IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 4, 2020, accepted February 21, 2020, date of publication March 3, 2020, date of current version March 13, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2977945

Similarity Analysis of 3D Structures of Proteins Based Tile-CNN

SHENGWEI QIN[1,3], (Member, IEEE), ZHONG LI[2], LEXUAN HE[3], AND WANMIN LIN[3]
[1]Faculty of Mechanical Engineering and Automation, Zhejiang Sci-Tech University, Zhejiang 310018, China
[2]School of Science, Zhejiang Sci-Tech University, Zhejiang 310018, China
[3]South China Institute of Software Engineering, Guangzhou University, Guangzhou 510990, China

Corresponding author: Zhong Li (lizhong@zstu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 11671009, in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LZ19A010002, in part by the Department of Education of Guangdong Province under Grant 2017KQNCX275, and in part by the South China Institute of Software Engineering of Guangzhou University under Grant ST201902.

ABSTRACT The 3D structure of a protein is closely related to its function, and the similarity analysis between their structures can help reveal the function of proteins. However, there exist two problems arising from the analysis of 3D structures of proteins. The proteins with a similar sequence may have different structures, while the proteins with a similar structure may have different sequences. In the analysis of similarity in 3D structures of proteins, it remains difficult for the traditional methods using the spatial feature distribution and geometry or topology features of proteins to solve these problems. In this paper, a Tile-CNN network is proposed to analyze the similarity of proteins in 3D structure. In order to capture the overall and the local features as exhibited by the 3D structures of proteins, it projects 3D protein models into 2D protein images from different views and then cuts these 2D projected images using the tile strategy. After the training of proteins with these images in the Tile-CNN, the test protein model can be expressed by an analysis matrix, and then the similarity between 3D structures of proteins is computed using the root mean square distance (RMSD) for the benchmark matrix and the analysis matrix. As revealed by the experimental results, the proposed algorithm is more robust in analyzing the similarity of 3D structures of proteins and produces a satisfactory performance in solving the two aforementioned problems.

INDEX TERMS 3D structures, similarity, Tile-CNN, protein.

I. INTRODUCTION

Bioinformatics is an interdisciplinary subject, which analyzes the biological information from such perspectives as computer science, biology, physics and mathematics [1], [2]. With the completion of sequencing of the human genome, the development of biological science has moved into the post-gene era and the focus of research has shifted to the regulation of proteins expression and their functions. In the study of protein functions, it mainly starts with the 3D structures of proteins. Besides, the analysis of 3D structures of proteins is essentially the study on its shape similarity. The development of bioinformatics can enable researchers to analyze the function of 3D structures of proteins more intuitively and easily. The similarity analysis of protein structures is actually a comparison of the 3D structures of proteins in space.

The associate editor coordinating the review of this manuscript and approving it for publication was Vincenzo Conti.

The similarity of protein structures can be established according to various distance measures for proteins represented by the feature vector, matrix and tensor [1], [3]. Normally, if the distance between two proteins is closer to zero, it suggests that two proteins are more similar. However, it is inevitable to encounter two problems (see in Fig 1). One is that the sequences of proteins are similar, but their 3D structures may be different, which requires the whole and local features of the 3D structure of protein to be described sufficiently to ensure the accuracy of similarity analysis. The other is that, since the 3D structure of a protein normally determines its function, the proteins with similar functions are possible to have similar structures, but their sequences may be clearly different. These kinds of proteins would have impact on the similarity analysis of proteins. Accordingly, how to obtain the accurate result for the two problems to be solved is a major difficulty facing the analysis of 3D structures of proteins. Currently, the similarity analysis of 3D structures of proteins

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

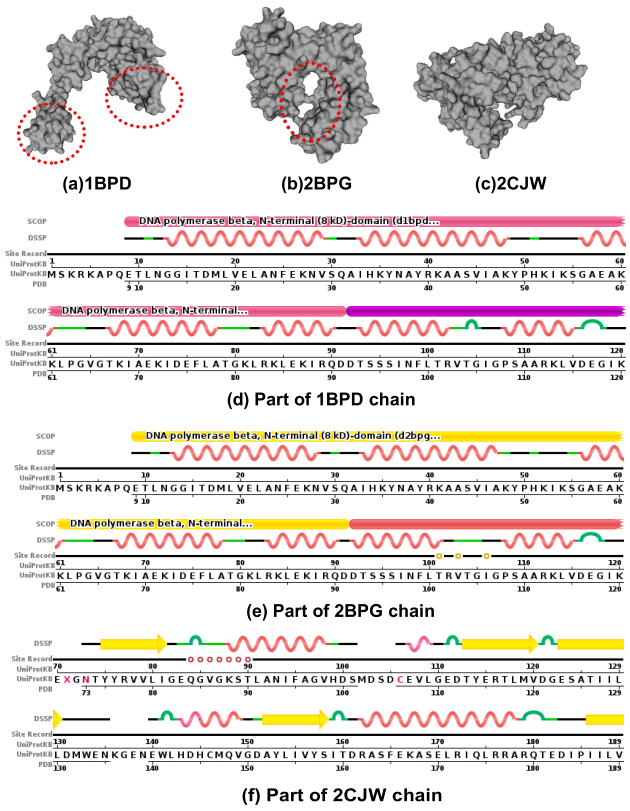


FIGURE 1. 3D Structures and sequences of proteins. Note that the protein 2BPG is considered to be similar to protein 1BPD as they have similar sequences (see (d) and (e)), but their 3D structure models are quite different (2BPG is largely deformed from 1BPD). For the protein 2CJW, it is also regarded to be similar to protein 1BPD as its 3D structure is similar to 1BPD, but their sequences (d) and (f) are quite different.

can be roughly divided into three types, which are the structural analysis based on spatial feature distribution, geometric feature-based analysis and topology-based analysis.

1) STRUCTURAL ANALYSIS BASED ON SPATIAL FEATURE DISTRIBUTION

The 3D structures of proteins are determined by the spatial positions of the atoms which can be used to analyze the similarity of structures of proteins. In order to ensure the accuracy of similarity analysis, it is necessary to maintain the rotation and translation invariance of the 3D structures of proteins. Carugo and Pongor [4] compared the similarity of proteins by using the distance distribution, which is computed by the coordinates of the skeletons of proteins. Hu and Peng [5] proposed a volume fractal dimensionality method to analyze the similarity shown by the 3D structures of proteins, which can keep the rotation and translation invariance. This method demonstrates a strong adaptive capacity when the amino acids mutate with no functional changes. However, it is based on the relevant statistics, which is inaccurate for the search of similar proteins in the massive database. Moreover, it is low in adaptability to functionally mutated proteins.

2) GEOMETRIC FEATURE-BASED ANALYSIS

Since the different rotation of the protein can cause the structure of the protein to be complex and diverse, it is difficult to describe the details of the structure with spatial distribution features. Another method is to rely on the geometric features to determine the consistency of structures of proteins from the geometric relations of proteins. Considering the skeleton of *Ca* as a continuous curve, Kotlovyi *et al.* [6] extracted shape features from this curve, such as curvature and torsion. By computing the deviation degree of these shape features, proteins can be analyzed for their structural similarity. However, this method is suitable exclusively for the analysis of local protein chains. The proteins with a long chain would have a low retrieval efficiency. Li *et al.* [7] proposed a 3D protein shape similarity analysis based on hybrid features. They constructed an analysis tensor based on local diameter (LD), heat kernel signature (HKS) and salient geometry features (SGF). Subsequently, similarity was measured by the norm of tensors between proteins. Though this method is capable of describing the structures of proteins with detailed features, different feature selections in advance would affect the robustness of similarity analysis of proteins.

3) TOPOLOGY-BASED ANALYSIS

As the spatial positions of the same protein still have a difference caused by the continuous movement of protein atoms, it may cause the wrong similarity analysis of proteins [8]. Bostick and Vaisman [9] analyzed the similarity in the topological relationship of protein structures with sequence similarity being less than 30%. It was found out that the topology of proteins can well overcome the errors of geometric analysis methods as caused by the frequent atomic motion. Hu *et al.* [10] demonstrated the structure of proteins as a graph, where the vertices of the graph represent the atoms of the skeleton of the protein chain, while the edge is used for the connection of the adjacent vertices. Then, the protein was mapped into a symmetric adjacency matrix and the similarity result is compared by analyzing the adjacency matrix between proteins. Li *et al.* [3] suggested an approach to 3D model similarity analysis for the proteins based on the skeleton. A local diameter (LD) was constructed as the analysis vector by extracting the skeleton of the 3D protein model, and then the LD between proteins was compared to determine their similarity. Both of the above-mentioned methods are premised on the local shape of the 3D structures of proteins, and the global feature of proteins is excluded from consideration. Consequently, it remains a challenge on how to reveal the similarity of 3D structures of proteins at different feature levels and scales.

This paper proposes a similarity analysis method for the 3D structures of proteins based on the neural network. For the 3D structures of proteins with the triangular mesh models, it first colors the protein model using Heat Kernel Signature (HKS) to ensure validity for the topological deformation

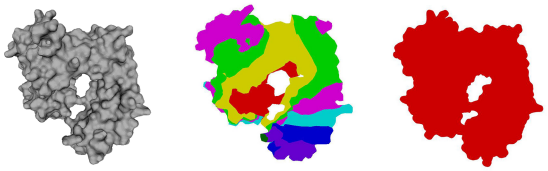


FIGURE 2. Protein 2BPG in different expression. ((a) is an model for input. (b) represents different color distributions and the color distributions always starts with red color. (c) is recolored to one kind of color in order to eliminate the effects of the same color distribution.)

of the 3D structures of proteins. Then, it maps the data on the 3D structures into 2D images from different views, based on which the overall features of proteins can be obtained. In addition, it cuts each 2D image to describe the details of the proteins. These tiled images are inputted into CNN for training and testing, which is known as Tile-CNN. Finally, an analysis matrix is constructed by the output of Tile-CNN, and the RMSD is computed to determine the similarity of proteins. As demonstrated by the experimental results, the proposed algorithm is capable of eliminating the impact of invalid features, and of achieving satisfactory performance in the similarity analysis of 3D structures of proteins.

II. METHODS

A. DATA SET CONSTRUCTION

In order to establish an effective and robust Tile-CNN network, there is a need to construct the 2D image data from the 3D protein structures. As for a 3D protein with the triangular mesh model, the 3D protein model is first colored using Heat Kernel Signature (HKS), as shown Fig. 2(a) and (b). Although the isometric invariance of HKS can be effective in analyzing similarity when the topology of protein is subject to deformation, the distribution of different colors will produce more invalid features when Tile-CNN is trained. Consequently, every 3D protein model is converted into a unique color representation according to the following formula, as shown in Fig.2(c).

$$\mathrm{RGB}_{final} = \frac{\sum_i \frac{A_i}{C_{total}} \times RGB_i}{m} \tag{1}$$

where A_i indicates the sum of the number of points in the i^{th} group with the same color. C_{total} denotes the total number of points of the 3D protein model. RGB_i represents the RGB value in the i^{th} group ($i = 1, 2, \ldots, m$) and m is defined as the total number of colors (groups) of the initial protein model.

To well describe the detail of the 3D structure of a protein, a tile data set is constructed by taking the following steps.

Step 1. A 3D structure of a protein model m^i is inputted into the global coordinate system.

Step 2. The 2D color images $x_j^i (j = 1, 2, \ldots, v)$ of the 3D mesh structure of a protein model m^i is obtained, where v represents the different views, as shown in Fig.3. Then, the corresponding width w and height h of the image x_j^i are determined, and the initial cut $x_j^{i_{cut}}$ can be computed by detecting the non-white pixels, as shown in Fig.4.

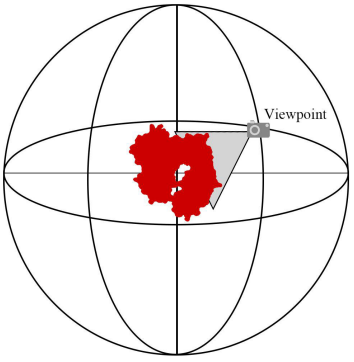


FIGURE 3. Illustration showing viewpoint for obtaining the 2D Images of 3D structure of protein.

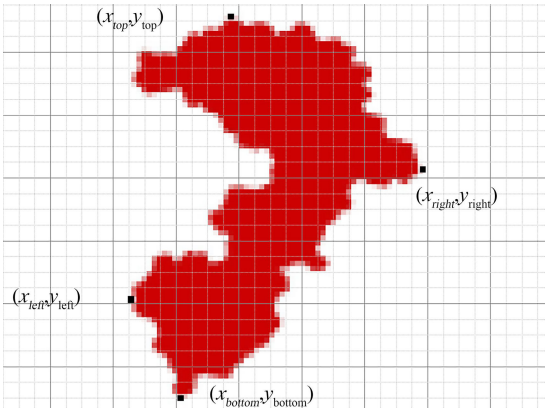


FIGURE 4. Pixel detection.

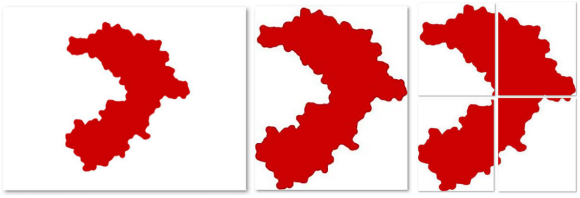


FIGURE 5. The procedure of image processing.

Step3. The central position of the initial cut image $x_j^{i_{cut}}$ is found and the number of tiles t is determined for the cutting of each image. Then the total $v \times t$ tile-images $x_{jk}^{i_{cut}} (j = 1, 2, \ldots, v; k = 1, 2, \ldots t)$ are acquired from the central position. The initial data sets are shown in Fig.5, where the number of tiles of each image is set to 4.

Step4. The proportions of the white pixels $p_{jk_{white}}^{i_{cut}}$ and the non-white pixels $p_{jk_{non-white}}^{i_{cut}}$ in the tile-image $x_{jk}^{i_{cut}}$ are computed. Each image of the initial dataset $\left\{x_{jk}^{i_{cut}} | j = 1, 2, \ldots, v; k = 1, 2, \ldots, t\right\}$ of a protein is judged using the following function

$$f\left(x_{jk}^{i_{cut}}\right) = \begin{cases} 0, & if \left| p_{jk_{non-white}}^{i_{cut}} - p_{jk_{white}}^{i_{cut}} \right| > \gamma^{i_{cut}} \\ 1, & if \left| p_{jk_{non-white}}^{i_{cut}} - p_{jk_{white}}^{i_{cut}} \right| \leq \gamma^{i_{cut}} \end{cases} \tag{2}$$

FIGURE 6. Excluded images from final data sets (The left is white with 94.7%, and the right is colored with 100%).

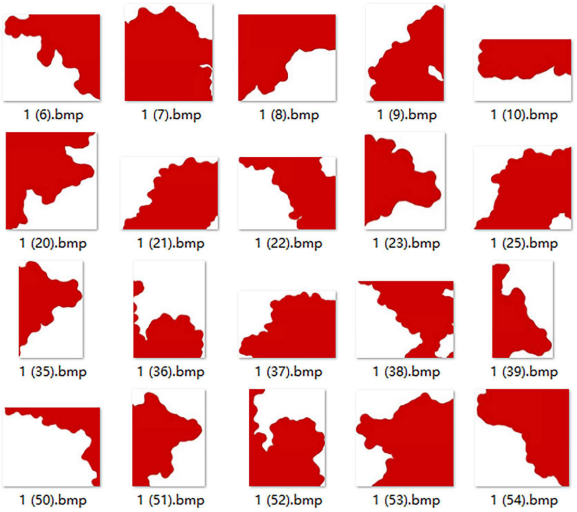


FIGURE 7. Data set of protein 1BPD.

TABLE 1. A Comparison using different images of protein datasets.

Benchmark model	Analyzed model	Tile image	Uncut image
1BPD	2BPG	0.00000079	0.00006754
	2CJW	0.00000136	0.00005407

where $\gamma^{i_{cut}}$ represents a positive threshold which is set as $1-\frac{\sum_j |p^{i_{cut}}_{j_{non-white}} - p^{i_{cut}}_{j_{white}}|}{v}$. If the function value is zero, the tile-image $x^{i_{cut}}_{jk}$ is excluded from the data set, and if the value is one, the tile-image $x^{i_{cut}}_{jk}$ is assigned to the final data sets, as shown in Fig.6.

By repeating step 2 to step 4 until all the tile images are processed, the final data set of a protein can be acquired, as shown in Fig.7.

A similarity analysis comparison is performed when the uncut images and the tile-images are input into the same CNN network, which is presented in Table 1. A similarity measure value that is closer to 0 suggests that the protein model is more similar to the benchmark model. It is already known that 2BPG, 2CJW and 1BPD are similar proteins. Besides, in comparison with 2CJW, 2BPG is more similar to 1BPD [1]. From Table 1, it can be seen that the value 0.00005407 of protein 2CJW is smaller than the value 0.00006754 of protein 2BPG, which indicates that it is difficult for the uncut image input to obtain satisfactory similarity result, while the tile image input can ensure the correctness of similarity result.

TABLE 2. Ratio of invalid images of date sets by cutting different tiles.

model	4-Tiles	8-Tiles	16-Tiles
1BPD	13.97%	36.25%	51.89%
1WRP	2.20%	23.17%	40.19%
2BPG	0.00%	17.45%	34.05%
3WRP	0.27%	14.97%	34.80%

In addition, the ratio of invalid images of data set cut by different degrees in the same CNN network is also compared, as shown in Table 2. When the images are cut into 4-tiles, 8-tiles and 16-tiles, the average proportions of invalid images of the data sets of proteins are 4.11%, 22.96% and 40.2325%, respectively. The more tiles of the image are cut, the higher the ratio of invalid images is. It is discovered that the local and overall features of the protein model can be well described when cutting 4-tiles. Therefore, a Tile-CNN with 4-tiles of data set is trained in our experiment.

B. SIMILARITY MEASUREMENT

The outline of the proposed Tile-CNN is shown in Fig.8, which includes the complete process of similarity analysis of 3D protein structures. It is divided into data set construction generating 2D images from 3D proteins by the multi-view and tile strategy, Tile-CNN training for obtaining the probability matrix for each category and testing by the RMSD computation between the benchmark matrix and the analysis matrix. The proposed Tile-CNN is a special form of the standard CNN, which is composed of five convolutional layers, five pooling layers, two fully connected layers and one output layers. The convolution layers and pooling layers mainly focus on the feature extraction and feature compression. The fully-connected layer and output layer perform the classification. The difference of Tile-CNN proposed in this paper is that the number of layers of Tile-CNN is less than that of GoogleNet and other CNNs, which can save lots of training time. In order to achieve better classification accuracy in the testing phase than other methods, the tile data set is constructed in the training phase by this Tile-CNN with shallow layers.

The similarity analysis in our algorithm includes two steps: (1) Probability matrix output based on a trained Tile-CNN. (2) Similarity determination between the test protein and the benchmark protein based on an RMSD computation by the probability vectors.

Let $\{m^i\}_{i=1}^n$ and $\{y^l\}_{l=1}^n$ be a set of proteins and a set of known labels. A set of multi-views of 2D protein tile-images is denoted by $\{x^{i_{cut}}_{jk}\}$. The main steps of our similar analysis are

(1) Training Phase

Step 1. A protein m^i from each category of a protein data set is randomly chosen as the benchmark protein and a tile-image data set $\{x^{i_{cut}}_{jk}\}$ is generated.

Step 2. Each image of data set $\{x^{i_{cut}}_{jk}\}$ is inputted into the Tile-CNN for the training.

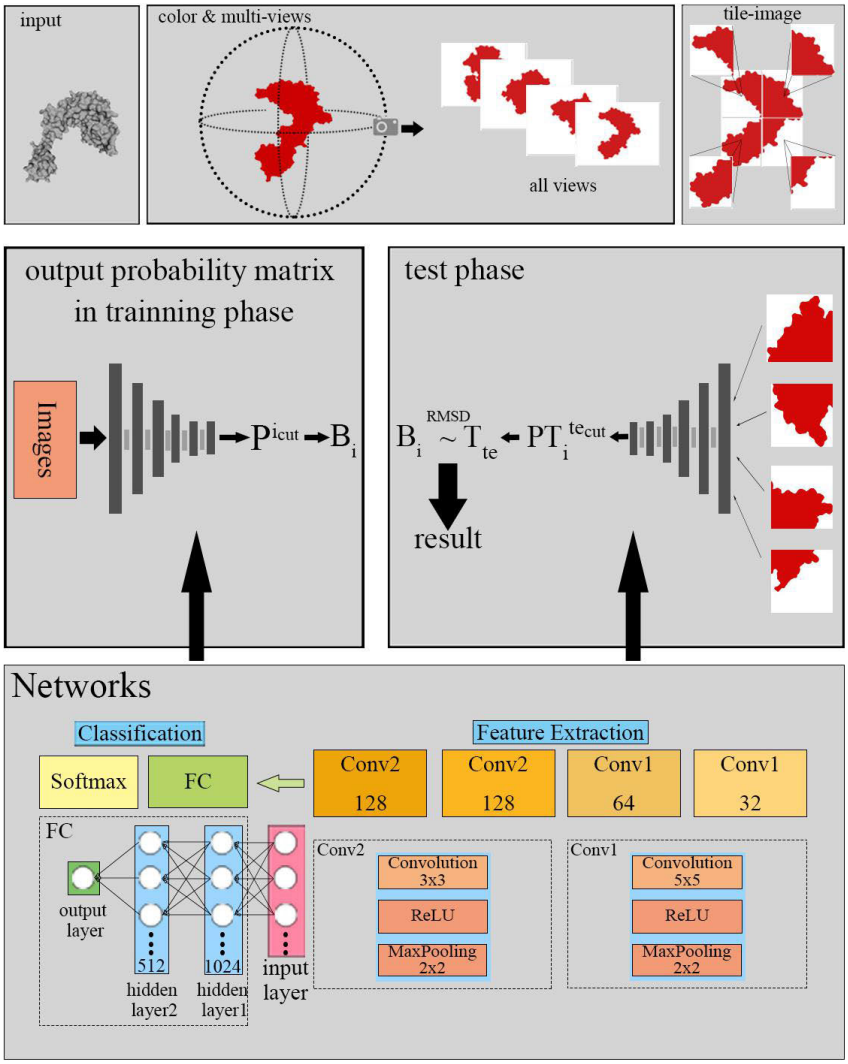


FIGURE 8. Schematic of the proposed Tile-CNN.

Step 3. A probability matrix $P^{i_{cut}}_{(l\times n)}$ in relation to the i^{th} category is constructed

$$P^{i_{cut}} = \begin{pmatrix} norm(p_1^{i_{cut}}) \\ norm(p_2^{i_{cut}}) \\ \vdots \\ norm(p_l^{i_{cut}}) \end{pmatrix} \tag{3}$$

where $p_k^{i_{cut}} = \left(p_{k1}^{i_{cut}}, p_{k2}^{i_{cut}}, \cdots, p_{kn}^{i_{cut}}\right), k = 1, 2, \ldots, l, norm$ represents the normalization of the vector $p_k^{i_{cut}}$, $l = v\times t$ indicates the total number of images of the protein m^i, and n refers to the total number of classification of trained proteins. For the total n categories of protein dataset, n probability matrices are obtained.

(2) Testing Phase

Step 1. A tile-image data set of test protein m^{te} is inputted into a trained Tile-CNN and a probability matrix $PT_i^{te_{cut}}$ corresponding to the classification category i is obtained.

$$PT_i^{te_{cut}} = \begin{pmatrix} norm\left(pt_1^{te_{cut}}\right)_i \\ norm\left(pt_2^{te_{cut}}\right)_i \\ \vdots \\ norm(pt_{l'}^{te_{cut}})_i \end{pmatrix} \tag{4}$$

where $p_{k'}^{te_{cut}} = \left(p_{k'1}^{te_{cut}}, p_{k'2}^{te_{cut}}, \cdots p_{k'n}^{te_{cut}}\right), k' = 1, 2, \ldots, l', 0 < l' \leq l.l$ denotes the total number of images of the protein $m^i.l'$ indicates the total number of images which be classified into the category i using the Tile-CNN. If l' equals 0, the matrix $PT_i^{te_{cut}}$ is non-existent, which implies that no image is classified into category i.

Step 2. The test protein is required to identify the similar shape features to the benchmark protein, which means it is necessary for each image of the test protein to find a similar image of the benchmark protein. Consequently, it defines a benchmark matrix B_i and a test matrix T_{te}, which are shown

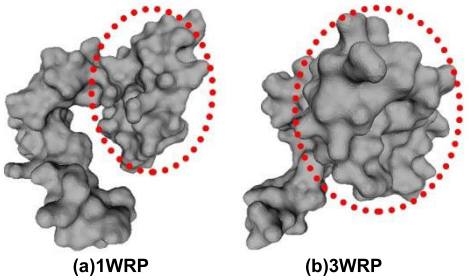


FIGURE 9. Deformation of similar 3D Structures of proteins. Note that the protein 1WRP is considered to be similar to protein 3WRP, but their 3D structure models are quite different (3WRP is deformed from 1WRP).

as

$$B_i = (B_{i1}, B_{i2}, \ldots, B_{il'})^T \tag{5}$$

$$T_{te} = PT_i^{te_{cut}} \tag{6}$$

where $B_{ik'}(k' = 1, \cdots, l')$ is set as the corresponding vector $p_k^{i_{cut}}$ when the condition $min\{\left\|(p_{k'}^{te_{cut}})_i - p_k^{i_{cut}}\right\|, k = 1, \ldots, l\}$ is satisfied.

Step 3. The final similarity result is referred to as the RMSD formula [11] combined with the benchmark matrix and the test matrix. As the images of test protein may be classified into multi-categories, it defines a formula to compute RMSD about the category i by

$$R_i = \frac{l'}{l} \times \sqrt{\frac{\sum_{k'=1}^{l'} (p_{k'}^{te_{cut}} - B_{ik'})}{l'}} \tag{7}$$

When the images are divided into different categories, the minimum RMSD value $\min\{R_i, i = 1 \cdots, n\}$ corresponding to the category i will be regarded as the final category result and the sum RMSD value $\sum R_i$ will be treated as the final RMSD result about the test protein in relation to the final category result. If this sum RMSD value is close to 0, it means that the 3D structure of this protein to be identified is similar to category i. It is noteworthy that if the RMSD results of different categories are identical, the test proteins will be classified into each category, which does not occur in our experiment.

III. EXPERIMENTAL RESULT

Our method is implemented on a Intel(R) Corm(TM) i7-7700 CPU @3.6 Ghz with 32 GB RAM running Windows 10. The environment of experiment is based on python and performed on Titan Xp NVIDIA. The protein models are selected from the protein data bank (http://www.rcsb.org/pdb), the Skolnick Protein Datasets [12] and SCOP [13].

Firstly, an analysis is conducted of the similarity between two groups of similar proteins, 1BPD and 2BPG (Fig. 1 (a-b)), 1WRP and 3WRP (Fig.9), which are frequently tested in other methods [1], [3], [7] as they have the similar sequences but different structures. In order to validate the proposed method for the proteins with similar structures in different sequences, another group of similar proteins 2CJW and 1BPD is tested as well in this experiment. The results

TABLE 3. The results of similar proteins.

model	1BPD	2BPG	1WRP	3WRP	2CJW
1BPD	0.00	0.00000079	NaN	NaN	0.00000136
1WRP	NaN	NaN	0.00	0.00000012	NaN

Note: the value "NaN" indicates that the protein has not been classified into this category by Tile-CNN.

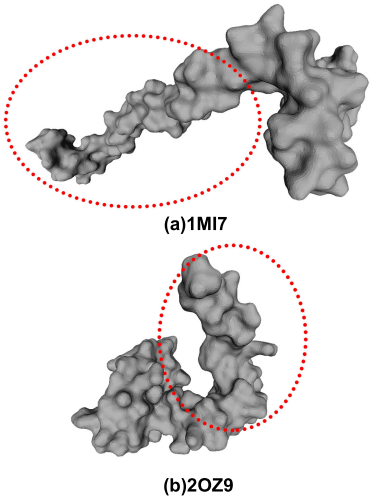


FIGURE 10. Proteins with large deformation. Note that the protein 1MI7 is considered to be similar to protein 2OZ9, but their 3D structure models are quite different (2OZ9 is largely deformed from 1MI7).

are indicated in Table 3. It is found out that the result of protein 2BPG to 1BPD is 0.00000079, while the result of protein 2CJW to 1BPD is 0.00000136. These values reveal that the structures of proteins 2BPG and 2CJW are similar to protein 1BPD. Meanwhile, the result of protein 2BPG is less compared to protein 2CJW, which implies that protein 2BPG is more similar to protein 1BPD. Similarly, the result of protein 3WRP to 1WRP is 0.00000012, which indicates that protein 3WRP is similar to protein 1WRP.

Furthermore, a pair of proteins 1MI7 and 2OZ9 are selected for testing. Reference [1] indicated that despite a significant deformation between two proteins which is more severe than proteins 1BPD and 2BPG (as shown in Fig.10), they are regarded as similar proteins. In our experiment, the result is 0.0000036, which is very close to 0, suggesting that protein 2OZ9 can be classified into the same category as 1MI7 by the Tile-CNN, which means the proposed algorithm is feasible for the similarity analysis of 3D protein structures with a significant deformation.

Secondly, the proposed algorithm (A) is compared with other popular algorithms, including algorithms B(FATCAT [14]), C(CE [14]), D(TM-align [15]), E (Superpose [16]), F(iPBA [17]), G(TM-Score [18]) and H (Skeleton based shape analysis [3]). A similar group of proteins (1WRP and 3WRP) and a dissimilar group of proteins (1WRP and 2CJW) are selected for validation by the RMSD results, as shown in Fig.11. In most cases, dissimilar proteins will challenge the algorithm with respect to the aforementioned

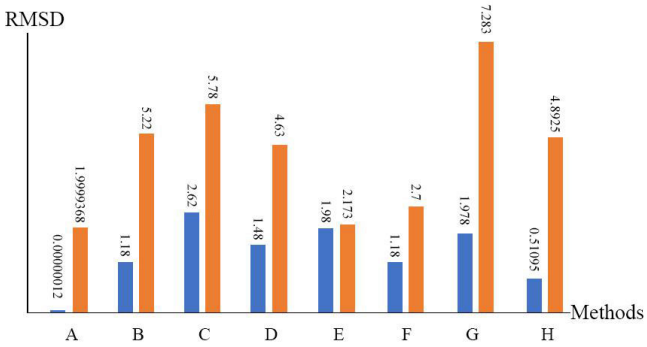
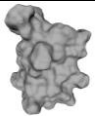
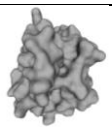
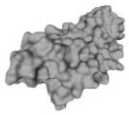
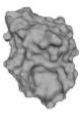
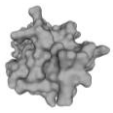


FIGURE 11. Results among the different algorithms. (The left(blue) is the results of similar proteins; the right(orange-yellow) is the result of dissimilar proteins).

TABLE 4. Ratio results of similar and dissimilar structures.

	A	B	C	D	E	F	G	H
ratio	0.00000006	0.23	0.45	0.32	0.91	0.44	0.27	0.10

TABLE 5. Three groups of similar proteins.

A	B	C
 1RCD	 2B3I	 1DBW
 1IER	 1NIN	 1B00

two problems. The ratio results of similar and dissimilar structures of proteins are presented in Table 4, from which it can be seen that our results are closer to 0, which evidences that this algorithm has a clear distinction in the similar and dissimilar structures of proteins.

Thirdly, the algorithm is verified in the different protein data sets. Three groups of similar proteins are selected from Skolnick protein data sets, as shown in Table 5, where the proteins in the first line are treated as the benchmark proteins. For proteins 1IER and 1NIN, they fall into the right categories which are similar to 1RCD and 2B3I respectively, because their corresponding RMSD results are 0.002 and 0.00033, respectively, which are close to zero. For protein 1B00, it is also classified into the correct category (protein 1DBW) since the RMSD result is 0.01048, which is also close to 0. Overall, our results are consistent with those of R [12].

Reference [19] proposed a similarity analysis method based on the different styles of proteins, as shown in Fig.12, and obtained the comparative results using the multi-view convolutional neural networks and multi-model joint networks in two protein data sets, including Fold95 and

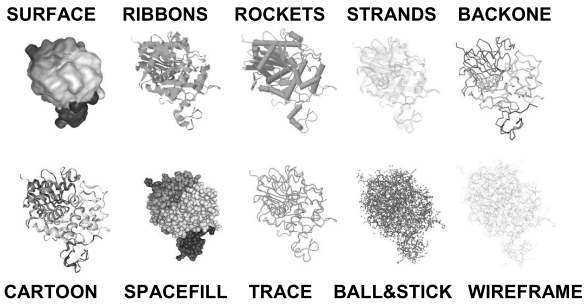


FIGURE 12. Different styles of protein 1WYB.

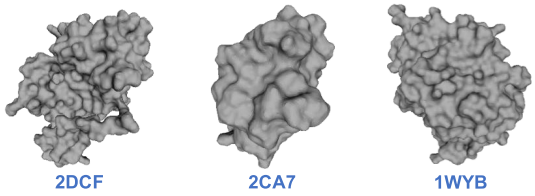


FIGURE 13. Proteins with wrong classification.

Class700. Fold95 involves 95 protein structures that have no greater than 10% sequence identity. There are five classes depending on their fold similarity: (1) f.1 toxins' membrane translocation domains; (2) f.17 transmembrane helix hairpin; (3) f.21 heme-binding four-helical bundle; (4) f.23 single transmembrane helix; and (5) f.4 transmembrane beta-barrels. Class700 contains 700 proteins with at most 20% sequence identity, for which these proteins are equally divided into seven classes [13]: (1) α-proteins; (2) β-proteins; (3) α/β-proteins; (4) $\alpha + \beta$-proteins; (5) multi-domain proteins that have multi-functions; (6) membrane and cell surface proteins; (7) small proteins. Meanwhile, the Fold95 data set is part of the sixth category of the Class700 data set.

According to our method, 5 and 10 proteins are randomly selected as the benchmark proteins separately from each category of two protein data sets to train the Tile-CNN, and then the remaining proteins are inputted into the Tile-CNN for testing. It is found out by RMSD that most of the experimental results are accurate, despite some wrong classifications that are insignificant. In Fig.14, the red box including the value shows that the protein is classified into the wrong category. For example, protein 2E75 is known to be similar to protein 1VF5 in the Fold95 data set. However, it is classified into the category of protein 1BXW according to the RMSD value. In the Class 700 data set, proteins 2DCF and 1WYB are known to be of the same category. In our experiment, however, protein 2DCF is misclassified into the category of protein 2CA7, since they are the most similar structure images in different views during our processing of Tile-CNN, as shown in Fig. 13.

Next, the average accuracy is computed by selecting different benchmark proteins on the two protein datasets. Consequently, our results are compared with other algorithms based on these proteins, including A(Our method), B(AC method [20]), C(QRC method [21]), D(TXT method [22]),

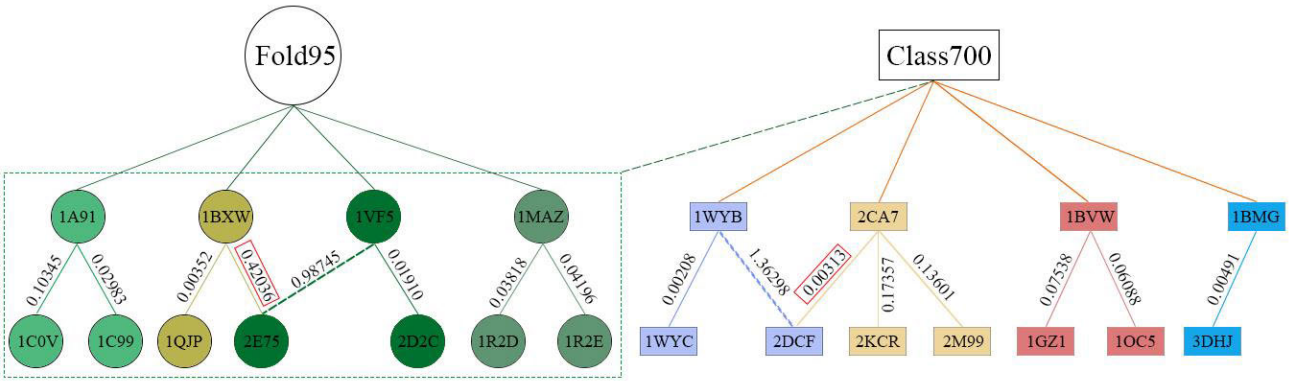


FIGURE 14. Part of experimental results (RMSD results are shown on the line between two proteins).

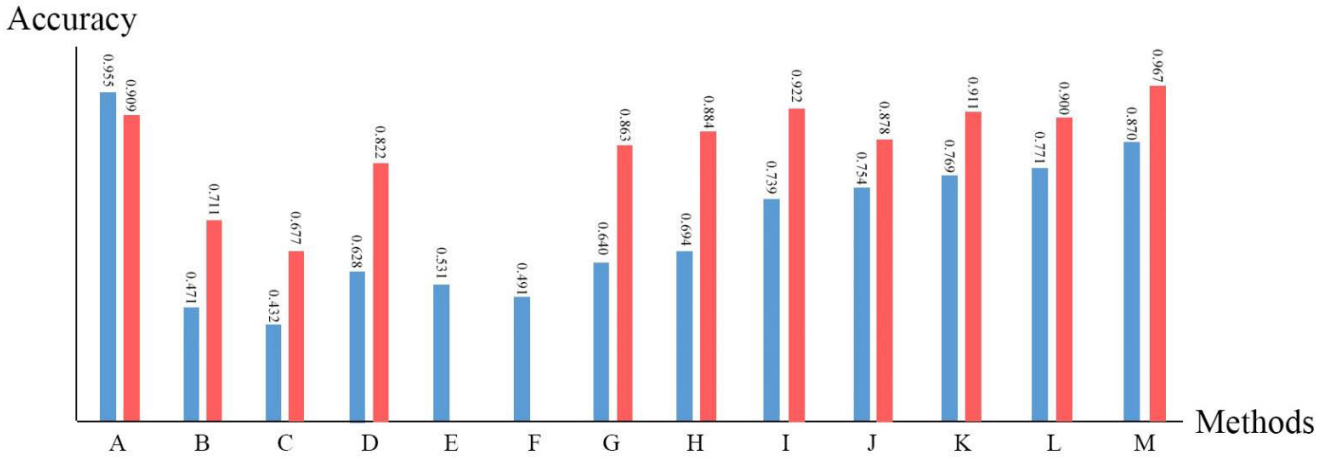


FIGURE 15. Comparison with other methods in Fold95 and Class700 Datasets (The blue represents the average accuracy in Class700 data set, the red is in Fold95 data set).

E (FATCAT method [23]), F(CE method [24]), G(TM method [15]), H(GDA method [25]), I(TOP3-Alex Net [19]), J(TOP3-GoogleNet [19]), K(TOP3-ResNet [19]), L(TOP3-G+R [19]), M(ORACLE [19]), as shown in Fig.15. The algorithms I, J and K are different neural networks for classification with different protein representation types, where TOP3 means the fusion of RIBBONS, ROCKETS, and STRANDS representations of proteins. In algorithm L, TOP3-G+R means the combination of two best ensembles, TOP3-GoogleNet and TOP3-ResNet. In algorithm M, it is obtained by an abstract fusion model known as ORACLE [26]. For the Class 700 data set, the proposed method outperforms all the other approaches in the literature. For the Fold95 data set as a part of the sixth category of the Class700 data set, the classification accuracy of the proposed method is higher than 90% and is superior to most of the other approaches except methods I, K and M. It demonstrates that our method could produce a satisfactory result in rough classification. For fine-gained recognition, however, it would be affected by the capacity of extracting features and the number of layers of CNN. Overall, the average percentage of classification accuracy of two data sets obtained by the Tile-CNN is higher compared to all of the other methods.

Finally, the training and testing time is compared between our Tile-CNN and other CNN models as illustrated in Fig.16. The training time in the first row is acquired by running our algorithm to train and test the Tile-CNN for two data sets. The test time consists of the time of classification and computation RMSD results for each test protein. The last three rows indicate the training and testing time using other CNN methods for two data sets. For these CNN methods, the test time is only given for the classification of 125 images of a test protein. Nevertheless, it is necessary that a series of layers are set for these CNN models to achieve the correct classification. With regard to our proposed Tile-CNN method, the test time is related to the classification of about 8000 images for a test protein, which is truly more time-consuming than other CNN models. However, our Tile-CNN model is capable of using the fewer layers in the network for training and testing, the overall time including the training and testing of our Tile-CNN model is reduced compared to other CNN methods. Besides, it can achieve higher accuracy than the other CNN

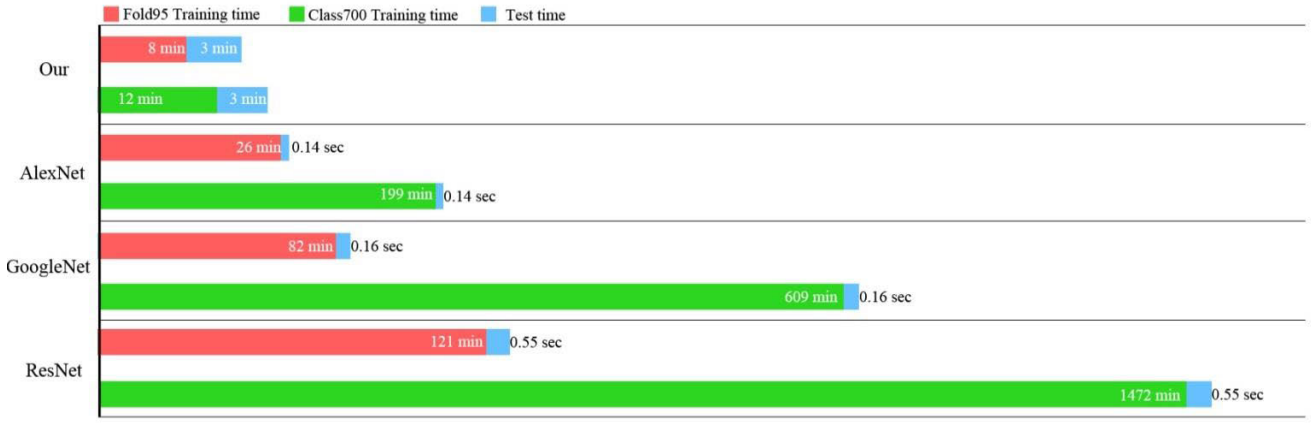


FIGURE 16. Training and testing time.

models since it applies the tile strategy to capture the detailed features of protein models.

IV. CONCLUSION

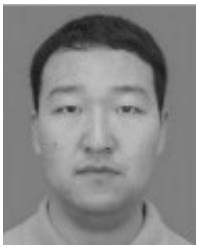
In this paper, a similarity analysis of 3D structures of proteins method based on the Tile-CNN is conducted, in which the constructed features can describe the overall and local features of the protein using the tile strategy. Meanwhile, an analysis matrix combined with the RMSD formula is carried out to determine the similarity analysis between proteins. This algorithm is also compared with other popular algorithms, which leads to the experimental results suggesting that it achieves a desirable performance in similarity analysis for different data sets. Besides, the proposed method is validated by the protein models with significant topological deformation.

In the current work, when the overall and local features of protein images are captured, the different selections for the number of viewpoints and tiles will affect both accuracy and speed during the similarity analysis of 3D protein models. Moreover, there is yet to be a standard to determine the number of layers and other parameter settings in the existing CNN. Therefore, our future work would focus on how to select the suitable feature input for the neural network while reducing the testing time, and on how to choose the appropriate parameters for the purpose of further improving the similarity accuracy of proteins in CNN or other neural networks, such as ResNet, LSTM and Attention Net. Besides, the function analysis of proteins based our 3D structure similarity will also be our future work.

REFERENCES

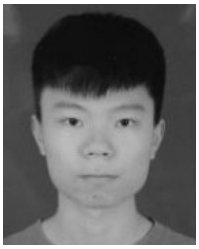
[1] Y. Fang, Y.-S. Liu, and K. Ramani, "Three dimensional shape comparison of flexible proteins using the local-diameter descriptor," *BMC Struct. Biol.*, vol. 9, no. 1, p. 29, 2009.
[2] Y.-S. Liu, Q. Li, G.-Q. Zheng, K. Ramani, and W. Benjamin, "Using diffusion distances for flexible molecular shape comparison," *BMC Bioinf.*, vol. 11, no. 1, pp. 480–494, Sep. 2010.
[3] Z. Li, S. Qin, Z. Yu, and Y. Jin, "Skeleton-based shape analysis of protein models," *J. Mol. Graph. Model.*, vol. 53, pp. 72–81, Sep. 2014.
[4] O. Carugo and S. Pongor, "Protein fold similarity estimated by a probabilistic approach based on cα-cα distance comparison," *J. Mol. Biol.*, vol. 315, no. 4, pp. 887–898, Jan. 2002.
[5] M. Hu and Q. Peng, "Volume fractal dimensionality: A useful parameter for measuring the complexity of 3D protein spatial structures," in *Proc. ACM Symp. Appl. Comput. (SAC)*, 2005, pp. 172–176.
[6] V. Kotlovyi, W. L. Nichols, and L. F. T. Eyck, "Protein structural alignment for detection of maximally conserved regions," *Biophys. Chem.*, vol. 105, nos. 2–3, pp. 595–608, Sep. 2003.
[7] Z. Li, J. Yu, H. Hu, and S. Ji, "Three-dimensional protein shape similarity analysis based on hybrid features," *Gene*, vol. 663, pp. 138–147, Jul. 2018.
[8] K. Mizuguchi and N. Go, "Seeking significance in three-dimensional protein structure comparisons," *Current Opinion Struct. Biol.*, vol. 5, no. 3, pp. 377–382, Jun. 1995.
[9] D. Bostick and I. I. Vaisman, "A new topological method to measure protein structure similarity," *Biochem. Biophys. Res. Commun.*, vol. 304, no. 2, pp. 320–325, May 2003.
[10] J. Hu, "Mining protein contact maps," in *Proc. BIOKDD*, Edmonton, AB, Canada, 2002, pp. 3–10.
[11] M. R. Betancourt and J. Skolnick, "Universal similarity measure for comparing protein structures," *Biopolymers*, vol. 59, no. 5, pp. 305–309, 2001.
[12] D. A. Pelta, J. R. González, and M. M. Vega, "A simple and fast heuristic for protein structure comparison," *BMC Bioinf.*, vol. 9, no. 1, p. 161, Mar. 2008.
[13] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: A structural classification of proteins database for the investigation of sequences and structures," *J. Mol. Biol.*, vol. 247, no. 4, pp. 536–540, Apr. 1995.
[14] S. K. Burley, "RCSB protein data bank: Biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D464–D474, Jan. 2019.
[15] Y. Zhang, "TM-align: A protein structure alignment algorithm based on the TM-score," *Nucleic Acids Res.*, vol. 33, no. 7, pp. 2302–2309, Apr. 2005.
[16] R. Maiti, G. H. Van Domselaar, H. Zhang, and D. S. Wishart, "SuperPose: A simple server for sophisticated structural superposition," *Nucleic Acids Res.*, vol. 32, pp. W590–W594, Jul. 2004.
[17] J.-C. Gelly, A. P. Joseph, N. Srinivasan, and A. G. de Brevern, "IPBA: A tool for protein structure comparison using sequence alignment strategies," *Nucleic Acids Res.*, vol. 39, no. 2, pp. W18–W23, May 2011.
[18] J. Xu and Y. Zhang, "How significant is a protein structure similarity with TM-score= 0.5?" *Bioinformatics*, vol. 26, no. 7, pp. 889–895, Feb. 2010.
[19] L. Nanni, A. Lumini, F. Pasquali, and S. Brahnam, "iProStruct2D: Identifying protein structural classes by deep learning via 2D representations," *Expert Syst. Appl.*, vol. 142, Mar. 2020, Art. no. 113019.
[20] Y.-H. Zeng, Y.-Z. Guo, R.-Q. Xiao, L. Yang, L.-Z. Yu, and M.-L. Li, "Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach," *J. Theor. Biol.*, vol. 259, no. 2, pp. 366–372, Jul. 2009.

[21] L. Nanni and A. Lumini, "An ensemble of K-local hyperplanes for predicting protein-protein interactions," *Bioinformatics*, vol. 22, no. 10, pp. 1207–1210, Feb. 2006.
[22] L. Nanni, A. Lumini, and S. Brahnam, "An empirical study of different approaches for protein classification," *Sci. World J.*, vol. 2014, pp. 1–17, Jun. 2014.
[23] Y. Ye and A. Godzik, "Flexible structure alignment by chaining aligned fragment pairs allowing twists," *Bioinformatics*, vol. 19, no. 2, pp. ii246–ii255, Oct. 2003.
[24] I. N. Shindyalov and P. E. Bourne, "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path," *Protein Eng. Des. Selection*, vol. 11, no. 9, pp. 739–747, Sep. 1998.
[25] C. H. Suryanto, H. Saigo, and K. Fukui, "Structural class classification of 3D protein structure based on multi-view 2D images," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 1, pp. 286–299, Jan. 2018.
[26] L. I. Kuncheva, "A theoretical study on six classifier fusion strategies," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 281–286, Aug. 2002.



SHENGWEI QIN (Member, IEEE) received the M.S. degree in mathematics from Zhejiang Sci-Tech University, Zhejiang, China, in 2015, where he is currently pursuing the Ph.D. degree. From 2017 to 2019, he was a Research Assistant with the Institute of Games, South China Institute of Software Engineering, Guangzhou University, Guangzhou, China. His research interests include bioinformatics and computer graphics.

ZHONG LI received the Ph.D. degree in mathematics from Zhejiang University, in 2003. He was a Postdoctoral Fellow with Shanghai Jiao Tong University, from 2004 to 2006. He was a Visiting Scholar with the University of California at Berkeley, Stanford University, and Harvard University, USA. He is currently a Professor with Zhejiang Sci-Tech University. His current research interests are in bioinformatics and computer graphics. He is a member of the IEEE Computer Society.



LEXUAN HE was born in Qingyuan, Guangdong, China, in 1999. He is currently studying in engineering at the South China Institute of Software Engineering, Guangzhou University, China. From 2018 to 2019, he was with the Bioinformatics and Graphics Team.

WANMIN LIN was born in Guangzhou, Guangdong, China, in 1999. She is currently studying in engineering at the South China Institute of Software Engineering, Guangzhou University, China. From 2018 to 2019, she was with the Bioinformatics and Graphics Team.