OXFORD

Briefings in Bioinformatics, 2022, **23**, 1–12
https://doi.org/10.1093/bib/bbab463
Problem Solving Protocol

Multi contact-based folding method for *de novo* protein structure prediction

Minghua Hou, Chunxiang Peng, Xiaogen Zhou, Biao Zhang and Guijun Zhang

Corresponding author. G. Zhang, College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China. E-mail: zgj@zjut.edu.cn

Abstract

Meta contact, which combines different contact maps into one to improve contact prediction accuracy and effectively reduce the noise from a single contact map, is a widely used method. However, protein structure prediction using meta contact cannot fully exploit the information carried by original contact maps. In this work, a multi contact-based folding method under the evolutionary algorithm framework, MultiCFold, is proposed. In MultiCFold, the thorough information of different contact maps is directly used by populations to guide protein structure folding. In addition, noncontact is considered as an effective supplement to contact information and can further assist protein folding. MultiCFold is tested on a set of 120 nonredundant proteins, and the average TM-score and average RMSD reach 0.617 and 5.815 Å, respectively. Compared with the meta contact-based method, MetaCFold, average TM-score and average RMSD have a 6.62 and 8.82% improvement. In particular, the import of noncontact information increases the average TM-score by 6.30%. Furthermore, MultiCFold is compared with four state-of-the-art methods of CASP13 on the 24 FM targets, and results show that MultiCFold is significantly better than other methods after the full-atom relax procedure.

Keywords: protein structure prediction, multi contact-based, noncontact information, evolutionary algorithm

Introduction

Protein structure prediction represents an important unsolved problem in computational biology, with the major challenge on distant-homology modeling (or *de novo* structure prediction; [1–3]). Thus, obtaining residue–residue contact information has been proved to be helpful for the improvement of the accuracy, as shown by recent encouraging advances in protein contact prediction in CASP experiments [4–7].

The accuracy of contact prediction has been significantly improved since 2012 in the 10th Critical Assessment of Protein Structure Prediction. At present, some groups have achieved remarkable success in contact prediction, such as Xu (RaptorX-Contact [8, 9]), DT. Jones (DeepMetaPSICOV [10]), Zhang (TripletRes [3]), Zhou (SPOT-Contact [11]), Yang (trRosetta [12]), Cheng (DNCON2 [13]), Gong (AmoebaContact [14]) and A7D (AlphaFold [15]). RaptorX-Contact applies an ultra-deep convolutional residual neural network to predict contacts and distance matrix as a whole instead of predicting one residue pair independent of the others, and structure motifs and long-range correlation can be used to greatly improve accuracy, which work well on proteins without many sequence homologs. TripletRes uses DeepMSA [16] to obtain multiple sequence alignment (MSA), and a precision matrix of the MSAs is then derived by maximum likelihood and used as the only input feature for contact model construction through deep residual Convolutional Neural Networks training. SPOT-Contact coupled of both residual Convolutional Neural Networks (ResNets [17]) and Residual two-dimensional Bidirectional Long Short-Term Memory Networks (2D-BRLSTM's [18, 19]) to improve the prediction accuracy of the secondary structure backbone angle and residue contact. The development of contact prediction has also improved the prediction accuracy of protein structure models and spawned a number of protein structure prediction methods based on contact prediction.

The idea of developing sequence-based contact-map prediction to assist de novo protein structure prediction is, however, not new, which can be traced back to at least 25 years ago [3]. In general, existing approaches for template-free model generation using predicted contacts fall into two main categories: fragment assembly and

Minghua Hou is a PhD candidate in the College of Information Engineering, Zhejiang University of Technology. His research interests include bioinformatics, intelligent information processing and optimization theory.
Chunxiang Peng is a PhD candidate in the College of Information Engineering, Zhejiang University of Technology. His research interests include bioinformatics, intelligent information processing and optimization theory.
Xiaogen Zhou is a postdoctoral fellow in the Department of Computational Medicine and Bioinformatics, University of Michigan. His research interests include bioinformatics, intelligent information processing and optimization theory.
Biao Zhang is a lecturer in the College of Information Engineering, Zhejiang University of Technology. His research interests include bioinformatics, intelligent information processing and optimization theory.
Guijun Zhang is a professor in the College of Information Engineering, Zhejiang University of Technology. His research interests include bioinformatics, intelligent information processing and optimization theory.
Received: July 14, 2021. **Revised:** September 21, 2021. **Accepted:** October 10, 2021

geometric optimization. Well established fragment-based methods such as C-QUARK [7], CoDiFold [20], CGLFold [21], SCDE [22] and FRAGFOLD [23] add constraints from predicted contacts to an existing fragment assembly pipeline. C-QUARK is extended from QUARK by the integration of contact-map predictions (ResPRE [24] and NeBcon [25]) into the fragment assembly simulations. In CoDiFold, contacts and distance-profiles are organically combined into the Rosetta low-resolution energy function to improve the accuracy of energy function. CGLFold designs a contact-assisted global exploration and loop perturbation cooperative to guide protein structure prediction. In SCDE, an improved DE involving the predicted secondary structure and contact information is proposed to reduce the search space and alleviate the prediction bias caused by the inaccurate energy function. FRAGFOLD combines fragment assembly with statistical potentials and predicted contacts to construct the tertiary structure. Another approach to *de novo* modeling is the use of distance geometry to project contact information into 3D space, such as RaptorX [9], CONFOLD [26, 27] and EVfold [28]. RaptorX feeds the top predicted contacts as restraints into the CNS suite [29] to generate 3D models. CONFOLD translates contacts and secondary structures into distance, dihedral angle and hydrogen bond restraints according to a set of new conversion rules and then provides these restraints as input for a distance geometry algorithm to build tertiary structure models. EVfold uses MSA and maximum entropy model to infer distance constraints from evolutionary sequence variations, and then utilizes the distance constraints to predict protein structure. However, the accuracy of structural prediction is limited by the contact accuracy to a certain extent, and further increased precision will allow for the production of more accurate models [30].

Recently, the incorporation of two different coevolution methods, based on distinct underlying principles, increases the accuracy of contact prediction [30]. Some meta contact-based approaches, are proposed, such as MetaPSICOV [30], NeBcon [25], DNCON2 [13] and DeepMetaPSICOV [10], using an optimal combination of the advantages of different contact maps to improve contact prediction accuracy [25]. MetaPSICOV is a hybrid method that combines a classical neural network-based contact prediction method with three different coevolution methods (PSICOV [31], CCMpred [32] and FreeContact [33]) to improve the accuracy of predicted contacts from MSAs. NeBcon is a hierarchical algorithm for sequence-based protein contact-map prediction. It first uses the naive Bayes classifier theorem to calculate the posterior probability of eight machine learning and coevolution-based contact prediction programs (SVMSEQ [34], BETACON [35], SVMcon [36], PSICOV [31], CCMpred [32], FreeContact [33], MetaPSICOV [30] and STRUCTCH [37]). Final contact maps are then created by neural network machine that trains the posterior probability scores with intrinsic structural features from secondary structure, solvent accessibility and Shannon entropy of MSAs. DNCON2, the MSAs are used by CCMPred [32], FreeContact [33] and PSICOV [31] to generate residue–residue coevolution features, and predict the contact map of a protein of any length by integrating both residue–residue coevolution features and other features such as secondary structures, solvent accessibility and pairwise contact potentials. DeepMetaPSICOV combines MetaPSICOV [30] and DeepCov [38] using full convolution residual network and data enhancement strategies to improve the prediction accuracy of the inter-residue contact. The accuracy of meta contact predictor is significantly improved; however, it cannot fully exploit the information carried by every input contact map.

Despite considerable progress in the accuracy of predicted contacts from sequence by using meta-type, a single contact map is not enough to derive all of the contacts in a protein structure, and even a small numbers of incorrect predictions are drastically detrimental to the modeling process [39]. The Jaccard indicator of CASP13 shows that different contact prediction servers capture quite different sets of contacts [6]; thus, multiple contact maps directly used to guide protein structure folding may be a feasible way to improve prediction accuracy. In addition, the accuracy of distance predictors has been significantly improved in recent years, and some methods gradually begin to use distance map to guide the conformation update. However, the evaluation standard of distance prediction has not been completely unified, and contact prediction is still adopted in CASP competitions to evaluate the accuracy of contact/distance map. Therefore, a multi contact-based folding method (MultiCFold) is proposed on the basis of the framework of evolutionary algorithms [40–42] to explore the use of thorough information and the mitigation of noise from the contact map. In MultiCFold, contact maps from different contact predictors are directly used to design an energy function, in which full information of every contact map is utilized to guide protein conformation sampling. The proposed algorithm adopts a population-based optimization selection strategy to select the model that best satisfies the information of several contact maps simultaneously. The experimental results show that the proposed MultiCFold can significantly improve the accuracy of the prediction model.

Materials and methods

Overview

Reliable contacts are important to improve the accuracy of *de novo* protein structure prediction, and the CASP13 results show that different methods can capture different sets of contacts. As illustrated in Figure 1(A), two different contact maps are used to guide protein folding. The dots in the picture represent different conformations. Among them, dots 1 and 2 have the lowest energy on a single contact map, whereas the corresponding energy of the other contact map is higher. Therefore,

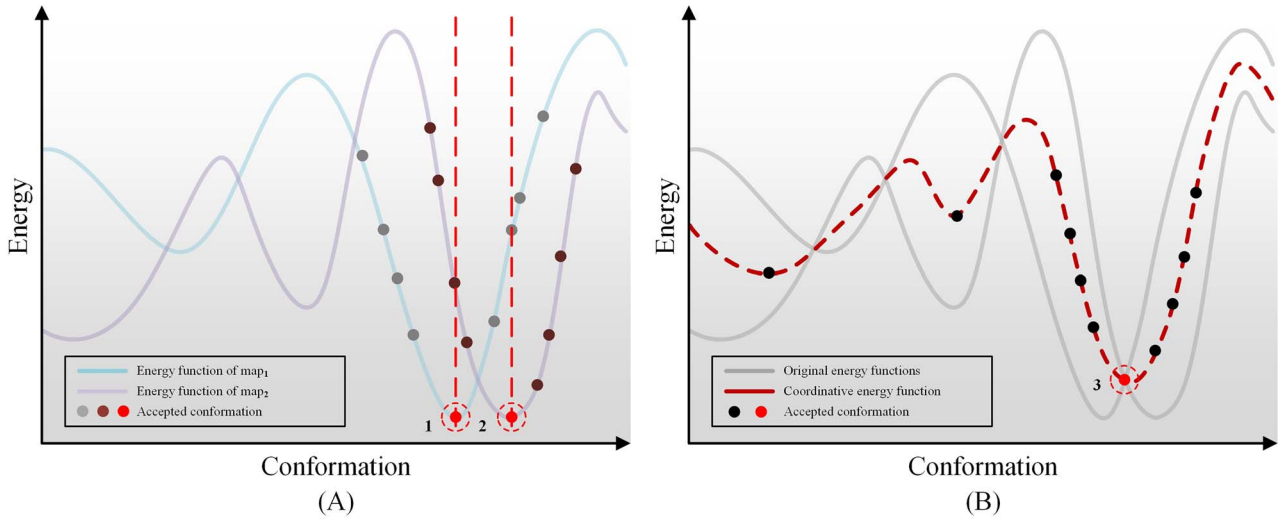

Figure 1. Schematic in conformational distribution under different energy functions. (**A**) The blue and purple lines represent the energy function curve corresponding to contact map 1 and contact map 2, respectively. (**B**) The gray lines represent the original energy functions showed in (A), and red dotted line represents their coordinative energy function. The dot of each line represents the accepted conformation and the red dot represents the best conformation in an energy function.

conformation 1 is consistent with the contact information of contact map 1, but it differs greatly from the contact information of contact map 2. Conformation 2 is just the opposite. Exploring a method to detect the optimal conformation satisfying different contact map constraints without losing each contact information may be important for protein structure prediction methods based on the predicted contact. In this work, an energy function is designed by using the full information of different contact maps to guide protein folding. As illustrated in Figure 1(B), although conformation 3 is not the local minimum of the two original energy functions, it satisfies the residue contact information of both two contact maps by transforming the multi-objective energy function into a single-objective function.

In addition, the protein structure prediction methods based on contact usually only uses residue contact information, but in fact residue noncontact is also an important information. There are more residue–residue pairs are noncontact instead of contact in a protein, and it may result in a greater number of reliable noncontacts being obtained in the contact map than contacts that are successfully predicted. Therefore, the noncontact information is an effective supplement to contact information, and it can further assist protein folding, which has also been introduced into the design of energy function. In MultiCFold, population-based optimization is used to sample conformational space under the guidance of the designed function, selecting the best solution of several energy functions to achieve the trade-off, that is, the conformation that satisfies multiple contact information for population update.

The pipeline is illustrated in Figure 2. The sequence and several residue–residue contact maps are used as inputs, and MultiCFold finally outputs the predicted three-dimensional structure of a target sequence. MultiCFold is developed on the framework of evolutionary algorithm. First, the initial population is generated through random fragment assembly. The decoys in the population will be scored and ranked by the designed function, and then the better half of them are selected as decoys for the next generation. Fragment recombination is used to generate new decoys through information interaction among the selected decoys in the population, and fragment assembly is performed to further improve the diversity of new conformations. Finally, the lowest energy conformation is selected from the final population as the prediction model. The Framework and specific details of MultiCFold are in Supplementary Materials Section S1.

Contact-based energy model

The information of the contact map can help guide protein folding, which has certain scientific significance for the correct prediction of protein structure. In this work, for each test protein in the benchmark dataset, four servers (TripletRes [3], RaptorX-Contact [9], DeepMetaPSICOV [10] and SPOT-Contact [11]) are used to predict the contact maps. In building an energy model based on contact map, residue contacts with confidence > 0.5 are selected as the contact information. In addition, the residue pairs in the predicted contact map with confidence below 0.1 are selected as noncontact information to supplement the energy model. Contact and noncontact information will be used as the prior knowledge in the algorithm. For each contact map, the energy model is defined as follows:

$$E_c = 2\sum_{k_1=1}^{N_c} e_c + \frac{N_c}{N_{nonc}}\sum_{k_2=1}^{N_{nonc}} e_{nonc} \quad (1)$$

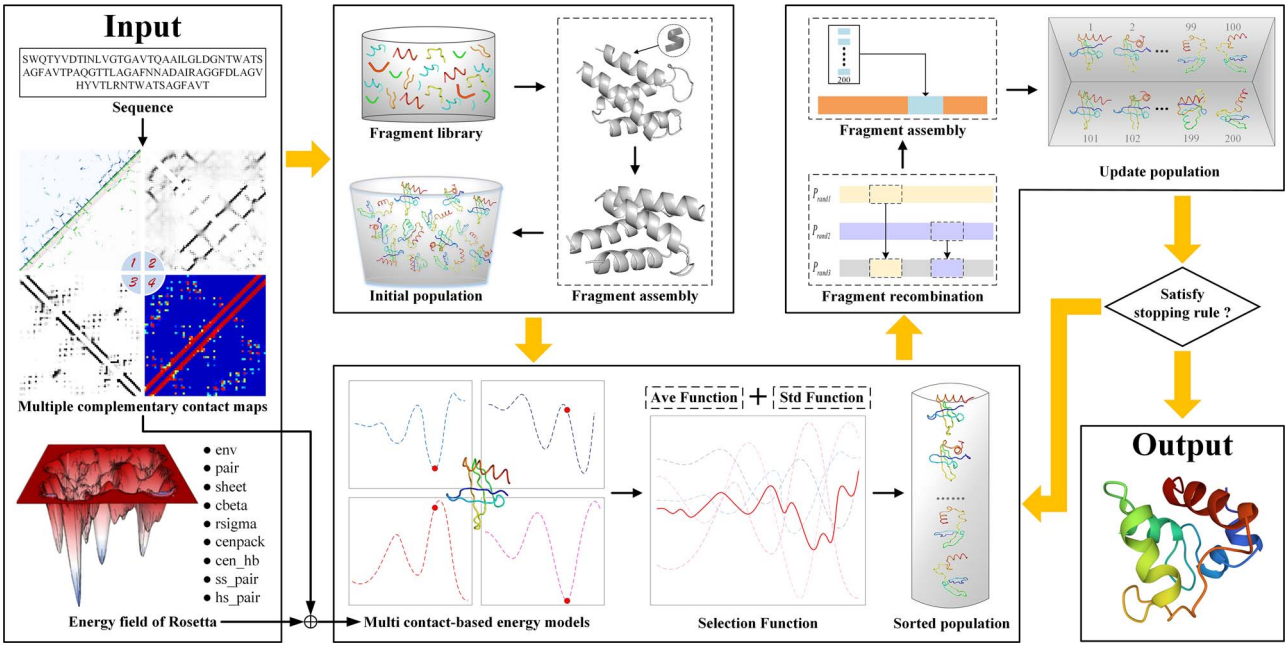

Figure 2. Pipeline of MultiCFold. (1) Initialization. The initial population and the contact-based function are constructed. (2) Sorting population. Low-energy conformations are selected by multi contact-based energy model. (3) Recruit population. New conformations are generated on the basis of the selected conformations to replenish the population to the set size. (4) Output. The final prediction model is obtained.

where N_c and N_{nonc} are the number of contact and non-contact residue pairs, respectively. e_c is the energy of contact, and e_{nonc} is the energy of noncontact. First, each residue pair of the decoy can be classified according to the confidence of residue–residue contact maps (P_{ij}): if $P_{ij} > 0.5$, then the residue pairs will be marked as contactable and scored by e_c; if $P_{ij} < 0.1$, then the residue pair will be marked as contactless and scored by e_{nonc}; otherwise, the residue pair will not be used for scoring. e_c and e_{nonc} are the scores for each residue–residue pair, which can be calculated as follows:

$$e_c = \begin{cases} e^{P_{ij}}\left(d_{clash} - d_{ij}\right), & d_{ij} \le d_{clash} \\ e^{P_{ij}}\left(d_{ij} - d_{con}\right), & d_{clash} < d_{ij} \le d_{con} \\ \ln\left(1 + P_{ij}\right)\left(d_{ij} - d_{con}\right), & d_{con} < d_{ij} \le 10 \\ \left(e^{P_{ij}} - 1\right)\left(d_{ij} - d_{con}\right), & \text{otherwise} \end{cases} \quad (2)$$

$$e_{nonc} = \begin{cases} e^{1-P_{ij}}\left(d_{con} - d_{ij}\right), & d_{ij} \le d_{con} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where i and j are the residue indices; P_{ij} is the confidence score that the ith and jth residue are in contact; d_{ij} is the true distance between residue i and residue j of the trial conformation and $d_{clash} = 3.8$Å is the minimum distance [43], indicating that the two residues have no spatial clash, and $d_{con} = 8$Å is the maximum distance between the two residues that are in contact.

In this contact-based energy model, the smaller E_c is, the more satisfying the corresponding contact constraint will be, and the closer the decoy is to the native structure.

Multi contact-based population optimization

The quality and quantity of the residue–residue contact is important to the prediction of the protein structure. This paper proposes a selection strategy based on population optimization to effectively exploit thorough information of four contact maps, which can mitigate the shortcomings of evaluating conformations with a single contact map by providing richer information.

In MultiCFold, the Rosetta low-resolution energy function E_{score3} [44] is added to the designed E_c to form an improved total energy function. The total energy function is as follows:

$$E_{sc} = E_{score3} + E_c \quad (4)$$

where E_c is calculated by Equation (3) and E_{score3} is calculated by Rosetta low-resolution energy function score3. For each target, several contact maps are predicted by different contact servers, and several different E_{sc} are calculated by Equation (4) using the contact maps obtained by prediction.

The servers contain different contact information, resulting in several different energies for each decoy in the population. When all the energies of the decoy are consistently low, the decoy has a high probability to close to the native structure. However, in the multi-objective optimization problem, a decoy that makes all the goals achieve the lowest energy may not be found. Therefore, a decoy that makes all the goals achieve the lowest energy value as much as possible is selected as the optimal decoy. In MultiCFold, the weighted average $\overline{E_{sc}}$ and variance E_{std} of the energies of each decoy are calculated, and the decoy with the smallest weighted sum of variance and average is selected as the optimal decoy. The weighted average energy function and variance function are calculated by the following formulas:

$$\overline{E_{sc}} = E_{score3} + \overline{E_c} = E_{score3} + \frac{\omega_1 E_{c1} + \omega_2 E_{c2} + \omega_3 E_{c3} + \cdots + \omega_n E_{cn}}{\omega_1 + \omega_2 + \omega_3 + \cdots + \omega_n} \quad (5)$$

[[DmEquation5]]

$$\Delta E_{scn} = E_{scn} - \overline{E_{sc}} = E_{cn} - \overline{E_c} \quad (6)$$

$$E_{std} = \sqrt{\frac{(\Delta E_{sc1})^2 + (\Delta E_{sc2})^2 + (\Delta E_{sc3})^2 + \cdots + (\Delta E_{scn})^2}{n}} \quad (7)$$

where ω_n and E_{scn} are the weight and energy of the nth contact map of several servers. When both $\overline{E_{sc}}$ and E_{std} are indicators to judge decoy quality, $\overline{E_{sc}}$ indicates the accuracy of the decoy, and E_{std} indicates the dissimilitude degree of the decoy in four contact maps. Use the weighted sum of the normalized average and variance as an improved scoring function to judge the decoy. The improved scoring model is as follows:

$$E_{score} = \lambda_1 \overline{E^*_{sc}} + \lambda_2 E^*_{std} \quad (8)$$

where λ_1 and λ_2 are the specific weight values; $\overline{E^*_{sc}}$ and E^*_{std} are the average $\overline{E_{sc}}$ and variance E_{std} of one decoy after normalization according to the $\overline{E_{sc}}$ and E_{std} of all decoys in the current iteration population, respectively. The lower the E_{score}, the more the decoy conforms to the comprehensive information predicted by the servers. In MultiCFold, four servers (TripletRes, RaptorX-Contact, MetaPSICOV and SPOT-Contact) are used to provide contact information, and the weight refers to the F-score of these servers in CASP13, which is used to measure the performance of the contact server.

Result

Experiment settings and performance evaluation

In this study, the performance of MultiCFold is tested in 120 benchmark proteins, which are systematically selected from the PDB. The length of these proteins ranges from 52 to 199 residues, with <30% sequence identity to each other. First, 243 819 proteins with known structures from the SCOPe 2.07–2020–07–16 [45, 46] are clustered by CD-HIT [47, 48] with a 30% sequence identity cutoff, which result in 11 198 proteins. MultiCFold is the predictor mainly for single-domain protein structure. Thus, according to SCOPe, 2481 proteins are obtained after excluding the multidomain proteins and the proteins with a length of <50 and >200 from the 11 198 proteins. Finally, 120 proteins are selected from the 2481 remaining proteins according to their length diversity as the benchmark set [49, 50]. MultiCFold is compared with four state-of-the-art servers in 24 CASP13 FM targets to further test its performance. The length of the 24 FM targets varies from 41 to 354 residues.

The parameters of MultiCFold in all experiments are set as follows: population size $NP = 200$, crossover rate $CR = 0.5$, temperature scaling factor $beta = 1$, maximum generation $G = 500$ and the weights λ_1=0.75, λ_2=0.25. Details about parameter settings are shown in Supplementary Materials Section S2. For all test proteins, the fragment library with homologous fragments removed, is downloaded from the Robetta server (https://robetta.bakerlab.org).The performance of MultiCFold is evaluated by two main measures, including RMSD and the template modeling score (TM-score) [51, 52].

Comparison with a meta contact-based method (MetaCFold)

To study the performance between multi contact-based method (MultiCFold) and meta contact-based method (MetaCFold) and investigate which method of using several contact maps is instrumental for better results, MultiCFold is compared with MetaCFold on the 120 proteins of a benchmark set. Population-based optimization and contact information from the same severs are used in both two methods to predict protein tertiary structure without a homologous structure. For MultiCFold, the decoy with the lowest E_{score} is considered as the final model. For the fairness of comparison, the contact-based energy model containing noncontact information (Equations 1–3) is used in MetaCFold, and the contact is a meta contact map, which is obtained by combining four predicted contact maps according to the confidence level. The first decoy of MetaCFold is selected as the final model. Details of MetaCFold are shown in Supplementary Materials Section S3.

The predicted results of MultiCFold and MetaCFold on the benchmark set are summarized in Table 1, and the detailed results of each protein are presented in Table S11, Supplementary Materials S6. The average TM-score of the final model predicted by MultiCFold (0.617) is 6.62% higher than that of MetaCFold (0.579), and the average RMSD of MultiCFold (5.815 Å) are reduced by 8.82% compared with the MetaCFold (6.377 Å). MultiCFold and MetaCFold obtain models with TM-score > 0.5 in 107 and 90 out of 120 proteins, accounting for 89.2 and 75.0% of the total protein, respectively.

For a more intuitive comparison of the RMSD and TM-score between MultiCFold and MetaCFold, their results are depicted in Figure 3. As shown in the figure, MultiCFold achieves a lower RMSD in 80 proteins and a higher TM-score in 93 proteins than MetaCFold.

In order to analyze the proposed MultiCFold more objectively, the Wilcoxon signed-rank test is used to analyze the significant difference among the results with a significance level of 0.05. In the significance test result, the P-value of MultiCFold with MetaCFold is 2.01E−03, indicating that the performance of MultiCFold is significantly better than that of MetaCFold. In addition, this result demonstrates that directly using multiple contact maps during the folding process may be more effective than using a meta contact map in exploiting thorough information of every input contact map.

Table 1. Predicted results of MultiCFold and MetaCFold

Method	RMSD	TM-score	#TM ≥ 0.5	#TM ≥ 0.6	#TM ≥ 0.7	#TM ≥ 0.8	P-value	Sig
MultiCFold	5.815	0.617	107	65	25	4	NA	NA
MetaCFold	6.377	0.579	90	52	15	3	2.01E−3	+

#TM ≥ 0.5, ≥0.6, ≥0.7, ≥0.8 are the number of the predicted model with TM-score ≥ 0.5, ≥0.6, ≥0.7 and ≥0.8, respectively. *P*-value and Sig (significance) are the results of the Wilcoxon signed rank test. NA represents that there has no data.

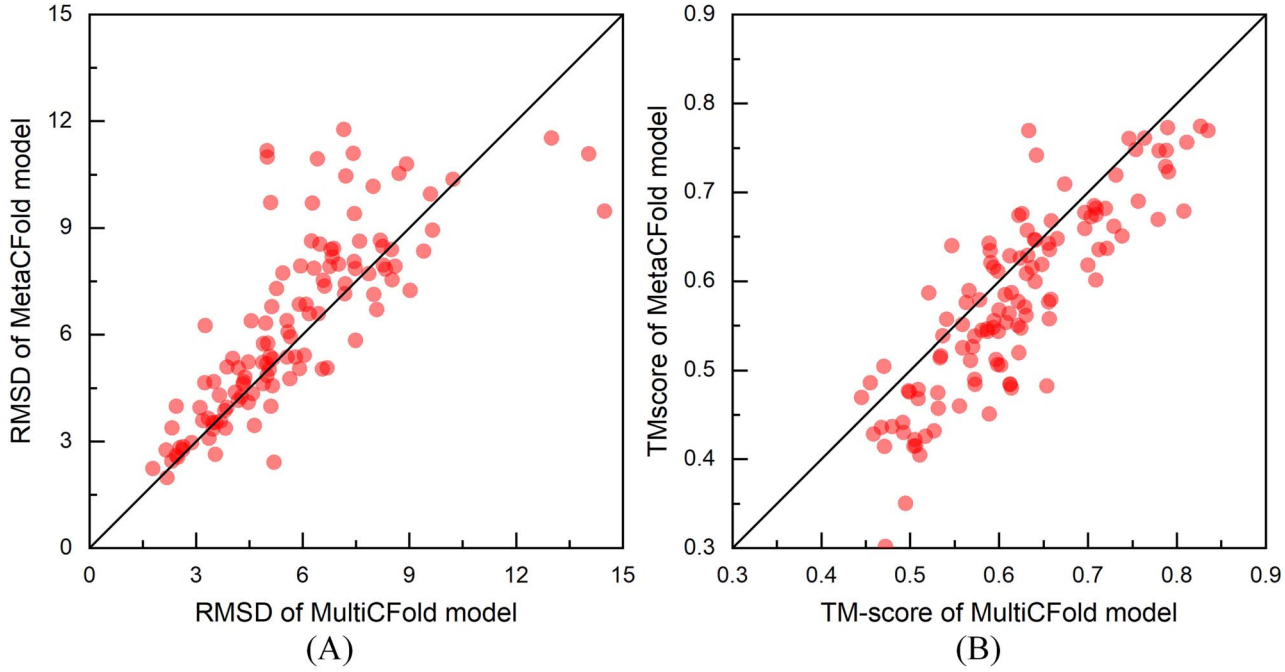

Figure 3. Comparison of the predicted models with the final models of MetaCFold. (**A**) represents the comparison of the predicted models in RMSD, where the *y* coordinate of red circle represents RMSD of the final model obtained by MetaCFold and the *x* coordinate represents RMSD of the model obtained by MultiCFold. Similarly, (**B**) represents the comparison of the predicted models in TM-score.

Effect of multi contact-based strategy

In MultiCFold, four contact maps, predicted from DeepMetaPSICOV, RaptorX-Contact, SPOT-Contact and TripletRes, are used to construct the total scoring function E_{score} and guide conformational folding. A single contact map is used in MultiCFold to discuss the effect of multi contact-based strategy and various experiments are conducted in 120 benchmark proteins. In addition, given that the current methods only use contact information when using a contact map. The MultiCFold method without noncontact information, named MultiCFold-C^4, is used in conducting comparative experiments to explore its potential application in other methods or servers. Here, MultiCFold-$\mathbf{C}_1^4$, MultiCFold-$\mathbf{C}_2^4$, MultiCFold-$\mathbf{C}_3^4$ and MultiCFold-$\mathbf{C}_4^4$ only use one of the contact maps used in MultiCFold-C^4 predicted by DeepMetaPSICOV, RaptorX-Contact, SPOT-Contact and TripletRes, respectively.

The predicted results of MultiCFold-C^4, MultiCFold-$\mathbf{C}_1^4$, MultiCFold-$\mathbf{C}_2^4$, MultiCFold-$\mathbf{C}_3^4$ and MultiCFold-$\mathbf{C}_4^4$ on the benchmark are summarized in Table 2, and the detailed results of each protein can be found in Table S12, Supplementary Material S6. Based on the results, MultiCFold-C^4 achieves better results than MultiCFold-$\mathbf{C}_1^4$, MultiCFold-$\mathbf{C}_2^4$, MultiCFold-$\mathbf{C}_3^4$ and MultiCFold-$\mathbf{C}_4^4$ in 86, 92, 78 and 94 out of 120 proteins with regard to RMSD, respectively. The average RMSD of MultiCFold-C^4 is 6.363 Å, which is reduced by 14.9, 20.3, 11.0 and 15.0% compared with that of MultiCFold-$\mathbf{C}_1^4$ (7.475 Å), MultiCFold-$\mathbf{C}_2^4$ (7.980 Å), MultiCFold-$\mathbf{C}_3^4$ (7.149 Å) and MultiCFold-$\mathbf{C}_4^4$ (7.485 Å). The average TM-score of MultiCFold-C^4 is 0.581, which is 14.2, 18.2, 11.7 and 12.8% higher than that of MultiCFold-$\mathbf{C}_1^4$ (0.509), MultiCFold-$\mathbf{C}_2^4$ (0.492), MultiCFold-$\mathbf{C}_3^4$ (0.520) and MultiCFold-$\mathbf{C}_4^4$ (0.515). MultiCFold-C^4 obtains models with TM-score > 0.5 in 90 out of 120 proteins and accounts for 75.0% of the total protein, whereas MultiCFold-$\mathbf{C}_1^4$, MultiCFold-$\mathbf{C}_2^4$, MultiCFold-$\mathbf{C}_3^4$ and MultiCFold-$\mathbf{C}_4^4$ obtain models with TM-score > 0.5 in 62, 57, 73, 66 out of 120 proteins and accounts for 51.7, 47.5, 60.8 and 55.0% of the total protein, respectively. The significance test results (*P*-value = 1.79E−06, 1.16E−08, 8.60E−05 and 5.66E−06) show that the performance of MultiCFold-C^4 is significantly better than that of MultiCFold-$\mathbf{C}_1^4$, MultiCFold-$\mathbf{C}_2^4$, MultiCFold-$\mathbf{C}_3^4$ and MultiCFold-$\mathbf{C}_4^4$.

Table 2. Predicted results of MultiCFold-C^4, MultiCFold-C_1^4, MultiCFold-C_2^4, MultiCFold-C_3^4 and MultiCFold-C_4^4

Method	RMSD	TM-score	#TM ≥ 0.5	#TM ≥ 0.6	#TM ≥ 0.7	#TM ≥ 0.8	*P*-value	Sig
MultiCFold-C^4	6.363	0.581	90	49	17	4	NA	NA
MultiCFold-C_1^4	7.475	0.509	62	26	11	0	1.79E−06	+
MultiCFold-C_2^4	7.980	0.492	57	25	6	0	1.16E−08	+
MultiCFold-C_3^4	7.149	0.520	73	34	9	0	8.60E−05	+
MultiCFold-C_4^4	7.485	0.515	66	27	7	0	5.66E−06	+

#TM ≥ 0.5, ≥0.6, ≥0.7, ≥0.8 are the number of the predicted model with TM-score ≥ 0.5, ≥0.6, ≥0.7, ≥0.8, respectively. *P*-value and Sig (significance) are the results of the Wilcoxon signed rank test. NA represents that there has no data.

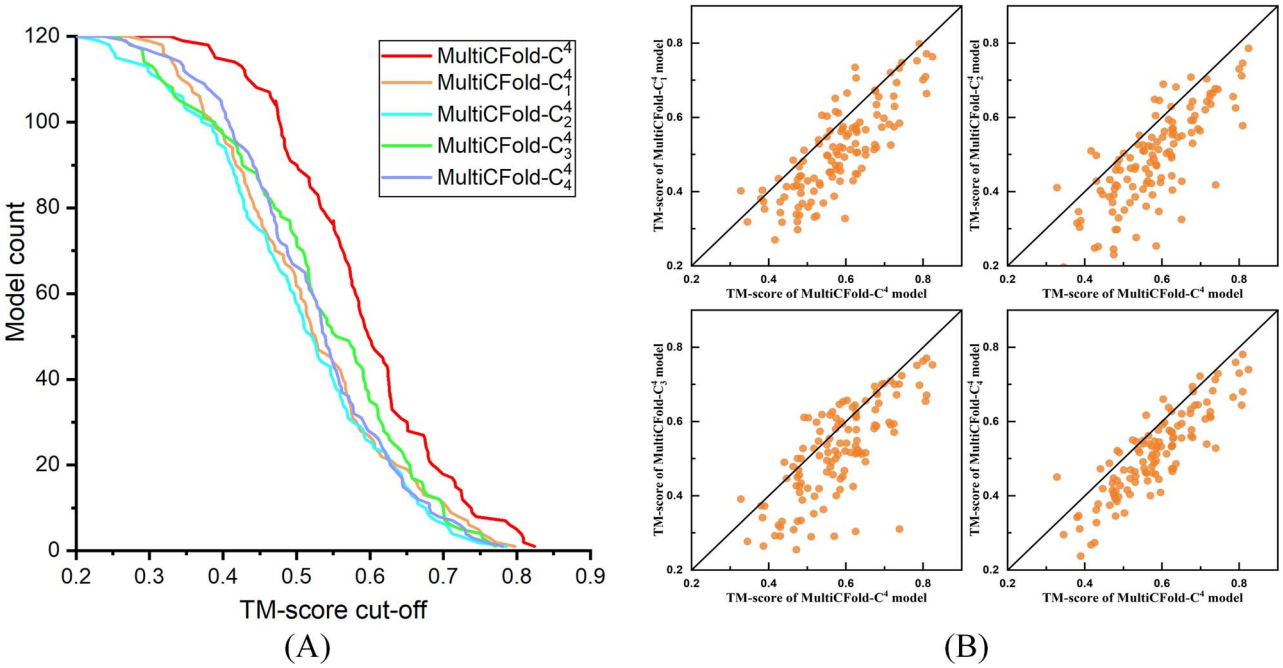

Figure 4. Comparison of MultiCFold-C^4 with MultiCFold-C_1^4, MultiCFold-C_2^4, MultiCFold-C_3^4 and MultiCFold-C_4^4. (**A**) comparison of the predicted models in TM-score, where the *x* and *y* coordinate of different colored lines represents the TM-score cut-off and the number of models obtained by different approaches, respectively. (**B**) head-to-head comparison of the predicted models by MultiCFold-C^4, MultiCFold-C_1^4, MultiCFold-C_2^4, MultiCFold-C_3^4, and MultiCFold-C_4^4 in TM-score.

For a more intuitive comparison of the quality of tertiary structures predicted by several test methods for benchmark target proteins, the cut-off threshold of their results is depicted in Figure 4A. As illustrated in Figure 4A, MultiCFold-C^4 has predicted more high-quality structures than the other approaches. When using the popular cut-off threshold for high-quality structures (TM-score > 0.7), MultiCFold-C^4 has predicted high-quality structures for 17 out of 120 proteins, whereas MultiCFold-C_1^4, MultiCFold-C_2^4, MultiCFold-C_3^4 and MultiCFold-C_4^4 have predicted high-quality structure for only 11, 6, 9 and 7 proteins, respectively. Figure 4B, head-to-head comparison, clearly demonstrates the advantages of MultiCFold-C^4 over other methods: for 101, 108, 90 and 107 out of 120 proteins, MultiCFold-C^4 outperformed MultiCFold-C_1^4, MultiCFold-C_2^4, MultiCFold-C_3^4 and MultiCFold-C_4^4 with regard to the TM-score.

Figure 5 shows the comparison of the native structure (protein ID: 1FCA_A) with the predicted model of MultiCFold-C^4, MultiCFold-C_1^4, MultiCFold-C_2^4, MultiCFold-C_3^4 and MultiCFold-C_4^4. The TM-score of MultiCFold-C^4 (0.808) is significantly higher than that of MultiCFold-C_1^4 (0.664), MultiCFold-C_2^4 (0.578), MultiCFold-C_3^4 (0.671) and MultiCFold-C_4^4 (0.681). By calculating residue–residue contacts in the final model predicted by MultiCFold-C^4 and comparing them with the residual contact information of four different contact maps and the nature structure, this result may indicate that the full information of multiple contact maps can be used to mitigate the noise caused by the mispredicted residue pairs in the single contact map during structure folding. For example, the distance between the 8th residue and 49th residue is incorrectly predicted as noncontact in the second contact map, whereas the right prediction of the other three contact maps has corrected this misinformation; therefore, the distance between the residue pair in the final model is consistent with the contact state of the natural structure. In addition, all five contact maps: four single contact maps predicted by DeepMetaPSICOV,

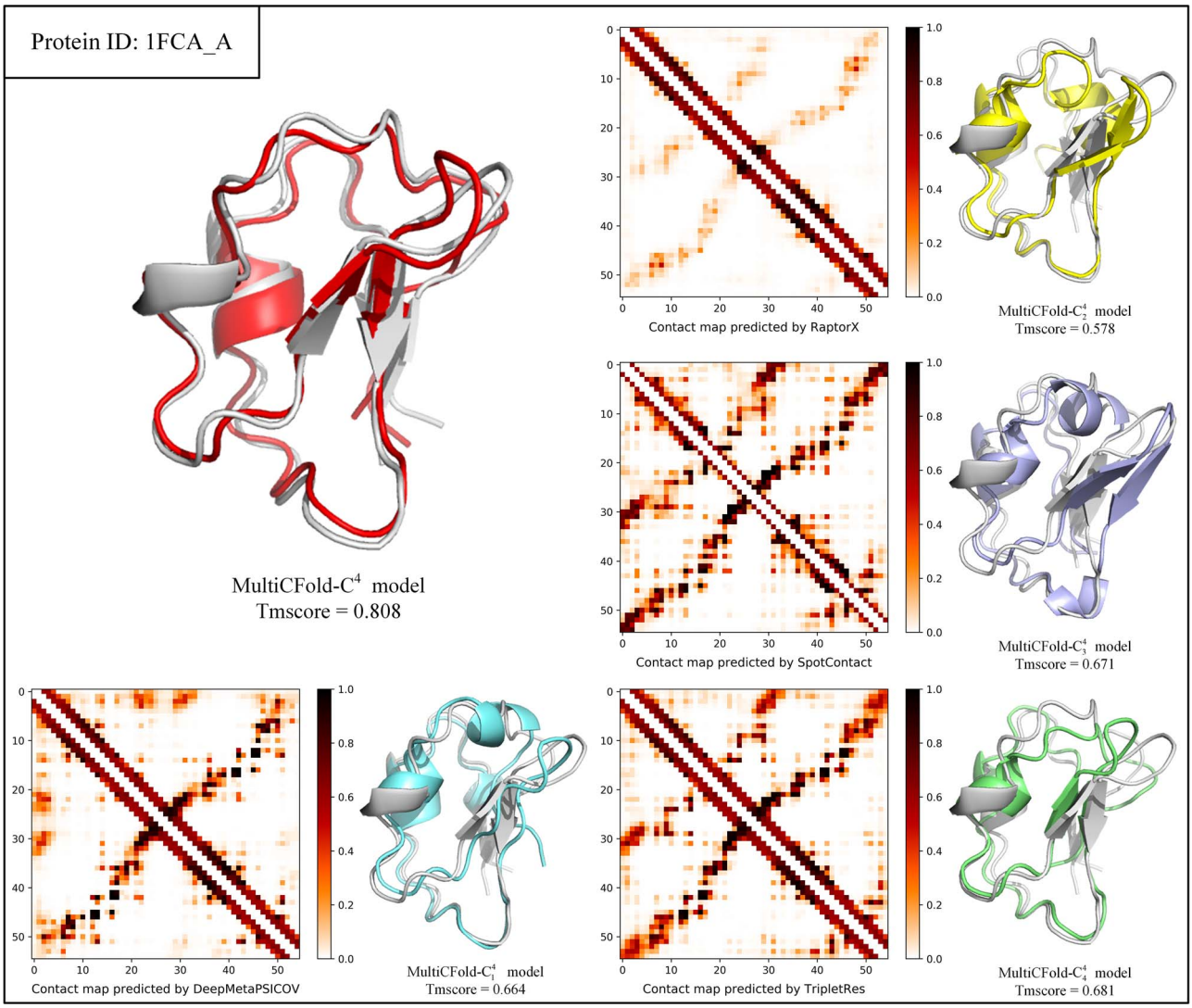

Figure 5. Contact maps predicted by four severs (DeepMetaPSICOV, RaptorX-Contact, SPOT-Contact and TripletRes) and superimposition between the predicted model by MultiCFold-C^4 (red), MultiCFold-C_1^4 (aquamarine), MultiCFold-C_2^4 (yellow), MultiCFold-C_3^4 (lightblue), MultiCFold-C_4^4 (aquamarine) and the native structure (gray) for target (protein ID: 1FCA_A).

RaptorX-Contact, SPOT-Contact and TripletRes; one meta contact map obtained by the four above-mentioned maps, are plotted for each target protein, and their precision is tabulated in the Supplementary Materials.

Effect of using noncontact information

In MultiCFold, for each contact map, the residue–residue pairs with contact confidence > 0.5 are used as contact information, and the residue–residue pairs with contact confidence < 0.1 are used as noncontact information. Only contact information is used in MultiCFold to discuss the effect of noncontact information, and various experiments are conducted in 120 benchmark proteins to investigate whether the use of noncontact information obtains remarkable results. Here, MultiCFold-C^4 represents MultiCFold without noncontact information.

The predicted results of MultiCFold and MultiCFold-C^4 on the benchmark are summarized in Table 3, and the detailed results of each protein can be found in Table S13, Supplementary Materials S6. The average RMSD of MultiCFold (5.815 Å) is reduced by 8.6% compared with that of MultiCFold-C^4 (6.363 Å) and the average TM-score of MultiCFold (0.617) is 6.3% higher than that of MultiCFold-C^4 (0.581). MultiCFold obtains models with TM-score > 0.5 in 107 out of 120 proteins and accounts for 89.1% of the total protein, whereas MultiCFold-C^4 obtains models with TM-score > 0.5 in 90 out of 120 proteins and accounts for 75.0% of the total protein. The significance test results (*P*-value = 5.22E−03) show that the performance of MultiCFold is significantly better than that of MultiCFold-C^4.

For a more intuitive comparison of the TM-score between MultiCFold and MultiCFold-C^4, their results are depicted in Figure 6. As illustrated in Figure 6A, with the introduction of noncontact information, the overall accuracy of the prediction results is improved and its distribution is also significantly improved. The comparison of each target protein in TM-score is depicted in Figure 6B, and MultiCFold achieves a higher TM-score in 90 proteins than MultiCFold-C^4. Based on the results

Table 3. Predicted results of MultiCFold and MultiCFold-C^4

Method	RMSD	TM-score	#TM ≥ 0.5	#TM ≥ 0.6	#TM ≥ 0.7	#TM ≥ 0.8	P-value	Sig
MultiCFold	5.815	0.617	107	65	25	4	NA	NA
MultiCFold-C^4	6.363	0.581	90	49	17	4	5.22E−03	+

#TM ≥ 0.5, ≥0.6, ≥0.7, ≥0.8 are the number of the predicted model with TM-score ≥ 0.5, ≥0.6, ≥0.7, ≥0.8, respectively. *P*-value and Sig (significance) are the results of the Wilcoxon signed rank test. NA represents that there has no data.

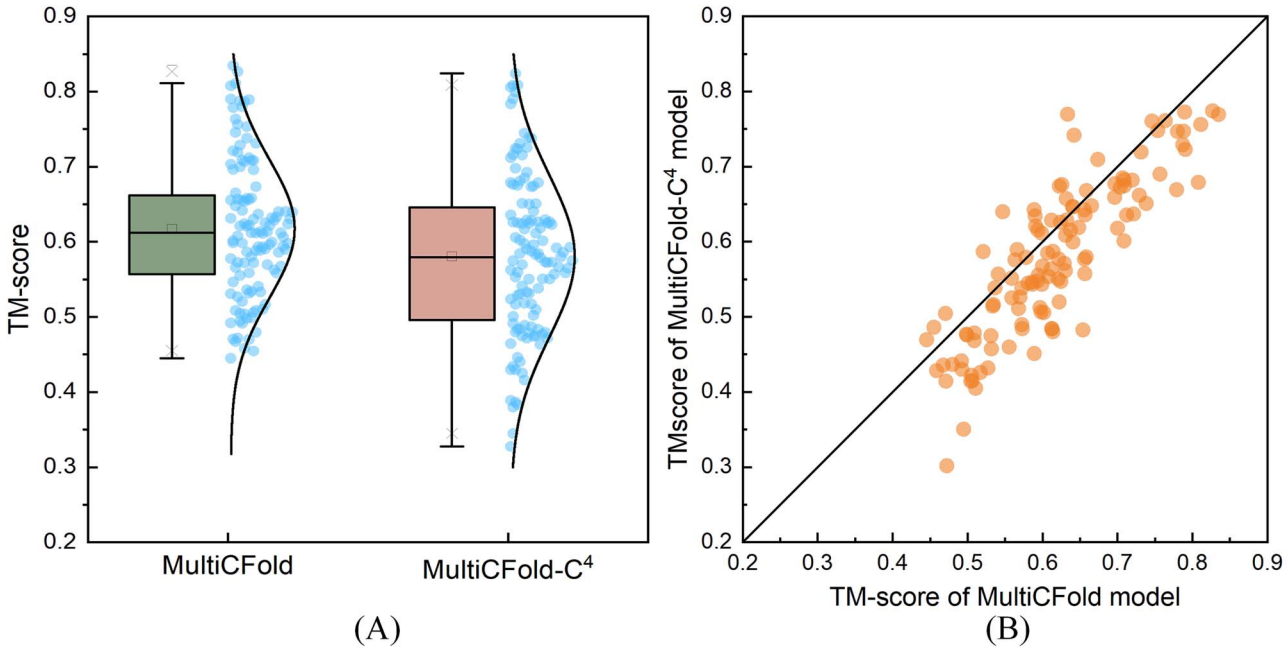

Figure 6. Comparison of the prediction performance between MultiCFold and MultiCFold-C^4 in all proteins. (**A**) Boxplot for the TM-score of the predicted models by MultiCFold and MultiCFold-C^4. (**B**) Head-to-head comparison between the TM-score of the predicted models by MultiCFold and MultiCFold-C^4.

of the above-mentioned comparative experiments, noncontact information is an effective supplement to contact information predicted in the same contact map, which can mitigate the noise in contacts and further assist protein folding. Figure 7 illustrates the clustered histogram of the average TM-score of all methods and reflects the contribution of different components of MultiCFold. In addition, we investigate the relationship between the map quality and folding accuracy with a figure (show in Supplementary Materials Section S4). It is not difficult to find that the folding accuracy of the model is improved with the improvement of the contact quality.

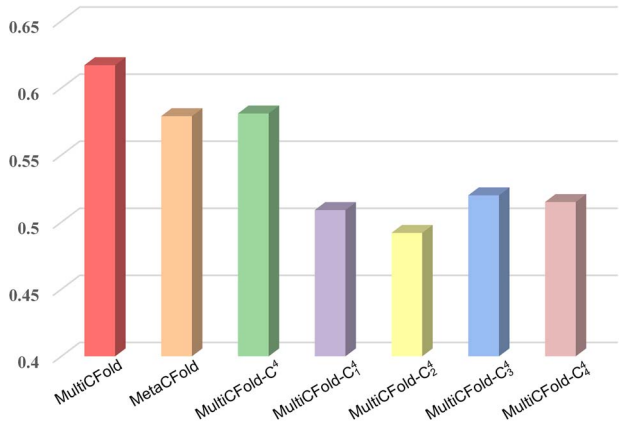

Figure 7. Average TM-score of the final model predicted by MultiCFold, MetaCFold, MultiCFold-C^4, MultiCFold-C_1^4, MultiCFold-C_2^4, MultiCFold-C_3^4 and MultiCFold-C_4^4 on the 120 benchmark test proteins.

Results of CASP13 targets

MultiCFold is compared with four methods of the server group in CASP 13, that is, QUARK [7], RaptorX-DeepModeller [53], BAKER-ROSETTASERVER and MULTICOM_CLUSTER [54], which are state-of-the-art methods in the FM category of CASP experiments (Table 4 and Table S14, Supplementary Materials S6). The results of the above-mentioned four methods are obtained from the CASP official website (https://predictioncenter.org/casp13/results.cgi), and such methods use contact information. Given that the above-mentioned four methods are full-version with a relax protocol to generate full-atom models, we use the Rosetta *Fastrelax* protocol to generate the full-atom model (MultiCFold-relax) for better comparison. In 24 CASP13 FM targets, the average TM-score of the final model predicted by MultiCFold and MultiCFold-relax is 0.431 and 0.567, respectively, which is comparable with QUARK (0.496), RaptorX-DeepModeller (0.491), BAKER-ROSETTASERVER (0.418) and MULTICOM_CLUSTER (0.371). In addition, we tested the

Table 4. Predicted results of MultiCFold, MultiCFold-relax, QUARK, RaptorX-DeepModeller, BAKER-ROSETTASERVER and MULTICOM_CLUSTER on the 24 CASP13 FM targets

Methods in CASP13	TM-score	Correctly fold
MultiCFold-relax	0.567	18
MultiCFold	0.431	7
QUARK	0.496	12
RaptorX-DeepModeller	0.491	13
BAKER-ROSETTASERVER	0.418	6
MULTICOM_CLUSTER	0.371	5

Correctly fold is the number of the predicted model with TM-score ≥ 0.5.

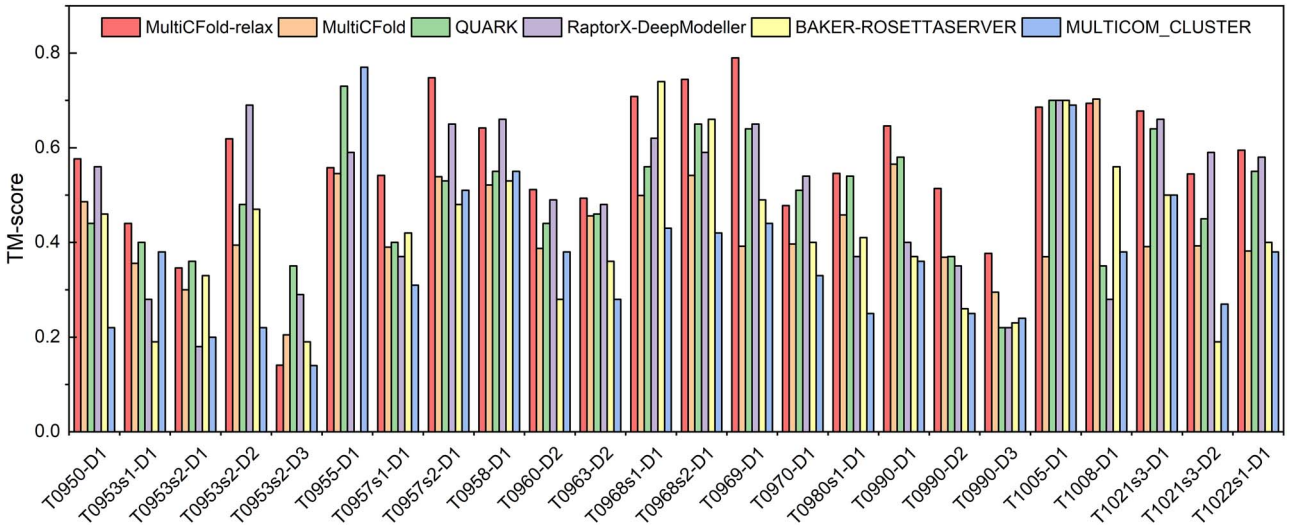

Figure 8. TM-score of the predicted model by MultiCFold, MultiCFold-relax, QUARK, RaptorX-DeepModeller, BAKER-ROSETTASERVER and MULTICOM_CLUSTER on the 24 FM targets of CASP13.

performance of MultiCFold-relax on 18 CASP14 and 20 hard recent CAMEO targets, and the comparison results are shown in Supplementary Materials Section S5. On the 18 FM targets of CASP14, the performance of MultiCFold-relax is comparable with the four methods. Meanwhile, in T1038-D2, T1074-D1, T1080-D1 and T1090-D1, we found that there are models generated by MultiCFold-relax closer to the native structure. On the 20 hard targets of CAMEO, the performance of MultiCFold is significantly better than the four methods.

The TM-score of the final model for each target using the six above-mentioned methods is illustrated in Figure 8. MultiCFold and MultiCFold-relax correctly folds 7 and 18 models with TM-score > 0.5 among 24 FM targets, respectively. An example of the improved performance of our method for target T0957s2-D1 is shown as follows: the TM-score of MultiCFold is 0.703 and MultiCFold-relax is 0.694. The values obtained by other methods are as follows: QUARK (0.53), BAKER-ROSETTASERVER (0.48), RaptorX-DeepModeller (0.65) and MULTICOM_CLUSTER (0.51). In long proteins, however, MultiCFold does not perform well, and the accuracy of MultiCFold on the targets T0969-D1 (354 AAs) and T1005-D1 (326 AAs) is relatively poor compared with other methods. However, there is an exception to the target T0950-D1 (342 AAs), TM-score of MultiCFold is 0.486, except for RaptorXDeepModeller (0.56), which is higher than the other three methods: QUARK (0.44), BAKER-ROSETTASERVER (0.46) and MULTICOM_CLUSTER (0.22). In addition, the accuracy of the MultiCFold-relax has improved significantly than MultiCFold, which might be due to the use of a more informative distance map than the contact map in the latest *Fastrelax* protocol released from Rosetta.

Conclusion

We propose a multi contact-based folding method under the evolutionary algorithm framework, MultiCFold, to directly use the thorough information of different contact maps using a population-based strategy to guide protein structure folding. In MultiCFold, the information of several contact maps predicted by different predictors is used to design an energy model. In this model, residue–residue noncontact information is used as an effective supplement to contact information to further assist protein folding. In addition, this algorithm adopts a multi contact-based population optimization strategy to select conformation that best satisfies four contact maps simultaneously, mitigating the shortcoming of one single contact map by fully exploiting the information carried by every input contact map. MultiCFold is tested on a set of 120 nonredundant proteins, and the results show that the two above-mentioned strategies can improve the accuracy of the model.

MultiCFold is compared with MetaCFold in 120 benchmark proteins and with QUARK, RaptorX-DeepModeller, BAKER-ROSETTASERVER and MULTICOM_CLUSTER in 24 FM targets from CASP13. The experimental results show that the accuracy of MultiCFold is comparable with these existing popular approaches. MultiCFold correctly folds 107 out of 120 benchmark proteins and 7 out of 24 CASP13 FM targets. Notably, the multi-folding strategy is an independent protocol, which can be used for improving the model accuracy on other advanced folding simulation programs. It has been applied to the Rosetta platform in this study as an illustrative example. Distance prediction has become more mature in recent years, and distance maps contain richer information than contact maps, we may then try to improve further based on distance maps. In addition, effectively combining the physical and chemical knowledge with the information obtained by deep learning prediction to improve the accuracy of structure prediction is also a direction of our follow-up exploration.

Key Points

- We propose a multi contact-based folding method under the evolutionary algorithm framework (MultiCFold), which is a valuable attempt to directly use the thorough information of different contact maps by populations to guide protein structure folding.
- We design a contact-based model using contact and noncontact information and a multi contact-based population optimization to fully exploit the information of every input contact map and reduce its noise. Both attempts effectively improve the accuracy of the prediction model.
- The results on the comparison of benchmark proteins suggest that our proposed protein folding algorithm significantly outperforms MetaCFold (meta contact-based method). On CASP 13 FM targets, MultiCFold is comparable with four state-of-the-art full-version methods.

Supplementary data

Supplementary data are available online at *Briefings in Bioinformatics*.

Data and code availability

All data needed to evaluate the conclusions are present in the paper and the Supplementary Materials. The additional data and code related to this paper can be downloaded from https://github.com/iobio-zjut/MultiCFold.

Funding

This work has been supported by the National Nature Science Foundation of China (grant numbers 62173304 and 61773346), the Key Project of Zhejiang Provincial Natural Science Foundation of China (grant number LZ20F030002) and the National Key Research and Development Program of China (grant number 2019YFE0126100).

References

1. Baker D, Sali A. Protein structure prediction and structural genomics. *Science* 2001;**294**:93.
2. Zhang Y. Progress and challenges in protein structure prediction. *Curr Opin Struct Biol* 2008;**18**:342–8.
3. Li Y, Zhang C, Bell EW, *et al.* Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks. *PLoS Comput Biol* 2021;**17**:e1008865.
4. Abriata LA, Tamò GE, Monastyrskyy B, *et al.* Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. *Proteins* 2018;**86**: 97–112.
5. Schaarschmidt J, Monastyrskyy B, Kryshtafovych A, *et al.* Assessment of contact predictions in CASP12: coevolution and deep learning coming of age. *Proteins* 2018;**86**: 51–66.
6. Shrestha R, Fajardo E, Gil N, *et al.* Assessing the accuracy of contact predictions in CASP13. *Proteins* 2019;**87**:1058–68.
7. Zheng W, Li Y, Zhang C, *et al.* Deep-learning contact-map guided protein structure prediction in CASP13. *Proteins* 2019;**87**: 1149–64.
8. Ma J, Wang S, Wang Z, *et al.* Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. *Bioinformatics* 2015;**31**:3506–13.
9. Wang S, Sun S, Li Z, *et al.* Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput Biol* 2017;**13**:e1005324.
10. Kandathil SM, Greener JG, Jones DT. Prediction of interresidue contacts with DeepMetaPSICOV in CASP13. *Proteins* 2019;**87**: 1092–9.
11. Hanson J, Paliwal K, Litfin T, *et al.* Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. *Bioinformatics* 2018;**34**: 4039–45.
12. Yang J, Anishchenko I, Park H, *et al.* Improved protein structure prediction using predicted interresidue orientations. *Proc Natl Acad Sci* 2020;**117**:1496–503.
13. Adhikari B, Hou J, Cheng J. DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics* 2018;**34**:1466–72.
14. Mao W, Ding W, Xing Y, *et al.* AmoebaContact and GDFold as a pipeline for rapid de novo protein structure prediction. *Nat Mach Intell* 2020;**2**:25–33.
15. AlQuraishi M. AlphaFold at CASP13. *Bioinformatics* 2019;**35**: 4862–5.
16. Zhang C, Zheng W, Mortuza SM, *et al.* DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics* 2020;**36**:2105–12.

17. He K, Zhang S, Ren S and Sun J. Deep residual learning for Image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016; pp. 770–8.
18. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;**9**:1735–80.
19. Schuster M, Paliwal KK. Bidirectional recurrent neural networks. *IEEE Trans Signal Process* 1997;**45**:2673–81.
20. Peng C, Zhou X, Zhang G. De novo protein structure prediction by coupling contact with distance profile. *IEEE/ACM Trans Comput Biol Bioinform* 2020; doi: 10.1109/TCBB.2020.3000758.
21. Liu J, Zhou X-G, Zhang Y, *et al.* CGLFold: a contact-assisted de novo protein structure prediction using global exploration and loop perturbation sampling algorithm. *Bioinformatics* 2020;**36**: 2443–50.
22. Zhang GJ, Ma LF, Wang XQ, *et al.* Secondary structure and contact guided differential evolution for protein structure prediction. *IEEE/ACM Trans Comput Biol Bioinform* 2020;**17**: 1068–81.
23. Kosciolek T, Jones DT. De novo structure prediction of globular proteins aided by sequence variation-derived contacts. *PLoS One* 2014;**9**:e92197.
24. Li Y, Hu J, Zhang C, *et al.* ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. *Bioinformatics* 2019;**35**:4647–55.
25. He B, Mortuza SM, Wang Y, *et al.* NeBcon: protein contact map prediction using neural network training coupled with naïve Bayes classifiers. *Bioinformatics* 2017;**33**:2296–306.
26. Adhikari B, Bhattacharya D, Cao R, *et al.* CONFOLD: residue-residue contact-guided ab initio protein folding. *Proteins* 2015;**83**: 1436–49.
27. Adhikari B, Cheng J. CONFOLD2: improved contact-driven ab initio protein structure modeling. *BMC Bioinformatics* 2018;**19**: 1–5.
28. Marks DS, Colwell LJ, Sheridan R, *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS One* 2011;**6**:e28766.
29. Brünger AT, Adams PD, Clore GM, *et al.* Crystallography & NMR system: a new software suite for macromolecular structure determination, Acta crystallographica. Section D. *Biol Crystallogr* 1998;**54**:905–21.
30. Jones DT, Singh T, Kosciolek T, *et al.* MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics* 2015;**31**: 999–1006.
31. Jones DT, Buchan DWA, Cozzetto D, *et al.* PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 2012;**28**:184–90.
32. Seemayer S, Gruber M, Söding J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics* 2014;**30**:3128–30.
33. Kaján L, Hopf TA, Kalaš M, *et al.* FreeContact: fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinformatics* 2014;**15**:85.
34. Cheng J, Baldi P. Three-stage prediction of protein β-sheets by neural networks, alignments and graph algorithms. *Bioinformatics* 2005;**21**:i75–84.
35. Cheng J, Baldi P. Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics* 2007;**8**:113.
36. Wu S, Zhang Y. A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics* 2008;**24**:924–31.
37. Yang J, Shen HB. An ensemble predictor by fusing multiple base predictors composed by both coevolution-based and machine learning-based approaches. *Abstract of CASP11 Experiment*, 2014, 209.
38. Jones DT, Kandathil SM. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics* 2018;**34**: 3308–15.
39. Konopka BM, Ciombor M, Kurczynska M, *et al.* Automated procedure for contact-map-based protein structure reconstruction. *J Membr Biol* 2014;**247**:409–20.
40. Zhou X-G, Peng C-X, Liu J, *et al.* Underestimation-assisted global-local cooperative differential evolution and the application to protein structure prediction. *IEEE Trans Evol Comput* 2019;**24**: 536–50.
41. Zhou X-G, Zhang G-J. Abstract convex underestimation assisted multistage differential evolution. *IEEE Trans Cybern* 2017;**47**: 2730–41.
42. Zhou X-G, Zhang G-J. Differential evolution with underestimation-based multimutation strategy. *IEEE Trans Cybern* 2018; **49**:1353–64.
43. Zhou X, Hu J, Zhang C, *et al.* Assembling multidomain protein structures through analogous global structural alignments. *Proc Natl Acad Sci* 2019;**116**:15930–8.
44. Rohl CA, Strauss C, Misura K, *et al.* Protein structure prediction using Rosetta. *Methods in Enzymology* 2004;**383**:66–93.
45. Fox NK, Brenner SE, Chandonia J-M. SCOPe: structural classification of proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res* 2014;**42**:D304–9.
46. Chandonia J-M, Fox NK, Brenner SE. SCOPe: classification of large macromolecular structures in the structural classification of proteins—extended database. *Nucleic Acids Res* 2019;**47**: D475–81.
47. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;**22**:1658–9.
48. Huang Y, Niu B, Gao Y, *et al.* CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 2010;**26**:680–2.
49. Xia Y, Peng C, Zhou X, *et al.* A sequential niche multimodal conformational sampling algorithm for protein structure prediction. *Bioinformatics*, 2021; doi: 10.1093/bioinformatics/btab500.
50. Zhao K, Liu J, Zhou X, *et al.* MMpred: a distance-assisted multimodal conformation sampling for de novo protein structure prediction. *Bioinformatics* 2021; doi: 10.1093/bioinformatics/btab484.
51. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins* 2004;**57**: 702–10.
52. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score= 0.5? *Bioinformatics* 2010;**26**:889–95.
53. Xu J, Wang S. Analysis of distance-based protein structure prediction by deep learning in CASP13. *Proteins* 2019;**87**: 1069–81.
54. Hou J, Wu T, Cao R, *et al.* Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. *Proteins* 2019;**87**:1165–78.