Enhancing Protein Conformational Space Sampling Using Distance Profile-Guided Differential Evolution

Gui-Jun Zhang, Xiao-Gen Zhou, Xu-Feng Yu, Xiao-Hu Hao, and Li Yu

Abstract—De novo protein structure prediction aims to search for low-energy conformations as it follows the thermodynamics hypothesis that places native conformations at the global minimum of the protein energy surface. However, the native conformation is not necessarily located in the lowest-energy regions owing to the inaccuracies of the energy model. This study presents a differential evolution algorithm using distance profile-based selection strategy to sample conformations with reasonable structure effectively. In the proposed algorithm, besides energy, the residue-residue distance is considered another measure of the conformation. The average distance errors of decoys between the distance of each residue pair and the corresponding distance in the distance profiles are first calculated when the trial conformation yields a larger energy value than that of the target. Then, the distance acceptance probability of the trial conformation is designed based on distance profiles if the trial conformation obtains a lower average distance error compared with that of the target conformation. The trial conformation is accepted to the next generation in accordance with its distance acceptance probability. By using the dual constraints of energy and distance in guiding sampling, the algorithm can sample conformations with lower energies and more reasonable structures. Experimental results of 28 benchmark proteins show that the proposed algorithm can effectively predict near-native protein structures.

Index Terms—Protein structure prediction, de novo, distance profile, differential evolution, distance error, fragment assembly

1 Introduction

In biological cells, there are a number of proteins which are formed by union of more than 20 kinds of amino acids. These macromolecules play an important role in living organisms and are vital to the completion of biological functions. Protein molecules reflect the significant relationship between structure and function at the molecular level. Different proteins perform different functions in the living body. The function of a protein is determined to a great extent by its spatial structure. In particular, the three-dimensional (3D) structure (native structure) of proteins is key to understanding and transforming biological and cellular functions. Predicting the 3D structure of the protein, therefore, is essential in engineering novel proteins [1], designing drugs [2], predicting protein stability [3], and modeling protein-protein interactions [4], [5].

The 3D structures of proteins can be experimentally predicted by X-ray crystallography and nuclear magnetic methods. However, these classical experimental methods are usually time-consuming and costly [6]. De novo methods, which predict the 3D structures of proteins by only using amino acid sequences, have been successfully used for numerous proteins [7], [8]. In de novo methods, an energy function is used to evaluate conformations, and an algorithm is employed to search through conformations. Based on the thermodynamics hypothesis [9], the energy function guides the search toward low-energy conformations of the query sequence because lower-energy conformations are expected to be more native-like than higher-energy ones. In other words, de novo methods involve the optimization of an energy function. Therefore, one important problem in de novo protein structure prediction is finding the global minimum of energy in the conformational space.

Numerous approaches have been proposed to sampling low-energy conformations in the conformational space. These techniques include evolutionary algorithms (EAs) [10], [11], [12], [13], [14], Monte Carlo (MC) [15], [16], [17], Molecular Dynamics (MD) [18], [19], and conformational space annealing (CSA) [20], [21], [22]. However, as the amino acid sequence increases, the degrees of freedom (dofs) of protein molecules also increase, which makes the protein structure prediction at the full-atom level a challenging problem. In order to explore the vast conformational space, coarse-grained energy models such as SICHO [23] and UNRES [24] which reduce the number of dofs, are widely used in protein structure prediction. Unfortunately, these coarse-grained energy models cannot sufficiently capture the long-range interactions between residues. While these interactions provide important information on the protein structure that is vital to prediction accuracy. Owing to the inaccuracies inherent in coarse-grained energy models, some conformations with higher energy but more reasonable structure cannot survive during the sampling, thereby affecting the prediction accuracy.

- *The authors are with the College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China.*
  E-mail: {zgj, zxg, yuxf, xhhao, lyu}@zjut.edu.cn.

Manuscript received 25 Jan. 2016; revised 5 May 2016; accepted 6 May 2016. Date of publication 11 May 2016; date of current version 6 Dec. 2017.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2016.2566617

In order to alleviate the impact of the inaccuracies of the energy function on prediction accuracy, Baker's research group [25], [26], [27] use the residue-residue distance, dihedral angle and other local geometry data of the target protein which obtained by the experimental method to guide the conformation search in the protein structure prediction. Experimental results indicated that the technique can improve prediction accuracy. Inspired by the concept of residue-residue potentials proposed by Miyazawa and Jernigan [28], various methods are presented to predict the residue-residue distance and are adopted in the protein structure prediction [29], [30], [31]. Zhang's research group [32], [33] propose a distance profile consisting of a histogram distribution of pairwise distances extracted from unrelated experimental structures based on the occurrence of fragments at different positions but from the same templates. The distance profile was added to the QUARK [32] energy model as an energy term.

Differential evolution (DE), proposed by Storn and Price [34], has been proven to be a simple yet effective stochastic global optimization algorithm in EAs [35]. Inspired by the mechanisms of natural evolution and survival of the fittest, DE guides the optimization search process based on the swarm intelligence generated by the cooperation and competition of individuals in a population. In DE, new candidate solutions are generated by combining the target solution and the corresponding mutant which are created by randomly distributing all solutions in the current population. A candidate solution replaces its parent only if it is better. DE has demonstrated superior robustness and fast convergence in solving single objective global optimization problems [34], [36]. Owing to DE's simple structure, ease of use, robustness and speed, DE has been successfully used in protein structure prediction [12], [37], [38].

In this study, we propose a distance profile-guided differential evolution (DPDE) for protein conformational space sampling. In DPDE, a distance profile-based selection strategy is designed to guide conformational space sampling. In the distance profile-based selection strategy, besides energy, the residue-residue distance is considered as an auxiliary conformation measure to compensate for the inaccuracies of the energy function, and a distance acceptance probability is designed based on the distance profiles for selecting the conformations. When the energy of the trial conformation is lower than that of the target conformation, the trial conformation is directly accepted for the next generation. Otherwise, the average distance error between the actual residue-residue distance and the predicted residue-residue distance obtained from the distance profiles of all residue pairs in the conformation is first calculated. If the trial conformation is better than the target conformation in terms of the average distance error, the distance acceptance probability of the trial conformation is calculated, and the trial conformation is accepted according to the distance acceptance probability. Thus, conformations with higher energies but more reasonable structures are retained. By using the distance profile-based selection strategy in guiding sampling, the capability of escaping the local minimum and search efficiency are improved. In addition, the algorithm converges to the region with lower energy and more reasonable structure. Experimental results show that the proposed DPDE can efficiently sample more near-native protein conformations.

The remainder of this paper is organized as follows. Section 2 briefly describes the knowledge-based coarse-grained energy model, fragment assembly, and distance profiles. In Section 3, the proposed algorithm are described at a finer level of detail. Experimental results demonstrating the performance of the proposed algorithm compared with other methods are presented in Section 4. Section 5 concludes this paper and points out some future works.

2 Preliminary

2.1 Knowledge-Based Coarse-Grained Energy Model

Protein structure prediction involves an extremely expensive to evaluate energy model which has thousands of dofs, and a highly degenerate energy hyperspace owing to its massive local minima and large regions of unfeasible conformations. To explore the vast conformational space, a coarse-grained model is usually employed to reduce the dofs, without losing important information of the amino acid sequence. The Rosetta score3 knowledge-based energy function [39], which includes all possible energy terms except for hydrogen bonding is employed in the proposed DPDE. The Rosetta score3 knowledge-based energy function is defined as follows:

$$\begin{aligned} E = {} & W_{\text{repullsion}} E_{\text{repullsion}} + W_{\text{attraction}} E_{\text{attraction}} \\ & + W_{\text{slovation}} E_{\text{slovation}} + W_{\text{bb-schb}} E_{\text{bb-schb}} \\ & + W_{\text{bb-bbhb}} E_{\text{bb-bbhb}} + W_{\text{sc-schb}} E_{\text{sc-schb}} \\ & + W_{\text{pair}} E_{\text{pair}} + W_{\text{dunbrack}} E_{\text{dunbrack}} \\ & + W_{\text{rama}} E_{\text{rama}} + W_{\text{reference}} E_{\text{reference.}} \end{aligned} \tag{1}$$

A detailed description of energy items and their corresponding parameter settings in Eq. (1) can be found in [39].

2.2 Fragment Assembly

Fragment assembly techniques currently present the state-of-the-art for de novo prediction [40], [41], [42]. In fragment assembly techniques for protein structure prediction, the query sequence is divided into a number of continuous segments. The fragments, which are locally similar to known protein structures, are then selected from the fragment library by sequence alignment method. The selected fragments are used to replace the specific fragments of the query sequence to assemble the target structure. Fragment assembly techniques often proceed in three stages: the insert position is first selected; then the corresponding fragment is selected from the fragment library at the second stage; the insert length is finally determined from the selected fragments, and the corresponding fragment of the query sequence is replaced. By using fragments of the known protein structures in the Protein Data Bank (PDB) [43] to construct the target conformation, the local propensities of the amino-acid chain are captured. Therefore, not only is the search conformational space significantly reduced, the computation speed is also remarkably improved. The construction of the fragment database used in this study is the same as [33]. However, as the Rosetta score3 knowledge-based energy function is employed in our proposed algorithm, and the fragment length of 3 or 9 residues is used in Rosetta. Therefore, to compare with Rosetta, we chose the fragment length of 9 residues in this study.

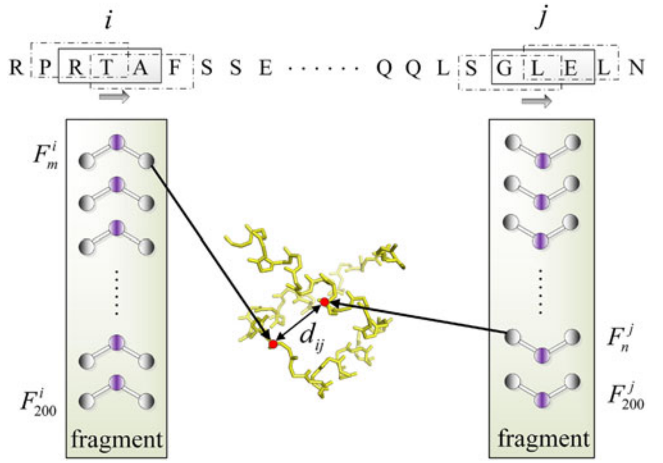


Fig. 1. The distance d_{ij} between two fragments from different residues (i and j) in the query sequence.

2.3 Distance Profile

The distance profile, proposed by Zhang's research group [32], [33], is a histogram distribution of pair-wise distances extracted from unrelated experimental structures based on the occurrence of fragments at different positions but from the same templates. A distance profile reflects the long-range interactions between each pair of residues.

Consider the ith and jth residue in the query sequence shown in Fig. 1, where $F^i_m(m = 1, \ldots, 200)$ and $F^j_n(n = 1, \ldots, 200)$ are the top 200 fragments of the positions i and j generated based on a scoring function in [33], respectively. d_{ij} represents the central distance between the fragments F^i_m and F^j_n. After calculating the distance between each pair of fragments from the same template, the number of the distances below 9 Å are counted to depict the line graph with the 0.5 Å distance bin, and this graph is called a distance profile. In the distance profile shown in Fig. 2, the distance $d_{ij} = 6$ Å corresponding to the peak of the line graph indicates that the distance between the ith and the jth residue has a high probability of being 6 Å. Suppose the above distance profile is the kth one, then the predicted distance between the ith and the jth residue in the distance profile can be denoted as $D^{k(ij)}_{\text{profile}} = 6$ Å. $k(ij)$ means the kth distance profile is built based on the distance between the ith and the jth residue.

3 Distance Profile-Guided DE Algorithm

In order to reduce the prediction error caused by the inaccuracies of the energy function, an improved DE using distance profile-based selection strategy referred to as DPDE is proposed for protein conformational space sampling in this study. In the distance profile-based selection strategy, the knowledge-based coarse-grained energy function Rosetta score3 [39], [44] and the residue-residue distance are used to measure the conformations. If the trial conformation obtains a lower value than that of the target conformation of the energy function, the target conformation is replaced by the trial conformation. Otherwise, the average distance error of the trial conformation and the target conformation are calculated based on the distance profile. If the trial conformation has a lower average distance error than that of the target conformation, the distance acceptance probability, which is based on the distance profile is first calculated. The trial conformation is then accepted to the next generation with the distance acceptance probability. With the distance and energy guiding selection operation, conformations with high energies but reasonable structures are retained, and it allows the algorithm to escape a local minimum. As the energy is the main measure of the conformation, and the residue-residue distance is the auxiliary measure, the algorithm can converge to a region with lower energy and more reasonable structure. Thus, the proposed DPDE can obtain the native-like conformation of the query sequence.

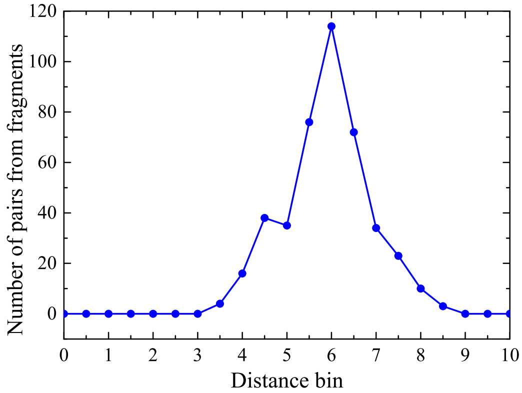


Fig. 2. A distance profile for a residue pair.

3.1 Average Distance Error

The distance profile reflects the statistical distribution of the distance between each residue pair after the amino acid folds into the 3D structure. We can construct multiple distance profiles by statistic analysis of all residue pairs of an amino acid sequence. Every distance profile reflects only the local structural feature between the two residues, and the 3D structure of the protein is a complex structure with numerous dofs; thus, a single distance profile cannot significantly guide the prediction. However, the global structural feature, which is made up of the local structural features of multiple distance profiles, presents a certain guiding significance in 3D predictions of the protein. Assume that the prediction distance between the residue pairs in the distance profiles is sufficiently accurate, we can infer the structural similarity of the protein by the difference between the residue-residue distances and the corresponding predicted distances in the distance profiles. Thus, the average distance error between the residue-residue distance and its corresponding predicted distance in the distance profiles for all residue pairs can be used to guide the conformational space sampling. The average distance error of a decoy conformation is calculated as follows:

$$\Delta D_{\text{decoy}} = \frac{1}{N}\sum_{k=1}^{N}\left|D^{k(ij)}_{\text{decoy}} - D^{k(ij)}_{\text{profile}}\right|, \tag{2}$$

where k is the index of the distance profile, and N is the size of the distance profile. $D^{k(ij)}_{\text{profile}}$ represents the predicted distance between the ith and jth residue in the kth distance profile, and $D^{k(ij)}_{\text{decoy}}$ is the corresponding Euclidean distance between the ith and jth residue in the decoy conformation. The smaller the average distance error ΔD_{decoy}, the larger the probability that the decoy conformation is close to the native structure.

Fragment assembly may cause a significant increase in the distance between certain residues in the newly generated decoy conformation, and these structures are often unreasonable. In order to obtain a reasonable average distance error to guide the algorithm converging to a more reasonable region, a cutoff value D_{cut}=6.5 Å is set according to [28]. If the error between the Euclidean distance and the predicted distance in the distance profiles is larger than D_{cut}, then truncated it to 6.5 Å.

3.2 Distance Profile-Based Selection Strategy

De novo methods are based on the hypothesis that the conformation with the lowest energy is considered the native structure. However, due to the inaccuracy of the energy model, the conformation associated with the global minimum of the energy function may not be the one closest to the native state. Sampling the conformation only according to the energy model is the key factor to restrict the accuracy of protein structure prediction. In our proposed distance profile-based selection strategy, the average distance error is also used to guide sampling. For each decoy conformation, it has its own distance acceptance probability calculated based on distance profiles. The distance acceptance probability of a decoy conformation is defined as follows:

$$P_{\text{accept}} = \frac{1}{N}\sum_{k=1}^{N}\exp\left(\frac{-\left|D_{\text{decoy}}^{k(ij)} - D_{\text{profile}}^{k(ij)}\right|}{\beta}\right), \tag{3}$$

where β is a temperature scaling factor. The other parameters in this equation are identical to those in Eq. (2). In this study, β is set to 4.

In the proposed distance profile-based selection strategy, the energy is the main measure of conformation, and the average distance error is the auxiliary measure. When a trial conformation yields a lower value than the target conformation of the energy function, it is directly accepted as a replacement for its corresponding target conformation. Otherwise, the average distance errors are first calculated according to Eq. (2). If the trial conformation has a lower average distance error than that of the target conformation, its distance acceptance probability is calculated according to Eq. (3), and the trial conformation is accepted according to its distance acceptance probability. By using the distance profile-based selection strategy, conformations with higher energies but more reasonable structures are retained. Thus, the algorithm can converge to the more promising region.

3.3 DPDE Algorithm Description

As shown in the flowchart in Fig. 3, the proposed DPDE algorithm is described as follows:

1. *Initialization*. Set the value of population size NP, crossover rate CR, and maximum generations G. Randomly initialize the conformation population P= $\{\mathrm{C}_1, \ldots, \mathrm{C}_{NP}\}$. Set the index counter $i = 1$ and the generation counter $g = 1$.

2. *Mutation*. Set the ith conformation C_i in the population of the gth generation to the target conformation $\mathrm{C}_{\text{target}}$, and perform the following operations:

2.1. Randomly select three conformations C_{rand1}, C_{rand2}, and C_{rand3} from the entire population, where $rand1 \neq rand2 \neq rand3 \in [1, NP]$.

2.2. Randomly select one fragment from C_{rand1} and C_{rand2} and replace the corresponding position in C_{rand3} to generate a mutant conformation $\mathrm{C}_{\text{mutant}}$ according to the fragment assembly.

3. *Crossover*. Perform the following operations to generate a trial conformation:

3.1. Generate a uniformly distributed random number $rand4$ between 0 and 1. If $rand4 > CR$, go to *3.2*; otherwise, go to *3.3*.

3.2. Take the mutant conformation $\mathrm{C}_{\text{mutant}}$ as a trial conformation $\mathrm{C}_{\text{trial}}$.

3.3. Randomly select one fragment from the target conformation $\mathrm{C}_{\text{target}}$, and replace the corresponding position in $\mathrm{C}_{\text{mutant}}$ to generate a trial conformation $\mathrm{C}_{\text{trial}}$.

4. *Selection*. Perform the proposed distance profile-based selection strategy to select the trial conformation:

4.1. Calculate the energy $E(\mathrm{C}_{\text{trial}})$ and $E(\mathrm{C}_{\text{target}})$ of $\mathrm{C}_{\text{trial}}$ and $\mathrm{C}_{\text{target}}$, respectively, according to Eq. (1). If $E(\mathrm{C}_{\text{trial}}) < E(\mathrm{C}_{\text{target}})$ then replace $\mathrm{C}_{\text{target}}$ by $\mathrm{C}_{\text{trial}}$ and go to *Step 5;* otherwise go to *4.2*.

4.2. Calculate the average distance errors of the decoy conformation ($\mathrm{C}_{\text{trial}}$ and $\mathrm{C}_{\text{target}}$) according to the following steps:

4.2.1. Set the index of the distance profile $k = 1$.

4.2.2. Calculate the distance error ΔD_k between the Euclidean distance D_{decoy}^{k} of the residue pair in the kth distance profile and the corresponding predicted distance D_{profile}^{k} in the distance profiles.

4.2.3. If $\Delta D_k > 6.5$ then truncate the value to 6.5 Å; otherwise retain ΔD_k.

4.2.4. If k is equal to the size of the distance profile N then calculate the average distance error ΔD_{decoy} of $\Delta D_k, k = 1, \ldots, N$ and go to *4.3*. Otherwise $k = k + 1$ and go to *4.2.2*.

4.3. If $\Delta D_{\text{trial}} < \Delta D_{\text{target}}$ then go to *4.4*; otherwise retain $\mathrm{C}_{\text{target}}$ and go to *Step 5*.

4.4. Calculate the distance acceptance probability P_{accept} of $\mathrm{C}_{\text{trial}}$ according to the following steps:

4.4.1. Set the index of the distance profile $k = 1$.

4.4.2. Calculate $P_k = \exp(-\Delta D_k/\beta)$ of the kth distance profile, where β is the temperature scaling factor.

4.4.3. If k is equal to the size of the distance profile N then calculate the average distance error P_{accept} of $P_k, k = 1, \cdots, N$ and go to *4.5*; otherwise $k = k + 1$ and go to *4.4.2*.

4.5. Generate a uniformly distributed random number $rand5$ between 0 and 1. If $rand5 < P_{\text{accept}}$ then replace $\mathrm{C}_{\text{target}}$ by $\mathrm{C}_{\text{trial}}$. Otherwise retain $\mathrm{C}_{\text{target}}$.

5. *Iteration*.Perform the following steps to check iteration:

5.1. If i is equal to the population size NP then go to *5.2*; otherwise $i = i + 1$ and go to *Step 2*.

5.2. If g is equal to the maximum generation G then output the result and stop the iteration; otherwise $g = g + 1$ and go to *Step 2*.

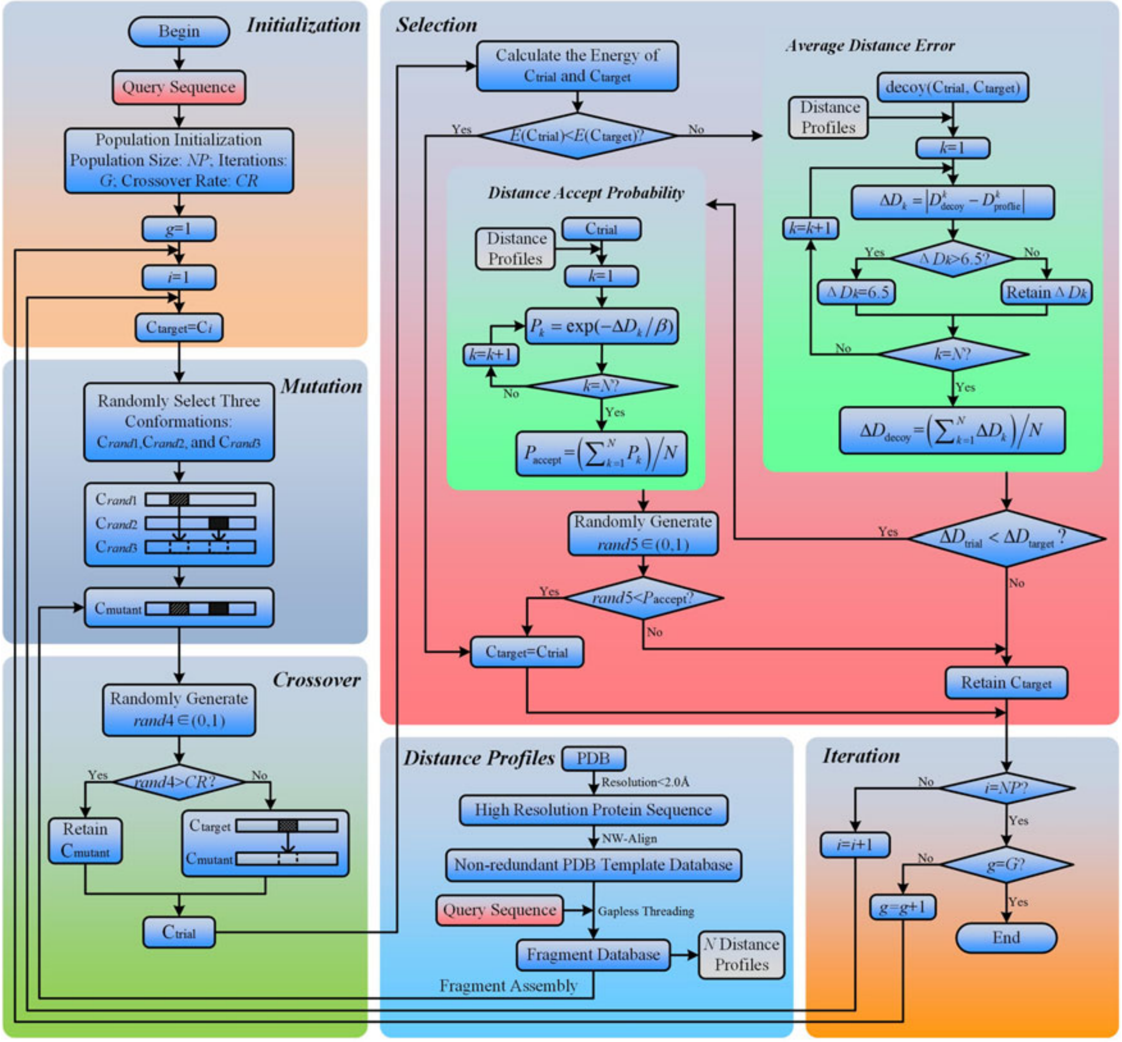

Initialization
Begin
Query Sequence
Population Initialization Population Size: NP; Iterations: G; Crossover Rate: CR
g=1
i=1
Ctarget=Ci
Mutation
Randomly Select Three Conformations: Crand1,Crand2, and Crand3
Crand1
Crand2
Crand3
Cmutant
Crossover
Randomly Generate rand4∈(0,1)
rand4>CR?
Yes
No
Retain Cmutant
Ctarget
Cmutant
Ctrial
Selection
Calculate the Energy of Ctrial and Ctarget
E(Ctrial)<E(Ctarget)?
Yes
No
Distance Accept Probability
Distance Profiles
Ctrial
k=1
$P_k = \exp(-\Delta D_k/\beta)$
k=k+1
k=N?
No
Yes
$P_{accept} = \left(\sum_{k=1}^{N} P_k\right)/N$
Randomly Generate rand5∈(0,1)
rand5<Paccept?
Yes
No
Ctarget=Ctrial
Average Distance Error
decoy(Ctrial, Ctarget)
Distance Profiles
k=1
$\Delta D_k = \left|D_{decoy}^{k} - D_{profile}^{k}\right|$
k=k+1
ΔDk>6.5?
Yes
No
ΔDk=6.5
Retain ΔDk
k=N?
No
Yes
$\Delta D_{decoy} = \left(\sum_{k=1}^{N} \Delta D_k\right)/N$
$\Delta D_{trial} < \Delta D_{target}$?
Yes
No
Retain Ctarget
Distance Profiles
PDB
Resolution<2.0Å
High Resolution Protein Sequence
NW-Align
Non-redundant PDB Template Database
Query Sequence
Gapless Threading
Fragment Database
N Distance Profiles
Fragment Assembly
Iteration
i=NP?
No
Yes
i=i+1
g=G?
No
Yes
g=g+1
End

Fig. 3. The flowchart of the DPDE.

The distance profile used in *4.2* and *4.4* are constructed according to [33]. Briefly, select proteins with a resolution higher than 2.0 Å from the protein-structure files obtained from the PDB website first. Then, remove homologous chains with sequence identities larger than 30 percent through NW-align [33] to construct a non-redundant template database. Subsequently, a fragment database that includes the top 200 fragments with highest scores of a scoring function in each position is constructed by a gapless-threading method presented in [33]. Finally, the number of distances <9 Å between each fragment pair from the same template is counted to construct the distance profiles.

4 Experimental Studies

In this section, various experiments on benchmark proteins are implemented to verify the performance of the proposed DPDE algorithm.

4.1 Materials and Experiment Settings

A total of 28 benchmark proteins with different sizes and folding types listed in Table 1 are used in the following experiments. Column 2 shows the PDB IDs of the proteins, column 3 shows their size, and column 4 shows their folding type. These benchmark proteins range from 36-146 amino acids in size. Proteins 1-13 are α folding, proteins 14-23 are α/β folding, and proteins 24-28 are β folding. These benchmark proteins all have experimentally-determined native structures deposited in the PDB and so allow verifying the effectiveness of the proposed DPDE. Moreover, many of these proteins have often been studied by other conformational search algorithms [10], [11], [13], [45], [46], thus allowing a direct comparison of DPDE with results published by other groups.

In the following experiments, homologous fragments for each target sequence are removed from the template database through the procedure described in [33]. The following parameters for DPDE are used unless a change is mentioned: population size $NP = 50$, crossover rate $CR = 0.5$, temperature scaling factor $\beta = 4$, fragment length $L = 9$. For each algorithm and for each benchmark proteins, 30 independent runs were conducted with 1,500,000 energy function evaluations as the stopping rule. To ensure fair comparison with Rosetta, we launched 50 MMC trajectories at one time in Rosetta, and every MMC trajectory is conducted with 30,000 iterations as the termination criterion. The proposed DPDE was implemented in the Python language, and tests were executed in a Windows Sever 2008

TABLE 1
Results of RMSD Obtained by DPDE and Rosetta for All Benchmark Proteins

No.	PDB ID	Size	Type	Minimum lowest RMSD(Å)		Mean ± Std lowest RMSD(Å)		p-value	Significance
				DPDE	Rosetta	DPDE	Rosetta		
1	1VII	36	α	2.08	**1.58**	2.37 ± 0.22	**1.75 ± 0.15**	0.0020	−
2	2IMU	46	α	**2.33**	4.23	**3.30 ± 0.38**	4.91 ± 0.36	0.0020	+
3	1ENH	54	α	**1.32**	1.37	**1.87 ± 0.18**	1.88 ± 0.31	0.7500	≈
4	1BBO	57	α	**3.94**	4.56	**4.58 ± 0.29**	5.44 ± 0.44	0.0059	+
5	1GYZ	60	α	1.98	**1.85**	2.31 ± 0.27	2.11 ± 0.13	0.0566	≈
6	1ISUa	62	α	**5.73**	6.81	**6.87 ± 0.50**	7.46 ± 0.37	0.0273	+
7	2MQK	65	α	2.43	**1.99**	3.06 ± 0.30	**2.46** ± 0.39	0.0059	−
8	1AIL	73	α	**1.58**	3.02	**3.01 ± 0.69**	4.51 ± 0.62	0.0020	+
9	4ICB	76	α	**2.50**	3.00	**2.92 ± 0.26**	3.46 ± 0.39	0.0020	+
10	1CC5	83	α	**4.70**	5.46	**5.43 ± 0.60**	6.26 ± 0.44	0.0039	+
11	4UEX	85	α	3.92	**3.50**	4.69 ± 0.45	**4.46 ± 0.68**	0.3750	≈
12	2EZK	99	α	3.67	**3.57**	4.54 ± 0.47	**4.03 ± 0.38**	0.1113	≈
13	3GWL	106	α	**3.68**	3.79	5.67 ± 1.00	**5.49 ± 0.93**	0.9219	≈
14	1FD4	41	α/β	**3.61**	4.81	**4.47 ± 0.41**	5.27 ± 0.32	0.0020	+
15	1GB1	56	α/β	**2.77**	4.10	**4.33 ± 0.61**	5.08 ± 0.53	0.0039	+
16	1DTDb	61	α/β	**5.14**	6.19	**5.70 ± 0.37**	6.79 ± 0.52	0.0020	+
17	1SAP	66	α/β	5.10	**4.57**	6.12 ± 0.38	**5.72 ± 0.58**	0.1934	≈
18	1HZ6a	67	α/β	**2.64**	2.75	**2.96 ± 0.29**	3.27 ± 0.30	0.0195	+
19	1DTJa	76	α/β	**2.16**	3.11	**4.12 ± 1.14**	5.16 ± 0.80	0.0645	≈
20	1AOY	78	α/β	**3.06**	4.07	**4.70 ± 0.65**	4.88 ± 0.53	0.9219	≈
21	1TIG	88	α/β	**2.82**	4.36	**3.95 ± 0.59**	5.65 ± 0.68	0.0020	+
22	1HHP	99	α/β	**8.90**	10.10	**9.32 ± 0.47**	10.84 ± 0.39	0.0020	+
23	2HG6	106	α/β	**8.48**	9.24	**9.11 ± 0.36**	9.94 ± 0.41	0.0039	+
24	2MIT	32	β	**2.67**	4.38	**3.39 ± 0.34**	4.84 ± 0.19	0.0020	+
25	1I6C	39	β	**2.82**	2.88	3.54 ± 0.34	**3.30 ± 0.24**	0.0371	−
26	1BQ9	53	β	**2.92**	4.32	**3.79 ± 0.47**	4.65 ± 0.30	0.0020	+
27	1WAPa	64	β	**6.08**	6.97	**6.70 ± 0.40**	7.40 ± 0.24	0.0020	+
28	1ALY	146	β	**11.60**	13.30	**12.58 ± 0.53**	14.66 ± 0.70	0.0020	+
	Overall results			1.32	1.37	4.83 ± 2.38	5.42 ± 2.79	0.0003	+

x86_64 environment of a Server with Intel Xeon CPU 2.4 GHz with 48 GB of RAM.

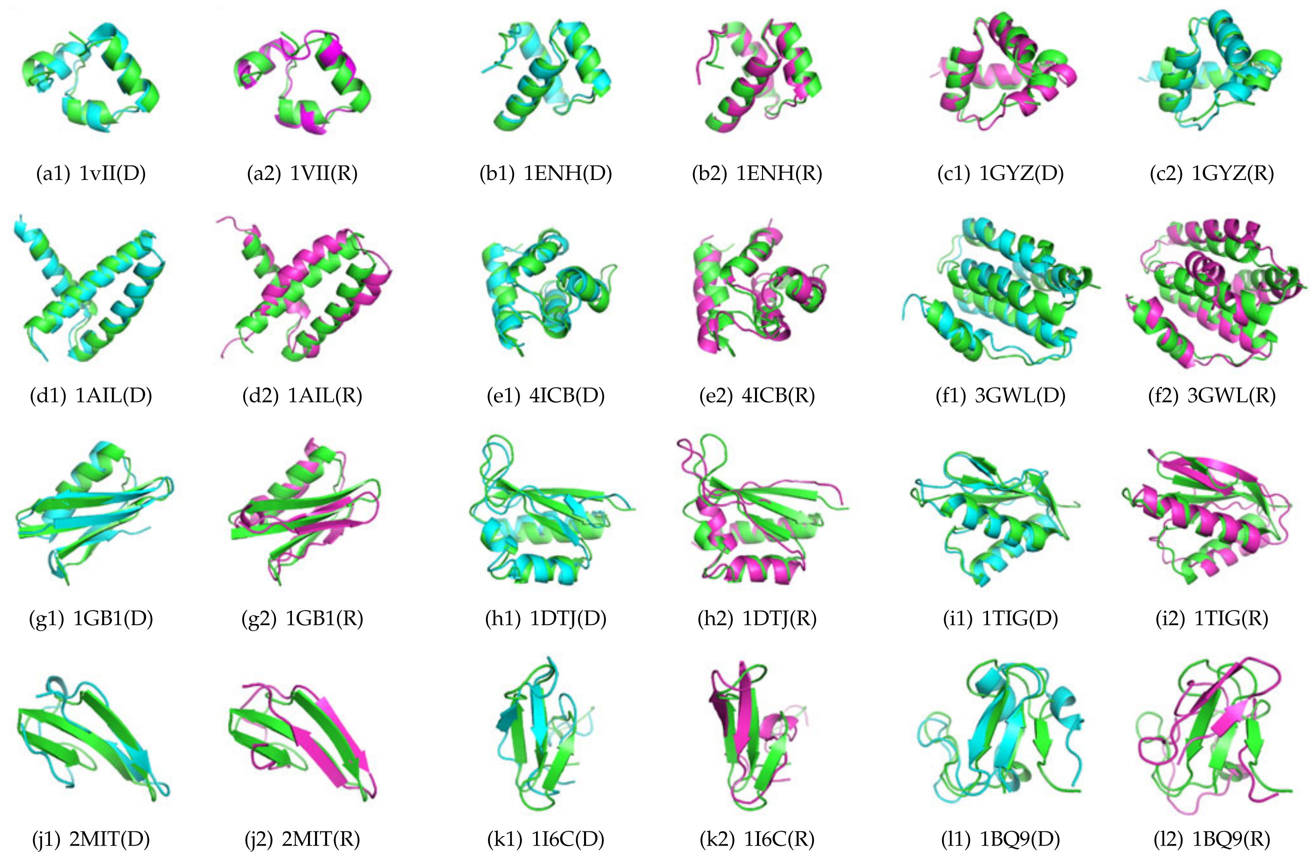
The main measure used in the analysis below is Root Mean Square Deviation (RMSD), which is the Euclidean distance between the corresponding atoms in the predicted conformation and the native structure after optimal rigid-body superimposition. In our experiments, RMSD is computed over only the C_α atoms.

4.2 Predicted Results Analysis

Table 1 summarizes the results of RMSD obtained by DPDE and Rosetta for all benchmark proteins, the best results are marked in *boldface*. The results listed in columns 5 and 6 indicated that the proposed DPDE achieves better results than Rosetta in 22 out of 28 proteins (with PDB IDs 2IMU, 1ENH, 1BBO, 1ISUa, 1AIL, 4ICB, 1CC5, 3GWL, 1FD4, 1GB1, 1DTDb, 1HZ6a, 1DTJa, 1AOY, 1TIG, 1HHP, 2HG6, 2MIT, 1I6C, 1BQ9, 1WAPa, and 1ALY) in terms of the minimum lowest RMSD. To be specific, DPDE outperforms Rosetta by 1.0 Å or more in 12 proteins, including 3 α folding proteins (with PDB IDs 2IMU, 1ISUa, and 1AIL), 6 α/β folding proteins (with PDB IDs 1FD4, 1GB1, 1DTDb, 1AOY, 1TIG, and 1HHP), and 3 β folding proteins (with PDB IDs 2MIT, 1BQ9, and 1ALY). For the four proteins with PDB IDs 2IMU, 1TIG, 2MIT, and 1ALY, DPDE is superior to Rosetta by more than 1.5 Å. Rosetta is better than DPDE in 6 proteins (with PDB IDs 1VII, 1GYZ, 2MQK, 4UEX, 2EZK, and 1SAP). The results also suggest that DPDE shows better performance than Rosetta in the majority of α and α/β folding proteins and all β folding proteins in terms of the minimum lowset RMSD. In addition, the overall results listed in the last row show that the lowest RMSD of the 28 benchmark proteins is 1.32 Å, while that of Rosetta is 1.37 Å.

A comparison of the 3D structure (with the minimum lowest RMSD) calculated by DPDE and Rosetta with the corresponding native structures for some selected benchmark proteins is given in Fig. 4. Comparisons between structures obtained by DPDE and native structures are marked with "D" at the back of the PDB ID, and comparisons between structures obtained by Rosetta with native structures are marked with "R". In this comparison, structures obtained by DPDE and Rosetta are depicted in blue and pink, respectively, and the native structures are depicted in green. As seen, DPDE obtains more accurate 3D structures than Rosetta for the majority of the benchmark proteins.

Columns 7 and 8 in Table 1 respectively present the mean of lowest RMSDs and the corresponding standard deviation ("Std") of DPDE and Rosetta. It is clear that the proposed DPDE always achieves better results than Rosetta for the majority of proteins. To be specific, DPDE performs better than Rosetta in 20 proteins including 7 α folding proteins (with PDB IDs 2IMU, 1ENH, 1BBO, 1ISUa, 1AIL, 4ICB, and 1CC5), 9 α/β folding proteins (with PDB IDs 1FD4, 1GB1, 1DTDb, 1HZ6a, 1DTJa, 1AOY, 1TIG, 1HHP, and 2HG6), and 4 β folding protein (with PDB IDs 2MIT, 1BQ9, 1WAPa, and 1ALY). For the other eight proteins, Rosetta shows


(a1) 1vII(D) (a2) 1VII(R) (b1) 1ENH(D) (b2) 1ENH(R) (c1) 1GYZ(D) (c2) 1GYZ(R)

(d1) 1AIL(D) (d2) 1AIL(R) (e1) 4ICB(D) (e2) 4ICB(R) (f1) 3GWL(D) (f2) 3GWL(R)

(g1) 1GB1(D) (g2) 1GB1(R) (h1) 1DTJ(D) (h2) 1DTJ(R) (i1) 1TIG(D) (i2) 1TIG(R)

(j1) 2MIT(D) (j2) 2MIT(R) (k1) 1I6C(D) (k2) 1I6C(R) (l1) 1BQ9(D) (l2) 1BQ9(R)

Fig. 4. Comparison of the structure obtained by DPDE (blue) and Rosseta (pink) with the native structure (green).

better performance than DPDE. In addition, DPDE outperforms Rosetta by at least 1.0 Å in 8 proteins (with PDB IDs 2IMU, 1AIL, 1DTDb, 1DTJa, 1TIG, 1HHP, 2MIT, and 1ALY). Rosetta is not superior to DPDE by more than 1.0 Å on any proteins. It is worth mentioning that DPDE outperforms Rosetta by 2.08 Å in the protein with PDB ID 1ALY. Moreover, the overall mean and std lowest RMSD of DPDE listed in the last row is also better than that of Rosetta.

In order to study the difference between the two algorithms in a more meaningful way, a nonparametric test named Wilcoxon Signed Rank Test [47] is also adopted in our study. The results are based on the lowest RMSDs obtained in each independent runs. As a null hypothesis, it is assumed that there is no significant difference between the best and/or mean values of two samples, whereas the alternative hypothesis is that there is significant difference in the best and/or mean values of the two samples, with 5 percent significance level. The p-value and the significance obtained according to the test are summarized in the columns 9 and 10 of Table 1, respectively. If the p-value is larger than 0.05, it indicates that there is significant difference between the two algorithms, otherwise there is no significant difference between them. Here, three signs $(+, -, \approx)$ are used to indicate the results, where the "+" means the first algorithm is significantly better than the second, the "−" means the first algorithm is significantly worse than the second, and the "≈" means that there is no significant difference between the two algorithms. As seen, DPDE is significantly better than Rosetta in 17 proteins (with PDB IDs 2IMU, 1BBO, 1ISUa, 1AIL, 4ICB, 1CC5, 1FD4, 1GB1, 1DTDb, 1HZ6a, 1TIG, 1HHP, 2HG6, 2MIT, 1BQ9, 1WAPa, and 1ALY), while Rosetta significantly outperforms DPDE only in 3 proteins (with PDB IDs 1VII, 2MQK, and 1I6C). For some other proteins, although DPDE is not significantly better than Rosetta, it performs better than Rosetta in these proteins. Furthermore, the overall results for all 28 benchmark proteins of DPDE is also significantly better than that of Rosetta, as DPDE obtains a p-value with 0.0003.

In addition, the performance profiles [48] are also depicted to demonstrate the superior performance of DPDE with respect to the RMSDs obtained over the 30 runs. Denoting the performance ration of the algorithm (a) for a protein (p) as follows:

$$r_{a,p} = \frac{\mathrm{RMSD}_{a,p}}{\min\{\mathrm{RMSD}_{a,p} : a \in A\}}, \tag{4}$$

where A is the set of comparison algorithms. Then the probability that the $r_{a,p}$ of the algorithm $a \in A$ is with a factor $\tau \geq 1$ of the best possible ratio can be calculated as follows:

$$\rho_a(\tau) = \frac{1}{n_p}|\{p \in P : r_{a,p} \leq \tau\}, \tag{5}$$

where P is the set of benchmark proteins. The $\rho_a(\tau)$ function is the distribution function of the $r_{a,p}$. Small values of τ and large values of $\rho_a(\tau)$ in the graph are preferred. That is, the first one that reaches the level of 1 with small values of τ is considered the best algorithm. The performance profiles of DPDE and Rosetta plotted in Fig. 5 indicates that DPDE

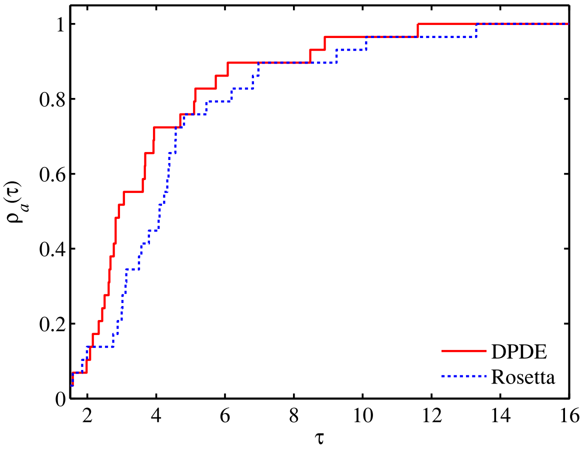


Fig. 5. Performance profiles of DPDE and Rosetta.

reaches the level of 1 first, thus DPDE is better than Rosetta with respect to the overall performance for the 28 proteins.

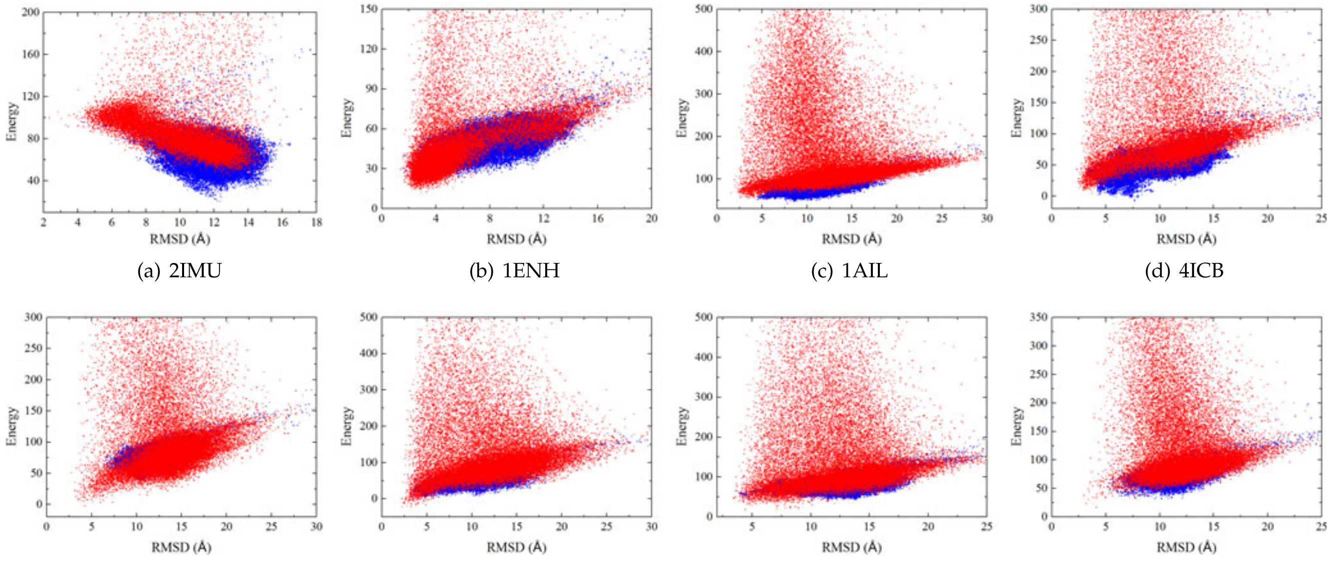
4.3 Convergence Analysis

Fig. 6 plots the energy of each conformation against its RMSD to the native structure on 8 representative benchmark proteins. The red dots are for the DPDE, and the blue dots are for the Rosetta. For proteins with PDB IDs 2IMU, 1AIL, 4ICB, and 1BQ9, DPDE is able to capture the conformations with lower RMSD than Rosetta, while Rosetta can reach much lower-energy levels than DPDE. This is expected, because Rosetta samples conformations with lower energies using only energy to measure the conformation. By contrast, DPDE tends to sample conformations with lower energies and more reasonable structures because it uses the energy and average distance errors of the residue pairs to measure the conformation. In particular, Rosetta converges to the region with low energy but high RMSD for protein with PDB ID 2IMU. This is because many low-RMSD conformations are not also populate the lowest-energy regions in the energy surface owing to the inaccuracy of the energy model. However, DPDE can capture low-RMSD conformations for the protein with PDB ID 2IMU because of its distance profile-based selection strategy. The minimum lowest RMSD for the protein with PDB ID 2IMU obtained by DPDE is 2.33 Å, whereas that obtained by Rosetta is 4.23 Å, and the mean lowest RMSD obtained by DPDE in the 30 runs is 3.3 Å, while that obtained by Rosetta is 4.91 Å. According to the k-means cluster for all conformations generated in the sampling, the RMSD of top model from DPDE cluster centroid is 8.08 Å, while that of Rosetta is 9.92 Å. Additionally, in the distance profile-based selection strategy, the energy is the main measure of conformation, and the average distance error of the residue pairs is merely the auxiliary measure. Therefore, DPDE can generate conformations with lower RMSDs and lower energies compared with Rosetta for proteins with PDB IDs 1ENH, 1FD4, 1HZ6a, and 1AOY. Fig. 6 also shows that the conformational ensemble generated by DPDE is more diverse than that generated by Rosetta, and DPDE contains more near-native conformations than Rosetta.

4.4 Near-Native Sampling Ability

In addition to the single lowest-RMSD conformation obtained from each algorithm, the distribution of decoys may be reasonably passed on to a second stage refinement in a complete de novo should be considered and compared.

To show the near-native sampling ability of DPDE, we employed the K-means cluster method for cluster analysis of all sampling conformations. In the K-means cluster, the RMSD is used to define the distance between points. The parameter K is set to 15, and the method is terminated when the maximum number of iterations reaches 30. The RMSDs of the top model in the top-populated cluster of DPDE and Rosetta for all 28 proteins are reported in Table 2. From the results, we can find that the RMSD of the DPDE cluster centroid is always lower than that of the Rosetta cluster centroid for the majority of benchmark proteins. DPDE outperforms Rosetta by 2.0 Å or more in 10 proteins



(a) 2IMU (b) 1ENH (c) 1AIL (d) 4ICB

(e) 1FD4 (f) 1HZ6a (g) 1AOY (h) 1BQ9

Fig. 6. The energy is plotted against RMSD to the native structure for each conformation on 9 representative proteins. The red and blue dots are for DPDE and Rosetta, respectively.

TABLE 2
Results of the Cluster Analysis for All Sampling Conformations

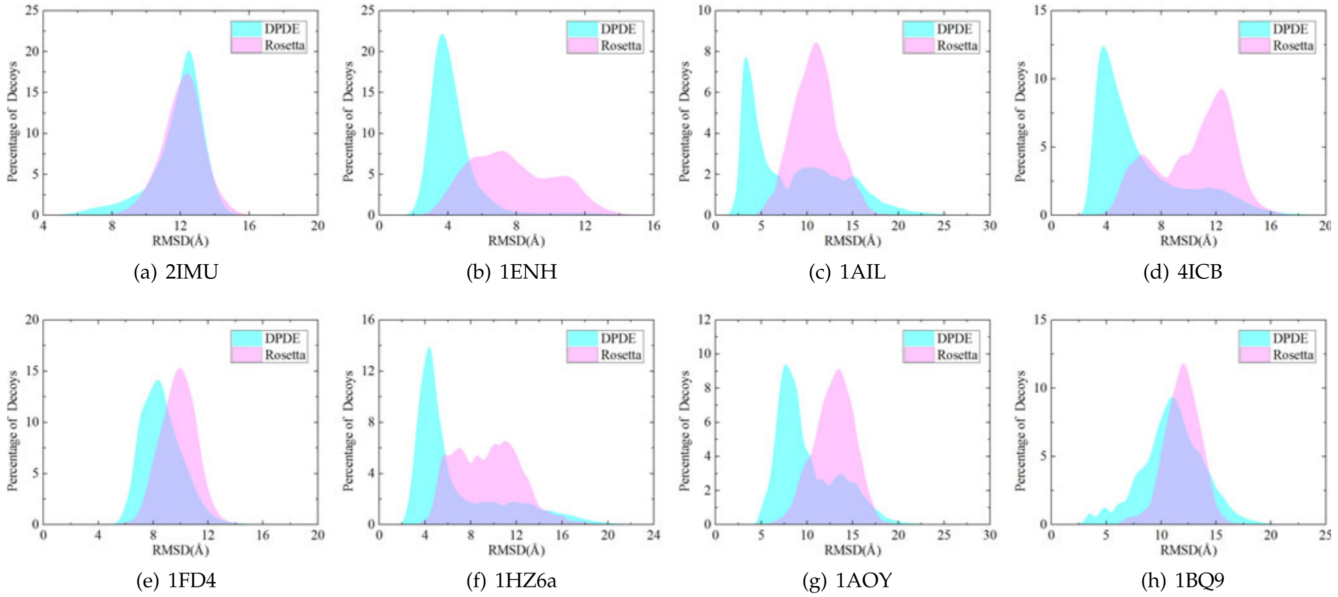
No.	PDB ID	Cluster RMSD(Å)		No.	PDB ID	Cluster RMSD(Å)	
		DPDE	Rosetta			DPDE	Rosetta
1	1VII	4.77	**4.08**	15	1GB1	**5.32**	6.75
2	2IMU	**8.08**	9.92	16	1DTDb	**10.84**	11.31
3	1ENH	**3.47**	5.16	17	1SAP	**11.07**	11.95
4	1BBO	**8.92**	10.83	18	1HZ6a	**3.71**	6.90
5	1GYZ	**3.40**	5.81	19	1DTJa	**10.28**	10.41
6	1ISUa	**9.62**	12.11	20	1AOY	**7.44**	11.86
7	2MQK	8.41	**7.66**	21	1TIG	**11.62**	12.32
8	1AIL	**4.34**	8.97	22	1HHP	**13.62**	16.65
9	4ICB	**4.23**	6.15	23	2HG6	**15.54**	16.40
10	1CC5	10.54	**9.97**	24	2MIT	**8.02**	8.73
11	4UEX	**5.82**	8.39	25	1I6C	**6.72**	8.72
12	2EZK	10.97	**9.75**	26	1BQ9	**7.90**	10.36
13	3GWL	**9.89**	13.29	27	1WAPa	**12.47**	13.50
14	1FD4	**7.57**	8.73	28	1ALY	23.17	**21.11**

(with PDB IDs 1GYZ, 1ISUa, 1AIL, 4UEX, 3GWL, 1HZ6a, 1AOY, 1HHP, 1I6C, and 1BQ9). Rosetta is superior to DPDE by more than 2.0 Å only in 1 protein with PDB ID 1ALY. For the two proteins with PDB IDs 1AIL and 1AOY, DPDE is better than Rosetta by more than 4.0 Å.

The histogram of distribution of predicted conformations with DPDE and Rosetta as a function of RMSD to native for the eight selected proteins shown in Fig. 6 is reported in Fig. 7. The distribution obtained by DPDE is superimposed in wathet blue over that obtained by Rosetta in pink. DPDE not only finds the lowest-RMSD conformation, but also samples significantly more low-RMSD conformations than Rosetta for most of these proteins.

From Fig 7, we can know that DPDE is able to sample decoys with lower RMSDs compared to Rosetta at remarkably higher proportions for most of the selected proteins. In more detail, distributions of the RMSD obtained by DPDE and Rosetta show almost the same shapes for the protein with PDB ID 2IMU, and the highest peak of them corresponding to the same RMSD. However, the percentage of decoys of DPDE is higher than that of Rosetta. The percentage of decoys corresponding to the highest peak for DPDE is 22.2 percent, while that of Rosetta is 18 percent. For the protein with PDB ID 1ENH, about 23 percent of the decoys sampled by DPDE reach 3.5 Å in RMSD, while only about 8 percent of the decoys sampled by Rosetta reach 7.0 Å in RMSD. For the protein with PDB ID 1AIL, the percentage of decoys of Rosetta (about 8.6 percent) is higher than that of DPDE (about 8 percent), but the highest peak of DPDE corresponding to 3.0 Å in RMSD, while that of Rosetta corresponds to 11 Å in RMSD. For protein with PDB ID 4ICB, DPDE can sample about 12.5 percent of the decoys with 3.5 Å in RMSD, while Rosetta can sample about 9.5 percent of the decoys with 12.5 Å in RMSD. For the protein with PDB ID 1FD4, although the percentage of decoys of Rosetta (15.7 percent) is higher than that of DPDE (14.6 percent), but the highest peak of DPDE corresponding to 8.5 Å in RMSD, while that of Rosetta corresponds to 10 Å in RMSD. For the protein with PDB ID 1HZ6a, DPDE can sample decoys with lower RMSDs than that of Rosetta in observably higher proportion. DPDE can sample about 14.8 percent of the decoys with 4.5 Å in RMSD, while Rosetta can sample about 6.5 percent of the decoys with 10 Å in RMSD. For the protein with PDB ID 1AOY, about 9.5 percent of the decoys sampled by DPDE reach 7.5 Å in RMSD, while about 9 percent of the decoys sampled by Rosetta reach 13.5 Å in RMSD. For the protein with PDB ID 1BQ9, the percentage of decoys of Rosetta (12.2 percent) is also higher than that of DPDE (9.5 percent), but the highest peak of DPDE corresponds to 11 Å in RMSD, while that of Rosetta corresponds to 12 Å in RMSD.

Based on the above results and analysis, we can know that the proposed DPDE is able to sample more decoys with lower RMSD compare to Rosetta, and the RMSD of the cluster centroid is also better than that of Rosetta for the majority of the benchmark proteins. We can conclude that the proposed distance profile-based selection strategy is also capable of improving the near-native sampling ability.



(a) 2IMU (b) 1ENH (c) 1AIL (d) 4ICB

(e) 1FD4 (f) 1HZ6a (g) 1AOY (h) 1BQ9

Fig. 7. Distribution of RMSD to the native structure on nine representative proteins. The wathet blue for the DPDE, and the pink for the Rosetta.

TABLE 3
Results of RMSD Obtained by DPDE, HEA, MOEA, and MOEA-PC for 16 Benchmark Proteins

No.	PDB ID	Size	Type	Minimum lowest RMSD(Å)			
				DPDE	HEA	MOEA	MOEA-PC
1	1ISUa	62	α	**5.73**	6.60	6.40	6.40
2	1AIL	73	α	1.58	1.40	1.90	**1.10**
3	1CC5	83	α	**4.70**	**4.70**	4.90	5.40
4	2EZK	99	α	3.67	3.40	3.20	**2.90**
5	3GWL	106	α	**3.68**	5.40	5.80	5.50
6	1DTDb	61	α/β	5.14	**4.40**	5.30	5.00
7	1SAP	66	α/β	5.10	3.70	3.70	**2.90**
8	1HZ6a	67	α/β	2.64	**1.90**	2.10	2.20
9	1DTJa	76	α/β	**2.16**	4.20	2.30	4.00
10	1AOY	78	α/β	**3.06**	3.90	3.70	3.90
11	1TIG	88	α/β	2.82	3.20	**2.50**	3.90
12	1HHP	99	α/β	8.90	8.80	**8.60**	8.90
13	2HG6	106	α/β	**8.48**	9.30	9.60	9.20
14	1BQ9	53	β	**2.92**	3.00	3.40	3.30
15	1WAPa	64	β	**6.08**	6.30	6.40	6.40
16	1ALY	146	β	11.60	**11.20**	11.40	**11.20**

4.5 Comparison with Three State-of-the-Art Methods

DPDE is also compared with three other state-of-the-art EA-based methods, namely, HEA [10], MOEA [11], and MOEA-PC [13]. HEA was proposed by Olson et al. [49]. In HEA, successful strategies used in state-of-the-art-initio protocols including a coarse-grained representation and molecular fragment replacement are incorporated into the EA, and a local search is employed to map each child conformation to a nearby local minimum. Experiment results show that HEA can enhance the sampling ability of low-energy regions in the protein conformational space. MOEA, which was proposed by Olson and Shehu [10], is a method which combines local and global searches in a population-based evolutionary algorithm. Through the guidance of Pareto analysis, a fixed-size population of decoy conformations is evolved through a series of generations. Experimental results indicate that MOEA can obtain lower energy conformations with near-native structure. MOEA-PC also was proposed by Olson and Shehu [13]. It is a HEA variant with Pareto rank and Pareto count. Experimental results show that MOEA-PC may not remarkably improve the overall lowest RMSD but can improve the funneling of the sampled landscape.

Table 3 summarizes the results obtained by DPDE, HEA [10], MOEA [11], and MOEA-PC [13] in terms of minimum lowest RMSDs to the native structure for 16 out of the 28 benchmark proteins. Results in columns 5 and 6 of Table 3 indicate that DPDE performs significantly better than HEA in 8 (with native PDB IDs 1ISUa, 3GWL, 1DTJa 1AOY, 1TIG, 2HG6, 1BQ9, and 1WAPa) out of 16 benchmark proteins in terms of minimum lowest RMSD to the native structure, while HEA is better than DPDE in the remaining seven benchmark proteins except for the protein with PDB ID 1CC5. Particularly, DPDE outperforms HEA by 0.5 Å or more in five proteins (with PDB IDs 1ISUa, 3GWL, 1DTJa, 1AOY, and 2HG6), while HEA is superior to DPDE by more than 0.5 Å in three proteins (with PDB IDs 1DTDb, 1SAP, and 1HZ6a). Particularly, DPDE is better than HEA by more than 1.5 Å in two proteins (with PDB IDs 3GWL and 1DTJa). From the table, we also find DPDE outperforms HEA in most of the α/β and β folding proteins. This suggests that DPDE is more effective at approaching the native structure of α/β and β folding proteins compared to HEA.

Comparison of columns 5 and 7 of Table 3 shows that DPDE achieves better minimum lowest RMSDs than MOEA in 10 (with PDB IDs 1ISUa, 1AIL, 1CC5, 3GWL, 1DTDb, 1DTJa, 1AOY, 2HG6, 1BQ9, and 1WAPa) out of 16 benchmark proteins. MOEA is superior to DPDE in the other 6 benchmark proteins. DPDE is significantly better than MOEA by over 0.5 Å in 3 benchmark proteins (with PDB IDs 1ISUa, 3GWL, and 2HG6), while MOEA outperforms DPDE by more than 0.5 Å in one benchmark proteins (with PDB ID 1HZ6a). For the protein with PDB ID 3GWL, DPDE is better than MOEA by 2.12 Å.

Comparison of the results listed in columns 5 and 8 of Table 3 shows that DPDE is superior to MOEA-PC in the majority of the benchmark proteins. To be specific, DPDE outperforms MOEA-PC in 9 (with PDB IDs 1ISUa, 1CC5, 3GWL, 1DTJa, 1AOY, 1TIG, 2HG6, 1BQ9, and 1WAPa) out of 16 benchmark proteins, while MOEA-PC performs better than DPDE in 6 proteins (with PDB IDs 1AIL, 2EZK, 1DTDb, 1SAP, 1HZ6a, and 1AIL). For the protein with PDB ID 1HHP, DPDE and MOEA-PC achieve the same minimum lowest RMSD. For the seven proteins with PDB IDs 1ISUa, 1CC5, 3GWL, 1DTJa, 1AOY, 1TIG, and 2HG6, DPDE is better than MOEA-PC by 0.5 Å or more. MOEA-PC is superior to DPDE by more than 0.5 Å only in two proteins (with PDB IDs 2EZK and 1SAP). In particular, DPDE significantly outperforms MOEA-PC by at least 1.5 Å in 2 proteins with PDB IDs 3GWL and 1DTJa. For the protein with PDB ID 2EZK, MOEA-PC performs better than DPDE by 2.2 Å. From the results, we can also find that DPDE outperforms MOEA-PC in 7 out of 11 α/β and β folding proteins. This indicate that DPDE is also more effective at predicting the structure of α/β and β folding proteins compared to MOEA-PC.

In addition, by comparing all results listed in Table 3 for the four methods, we can find that DPDE is superior to the other three methods in eight (with PDB IDs 1ISUa, 1CC5, 3GWL, 1DTJa, 1AOY, 2HG6, 1BQ9, and 1WAPa) out of 16 benchmark proteins. HEA achieves the best results in four proteins with PDB IDs 1CC5, 1DTDa, 1HZ6a, and 1ALY. MOEA outperforms the other three methods in two proteins with PDB IDs 1TIG and 1HHP. MOEA-PC performs better than the other competitors in four proteins with PDB IDs 1AIL, 2EZK, 1SAP, and 1ALY.

In order to compare the performance of the different methods in the benchmark set, the Friedman test [50] is conducted. The results are based on the minimum lowest RMSD listed in Table 3. Column 2 of Table 4 shows the mean rankings of the four methods. It is clear that the performance of the four methods can be sorted according to their mean rankings in the following order: DPDE, HEA, MOEA, and MOEA-PC. The highest mean ranking is obtained by the proposed DPDE, which is superior to the other three state-of-the-art methods for 8 benchmark proteins in terms of the minimum lowest RMSD.

In addition to the Friedman test, the Wilcoxon Signed Rank Test [47] is also conducted to recognize significant differences between the results of the six methods. The test results are based on the minimum lowest RMSD shown in

TABLE 4
Results Obtained by Friedman Test and Wilcoxon Signed Rank Test

Methods	Mean Ranking	p-value	Significance
DPDE	2.75	NA	NA
HEA	2.63	0.4796	≈
MOEA	2.22	0.3651	≈
MOEA-PC	2.41	0.3028	≈

TABLE 5
The Size of Distance Profile and Average Distance Profile Error of All Benchmark Proteins

No.	PDB ID	Size of Profile (N)	Profile Error	No.	PDB ID	Size of Profile (N)	Profile Error
1	1VII	32	2.74	15	1GB1	85	3.27
2	2IMU	38	4.12	16	1DTDb	213	3.80
3	1ENH	58	2.46	17	1SAP	169	3.53
4	1BBO	135	5.08	18	1HZ6a	160	1.81
5	1GYZ	64	1.81	19	1DTJa	214	3.75
6	1ISUa	240	3.66	20	1AOY	107	2.72
7	2MQK	53	1.89	21	1TIG	199	2.37
8	1AIL	56	0.43	22	1HHP	361	2.99
9	4ICB	148	2.44	23	2HG6	132	5.26
10	1CC5	135	3.33	24	2MIT	54	4.70
11	4UEX	135	4.86	25	1I6C	76	2.19
12	2EZK	68	1.91	26	1BQ9	139	2.44
13	3GWL	180	3.18	27	1WAPa	271	3.44
14	1FD4	98	2.39	28	1ALY	672	4.27

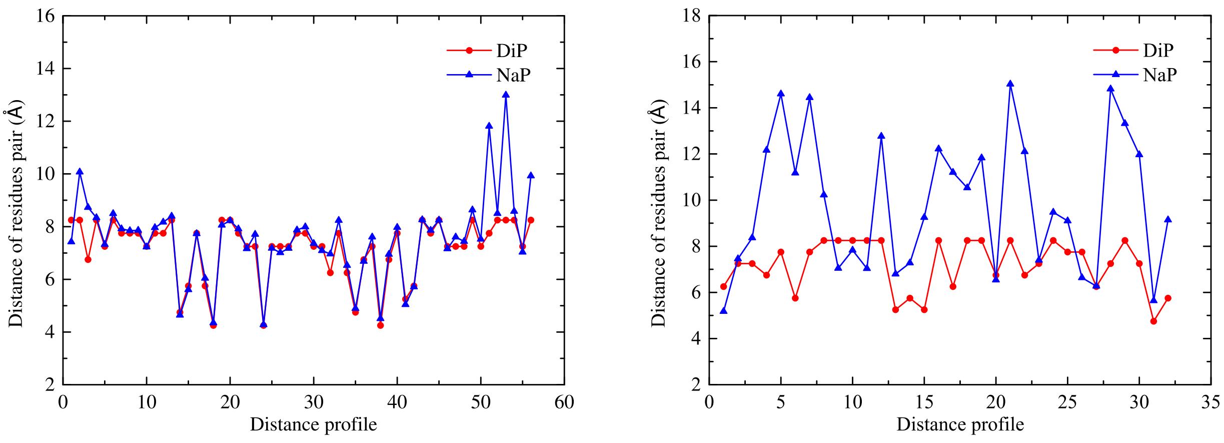
Table 3. Columns 3 and 4 of Table 4 respectively show the p-value and significance between DPDE and the three state-of-the-art methods obtained by the Wilcoxon Signed Rank Test. The results indicate that the results obtained by DPDE are similar with HEA, MOEA-PC, and MOEA, which are marked with "≈". Although DPDE doesn't perform significantly better than these three competitors, it is superior to them in the majority of the benchmark proteins in terms of minimum lowest RMSD.

The above results and analysis suggest that the proposed distance profile-based selection strategy of DPDE allows effective conformational sampling nearer the native structure in the majority of the benchmark proteins compared with the three state-of-the-art methods.

4.6 Distance Profiles Analysis

The size of distance profiles and the average distance profile error, which is calculated according to Eq. (2) for all 28 benchmark proteins, are given in Table 5. As seen in the table, the average distance profile errors of all proteins are below 6 Å. In addition, on the basis of the results listed in Table 1 and the profile error given in Table 5, we find that DPDE can obtain more accurate structures with lower RMSDs to the native for specific proteins if the distance profiles are sufficiently accurate. Here, 2 representative proteins (with PDB IDs 1VII and 1IAL) are selected to investigate this conclusion. Fig. 8 shows the comparison of the predicted distance profiles and the distance profiles of the native structure for the protein with PDB IDs 1AIL and 1VII, where "DiP" represents the predicted distance profile and is plotted in red, whereas "NaP" represents the distance profiles of the native structure and is plotted in blue. For the protein with PDB ID 1AIL, the predicted distance profiles nearly overlaps with the corresponding distance profiles of the native structure, and the average profile error is 0.43 Å. Thus DPDE obtains a good predicted structure with the minimum lowest RMSD of 1.58 Å. For the protein with PDB ID 1VII, the predicted distance profiles are completely different from that of its native structure, and the average profile error is 2.74 Å. Therefore, the minimum lowest RMSD achieved by DPDE is higher than 2.0 Å.

To investigate the effect of the accuracy of distance profiles on prediction accuracy further, DPDE using distance profiles obtained from the native structure (DPDE-NaP) is used to predict the structure of the six representative proteins with PDB IDs 1VII, 1GYZ, 2MQK, 4UEX, 2EZK, and 1SAP, the accuracies of which are lower than those of Rosetta, as shown in Table 1. Table 6 summarizes the minimum lowest RMSDs obtained by DPDE-NaP. As seen, DPDE-NaP clearly performs significantly better than either DPDE or Rosetta in all these 6 proteins in terms of minimum lowest RMSD and the Mean lowest RMSD. Specifically, DPDE-NaP outperforms DPDE by 0.5 Å or more in five out of six proteins (with PDB IDs 1VII, 2MQK, 4UEX, 2EZK, and 1SAP) in terms of minimum lowest RMSD and mean lowest RMSD. In particular, DPDE-NaP is superior better than DPDE by 1.16 Å for the protein with PDB ID 2MQK in



(a) 1AIL

(b) 1VII

Fig. 8. Comparison of the predicted distance profiles (red) with the distance profiles of the native structure (blue) on two representative proteins.

TABLE 6
Comparison of the Results Obtained by DPDE-NaP with DPDE and Rosetta for Six Representative Proteins

No.	PDB ID	Size	Type	Minimum lowest RMSD(Å)			Mean ± Std lowest RMSD(Å)		
				DPDE-NaP	DPDE	Rosetta	DPDE-NaP	DPDE	Rosetta
1	1VII	36	α	1.12	2.08	1.58	1.37 ± 0.20	2.37 ± 0.22	1.75 ± 0.15
2	1GYZ	60	α	1.61	1.98	1.85	1.95 ± 0.26	2.31 ± 0.27	2.11 ± 0.13
3	2MQK	65	α	1.27	2.43	1.99	1.53 ± 0.16	3.06 ± 0.30	2.46 ± 0.39
4	4UEX	85	α	3.20	3.92	3.50	4.00 ± 0.70	4.69 ± 0.45	4.46 ± 0.68
5	2EZK	99	α	2.85	3.67	3.57	3.51 ± 0.37	4.54 ± 0.47	4.03 ± 0.38
6	1SAP	66	α/β	4.38	5.10	4.57	5.44 ± 0.79	6.12 ± 0.38	5.72 ± 0.58

terms of minimum lowest RMSD. For the two proteins with PDB IDs 1VII, 2MQK, and 2EZK, DPDE-NaP outperforms DPDE by 1.0 Å or more in terms of mean lowest RMSD. Compared with Rosetta, DPDE-NaP performs better by more than 0.5 Å in two proteins (with PDB IDs 2MQK and 2EZK) in terms of minimum lowest RMSD and mean lowest RMSD. In addition, the energy of each conformation generated by DPDE-NaP against its RMSD to the native structure on the 4 representative proteins (1VII, 1GYZ, 2MQK, and 2EZK) are plotted in Fig. 9. It shows that DPDE-NaP yields more conformations with lower RMSD to the native structure compared with DPDE and Rosetta.

Based on the above results and analysis, it can be concluded that the proposed DPDE can obtain conformations with more reasonable structures which have lower RMSD if the distance profiles are sufficiently accurate.

5 Conclusion

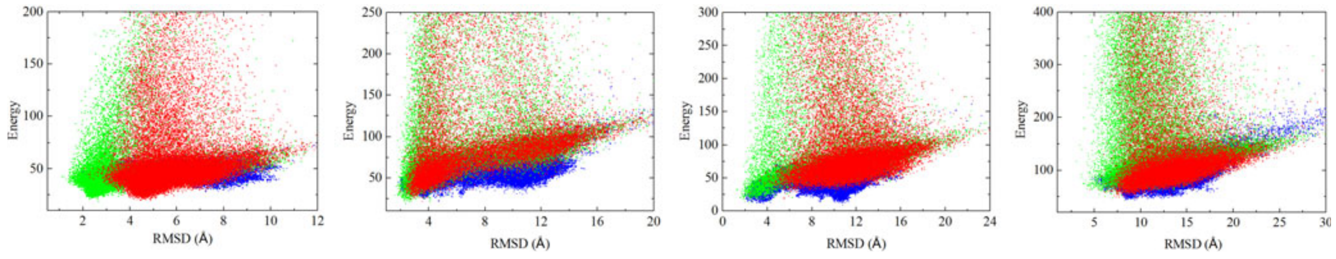
Due to the inaccuracies of the energy model, the lowest-energy regions in the energy surface are inconsistently associated with low-RMSD conformations. This study proposes a distance profile-guided differential evolution algorithm, DPDE, for protein conformational space sampling. In DPDE, a distance profile-based selection strategy is proposed to guide the conformational space sampling. The distance profile-based selection strategy uses not only the energy but also the average distance error between the distance of the residue-residue and its corresponding predicted distance in the distance profiles as measures of conformation. Using the average distance error as the auxiliary measure of the conformation, the prediction error caused by the inaccuracy of the energy model can be reduced, and it also prevents the algorithm from getting trapped into a local minimum. Thus, DPDE is able to sample conformations in the region closer to the native state and steadily obtain conformations with high accuracy.

Experimental results of 28 benchmark proteins show that the distance profile-based selection strategy allows DPDE to sample near-native conformations more effectively than Rosetta in the majority of the benchmark proteins. The results also indicate that DPDE tends to capture conformations with lower energies and more reasonable structures, while Rosetta only tends to generate conformations with low energy. The auxiliary measure distance error allows the algorithm to escape a local minimum and jump to a conformation of higher energy but more reasonable structure. Therefore, DPDE can obtain conformations with lower RMSD than Rosetta in most of the proteins. The proposed DPDE is also compared with three state-of-the-art EA-based methods in 16 benchmark proteins. Results indicate that the proposed DPDE is significantly better than the three state-of-the-art methods in the majority of benchmark proteins. In addition, the proposed distance profile-based selection strategy also can be introduced to other EAs.

The effect of the accuracy of the distance profiles on prediction accuracy is studied, and results reveal that more accurate distance profiles may lead to more accurate prediction structure. In future studies, we intend to construct more accurate distance profiles by using other techniques. As two criteria are employed to evaluate the conformation in our proposed DPDE, we intend to integrate the criteria into certain multi-objective optimization techniques, such as MOEA and MOEA-PC, which were proposed by Olson et al. [13]. Moreover, the setting of the temperature scaling factor β should be researched in detail. Guidelines for the adaptive selection of the parameter β in different stages will also be integrated into our future research.

Acknowledgments

This work was supported by the National Nature Science Foundation of China (No. 61075062, 61573317), Public



(a) 1VII (b) 1GYZ (c) 2MQK (d) 2EZK

Fig. 9. The energy is plotted against RMSD to the native structure for each conformation generated by DPDE-NaP on four selected proteins. The green, red, and blue dots are for DPDE-NaP, DPDE, and Rosetta, respectively.

Welfare Project of Science Technology Department of Zhejiang Province (No. 2014C33088). The authors would like to thank the anonymous reviewers for their useful comments and suggestions. G.-J. Zhang is corresponding author.

References

[1] P. Bradley, K. M. S. Misura, and D. Baker, "Toward high-resolution de novo structure prediction for small proteins," *Science*, vol. 309, no. 5742, pp. 1868–1871, 2005.

[2] M. E. M. Noble, J. A. Endicott, and L. N. Johnson, "Protein kinase inhibitors: Insights into drug design from structure," *Science*, vol. 303, no. 5665, pp. 1800–1805, 2004.

[3] L. Wickstrom, E. Gallicchio, and R. M. Levy, "The linear interaction energy method for the prediction of protein stability changes upon mutation," *Proteins: Struct., Function, Bioinf.*, vol. 80, no. 1, pp. 111–125, 2012.

[4] K. H. Ambert and A. M. Cohen, "K-information gain scaled nearest neighbors: a novel approach to classifying protein-protein interaction-related documents," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 9, no. 1, pp. 305–310, Jan./Feb. 2012.

[5] Y. Zhu, X. F. Zhang, D. Q. Dai, and M. Y. Wu, "Identifying spurious interactions and predicting missing interactions in the protein-protein interaction networks via a generative network model," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 10, no. 1, pp. 219–225, Jan./Feb. 2013.

[6] M. Dorn, M. B. e Silva, L. S. Buriol, and L. C. Lamb, "Three-dimensional protein structure prediction: Methods and computational strategies," *Comput. Biol. Chem.*, vol. 53, pp. 251–276, 2014.

[7] W. Zhang, J. Yang, B. He, S. E. Walker, H. Zhang, B. Govindarajoo, J. Virtanen, Z. Xue, H. B. Shen, and Y. Zhang, "Integration of QUARK and I-TASSER for AB initio protein structure prediction in CASP11," *Proteins: Struct. Function, Bioinf.*, to be published, DOI: 10.1002/prot.24930, 2015.

[8] D. Xu and Y. Zhang, "Ab initio structure prediction for escherichia coli: Towards genome-wide protein structure modeling and fold assignment," *Sci. Rep.*, vol. 3, p. 1895, 2013.

[9] C. B. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, pp. 223–230, 1973.

[10] B. Olson, K. De Jong, and A. Shehu, "Off-lattice protein structure prediction with homologous crossover," in *Proc. 15th Annu. Conf. Genetic Evol. Comput.*, 2013, pp. 287–294.

[11] B. Olson and A. Shehu, "Multi-objective stochastic search for sampling local minima in the protein energy surface," in *Proc. ACM Conf. Bioinf., Comput. Biol.*, 2013, pp. 430–439.

[12] S. Sudha, S. Baskar, S. M. J. Amali, and S. Krishnaswamy, "Protein structure prediction using diversity controlled self-adaptive differential evolution with local search," *Soft Computing*, vol. 19, no. 6, pp. 1635–1646, 2015.

[13] B. Olson and A. Shehu, "Multi-objective optimization techniques for conformational sampling in template-free protein structure prediction," in *Proc. 6th Int. Conf. Bioinf. Comput. Biol.*, 2014, pp. 143–148.

[14] M. T. Hoque, M. Chetty, A. Lewis, and A. Sattar, "Twin removal in genetic algorithms for protein structure prediction using low-resolution model," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 8, no. 1, pp. 234–245, Jan./Mar. 2011.

[15] Z. Li and H. A. Scheraga, "Monte Carlo-minimization approach to the multiple-minima problem in protein folding," *Proc. Nat. Academy Sci.*, vol. 84, no. 19, pp. 6611–6615, 1987.

[16] Y. Zhang, "Interplay of I-TASSER and QUARK for template-based and AB initio protein structure prediction in CASP10," *Proteins: Struct., Function, Bioinf.*, vol. 82, no. S2, pp. 175–187, 2014.

[17] Y. Zhang, D. Kihara, and J. Skolnick, "Local energy landscape flattening: Parallel hyperbolic Monte Carlo sampling of protein folding," *Proteins: Struct., Function, Bioinf.*, vol. 48, no. 2, pp. 192–201, 2002.

[18] Y. M. Cheng, S. M. Gopal, S. M. Law, and M. Feig, "Molecular dynamics trajectory compression with a coarse-grained model," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 9, no. 2, pp. 476–486, Mar./Apr. 2012.

[19] K. Lindorff-Larsen, N. Trbovic, P. Maragakis, S. Piana, and D. E. Shaw, "Structure and dynamics of an unfolded protein examined by molecular dynamics simulation," *J. Amer. Chemical Soc.*, vol. 134, no. 8, pp. 3787–3791, 2012.

[20] J. Lee, H. A. Scheraga, and S. Rackovsky, "New optimization method for conformational energy calculations on polypeptides: conformational space annealing," *J. Comput. Chemistry*, vol. 18, no. 9, pp. 1222–1232, 1997.

[21] J. Lee, J. Lee, T. N. Sasaki, M. Sasai, C. Seok, and J. Lee, "De novo protein structure prediction by dynamic fragment assembly and conformational space annealing," *Proteins: Struct., Function, Bioinf.*, vol. 79, no. 8, pp. 2403–2417, 2011.

[22] K. Joo, I. Joung, J. Lee, J. Lee, W. Lee, B. Brooks, S. Lee, and J. Lee, "Protein structure determination by conformational space annealing using NMR geometric restraints," *Proteins: Struct., Function, Bioinf.*, vol. 83, no. 12, pp. 2251–2262, 2015.

[23] A. Kolinski and J. Skolnick, "Monte Carlo simulations of protein folding. i. lattice model and interaction scheme," *Proteins: Struct. Function Bioinf.*, vol. 18, no. 4, pp. 338–352, 1994.

[24] A. Liwo, P. Arlukowicz, C. Czaplewski, S. Oldziej, J. Pillardy, and H. A. Scheraga, "A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape: Application to the UNRES force field," *Proc. Nat. Academy Sci.*, vol. 99, no. 4, pp. 1937–1942, 2002.

[25] P. M. Bowers, C. E. Strauss, and D. Baker, "De novo protein structure determination using sparse NMR data," *J. Biomolecular NMR*, vol. 18, no. 4, pp. 311–318, 2000.

[26] O. Morag, N. G. Sgourakis, D. Baker, and A. Goldbourt, "The NMR-Rosetta capsid model of m13 bacteriophage reveals a quadrupled hydrophobic packing epitope," *Proc. Nat. Academy Sci.*, vol. 112, no. 4, pp. 971–976, 2015.

[27] P. Rossi, L. Shi, G. Liu, C. M. Barbieri, H. W. Lee, T. D. Grant, J. R. Luft, R. Xiao, T. B. Acton, S. E. H., G. T. Montelione, D. Baker, O. F. Lange, and N. G. Sgourakis, "A hybrid NMR/SAXS-based approach for discriminating oligomeric protein interfaces using rosetta," *Proteins: Struct. Function Bioinf.*, vol. 83, no. 2, pp. 309–317, 2015.

[28] S. Miyazawa and R. L. Jernigan, "Residuecresidue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading," *J. Molecular Biol.*, vol. 256, no. 3, pp. 623–644, 1996.

[29] G. Z. Zhang and D. S. Huang, "Combing genetic algorithm with neural network technique for protein inter-residue spatial distance prediction," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2004, pp. 1687–1691.

[30] P. Kukic, C. Mirabello, G. Tradigo, I. Walsh, P. Veltri, and G. Pollastri, "Toward an accurate prediction of inter-residue distances in proteins using 2D recursive neural networks," *BMC Bioinf.*, vol. 15, no. 1, p. 6, 2014.

[31] S. Miyazawa, "Prediction of contact residue pairs based on co-substitution between sites in protein structures," *PloS One*, vol. 8, no. 1, p. e54252, 2013.

[32] D. Xu and Y. Zhang, "Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field," *Proteins: Struct. Function Bioinf.*, vol. 80, no. 7, pp. 1715–1735, 2012.

[33] D. Xu and Y. Zhang, "Toward optimal fragment generations for AB initio protein structure assembly," *Proteins: Struct. Function Bioinf.*, vol. 81, no. 2, pp. 229–239, 2013.

[34] R. Storn and K. Price, "Differential evolution—A simple and efficient heuristic for global optimization over continuous spaces," *J. Global Optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[35] S. Das and P. N. Suganthan, "Differential evolution: A survey of the state-of-the-art," *IEEE Trans. Evol. Comput.*, vol. 15, no. 1, pp. 4–31, Feb. 2011.

[36] M. Y. Cheng and D. H. Tran, "Two-phase differential evolution for the multiobjective optimization of time-cost tradeoffs in resource-constrained construction projects," *IEEE Trans. Eng. Manage.*, vol. 61, no. 3, pp. 450–461, Aug. 2014.

[37] J. Santos and M. Diéguez, *"Differential evolution for protein structure prediction using the HP model,"* in *Proc. 4th Int. Work-Conf. Interplay Between Natural Artif. Comput.*, 2011, pp. 323–333.

[38] H. S. Lopes and R. Bitello, "A differential evolution approach for protein folding using a lattice model," *J. Comput. Sci. Technol.*, vol. 22, no. 6, pp. 904–908, 2007.

[39] C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker, "Protein structure prediction using rosetta," *Methods Enzymology*, vol. 383, pp. 66–93, 2004.

[40] J. Handl, J. Knowles, R. Vernon, D. Baker, and S. C. Lovell, "The dual role of fragments in fragment-assembly methods for de novo protein structure prediction," *Proteins: Struct. Function Bioinf.*, vol. 80, no. 2, pp. 490–504, 2012.

[41] A. Shehu, L. E. Kavraki, and C. Clementi, "Multiscale characterization of protein conformational ensembles," *Proteins: Struct. Function Bioinf.*, vol. 76, no. 4, pp. 837–851, 2009.

[42] D. E. Kim, B. Blum, P. Bradley, and D. Baker, "Sampling bottlenecks in de novo protein structure prediction," *J. Molecular Biol.*, vol. 393, no. 1, pp. 249–260, 2009.

[43] J. M. Bujnicki, "Protein-structure prediction by recombination of fragments," *Chembiochem*, vol. 7, no. 1, pp. 19–27, 2006.

[44] A. Leaver-Fay, M. Tyka, S. M. Lewis, O. F. Lange, J. Thompson, R. Jacak, and I. W. Davis, "ROSETTA3: An object-oriented software suite for the simulation and design of macromolecules," *Methods Enzymology*, vol. 487, pp. 545–574, 2011.

[45] K. Molloy, S. Saleh, and A. Shehu, "Probabilistic search and energy guidance for biased decoy sampling in AB initio protein structure prediction," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 10, no. 5, pp. 1162–1175, Sep./Oct. 2013.

[46] B. Olson and A. Shehu, "Evolutionary-inspired probabilistic search for enhancing sampling of local minima in the protein energy surface," *Proteome Sci.*, vol. 10, no. Suppl 1, p. S5, 2012.

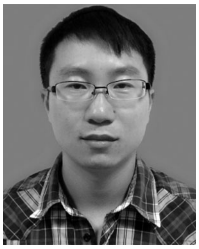
[47] G. W. Corder and D. I. Foreman, *Nonparametric Statistics Non-Statisticians: Step-By-Step Approach*. Hoboken, NJ, USA: Wiley, 2009.

[48] H. J. C. Barbosa, H. S. Bernardino, and A. M. S. Barreto, "Using performance profiles to analyze the results of the 2006 CEC constrained optimization competition," in *Proc. IEEE Congr. Evol. Comput.*, 2006, pp. 1–8.

[49] A. Shehu and B. Olson, "Guiding the search for native-like protein conformations with an AB-initio tree-based exploration," *Int. J. Robot. Res.*, vol. 29, no. 8, pp. 1106–1127, 2010.

[50] S. Garcĺa, F. A., J. Luengo, and F. Herrera, "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power," *Inf. Sci.*, vol. 180, no. 10, pp. 2044–2064, 2010.

Gui-Jun Zhang is currenlty a professor in the College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. His research interest include intelligent information processing, optimization theory and algorithm design, and bioinformatics.



Xiao-Gen Zhou is currently working toward the PhD degree in the College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. His research interest include intelligent information processing, optimization theory and algorithm design, and bioinformatics.

Xu-Feng Yu is currently working toward the master's degree in the College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. His research interest include intelligent information processing, intelligent optimization, and bioinformatics.

Xiao-Hu Hao is currently working toward the PhD degree in the College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. His research interest include intelligent information processing, intelligent optimization, and bioinformatics.

Li Yu is a professor in the College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. His research interests include intelligent control, decentralized control, and networked control systems.