the loop perturbation achieves a small adjustment of the topology, and this small adjustment continues to accumulate and eventually yields considerable benefits. The conformation selection model based on the filtered residue–residue contact can assist the inaccurate energy function to guide conformational search. GLFold obtains TM-score ≥ 0.5 models on 95 of 145 QUARK test proteins, 7 FM targets of CASP13, and 9 FM targets of CASP12.

Currently, the CGLFold just can predict single-domain proteins since the energy and simulation method are optimized for the single-domain proteins. For the multi-domain proteins, we can use the method in Zhou *et al.* (2019a,b) to generate the model by assembling the CGLFold predicted individual domain structures together in the future. Moreover, the information entropy metric and abstract convex underestimation method (Hao, 2016; Zhou and Zhang, 2017) maybe help CGLFold to make a good balance between global exploration and local exploitation adaptively, This is our ongoing work for improving performance of CGLFold.

Funding

This work was supported by the National Nature Science Foundation of China [No. 61773346] and the Key Project of Zhejiang Provincial Natural Science Foundation of China [No. LZ20F030002].

Conflict of Interest: none declared.

References

Abriata,L.A. *et al.* (2018) Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. *Proteins*, **86**, 97–112.

Adhikari,B. *et al.* (2015) CONFOLD: residue-residue contact-guided ab initio protein folding. *Proteins*, **83**, 1436–1449.

Anishchenko,I. *et al.* (2018) Improving Robetta by a broad usage of sequence data and coevolutionary restraints. In: Thirteenth Meeting of Critical Assessment of Techniques for Protein Structure Prediction, pp. 22, Riviera Maya, Mexico.

Arnautova,Y.A. *et al.* (2011) Development of a new physics-based internal coordinate mechanics force field and its application to protein loop modeling. *Proteins*, **79**, 477–498.

Bhattacharya,D. and Cheng,J. (2015) De novo protein conformational sampling using a probabilistic graphical model. *Sci. Rep.*, **5**, 16332.

Bradley,P. (2005) Toward high-resolution de novo structure prediction for small proteins. *Science*, **309**, 1868–1871.

Corder,G.W. and Foreman,D.I. (2009) *Nonparametric Statistics for Non-Statisticians: A Step-By-Step Approach*. Wiley.

Custodio,F.L. *et al.* (2014) A multiple minima genetic algorithm for protein structure prediction. *Appl. Soft Comput.*, **15**, 88–99.

De Oliveira,S.H.P. *et al.* (2018) Sequential search leads to faster, more efficient fragment-based de novo protein structure prediction. probabilistic sampling. *Bioinformatics*, **34**, 1132–1140.

Di Lena,P. *et al.* (2012) Deep architectures for protein contact map prediction. *Bioinformatics*, **28**, 2449–2457.

Dukka,B.K. (2017) Recent advances in sequence-based protein structure prediction. *Brief. Bioinform.*, **18**, 1021–1032.

Evans,R. *et al.* (2018) De novo structure prediction with deep-learning based scoring. In: Thirteenth Meeting of Critical Assessment of Techniques for Protein Structure Prediction, pp. 11, Riviera Maya, Mexico.

Favrin,G. *et al.* (2001) Monte Carlo update for chain molecules: Biased Gaussian steps in torsional space. *J. Chem. Phys.*, **114**, 8154–8158.

Garza-Fabre,M. *et al*. (2016) Generating, maintaining, and exploiting diversity in a Memetic algorithm for protein structure prediction. *Evol. Comput*., **24**, 577–607.

Han,K.F. and Baker,D. (1996) Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proc. Natl. Acad. Sci. USA*, **93**, 5814–5818.

Handl,J. *et al*. (2012) The dual role of fragments in fragment-assembly methods for de novo protein structure prediction. *Proteins*, **80**, 490–504.

Hao,X.H. *et al*. (2016) A novel method using abstract convex underestimation in ab-initio protein structure prediction for guiding search in conformational feature space. *IEEE/ACM Trans. Comput. Biol. Bioinform*., **13**, 887–900.

Hao,X.H. *et al*. (2017) Conformational space sampling method using multi-subpopulation differential evolution for de novo protein structure prediction. *IEEE Trans. NanoBiosci*., **16**, 618–633.

Heo,S. *et al*. (2017) Protein loop structure prediction using conformational space annealing. *J. Chem. Inf. Model*., **57**, 1068–1078.

Hou,J. *et al*. (2018) Improving protein tertiary structure prediction by deep learning, contact prediction and domain recognition. *Thirteenth Meeting of Critical Assessment of Techniques for Protein Structure Prediction, pp. 128, Riviera Maya, Mexico*.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kandathil,S.M. *et al*. (2018) Improved fragment-based protein structure prediction by redesign of search heuristics. *Sci. Rep*., **8**, doi: 10.1038/s41598-018-31891-8.

Kosciolek,T. and Jones,D.T. (2014) De novo structure prediction of globular proteins aided by sequence variation-derived contacts. *PLoS One*, **9**, e92197.

Krasnogor,N. *et al*. (1999) *Protein Structure Prediction with Evolutionary Algorithms. Conference on Genetic & Evolutionary Computation*, Morgan Kaufmann Publishers Inc.

Li,Y. *et al*. (2018) Contact prediction by stacked fully convolutional residual neural network using coevolution features from deep multiple sequence alignment. *Thirteenth Meeting of Critical Assessment of Techniques for Protein Structure Prediction, pp. 154, Riviera Maya, Mexico*.

Liang,S. *et al*. (2014) LEAP: highly accurate prediction of protein loop conformations by integrating coarse-grained sampling and optimized energy scores with all-atom refinement of backbone and side chains. *J. Comput. Chem*., **35**, 335–341.

Ma,J. *et al*. (2015) Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. *Bioinformatics*, **31**, 3506–3513.

Marks,C. *et al*. (2017) Sphinx: merging knowledge-based and ab initio approaches to improve protein loop prediction. *Bioinformatics*, **33**, 1346–1353.

Marks,C. *et al*. (2018) Increasing the accuracy of protein loop structure prediction with evolutionary constraints. *Bioinformatics*, doi: 10.1093/bioinformatics/bty996.

Marks,D.S. *et al*. (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS One*, **6**, e28766.

Metropolis,N. *et al*. (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys*., **21**, 1087–1092.

Michel,M. *et al*. (2014) PconsFold: improved contact predictions improve protein models. *Bioinformatics*, **30**, i482–i488.

Mortuza,S.M.G. *et al*. (2018) C-QUARK: ab initio protein structure folding simulation guided by deep-learning based contact predictions. In: *Thirteenth Meeting of Critical Assessment of Techniques for Protein Structure Prediction, pp. 144, Riviera Maya, Mexico*.

Moult,J., *et al*. (2018) Critical assessment of methods of protein structure prediction (CASP) - Round XII. *Proteins*, **86**, 7–15.

Olson,B. and Shehu,A. (2014) Multi-objective optimization techniques for conformational sampling in template-free protein structure prediction. In: *International Conference on Bioinformatics and Computational Biology (BICoB)*, Vol. 2. Las Vegas, NV.

Ovchinnikov,S. *et al*. (2018) Protein structure prediction using Rosetta in casp12. *Proteins*, **86**, 113–121.

Rohl,C.A. *et al*. (2004) Protein structure prediction using Rosetta. *Methods Enzymol*., **383**, 66–93.

Rotkiewicz,P. and Skolnick,J. (2008) Fast procedure for reconstruction of full-atom protein models from reduced representations. *J. Comput. Chem*., **29**, 1460–1465.

Schaarschmidt,J. *et al*. (2018) Assessment of contact predictions in CASP12: co-evolution and deep learning coming of age. *Proteins*, **86**, 51–66.

Shrestha,R. and Zhang,K.Y.J. (2014) Improving fragment quality for de novo structure prediction. *Proteins*, **14**, 1288–1301.

Simons,K.T. *et al*. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol*., **268**, 209–225.

Skwark,M.J. *et al*. (2013) PconsC: combination of direct information methods and alignments improves contact prediction. *Bioinformatics*, **29**, 1815–1816.

Soto,C.S. *et al*. (2010) Loop modeling: sampling, filtering, and scoring. *Proteins*, **70**, 834–843.

Spassov,V.Z. *et al*. (2008) LOOPER: a molecular mechanics-based algorithm for protein loop prediction. *Protein Eng. Des. Sel*., **21**, 91–100.

Storn,R. and Price,K. (1997) Differential evolution: a simple and efficient heuristic for global optimization over continuous spaces. *J. Global Optim*., **11**, 341–359.

Wang,S. *et al*. (2016) CoinFold: a web server for protein contact prediction and contact-assisted protein folding. *Nucleic Acids Res*., **44**, W361–W366.

Wang,S. *et al*. (2017a) Folding membrane proteins by deep transfer learning. *Cell Syst*., **5**, 202–211.

Wang,S. *et al*. (2017b) Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput. Biol*., **13**, e1005324.

Wang,S. *et al*. (2018) Analysis of deep learning methods for blind protein contact prediction in CASP12. *Proteins*, **86**, 67–77.

Xu,J. and Zhang,Y. (2010) How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, **26**, 889–895.

Xu,D. and Zhang,Y. (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins*, **80**, 1715–1735.

Xu,D. and Zhang,Y. (2013) Toward optimal fragment generations for, ab initio protein structure assembly. *Proteins*, **81**, 229–239.

Xu,J.B. (2018) Protein structure modeling by predicted distance instead of contacts. In: *Thirteenth Meeting of Critical Assessment of Techniques for Protein Structure Prediction, pp. 146, Riviera Maya, Mexico*.

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

Zhang,G.J. *et al*. (2017) Enhancing protein conformational space sampling using distance profile-guided differential evolution. *IEEE/ACM Trans. Comput. Biol. Bioinform*., **14**, 1288–1301.

Zhang,G.J. *et al*. (2018) Secondary structure and contact guided differential evolution for protein structure prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform*., doi: 10.1109/TCBB.2018.2873691.

Zhang,Y. and Skolnick,J. (2004a) SPICKER: a clustering approach to identify near-native protein folds. *J. Comput. Chem*., **25**, 865–871.

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

Zhou,X.G. *et al*. (2016a) A novel differential evolution algorithm using local abstract convex underestimate strategy for global optimization. *Comput. Oper. Res*., **75**, 132–149.

Zhou,X.G. *et al*. (2016b) Enhanced differential evolution using local lipschitz underestimate strategy for computationally expensive optimization problems. *Appl. Soft Comput*., **48**, 169–181.

Zhou, X,G. *et al*. (2016c) Differential evolution with multi-stage strategies for global optimization. In: *IEEE Congress on Evolutionary Computation*, Vol. **49**, pp. 2550–2557. Canada.

Zhou,X.G. and Zhang,G.J. (2017) Abstract convex underestimation assisted multistage differential evolution. *IEEE Trans. Cybern*., **47**, 2730–2741.

Zhou,X.G. and Zhang,G.J. (2019) Differential evolution with underestimation-based multimutation strategy. *IEEE Trans. Cybern*., **49**, 1353–1364.

Zhou,X.G. *et al*. (2019a) Underestimation-assisted global-local cooperative differential evolution and the application to protein structure prediction. *IEEE Trans. Evol. Comput*., doi: 10.1109/TEVC.2019.2938531.

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