Refined Contact Map Prediction of Peptides Based on GCN and ResNet

Jiawei Gu¹, Tianhao Zhang¹, Chunguo Wu¹,², Yanchun Liang¹,²,³ and Xiaohu Shi¹,²,³*

¹College of Computer Science and Technology, University of Jilin, Changchun, China, ²Key Laboratory of Symbolic Computation and Knowledge Engineering, Ministry of Education, Changchun, China, ³School of Computer Science, Zhuhai College of Science and Technology, Zhuhai, China

Predicting peptide inter-residue contact maps plays an important role in computational biology, which determines the topology of the peptide structure. However, due to the limited number of known homologous structures, there is still much room for inter-residue contact map prediction. Current models are not sufficient for capturing the high accuracy relationship between the residues, especially for those with a long-range distance. In this article, we developed a novel deep neural network framework to refine the rough contact map produced by the existing methods. The rough contact map is used to construct the residue graph that is processed by the graph convolutional neural network (GCN). GCN can better capture the global information and is therefore used to grasp the long-range contact relationship. The residual convolutional neural network is also applied in the framework for learning local information. We conducted the experiments on four different test datasets, and the inter-residue long-range contact map prediction accuracy demonstrates the effectiveness of our proposed method.

Keywords: peptide inter-residue contact map prediction, deep learning, graph convolutional network, residual convolutional neural network, multiple sequence alignment

1 INTRODUCTION

Peptides play an important role in computational and experimental biology (Torrisi et al., 2020), which motivates the development of accurate methods to predict their native conformations from the sequences. As a special kind of peptide, protein-related predictions from its amino acid sequence remain an open problem in the field of computational biology. Using biological experiments to determine the protein structure is very cumbersome and expensive. Therefore, it is very effective to use machine learning methods or deep learning methods to obtain a universal law from the amino acid sequence to the prediction of a protein’s three-dimensional structure. The inter-residue contact map (Lena et al., 2012) is a two-dimensional representation of a protein’s three-dimensional structure. The contact map constrains the conformation of protein structures; as a result, accurate prediction of the contact map can facilitate ab initio structure modeling, and the accuracy of the contact map affects the accuracy of the three-dimensional structure of the protein. Furthermore, contact maps have been widely used for model assessment and structure alignment.

The current contact map prediction methods are mainly based on direct coupling analysis (DCA) methods, machine learning methods, and deep learning methods. DCA-based methods mainly use multiple sequence alignment methods to determine the relationships between amino acid pairs. However, DCA-based methods assume that pairs of contacted residues are more likely to mutate simultaneously as the protein structure or function evolves and mainly use the multiple sequence
alignment (MSA) to determine the relationships between the amino acid pairs. Therefore, the accuracy of the DCA-based method depends on the number of homologous protein sequences in the protein sequence library. On the other hand, due to the existence of indirect evolutionary coupling information, the generated coupling information from the DCA might include “noise signal.” The common DCA-based methods include CCMpred (Seemayer et al., 2014), PSICOV (Jones et al., 2012), and GREMLIN (Kamisetty et al., 2013). CCMpred mainly uses Markov random field pseudo-likelihood maximization to learn the contacts between the protein inter-residues. When there are a large number of homologous proteins in the protein sequence, the accuracy of the contact prediction results is higher; however, when the sequences of the homologous protein are fewer, the accuracy is lower. On the other hand, machine learning-based and deep learning-based methods use a set of input features derived from multiple sequence alignments (MSAs) to predict the protein inter-residue contact map, including position-specific scoring matrices (PSSMs), secondary structure (SS) predictions, and solvent accessibility (SA) information. Machine learning-based methods are mainly based on support vector machines (SVMs) (Hearst et al., 1998) to learn the abovementioned features and common support vector machine (SVM) methods including SVMCon (Cheng and Baldi, 2007) and R2C (Yang et al., 2016). SVMCon used support vector machines (SVMs) and yields good performance on medium- to long-range contact predictions. In recent years, deep learning methods have been mainly used to predict the contact map between the protein inter-residues and are mainly based on the structure of the convolutional neural network (CNN) and residual neural network (ResNet) (He et al., 2016). The ResNet structure further improves the CNN structure and solves the problem of reduced accuracy when there are too many convolutional layers through the skip connection mechanism. RaptorX-Contact (Wang et al., 2017) was the first model that used the ResNet structure for protein inter-residue contact map prediction tasks. Zhong Li et al. (Li et al., 2020) used ResNet and DenseNet (Huang et al., 2017) structures and a new protein sequence feature (PSFM) to improve the contact map prediction accuracy. DeepCov (Jones and Kandathil, 2018) applied the CNN to predict contact maps when limited evolutionary information is available, which has been trained on a very limited set of input features: pair frequencies and covariance. It is noticed that there are several similar studies predicting the distance matrix instead of the contact map, such as RaptorX structure prediction (Xu, 2018), PG-GNN (Xia and Ku, 2020), and AlphaFold (Senior et al., 2020).

However, there are two main difficulties in obtaining accurate contact predictions. First, many amino acid sequences lack a large number of homologous sequences, which limits the level of accuracy of predictions. On the contrary, the target sequences with many

---

**FIGURE 1** | Framework of the RCMPM.
homologous sequences might generate “noise signals” from the evolutionary coupling information. Second, most methods use convolutional neural network (CNN)-based models for inter-residue contact map prediction, leading to over-learning of the local information, but under-learning of the long-range information, which is reflected by a low long-range accuracy.

Therefore, eliminating “noise signals” is necessary to improve the residue contact prediction. Improving the inter-residue contact prediction has been of interest for many years due to its critical importance in structure bioinformatics, with either the sequence or structure template information. R2C (Yang et al., 2016) used SVM and PSICOV methods and used a dynamic fusion strategy to predict the contact map between amino acids and applied Gaussian noise filters for further denoising. Amelia Villegas-Morcillo et al. (Villegas-Morcillo et al., 2018) applied K-SVD (Aharon et al., 2006) and deep convolutional neural network (DCNN) methods specially designed for image denoising to solve the problem of Gaussian noise. DNCON2 (Adhikari et al., 2017) adopted the structure of the two-stage convolutional neural networks (CNNs) to improve the contact map prediction, which divides the prediction into two parts. The first part trains five CNNs to predict the contact map between the distances of 6, 7.5, 8, 8.5, and 10, respectively. The second part takes the input feature as the output of the first part and then utilizes a CNN structure for further prediction.

In the past few years, the graph neural network (Zhou et al., 2018) was raised to represent the protein structure in various deep learning-based methods and had succeeded in the computational biology area, such as protein interface prediction, protein solubility prediction, and protein function prediction. Fout et al., (2017) proposed a type of architecture for the task of predicting protein interfaces between the pairs of proteins using a graph representation of the underlying protein structure. GraphSol (Chen et al., 2020) was used to predict the protein residue solubility by combining the predicted contact maps, graph neural networks, and attention mechanisms. DeepFRI (Gligorijević et al., 2021) used an LSTM (Hochreiter and Schmidhuber, 1997) and a graph convolutional network to predict protein functions. PG-GNN (Xia and Ku, 2020) used a new convolution kernel to perform deep convolution to obtain the distance map, which was used to construct an inter-residue graph between the residues for obtaining the dihedral information between residues, and finally constructed a three-dimensional protein structure.

Here, to focus on getting more accurate contact maps, especially on the long-range level, we developed a novel refined contact map prediction model (RCMP) to refine the rough contact map produced by the existing methods, which combines a graph convolution network (GCN) (Kipf and Welling, 2016) and residual convolution neural networks (ResNet) (He et al., 2016). The main contributions of the article are summarized as follows:

- The peptide contact map refinement task is modeled as a geometric 2D graph improvement, with nodes representing the amino acid residues and edges representing contacts between the residues. The rough results of other models such as CCMpred and RaptorX-Contact are used to construct the inter-residue contact graph.
- Aiming at the challenges previously mentioned, a novel deep neural network framework is proposed for the inter-residue contact prediction by combining a graph convolution network (GCN) and residual convolution neural networks (1D ResNet and 2D ResNet), of which the GCN has a strong global information extraction ability, and hence can better capture the long-range contact relationships among the complex sequence inter-residues.
- The experiments are conducted on four different test datasets, and the inter-residue long-range contact map prediction accuracy demonstrates the effectiveness of our proposed method due to the new network architecture.

The rest of the article is organized as follows. Section 2 details the materials and methods, including contact definition, graph construction, feature selection, and the proposed prediction model. Section 3 reports the datasets used in our method, evaluation metrics, and experiments on four test datasets. Section 4 concludes the article and discusses the directions for the future work.

2 MATERIALS AND METHODS

2.1 Contact Definition

In general, two residues are considered to be in contact if certain atoms are close enough to form a molecular interaction. In the Critical Assessment of protein Structure Prediction (CASP) experiment (Moult et al., 2014, 2016, 2018), the contact definition is based on the spatial distance of $C_\beta$ atoms. For instance, assuming that $v = \{v_1, v_2, \ldots, v_i, \ldots, v_j, \ldots, v_L\}$ is the residue sequence, where $L$ is the sequence length, and $(x_{vi}, y_{vi}, z_{vi})$ is the three-dimensional coordinates of amino acid residue $v_i$ then the equation for the distance between the residues $v_i$ and $v_j$ is

$$\text{Distance}(i,j) = \sqrt{(x_{vi} - x_{vj})^2 + (y_{vi} - y_{vj})^2 + (z_{vi} - z_{vj})^2}.$$  \hspace{1cm} (1)

If the Euclidean distance between the $C_\beta$ atoms ($C_\alpha$ for GLY) of two amino acids is less than a given threshold $\gamma$, then the two residues are said to be in contact.

2.2 Graph Construction

As mentioned in Section 2.1, we can use the other contact map prediction models, such as CCMpred (Seemayer et al., 2014) and RaptorX-Contact (Wang et al., 2017), to obtain a contact matrix $CM$. Assuming that the length of the peptide is $L$, then $CM$ is an $L \times L$ matrix, whose element $CM_{ij}$ denotes whether the pair of residues $i$ and $j$ is contacted or not (1 or 0). Denote $G = \{N, E\}$ is the contact graph of the peptide, where $N$ is the node set including $L$ amino acids, and $E$ is the edge set. Then, the contact graph could be constructed as follows:
Algorithm 1. Graph construction.

\[
E = \Phi \\
\text{for } i = 1 \text{ to } L-1 \text{ do} \\
\quad \text{for } j = i + 1 \text{ to } L \text{ do} \\
\quad \quad \text{if } Distance_{ij} < \gamma \text{ then} \\
\quad \quad \quad E = E \cup i_{ij} \\
\quad \quad \text{end if} \\
\quad \text{end for} \\
\text{end for}
\]

where \( l_{ij} \) is the edge between node \( i \) and node \( j \), and the threshold \( y \) is set as 8Å in this article.

### 2.3 Feature Selection

#### 2.3.1 Sequence Features

We devised three groups of sequence features to train our model, namely, the position-specific scoring matrix (PSSM), secondary structure (SS), and solvent accessibility (SA). The PSSM is a widely used sequence feature, which is produced by executing PSI-BLAST (Altschul et al., 1997) on the UniRef90 database (Suzek et al., 2015) with 0.001 e-value after the three iterations, which is a 20-dimensional profile feature for each residue. The secondary structure and solvent accessibility are predicted by the RaptorX-Property (Wang et al., 2016) program (http://raptorx.uchicago.edu/StructurePropertyPred/predict/). The secondary structure is divided into three categories, namely, helix (H), strand (E), and coil (C), and the solvent accessibility is also classified into three types, namely, buried, medium, and exposed. The PSSM is represented as a two-dimensional matrix of \( L \times 20 \), while both the secondary structure and solvent accessibility are represented as a two-dimensional matrix of \( L \times 3 \); therefore, the concatenation sequence embedding vector \( X_{seq} \) is obtained with the \( L \times 26 \) dimension, where the order of the splicing input is [PSSM, secondary structure, and solvent accessibility].

#### 2.3.2 Pairwise Features

Pairwise features are the information that characterizes the relationship between the pairs of residues, including the co-evolutionary information, statistical information, and so on. Four groups of pairwise features are used to train our model, namely, RaptorX-Contact prediction, CCMpred prediction, mutual information (Dunn et al., 2008), and contact potential (Betancourt and Thirumalai, 1999), which provide the co-evolutionary information for each pair of alignment columns. RaptorX-Contact and CCMpred prediction are mainly used as inter-residue scores. RaptorX-Contact prediction results can be obtained by model training, the source code of which can be downloaded from https://github.com/j3xugit/RaptorX-Contact. CCMpred prediction results can be obtained by the CCMpred program, which could be accessed at https://github.com/sodinglab/CCMpred. However, CCMpred requires the homologous sequence result of the multiple sequence alignment (MSA) as the input, which is produced by executing the HHblits program (Remmert et al., 2012) on the Uniclust30 database (Mirdita et al., 2017) with 0.001 e-value after three iterations. Both RaptorX-Contact and CCMpred output an inter-residue score for each residue pair. After the MSA profile is obtained, the mutual information could be defined by

\[
MI_{ij} = \sum_{x,y} p_{ij}(x,y) \ln \frac{p_{ij}(x,y)}{p_i(x)p_j(y)}
\]

where \( R \) is the set of amino acid types, \( x \) and \( y \) are the elements in column \( i \) and column \( j \), respectively, \( p_i(x) \) and \( p_j(y) \) indicate the probabilities of residue \( x \) in column \( i \) and residue \( y \) in column \( j \), and \( p_{ij}(x,y) \) is the probability that residue \( x \) is in column \( i \) and residue \( y \) is in column \( j \), respectively. Normalized mutual information, namely, average product correct (APC) mutual information is also used in our method, which is defined by

\[
MI_{ij}^{APC} = MI_{ij} - APC_{ij},
\]

\[
APC_{ij} = \frac{\sum_{i,j} MI_{ij} \sum_{i,j} MI_{ij}}{\sum_{i,j} MI_{ij}}.
\]

The contact potential is computed by averaging the contact potential terms across the two alignment columns. Mutual information and contact potential are generated by alnstats in the MetaPSICOV (Jones et al., 2015) program, which also requires the homologous sequence as the input. For RaptorX-Contact prediction, CCMpred prediction, mutual information, APC mutual information, and contact potential, all are represented as a three-dimensional matrix of \( L \times L \times 1 \); therefore, the concatenation pair-wise embedding features \( X_{pair} \) are obtained with the \( L \times L \times 5 \) dimension, where the order of the splicing input is [RaptorX-Contact prediction, CCMpred prediction, MI, APC MI, and contact potential].

### 2.4 Prediction Model

#### 2.4.1 The Framework of the RCMPM Model

Residual networks (ResNets) are very helpful for accurate peptide contact map prediction, which has been demonstrated in the RaptorX-Contact model (Wang et al., 2017). Therefore, ResNet architecture is retained in our proposed refined contact map prediction model (RCMPM). On the other hand, the rough contact map obtained by the other methods could be well utilized by transferring it into an amino acid graph, and therefore, the graph convolution network (GCN) could handle the graph topology very well. Hence, the proposed RCMPM model includes a GCN module, a 1D ResNet module, and a 2D ResNet module, respectively.

Figure 1 shows the framework of the RCMPM model, which has two types of features, namely, sequence features and pairwise features. The GCN module is used to learn the global structural features of the inter-residue contact graphs, whose input is the node representation of the sequence features, and the output is a dense global structural embedding vector for each amino acid node. 1D ResNet module is used to handle the one-dimensional sequence feature and output a sequence embedding vector for each amino acid. 2D ResNet module integrates the above two modules’ outputs and the pairwise features as well and finally generates the refined contact map.
The following part of this section will describe these three modules in detail.

### 2.4.2 GCN Module

Given a sequence with $L$ residues, the residue graph can be represented by a contact map, that is, the nodes of the graph are the residues of the peptide, and the features of the nodes are represented by the attributes of the residues. The edges of the contact graph indicate whether there are connections between the amino acid nodes, and the weight of the edge represents the probability of contact. We used the graph convolution network (GCN) to obtain the global structural features of the graph.

The graph convolutional layer in the prediction model uses the following equation:

$$ H^{(l+1)} = \sigma(\tilde{A}H^{(l)}W^{(l)}), $$

where $\tilde{A} = A + I_L$ is the variant of the adjacency matrix by adding the self-loop identity matrix $I_L$ on the original adjacency matrix $A$, and $H^{(l)}$ is the hidden matrix learned by the $l$th layer, initial of which is the hidden matrix $H^{(0)} = X_{seq}$. $W^{(l)}$ is a weight matrix of the layer-specific trainable parameters and is used to map the iterations to a low-dimensional rich information space, and $\sigma$ is a nonlinear activation function, which is taken as the ReLU function in our model. We also use normalization to map the input feature of each layer $H^{(l)}$ to $[0,1]$ to improve the data performance and reduce errors. Finally, we used a 2-layer graph convolutional network to learn the global structural features of the contact graph containing amino acid node features. Hence, the final output of the GCN module in the RCMPM model uses the following equation:

$$ X_{GCN} = \text{RELU}(\tilde{A}\text{ReLU}(\hat{A}X_{seq}W^{(0)})W^{(1)}). $$

### 2.4.3 1D ResNet Module

A 1D ResNet module is used to handle the one-dimensional sequence feature and outputs a sequence embedding vector for each residue, which is stitched together by the residual blocks. A residual block consists of two convolutional layers and two activation layers, which can be defined as follows:

$$ X_{l+1} = F(X_l, W_l) + X_l, $$

where $X_l$ and $X_{l+1}$ are the input and output vectors of the residual block, respectively, and the initial hidden matrix $X_0 = X_{seq}$. Here, $W_l$ is the weight matrix in convolutional layers of the $l$th block, and $F(X_l, W_l)$ represents the result after the action of the convolutional layer and activation function layer. Here, the operation of the convolutional layer is implemented by the conv1d function of the tensorflow framework. Here, we used the ReLU function as the activation function of our method and also used normalization to map the data to $[0,1]$ to improve data performance and reduce errors. We kept the dimension of $X_{l+1}$ larger than $X_l$ because the higher dimension can carry more information. For a residual block, the $F(X_l, W_l)$ function can be expressed as shown in Figure 2.

Finally, the output of the 1D ResNet module in the RCMPM model could be described as follows:

$$ X_{1DResNet} = \sum_{i=0}^{n} F(X_i, W_i) + X_i. $$

In our 1D ResNet module, the number of residual blocks is selected as 3.

### 2.4.4 2D ResNet Module

The 2D ResNet module is used to learn the final contact relationship for each residue pair by integrating the aforementioned two modules, namely, that it takes the input of the output feature $X_{GCN}$ of the GCN module and the output feature $X_{1DResNet}$ of the 1D ResNet module and the pairwise feature $X_{pair}$ as well. Different with the 1D ResNet module, the 2D ResNet module is dealing with two-dimensional feature maps. The pairwise features $X_{pair}$ is of $L \times L \times 5$ dimension, as described in section 2.3.2, while the output features $X_{GCN}$ and $X_{1DResNet}$ are the one-dimensional feature map with the same dimension $L \times n$, which should be converted to a two-dimensional feature map. Similarly with the method used in (Wang et al., 2017), $X_{GCN}$ and $X_{1DResNet}$ are first concatenated on the second dimension, obtaining an $L \times 2n$ feature map $X_{1DFinal}$:

$$ X_{1DFinal} = X_{1DResNet} \oplus X_{GCN}. $$

Then, it is converted to a 2-dimensional feature map. Redefined $X_{1DFinal}$ from $L \times 2n$ to $L \times 1 \times 2n$ dimension by adding a second-order dimension with 1, then duplicate $X_{1DFinal}$ $L$ times to extend the second order from 1 to $L$, getting an $L \times L \times 2n$ tensor $TX_{GCN,1D}$. $TX_{GCN}$ is denoted as the transpose of $TX_{GCN,1D}$ on the first two orders; $X_{GCN}$ and $X_{1DResNet}$ are finally integrated as $\text{Integrate}_{GCN}$:

$$ \text{Integrate}_{GCN} = TX_{GCN,1D} \oplus TX_{1DResNet}. $$
TABLE 1 | Contact map results by four different methods on the PDB25 testing dataset.

| Method                | Long-range | Medium-range | Short-range |
|-----------------------|------------|--------------|-------------|
|                        | L/10       | L/5          | L/2         | L           |
|                        |            |              |             |             |
| CCMpred                | 0.528      | 0.475        | 0.361       | 0.257       |
| R2C                   | 0.666      | 0.667        | 0.648       | 0.449       |
| RaptorX-Contact       | 0.774      | 0.739        | 0.633       | 0.497       |
| RCMPM (CCMpred)       | 0.718      | 0.685        | 0.582       | 0.446       |
| RCMPM (RaptorX-Contact)| 0.784      | 0.748        | 0.646       | 0.508       |

The training dataset is a subset of PDB25 extracted from the PDB database (http://www.rcsb.org/pdb/home/home.do) with homology reduction at 25% level of sequence identity, resulting in 6767 non-homologous protein sequences. The number of amino acids of each training protein ranges from 26 to 300. To avoid overfitting, 400 proteins are randomly chosen for validation and the remaining others for training. To evaluate the performance of our model, it is applied to four testing datasets. The first testing dataset is the PDB25 dataset, which contains 500 nonhomologous protein sequences. The training set, validation dataset, and testing dataset of the abovementioned PDB25 dataset can be downloaded from http://raptorx.uchicago.edu/ContactMap/. The other three datasets were obtained from three CASP (Critical Assessment of Structure Prediction) competitions (CASP10 (Moult et al., 2014), CASP11 (Moult et al., 2016), and CASP12 (Moult et al., 2018)). For the three CASP datasets, we used the same screening method as that used in the R2C method (Yang et al., 2016). For the CASP10 dataset, the sequence data could be accessed on the website of https://predictioncenter.org/download_area/CASP10/targets/.

The training dataset is a subset of PDB25 extracted from the PDB database (http://www.rcsb.org/pdb/home/home.do) with homology reduction at 25% level of sequence identity, resulting in 6767 non-homologous protein sequences. The number of amino acids of each training protein ranges from 26 to 300. To avoid overfitting, 400 proteins are randomly chosen for validation and the remaining others for training.

To evaluate the performance of our model, it is applied to four testing datasets. The first testing dataset is the PDB25 dataset, which contains 500 nonhomologous protein sequences. The training set, validation dataset, and testing dataset of the abovementioned PDB25 dataset can be downloaded from http://raptorx.uchicago.edu/ContactMap/. The other three datasets were obtained from three CASP (Critical Assessment of Structure Prediction) competitions (CASP10 (Moult et al., 2014), CASP11 (Moult et al., 2016), and CASP12 (Moult et al., 2018)). For the three CASP datasets, we used the same screening method as that used in the R2C method (Yang et al., 2016). For the CASP10 dataset, the sequence data could be accessed on the website of https://predictioncenter.org/download_area/CASP10/targets/.

The training dataset is a subset of PDB25 extracted from the PDB database (http://www.rcsb.org/pdb/home/home.do) with homology reduction at 25% level of sequence identity, resulting in 6767 non-homologous protein sequences. The number of amino acids of each training protein ranges from 26 to 300. To avoid overfitting, 400 proteins are randomly chosen for validation and the remaining others for training.

To evaluate the performance of our model, it is applied to four testing datasets. The first testing dataset is the PDB25 dataset, which contains 500 nonhomologous protein sequences. The training set, validation dataset, and testing dataset of the abovementioned PDB25 dataset can be downloaded from http://raptorx.uchicago.edu/ContactMap/. The other three datasets were obtained from three CASP (Critical Assessment of Structure Prediction) competitions (CASP10 (Moult et al., 2014), CASP11 (Moult et al., 2016), and CASP12 (Moult et al., 2018)). For the three CASP datasets, we used the same screening method as that used in the R2C method (Yang et al., 2016). For the CASP10 dataset, the sequence data could be accessed on the website of https://predictioncenter.org/download_area/CASP10/targets/.

The training dataset is a subset of PDB25 extracted from the PDB database (http://www.rcsb.org/pdb/home/home.do) with homology reduction at 25% level of sequence identity, resulting in 6767 non-homologous protein sequences. The number of amino acids of each training protein ranges from 26 to 300. To avoid overfitting, 400 proteins are randomly chosen for validation and the remaining others for training.

To evaluate the performance of our model, it is applied to four testing datasets. The first testing dataset is the PDB25 dataset, which contains 500 nonhomologous protein sequences. The training set, validation dataset, and testing dataset of the abovementioned PDB25 dataset can be downloaded from http://raptorx.uchicago.edu/ContactMap/. The other three datasets were obtained from three CASP (Critical Assessment of Structure Prediction) competitions (CASP10 (Moult et al., 2014), CASP11 (Moult et al., 2016), and CASP12 (Moult et al., 2018)). For the three CASP datasets, we used the same screening method as that used in the R2C method (Yang et al., 2016). For the CASP10 dataset, the sequence data could be accessed on the website of https://predictioncenter.org/download_area/CASP10/targets/.

The training dataset is a subset of PDB25 extracted from the PDB database (http://www.rcsb.org/pdb/home/home.do) with homology reduction at 25% level of sequence identity, resulting in 6767 non-homologous protein sequences. The number of amino acids of each training protein ranges from 26 to 300. To avoid overfitting, 400 proteins are randomly chosen for validation and the remaining others for training.

To evaluate the performance of our model, it is applied to four testing datasets. The first testing dataset is the PDB25 dataset, which contains 500 nonhomologous protein sequences. The training set, validation dataset, and testing dataset of the abovementioned PDB25 dataset can be downloaded from http://raptorx.uchicago.edu/ContactMap/. The other three datasets were obtained from three CASP (Critical Assessment of Structure Prediction) competitions (CASP10 (Moult et al., 2014), CASP11 (Moult et al., 2016), and CASP12 (Moult et al., 2018)). For the three CASP datasets, we used the same screening method as that used in the R2C method (Yang et al., 2016). For the CASP10 dataset, the sequence data could be accessed on the website of https://predictioncenter.org/download_area/CASP10/targets/.

The training dataset is a subset of PDB25 extracted from the PDB database (http://www.rcsb.org/pdb/home/home.do) with homology reduction at 25% level of sequence identity, resulting in 6767 non-homologous protein sequences. The number of amino acids of each training protein ranges from 26 to 300. To avoid overfitting, 400 proteins are randomly chosen for validation and the remaining others for training.

To evaluate the performance of our model, it is applied to four testing datasets. The first testing dataset is the PDB25 dataset, which contains 500 nonhomologous protein sequences. The training set, validation dataset, and testing dataset of the abovementioned PDB25 dataset can be downloaded from http://raptorx.uchicago.edu/ContactMap/. The other three datasets were obtained from three CASP (Critical Assessment of Structure Prediction) competitions (CASP10 (Moult et al., 2014), CASP11 (Moult et al., 2016), and CASP12 (Moult et al., 2018)). For the three CASP datasets, we used the same screening method as that used in the R2C method (Yang et al., 2016). For the CASP10 dataset, the sequence data could be accessed on the website of https://predictioncenter.org/download_area/CASP10/targets/.

The training dataset is a subset of PDB25 extracted from the PDB database (http://www.rcsb.org/pdb/home/home.do) with homology reduction at 25% level of sequence identity, resulting in 6767 non-homologous protein sequences. The number of amino acids of each training protein ranges from 26 to 300. To avoid overfitting, 400 proteins are randomly chosen for validation and the remaining others for training.

To evaluate the performance of our model, it is applied to four testing datasets. The first testing dataset is the PDB25 dataset, which contains 500 nonhomologous protein sequences. The training set, validation dataset, and testing dataset of the abovementioned PDB25 dataset can be downloaded from http://raptorx.uchicago.edu/ContactMap/. The other three datasets were obtained from three CASP (Critical Assessment of Structure Prediction) competitions (CASP10 (Moult et al., 2014), CASP11 (Moult et al., 2016), and CASP12 (Moult et al., 2018)). For the three CASP datasets, we used the same screening method as that used in the R2C method (Yang et al., 2016). For the CASP10 dataset, the sequence data could be accessed on the website of https://predictioncenter.org/download_area/CASP10/targets/.
TABLE 2 | Contact results by four different methods on the CASP10 testing dataset.

| Method                  | Long-range | Medium-range | Short-range |
|-------------------------|------------|--------------|-------------|
|                         | L/10       | L/5          | L/2         | L           | L/10       | L/5          | L/2         | L           |
| CCMpred                 | 0.533      | 0.477        | 0.355       | 0.242       | 0.512      | 0.417        | 0.272       | 0.185       |
| R2C                     | 0.413      | 0.306        | 0.198       | 0.143       | 0.540      | 0.425        | 0.278       | 0.191       |
| RaptorX-Contact         | 0.674      | 0.625        | 0.490       | 0.372       | 0.699      | 0.629        | 0.458       | 0.318       |
| RCMPM (CCMpred)         | 0.639      | 0.583        | 0.455       | 0.342       | 0.646      | 0.593        | 0.426       | 0.290       |
| RCMPM (RaptorX-Contact) | 0.673      | 0.611        | 0.496       | 0.371       | 0.681      | 0.612        | 0.452       | 0.312       |

3.3 Performance on PDB25 Testing Datasets and CASP Testing Datasets

In our experiment, we used top \( L/k \) (\( k = 10, 5, 2, 1 \)) in the long-range contact to evaluate the prediction accuracy of contact maps. Here, \( L \) is the length of the sequence, and the prediction accuracy rates are given in three kinds of contact, namely, long-range, medium-range, and short-range.

The datasets used in our experiment are PDB25, CASP10, CASP11, and CASP12 datasets. To examine the performance of our proposed RCMPM model, three state-of-the-art methods are used for comparison, namely, CCMpred (based on Markov random field pseudo-likelihood maximization, MSA), R2C (based on SVM), and RaptorX-Contact (based on ResNet), respectively. We have realized CCMpred and RaptorX-Contact models and trained them under the same environments with that of the RCMPM and hence obtained the experiment results of the two models by ourselves. The results of R2C on CASP10 and CASP11 datasets are cited from the reference (Villegas-Morcillo et al., 2018), while the results on the other two datasets are calculated through its webserver (http://www.csbio.sjtu.edu.cn/bioinf/R2C/). For comparison, two rough contact maps, produced by CCMpred and RaptorX-Contact models, are used to construct the amino acid graph in the proposed RCMPM, respectively.

Table 1 shows the comparison results on the PDB25 dataset. For the long-range contact type prediction, the results of the RCMPM by using the CCMpred outputs as the rough contact map (RCMPM (CCMpred)) are significantly better than those of CCMpred, with 19.9%, 22.2%, 38.0%, and 73.5% improvements on top \( L/10 \), \( L/5 \), \( L/2 \), and \( L \) levels, respectively. Compared to R2C, it improves by 7.8% and 2.7% on the top \( L/10 \) and \( L/5 \) levels, respectively, and decreased by 10.2% and 0.7% on the top \( L/2 \) and \( L \) levels, respectively. RaptorX-Contact is an excellent algorithm, the results of which are better than those of RCMPM (CCMpred). However, when the RCMPM model uses the output of RaptorX-Contact as the rough contact map (RCMPM (RaptorX-Contact)), it outperforms RaptorX-Contact on all the four top levels with 1.3%, 1.2%, 2.1%, and 2.2% improvements, respectively. The results of RCMPM (RaptorX-Contact) are also significantly better than CCMpred, R2C, and RCMPM (CCMpred), with the only exception being slightly below R2C at the top \( L/2 \) level.

Figure 3 shows the comparison results of five methods on the long-range contact type prediction. For the medium-range contact type, both RCMPM (CCMpred) and RCMPM (RaptorX-Contact) are significantly superior to CCMpred and RaptorX-Contact, both outperforming their opponents at the four top levels. Among all the five comparison methods, RCMPM (RaptorX-Contact) performs...
best at all the four levels. For the short-range contact type, RCMPM (CCMpred) is greatly better than CCMpred, while RCMPM (RaptorX-Contact) performs similarly with RaptorX-Contact, both significantly outperforming the other three methods.

**Table 2** shows the comparison results on the CASP10 dataset. For the long-range contact type prediction, the results of RCMPM by using CCMpred outputs as the rough contact map (RCMPM (CCMpred)) are significantly better than those of CCMpred, with 19.8%, 22.2%, 28.2%, and 41.3% improvements on top L/10, L/5, L/2, and L levels, respectively. Compared to R2C, it improved by 54.7%, 90.5%, 129.8%, and 139.2% at the four top levels, respectively. When the RCMPM uses the RaptorX-Contact outputs as the rough contact map (RCMPM (RaptorX-Contact)), it performs similarly with the RaptorX-Contact, with −0.1%, 1.0%, −2.2% and −0.2% variations at the top levels, respectively. Both of them significantly outperform CCMpred and R2C, and a little better than RCMPM (CCMpred). RCMPM (RaptorX-Contact) improves by 26.3%, 28.1%, 39.4%, and 53.3% compared to CCMpred and increases by 150.0%, and 159.4% compared to R2C at the top four levels, while compared to RCMPM (CCMpred), the improvements are 5.3%, 4.8%, 8.8%, and 8.5%, respectively. Figure 4 shows the comparison results of the five methods on the long-range contact type prediction. The results are similar for both the medium-range contact and short-range contact types, with RCMPM (CCMpred) being significantly superior to CCMpred, while RCMPM (RaptorX-Contact), despite its lower performance than RaptorX-Contact, had a weak gap.

**Table 3** shows the comparison results on the CASP11 dataset. For the long-range contact type prediction, the results of RCMPM by using the CCMpred outputs as the rough contact map (RCMPM (CCMpred)) are significantly better than those of CCMpred, with 40.8%, 50.9%, 72.1%, and 86.9% improvements at the four top levels. Compared to R2C, it outperforms at all the four top levels with 26.2%, 39.5%, 62.5%, and 72.6%. RaptorX-Contact is better than RCMPM (CCMpred). However, when the RCMPM uses the output of RaptorX-Contact as the rough contact map (RCMPM (RaptorX-Contact)), it improves by 0.8%, 1.8%, 1.4%, and 1.5% on the four top levels, respectively. The results of RCMPM (RaptorX-Contact) are also significantly better than CCMpred and R2C. The results are similar for both the medium-range contact and short-range contact types, with RCMPM (CCMpred) being significantly superior to CCMpred, while RCMPM (RaptorX-Contact), despite its lower performance than RaptorX-Contact, had a weak gap. Figure 5 shows the comparison results of the five methods on the long-range contact type prediction. For both the medium-range contact and short-range contact types’ results, we can draw the following conclusions: among the three existing state-of-the-art (SOTA) methods, RaptorX-Contact performs the best; RCMPM
(CCMpred) is significantly superior to CCMpred; RCMPM (RaptorX-Contact) obtains similar results with RaptorX-Contact, while CCMpred is much lower than RaptorX-Contact; and RCMPM (CCMpred) has yielded results comparable to RCMPM (RaptorX-Contact).

Table 4 shows the comparison results on the CASP12 dataset. For the long-range contact type prediction, the results of the RCMPM by RCMPM (CCMpred) are significantly better than those of CCMpred, with 24.8% 28.1%, 36.1%, and 41.5% improvements at the top L/10, L/5, L/2, and L levels, while RCMPM (RaptorX-Contact) outperforms RaptorX-Contact with 43.3%, 38.3%, 16.6%, and 6.6% improvements on the four top levels, respectively. The results of RCMPM (RaptorX-Contact) are also significantly better than those of CCMpred and RCMPM (CCMpred). Figure 6 shows the comparison results of the five methods on the long-range contact type prediction. For both the medium-range and short-range contact types' results, we can draw the following conclusions: among the three existing SOTA methods, R2C performs the best; RCMPM (CCMpred) is significantly superior to CCMpred; RCMPM (RaptorX-Contact) obtains similar results with RaptorX-Contact; CCMpred is much lower than RaptorX-Contact, but the gap between RCMPM (CCMpred) and RCMPM (RaptorX-Contact) is greatly reduced.

To summarize the long-range results of the four datasets, it could be found that our proposed RCMPM method is significantly superior to the other methods on PDB25 and CASP11. For CASP10, RCMPM performs much better than CCMpred and R2C, and although it does not perform as well as RaptorX, the gap is very small. For CASP12, the accuracy of RCMPM is higher than that of CCMpred and R2C, and is slightly lower than that of R2C. Therefore, it can be concluded that our proposed method performs best overall on the four datasets and is the most stable one as well.
3.4 Ablation Study

3.4.1 Evaluation of the GCN Module of the Model Structure

In order to examine the effectiveness of our proposed method, we used two network structures to construct different network structures, the original RCMPM and the RCMPM removal GCN module (RCMPM (without GCN)). Table 5 shows the comparison results on the PDB25 dataset. Compared to the RCMPM (without GCN), the RCMPM improves by 1.2%, 0.9%, 1.7%, and 2% on the top L/10, L/5, L/2, and L levels, respectively, while the RCMPM performs much similar with the RCMPM (without GCN) on both the short-range and medium-range levels. This is because the graph neural network module can utilize the output of the existing methods, especially on the global information level, and therefore reflected by improvements on the long-range level contact prediction.

3.4.2 Evaluation of Different Feature Combinations

In order to verify the effectiveness of the sequence features on the long-range contact map prediction, we used three different feature combinations as the input of the 1D ResNet module and GCN module, including PSSM, PSSM, and secondary structure (PSSM+SS), including the PSSM, secondary structure, and solvent accessibility (PSSM+SS+SA), a total of L×26 dimensional features. Table 6 shows the comparison results by using the RaptorX-Contact outputs as the rough contact map on the PDB25 dataset. From Table 6, it could be found that on the long-range contact type, RCMPM (PSSM+SS+SA) is improved by 0.9%, 0.8%, 1.7%, and 0.6% at the four top levels compared to RCMPM (PSSM+SS) and 1.6%, 0.9%, 1.1%, and 1.2% on the four top levels compared to RCMPM (PSSM). Meanwhile, on the medium-range contact type, although the trend is the same as the long-range type, the increase is very small. On the short-range contact type, the results of the three methods are even very close. Table 7 shows the comparison results by using the CCMpred outputs as the rough contact map on the PDB25 dataset. From Table 7, it could be found that on the long-range contact type, RCMPM (PSSM+SS+SA) is improved by 0.8%, 2.2%, 2.1%, and 2.1% at the four top levels compared to RCMPM (PSSM+SS) and 9.3%, 13.4%, 26.2%, and 44.8% at the four top levels compared to RCMPM (PSSM). On the medium-range and short-range contact types, the results of the RCMPM (PSSM+SS+SA) are also better than the other two types’ results. The results show that the PSSM is a very important feature for the contact prediction, and the secondary structure and solvent accessibility are also beneficial. When the initial contact map is used as RaptorX-Contact, the secondary structure and solvent accessibility have a limited effect on the medium- and short-range contact type predictions, in part because the RCMPM uses the output of RaptorX-Contact, which already contains the secondary structure and solvent accessibility information.

4 DISCUSSION

In this article, we formulated the peptide contact map refinement task as a geometric 2D graph improvement and proposed a novel refined contact map prediction model (RCMPM) to refine the protein inter-residue contact map predictions using graph convolutional neural networks (GCNNs) and one-dimensional and two-dimensional residual neural network (1D ResNet and 2D ResNet) architectures. Our method combines the residual neural networks for learning the local information with the graph convolutional neural networks for learning the global information, which can better capture the long-range contact relationship between the complex sequence inter-residues. The experimental results show that our method can refine the contact map greatly for the long-range contact type, that is to say, by using CCMpred outputs as the rough contact map, the RCMPM is significantly better than CCMpred, and by using the RaptorX-Contact outputs as the rough contact map, the RCMPM is significantly better than RaptorX-Contact as well. For the medium-range contact prediction, the degree of improvement is significantly reduced, and for the short-range contact prediction, there is not even a significant improvement. The main reason is that the GCN module of the RCMPM can utilize the outputs of the existing methods, which are highly reflected on the global information level, and therefore, the RCMPM model makes improvements mainly on the long-range contact types. By using a larger protein database in HHblits or PSI-BLAST to calculate the homology features of protein sequences and combining more effective features as inputs, we can expect to further improve the precision.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

XS, CW, and YL contributed to conception and design of the study. JG performed the statistical analysis. JG wrote the first draft of the manuscript. JG, TZ, and XS wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version. The handling editor (JW) declared a past co-authorship with the author (YL).

FUNDING

This research was funded by the National Natural Science Foundation of China (61972174), the Key Research and Development Project of Jilin Provincial Science and Technology Department (20210201080GX), Jilin Province Development and Reform Commission (2021C044-1), Guangdong Science and Technology Planning (2020A0505100018), Guangdong universities’ innovation team (2021KCXTD015), and Guangdong key discipline (2021ZDJS138) projects.
