Hindawi
BioMed Research International
Volume 2020, Article ID 4569037, 10 pages
https://doi.org/10.1155/2020/4569037

Research Article

IMPContact: An Interhelical Residue Contact Prediction Method

Chao Fang,[1] **Yajie Jia**,[1,2] **Lihong Hu**,[1] **Yinghua Lu**,[1,3] **and Han Wang**[1,2,3]

[1] *School of Information Science and Technology, Northeast Normal University, Changchun 130117, China*
[2] *Institute of Computational Biology, Northeast Normal University, Changchun 130117, China*
[3] *Department of Computer Science, College of Humanities & Sciences of Northeast Normal University, Changchun 130117, China*

Correspondence should be addressed to Yinghua Lu; luyh@nenu.edu.cn and Han Wang; wangh101@nenu.edu.cn

Received 1 March 2020; Accepted 9 March 2020; Published 27 March 2020

Guest Editor: Quan Zou

As an important category of proteins, alpha-helix transmembrane proteins (αTMPs) play an important role in various biological activities. Because the solved αTMP structures are inadequate, predicting the residue contacts among the transmembrane segments of an αTMP exhibits the basis of protein fold, which can be used to further discover more protein functions. A few efforts have been devoted to predict the interhelical residue contact using machine learning methods based on the prior knowledge of transmembrane protein structure. However, it is still a challenge to improve the prediction accuracy, while the deep learning method provides an opportunity to utilize the structural knowledge in a different insight. For this purpose, we proposed a novel αTMP residue-residue contact prediction method IMPContact, in which a convolutional neural network (CNN) was applied to recognize those interhelical contacts in a TMP using its specific structural features. There were four sequence-based TMP-specific features selected to descript a pair of residues, namely, evolutionary covariation, predicted topology structure, residue relative position, and evolutionary conservation. An up-to-date dataset was used to train and test the IMPContact; our method achieved better performance compared to peer methods. In the case studies, IHRCs in the regular transmembrane helixes were better predicted than in the irregular ones.

1. Introduction

Alpha-helical transmembrane protein (αTMP) is an important type of membrane protein (MP) widely existing in eukaryotic cells and carrying on the responsibility of transferring signals or small molecules between two sides of biological membranes. For this reason, αTMPs are involved in many vital biological processes [1], such as solute and ion transport, energy transduction in respiratory and photosynthetic systems, or sensory stimuli transduction and information processing [2]. Consequently, they are major drug targets accounting for approximately 70% of the known and tested drug targets [3]. Therefore, the study of αTMPs' structure and function is currently a popular topic in chemistry and biology fields [4].

However, due to the specificity of the function of membrane proteins, many efforts had been applied to derive their structures, but both biological experiment and prediction approach cannot satisfy the requirements in the balance of quantity and quality. In recent years, semistructural research of TMP became more practical instead of the entire 3D structure, which is based on computational prediction to make the balance, such as hydrophobicity, electrical polarity, or contact prediction. In an αTMP, interhelical residue contacts (IHRCs) bind its alpha-helixes as anchor points, while most αTMP family members have similar transmembrane structures, so that they possibly have a similar function. It is obvious that predicting IHRCs has a special meaning for αTMPs compared to that of solvent proteins, although those residue contacts are totally the same in view of the biochemical background.

For the solvent proteins, considerable improvements had been made to predict the residue contact in decades, corresponding to prediction methods mainly using correlated mutation analysis- (CMA-) based or machine learning- (ML-) based methods. The CMA-based methods, such as PSICOV [5] and CCMpred [6], take into consideration that fact the contacts mostly happened between those coevolutionary residues because residue contact is an important structural feature that remains between a pair of residues on a protein sequence in the evolution process. Statistical models were mainly used in CMA-based methods to find the coevolutionary relationships from multiple sequence alignment. ML-based methods, such as CMAPpro [7], identify the contacts by abstracting various structural features of a protein sequence to a classifier. It is apparent that the coevolutionary residues can be used in an ML-based predictor to improve the prediction accuracy, which was proved by R2C [8].

The residue contact prediction still remains an open problem in the field of structural bioinformatics [5], and the above methods cannot accurately identify the residue contacts in an αTMP. Therefore, many TMP-specific predictors have been raised in recent decades, which are still following a similar way like those methods used on solvent proteins. After the coevolving residues were noticed relevant to the residue contact in MPs [9], the CMA-based method was applied to MPs, namely, direct-coupling analysis (DCA). Although corresponding methods continuously improved the prediction accuracy, such as mfDCA [10], PSICOV [5], plmDCA [11], and GREMLIN [12], a pure CMA-based method still has its intrinsical limitation. A direct cause is that multiple sequence alignment does not work well against the MPs, because many MPs have less family member proteins. Consequently, it is found that only a small fraction of predicted correlations involved pairs of residues in physical contact, while a sizeable fraction of the correlations was found to be in close vicinity to interhelical contacts [9]. Thus, coevolution information may not accurately detect the residue contacts by itself, but it obviously is a clue to find where the contacts exist.

For the purpose of IHRC prediction, various sequence- or structure-based features and machine learning algorithms were utilized. Neural network (NN), support vector machines (SVMs), and random forest (RF) algorithms were, respectively, proven to be effective in predicting IHRCs in TMHcon [13], TMhit [14], MEMPACK [15], TMhhcp [4], COMSAT [16], and MemConP [17]. It is convincing that the coevolution relationship highly improved the prediction accuracy used in TMHcon [13], MEMPACK [15], and MemConP [17]. These methods had already tried the best to maximize the coevolution and machine learning algorithms, but it is still a challenging problem to improve the prediction accuracy to an acceptable level, where the key factor is how to find more IHRC-specific structural features for a proper machine learning algorithm. Our previous work on interbarrel residue contact prediction for outer membrane proteins accessed a high performance method by using MP-specific features [18]. It is possible to further improve the IHRC prediction accuracy when the IHRC-specific features meet the deep learning networks, as many successful cases did in the field of bioinformatics [19–22].

In this study, we abstracted four IHRC-specific features from the αTMP sequences, including the Evolution Conservation Feature, Evolutionary Covariation Feature, Topology Structure Feature, and Residue Relative Position Feature. Then, a CNN- (Convolutional Neural Networks-) based predictor was proposed to predict IHRC for αTMPs, named IMPContact. An up-to-date dataset was used to train and test the method; the results showed that deep learning can better utilize the IHRC-specific features than the SVM, RF, and NN methods and derive higher prediction accuracy than the other classical peer methods. Considering further application, the source code of IMPContact was published at https://github.com/NENUBioCompute/IMPContact.

2. Materials and Methods

2.1. Datasets. In order to improve the reliability and authenticity of the predicted results, the data were obtained from the redundant αTMP sequence ID from PDBTM [23], which is a comprehensive and up-to-date transmembrane protein selection extracted from the Protein Data Bank (PDB). The dataset includes 348 nonredundant sequences, where 313 sequences were used in training and validation and the rest of the 35 sequences were used for testing.

There are three ways to label the residue contact from a protein structure: (I) the minimal distance between the heavy atoms of the side chain or backbone was less than 5.5 Å [13]; (II) the sum of their van der Waals (VDW) radii plus a threshold of 0.6 Å [14] (the VDW radii were taken from Li and Nussinov in 1998) [24]; and (III) a maximal distance of 8 Å between their C-beta atoms (C-alpha for glycine) [15]. Since it is essentially the same as the three definitions, MEMPACK [15] found that the prediction accuracy will be less different regardless of what definition was selected, and we used the more restricted definition (I) to label the IHRC for an αTMP.

2.2. TMP-Specific Sequence Features. As known, distinctive features will be highly helpful for a machine learning method to make a better prediction. The IHRCs in an αTMP have many specific features observed in transmembrane segments. We selected four structural features to describe the existence of IHRCs from different perspectives, where an evolution conservation feature is widely used in protein research and proved to be available to describe the conservation sequence segments, a topology feature is used to describe the transmembrane segments, an evolutionary covariation feature is used to indirectly reveal the existing residue-residue contacts, and residue relative position features are used to enhance the structural information.

2.2.1. Evolution Conservation Feature. Multiple sequence alignment discovered the evolution conservation against the large-scale protein sequence database; it had already been wildly applied in various biological sequence researches [18, 25]. More than 30% of the homologous superfamilies described in CATH are composed mainly or entirely of α-helixes [26]. Transmembrane helixes are different from those in the soluble proteins because the environment in the lipid layer force those helixes that were stretched. But the evolution conservation becomes more distinct and special in these sequence segments, leading to a different way to descript the evolutionary conservation between a pair of residues that have IHRC. At this point, evolution conservation is a particular TMP-specific feature.

The evolution conservation of a residue can be described by the PSSM (Position-Specific Scoring Matrix) produced from a multiple alignment tool; it presents the frequency of a residue type that appeared in each position of the protein sequence. The raw PSSM profile is represented by a 20-dimensional score vector [16, 27]. We obtained the PSSM by running a stand-alone PSIBLAST [28] against NCBI's nonredundant sequence database (NR) with three iterations and the E value set to 1e-10.

2.2.2. Evolutionary Covariation Feature. Different from evolution conservation, evolutionary covariation is aimed at descripting the evolutionary correlation between two residues on a protein sequence. It was observed that the residues having contact possibly present highly evolutionary conservation, because the contacts are not randomly existing in the protein, and they are closely related to the protein structures. But on the contrary, not all the residue pairs with evolutionary conservation will have contact. Therefore, conservation and covariation present the evolutionary information of a protein sequence from discriminate perspectives.

All methods generate the evolutionary covariation from the multiple sequence alignments; we allow IMPContact to accept the evolutionary covariation feature abstracted by different methods under the condition that those methods can be integrated into IMPContact, where the selected method uses the multiple alignments generated from the process we did for the evolution conservation feature then calculates the evolutionary covariation for all the possible residue pairs on a TMP. As known, different evolutionary conservative tools may have different numerical spaces when predicting different protein sequences, so the evolutionary covariation feature finally inputting to IMPContact will be standardized using the z-score for each sequence.

2.2.3. Topology Structure Feature. As mentioned, the helixes are special in the transmembrane domain in the TMPs compared to soluble proteins; they are characterized by the topology structure to distinguish the transmembrane and nontransmembrane segments. There is no doubt that the topology structure will directly reduce the searching space for the IHRCs, and it is exactly a TMP-specific feature. According to our goal, only sequence-based features are suitable for use in the prediction and the predicted topology structure meets the requirement.

There are several methods available to predict the topological structure of intimal proteins, including some that are based on hydrophobicity analyses [29], statistical procedures [30], or machine learning-based methods [31, 32]. The accuracy is the most important factor in choosing a topology structure prediction method; in our previous work, DMCTOP [33] was used to abstract the topology structure features. Another consideration is that the DMCTOP is a deep learning-based method, and it was upgraded during the further improvement of IMPContact.

The predicted topological structure was output in a uniform format, in which the cross-diaphragm residues were identified by the character "H", the outer residues by the character "o/O", and the inner residues by the character "i/I". Eventually, all the characters form a sequence equal to the length of the original protein sequence. The predicted topology structures were digitized into a vector as one input feature.

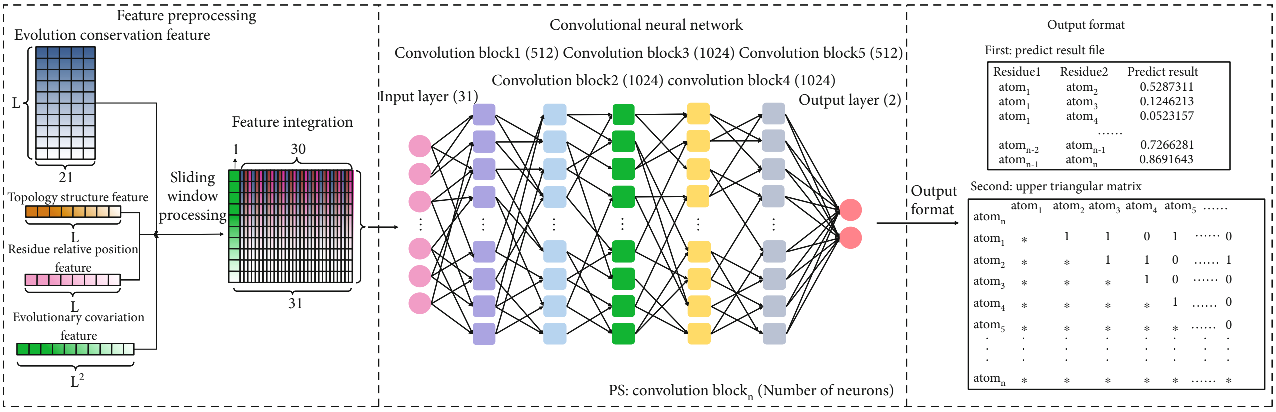
2.2.4. Residue Relative Position Feature. Relative position is a derivative feature from the topology feature. According to the observation, the transmembrane helixes on a sequence always alternately cross the membrane, resulting in IHRCs not occurring between the residues that separately close to the different sides of the membrane. Here, the relative position feature was used to descript whether the two residues on the neighboring helixes close to the same side of the membrane are close to each other.

The relevant position of a residue was assigned depending on the predicted topology structure. The process started from the N-terminal of the sequence. We assigned increasing integer values starting from 1 for each residue on each odd transmembrane helix that appeared and then did the same to the rest of the transmembrane helixes from the C-terminal; the other residues were assigned to -1 finally. After the above process, all the residues had a value representing their relative position; those residues who had closing positive values are considered more possibly closing to each other.

2.3. Sliding Windows. The IHRC is a local residue interaction on the TMP sequence, though it possibly happens between any pair of transmembrane helixes. The remote residues of the protein structure have fewer influences on IHRC, while the structural neighboring residues are mainly involved in forming the necessary surrounding environment. For this reason, two sliding windows with a size of 5 were used to characterize the features for both residues of a contact pair, where the 2 upstream and 2 downstream residues were included in the sliding window with the first candidate residue, and the same sliding window was applied to the other candidate residues. In each turn, the 4 TMP-specific features of the residues in the two sliding windows constituted a one-dimensional feature vector for the corresponding candidate pair of residues, in which an evolution covariation feature was used only between the two candidate residues. For the contact prediction between residues Ai and Bj, the features were abstracted from residues $(A_{i-2}A_{i-1}A_iA_{i+1}A_{i+2})$ and $(B_{j-2}B_{j-1}B_jB_{j+1}B_{j+2})$, where i and j are the sequence positions and $i \neq j$. According to the sliding window, an SVM input $V_{i,j}$ is given as follows:

$$\begin{aligned} V_{i,j} = \big(&C_{i,j}, (E,T,R)_{i-2}, (E,T,R)_{i-1}, (E,T,R)_i, (E,T,R)_{i+1}, \\ &(E,T,R)_{i+2}, (E,T,R)_{j-2}, (E,T,R)_{j-1}, \\ &(E,T,R)_j, (E,T,R)_{j+1}, (E,T,R)_{j+2}\big), \end{aligned} \tag{1}$$

where $(E,T,R)_i$ are the evolution conservation feature, topology feature, and relative position feature of residues in position i, respectively, and $\mathrm{C}_{i,j}$ is the predicted


Feature preprocessing
Evolution conservation feature
L
21
Topology structure feature
L
Residue relative position feature
L
Evolutionary covariation feature
L^2
Sliding window processing
Feature integration
1
30
31
Convolutional neural network
Convolution block1 (512) Convolution block3 (1024) Convolution block5 (512)
Convolution block2 (1024) convolution block4 (1024)
Input layer (31)
Output layer (2)
PS: convolution block$_n$ (Number of neurons)
Output format
Output format

First: predict result file

Residue1	Residue2	Predict result
atom_1	atom_2	0.5287311
atom_1	atom_3	0.1246213
atom_1	atom_4	0.0523157
	......	
atom_{n-2}	atom_{n-1}	0.7266281
atom_{n-1}	atom_n	0.8691643

Second: upper triangular matrix

atom_n	atom_1	atom_2	atom_3	atom_4	atom_5	......	
atom_1	*	1	1	0	1	......	0
atom_2	*	*	1	1	0	......	1
atom_3	*	*	*	1	0	......	0
atom_4	*	*	*	*	1	......	0
atom_5	*	*	*	*	*	......	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
atom_n	*	*	*	*	*	......	*

Figure 1: The neural network construction of IMPContact.

evolutionary covariation between residues i and j. Finally, a 31-dimensional eigenmatrix is formed. In this way, it became a simplified binary classification problem to predict IHRCs.

2.4. Deep Learning Network. As a representative branch of deep learning, CNN [34] had already made great achievements in various research fields, including protein sequence studies [35–39]. It can capture various nonlinear features by constructing neural networks consisting of convolution, pooling, and fully connected layers. The learning capability of CNN is strongly supported by diversification through the assembly of the different layers, activation functions, and ways of connecting the nodes into the networks. Therefore, lots of CNN-based applications have been raised to solve various kinds of researching issues in recent years; it has high adaptability in keeping improvements.

According to the sliding window feature, we constructed the IMPContact prediction process with a CNN-based model kernel, as shown in Figure 1. IMPContact allows users to input a whole TMP sequence; all the TMP-specific features will be generated at one time, then it scales the whole sequence from the N-terminal to the C-terminal using two sliding windows, until all the residue pairs are predicted by the CNN-based kernel. Finally, it outputs the prediction results into two text files: one is the list of residue pairs marked with a prediction conclusion and the other one is an upper triangular matrix recording the prediction conclusion accordingly.

The CNN kernel was designed as follows:

(1) Input layer: the input is a 31 × 1 feature vector, corresponding to the residue features in the two sliding windows; the middle residues in the windows are the candidates to be predicted

(2) Convolutional block layers: there are 5 layers of convolutional blocks—512, 1024, 512, 1024, and 1024. All the block layers are fully connected, and all the blocks use the same network, which is composed of a convolutional layer (kernel = 3), a rectified linear unit (ReLU) layer, and a max pooling layer

(3) Output layer: the output is the binary classification result for each input residue pair

3. Results and Discussion

3.1. Evolutionary Covariation Method Selection. Different from the other features used in this work, the evolutionary covariation feature is the key to accurately identify the existence of IHRCs. As mentioned previously, the evolutionary covariation relationship between a pair of residues was obtained from the multiple alignments against a particular protein sequence database, but the process cannot guarantee that those residue pairs with high evolutionary conservation must have residue contact, especially for the protein sequences that have fewer homology proteins in the sequence database. This is the reason why the evolutionary covariation is an indirect feature so that choosing the method to abstract this feature becomes important.

There were three evolutionary covariation methods taking into consideration IMPContact, namely, ELSC [40], MI [41], and OMES [42]. In spite of the many similar methods providing computational tools, e.g., EVFOLD [43], which obtained an even higher accuracy on covariation, those 3 methods are more convenient to realize an easy-to-use particular IHRC method for the users. For the purpose of discovering the covariation residue pairs from the multiple alignment result, ELSC (Explicit Likelihood of Subset Covariation) uses combinatorial arguments to realize a perturbative algorithm [44]. MI (mutual information) measures the codependency of two residues as random variables [45]. OMES iteratively calculates the score for each pair of residues by the frequency they observed in the sequence alignment based on a correlated analysis method [46].

The above three evolutionary covariation methods were discussed having different performances on different datasets [42], and we tested all the methods in our method to choose the best one with the highest compatibility to the IMP alignment results, where some TMPs have rare homology protein sequences, and IHRC has special distribution of the evolutionary covariation relative to the soluble proteins caused by standing in the lipid layer environment. In comparison, the evolutionary covariation features were abstracted using

Table 1: Comparison of candidate evolutionary covariation methods.

Methods	ACC	MCC
ELSC	0.8408	0.3371
MI	0.9742	0
OMES	0.7122	0.1335

Table 2: The prediction accuracy comparison of 4 classifiers.

Methods	ACC	MCC
RF	0.6337	0.1209
SVM	0.7451	0.1478
NN	0.8050	0.1745
CNN	0.8408	0.3371

the ELSC, MI, and OMES methods and, respectively, input into the IMPContact while the other features and the deep learning networks are completely the same. The experiments randomly selected 80% of the sequences from the training dataset to train the prediction models corresponding to the three covariation methods then used the other 20% of the sequences to test the three models.

As shown in Table 1, the three methods obtained obvious prediction performances. ELSC provided the highest accuracy both on ACC and MCC. MI obtained the highest ACC of 0.9742 but failed to get an acceptable MCC; it is obvious that this method could not help the IMPContact to distinguish the positive and negative samples, because most residue pairs in the transmembrane helixes were considered having a similar covariation, and it ignored those residue pairs with contact that should have a stronger covariation. Compared to the other two methods, the OMES obtained a middle accuracy on both standards. Therefore, the ELSC is more compatible with our model to descript the evolutionary covariation feature. All the following tests and publishing models used the ELSC method.

3.2. Classifier Comparison. Deep learning methods were proven to be inefficient for all the studies. They depend on many factors, such as data space, data distribution, and the researching problems; there is no guarantee that a deep learning method could be better at solving the IHRC prediction compared to traditional machine learning methods. Therefore, we did the prediction performance comparison among our CNN-based model and three widely used machine learning methods: SVM (Support Vector Machine) [47–49], RF (Random Forest) [50, 51], and NN (Neural Network) [52, 53].

The CNN model was built on the PyTorch platform [54]; the other three machine learning methods used the Scikit-learn toolkit [55]. All the models were developed using the Python language and trained and tested against part of our training datasets the same as that in Section 3.1. The parameters of each model were optimized to obtain the best prediction accuracy correspondingly. The same four TMP-specific features were input to each model for all the TMP sequences; among them, the evolutionary covariation feature was abstracted using the ELSC method, which was proven to be the most compatible to the IHRC prediction.

The prediction accuracy levels of the four methods are listed in Table 2; the CNN model achieved the best performance with the highest MCC and ACC, which were about 0.34 and 0.84, respectively. The other three machine learning methods obtained obviously lower accuracy levels on both criteria, where the RF model had the worst accuracy with an MCC of about 0.12 and an ACC of about 0.63; the SVM and NN models were better than the RF model with a little higher accuracy. Although the NN model obtained a closer ACC to the CNN model, and a better MCC value, none of the machine learning methods had accessed an acceptable MCC as that of the CNN model. The results showed that those three methods predicted more negative samples to be positive samples, casing the lower MCCs even when high ACCs appeared, not to mention that the RF and SVM models had much lower ACC values. Finally, the CNN model was chosen as a classifier used in the IMPContact.

3.3. Prediction Performance

3.3.1. Cross-Validation on the Training Dataset. Fivefold cross-validation was used for our method on our training dataset to show the training performance. In this experiment, all the 4 IMP-specific features and the prediction classifier had already been chosen according to the above processes. All the training sequences were randomly assigned to 5 subsets with similar members; in each fold, one subset was used as the testing dataset, while the other four subsets were used as the training datasets. The validation process was completed when each subset had been tested; there were a total of 5 models produced during the training and testing processes. At the same time, traditional machine learning methods were validated using the same steps to comprehensively represent that the deep learning method performed better than the other methods in this issue. Here, the MCC was used to evaluate the performance as a balanced criterion.

As known, fivefold is not an extremely strict cross-validation; it was adopted in this experiment determined mainly by the size of the TMP dataset. In our nonredundant training dataset, many TMPs have only a few homology proteins to obtain the comprehensive evolutionary investigation, which will mislead the testing when too many folds are used in the cross-validation. Moreover, fewer folds cannot descript the stability of the model.

The prediction accuracy is shown in Figure 2; the CNN-based model obviously surpasses the other three machine learning methods, while they had a similar performance. It can be found that the IMPContact could obtain a consistent performance in the cross-validation no matter what classifier was taken, where all the models were stable in the prediction. The phenomenon illustrates that TMP-specific features had strongly supported the prediction, but they had no such clear data space bundle for the

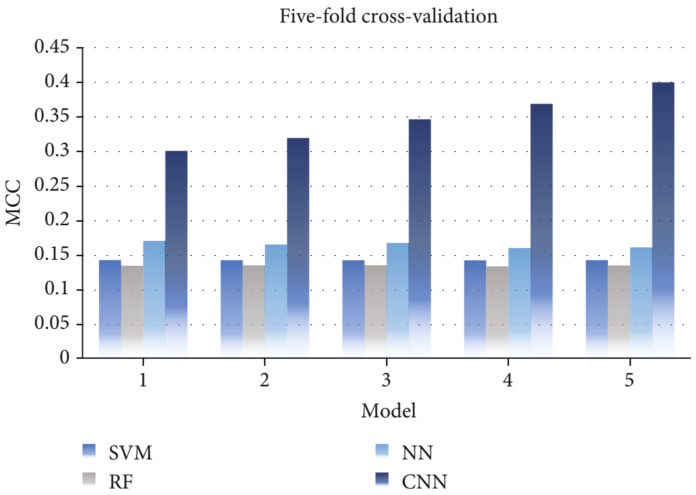


Figure 2: Prediction model comparison on fivefold cross-validation.

traditional machine learning methods to distinguish the IHRCs, while the CNN model found more details by deep learning.

3.3.2. Sequence Length Distribution in the Datasets. In previous studies, TMPs' protein family was found strongly relative to their sequence length; this is because the TMPs are mostly remotely homological to have the conservative transmembrane segments and the functional segments. Therefore, the distributions of the training dataset and testing dataset should be discussed before the final testing.

We counted the sequence lengths in both the training and testing datasets, as shown in Figures 3(a) and 3(b). The diagram in Figure 3(a)is the sequence length distribution of all the training datasets, and the one in Figure 3(b) represents that of the testing dataset. It can be found that both datasets have a similar distribution in sequence length descripted using the red lines, where most sequences are less than 600 amino acids, a few sequences are 600-800 amino acids in length, and a small number of sequences are larger than 800 amino acids. It is obvious that our testing results would be more accurate under the condition that the training and testing datasets have such similar distribution. There is another problem that those long sequences have more complex IHRCs than short sequences, and they may not obtain enough training. However, IMPContact used a sliding window of 5 residues to descript the features of the candidate contact pairs and it will be less affected by the sequence length as the local features.

3.3.3. Prediction Performance. By the above studies, the final IMPContact was built using a CNN-based deep learning model; accepting the 4 TMP-specific features as inputs, the ELSC method was chosen to abstract the evolutionary covariation feature, outputting the IHRC prediction results into two formats. Here, we excluded the peer methods which are unavailable or proved to have low accuracy, and the methods were not suitable for IHRC prediction; thus, PSICOV and CCMpred were selected to make the comparison. Both the two methods are CMA-based methods, but they did not use any TMP-specific features and machine learning models. We downloaded the stand-alone tools of both methods to input the multiple alignment results the same as that of the IMPContact. The experiment used our testing sequences. To comprehensively descript the prediction performance, ACC, Precision, Recall, and MCC evaluation criteria were, respectively, counted using the prediction results.

As shown in Table 3, a total of 35 testing sequences were input into the three methods. PSICOV obtained the prediction results for 13 sequences, while the other two methods were available for all the sequences, and the performances were calculated based on those predicted samples by each method. Additionally, precision was used instead of ACC in a few peer studies, so we list both values. CCMpred had not predicted any residue pairs having contact, and the negative samples were far more than the positive samples, so it obtained a high ACC, while Precision, Recall, and MCC could not be calculated. PSICOV was better than CCMpred, but still conservative in positive sample prediction; consequently, it obtained a high ACC, while the other evaluation criteria were low. In contrast, IMPContact achieved the best performance; although there were more false positive samples predicted causing a lower ACC, it had a much higher MCC than PSICOV.

The comparison above revealed that the evolutionary covariation analyses cannot efficiently predict the IHRC currently, because less homological protein sequences limited the multiple alignments to discover the evolutionary information for TMPs, both in conservation and covariation. TMP-specific features are necessary to improve the prediction by bringing in more structural features to the classifier. Meanwhile, deep learning can better capture those TMP-specific features to identify the IHRCs.

3.4. Case Studies. For the purpose of displaying the details of the prediction, 3UDC_A [56] and 2WSC_G [57] were chosen as representative causes shown in Figures 4 and 5, where the residues in a true positive sample were connected using a blue dotted line, as well as yellow and red dotted lines for the false positive and false negative samples, respectively.

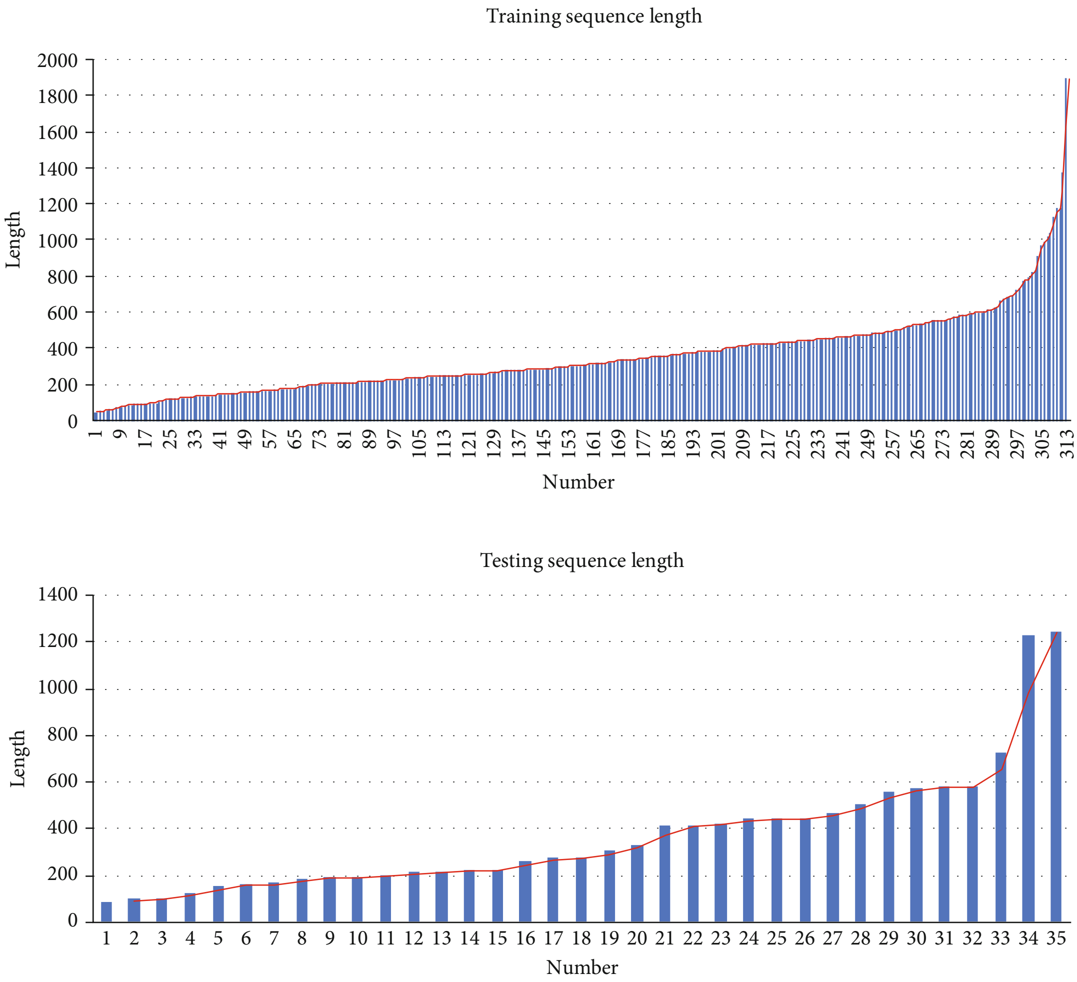


Figure 3: Sequence length distributions of datasets.

Table 3: Comparison with peer methods.

Methods	ACC	Precision	Recall	MCC	Predicted samples
CCMpred	0.9978	—	—	—	35
PSICOV	0.9878	0.0030	0.3842	0.0166	13
IMPContact	0.6293	0.0035	0.5920	0.0271	35

The 3UDC_A belongs to a mechanosensitive channel, which is composed of seven domains having the same sequences. Each domain has two transmembrane helixes along with one helix half crossing the membrane. 3UDC_A is a classical TMP, with two transmembrane helixes binding together through IHRCs. In this case, all positive contacts were correctly predicted, a few false positives were made close to the positive ones, and no false negative was predicted. It illustrated that the IMPContact is efficient for predicting the IHRCs between the regular transmembrane helixes.

In the other case, the 2WSC_G comes from a huge protein complex Plant Photosystem I. The transmembrane segments are partly formed as a helix; the rest of the parts mostly are flexible structures. The 2WSC_G has no such regular transmembrane helixes as in 3UDC_A; it directly weakened all the features helping IMPContact to make the classification, not only the evolutionary relevant features but also the topology relevant features, so that much more false positives and even false negatives appeared. It is extremely hard for our method to accurately predict the IHRCs for an irregular TMP, especially when the transmembrane segments are affected by the other segments in the protein complex.

4. Conclusions

In this study, we proposed an IHRC prediction method for inner transmembrane proteins. The TMP-specific features were used as inputs representing evolutionary and topology structure information, and a CNN model was used as the

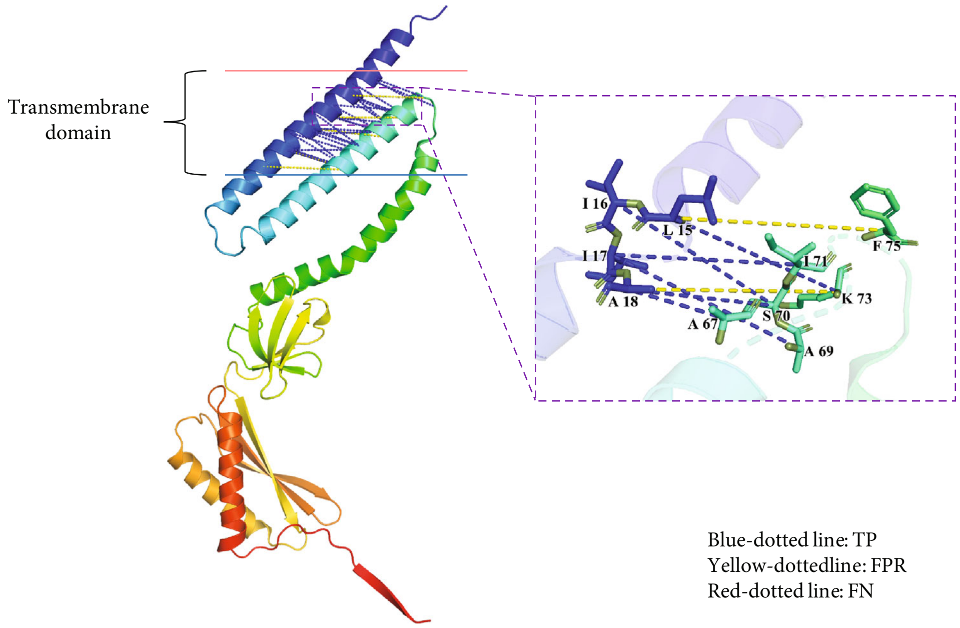


Figure 4: Prediction case of 3UDC_A.

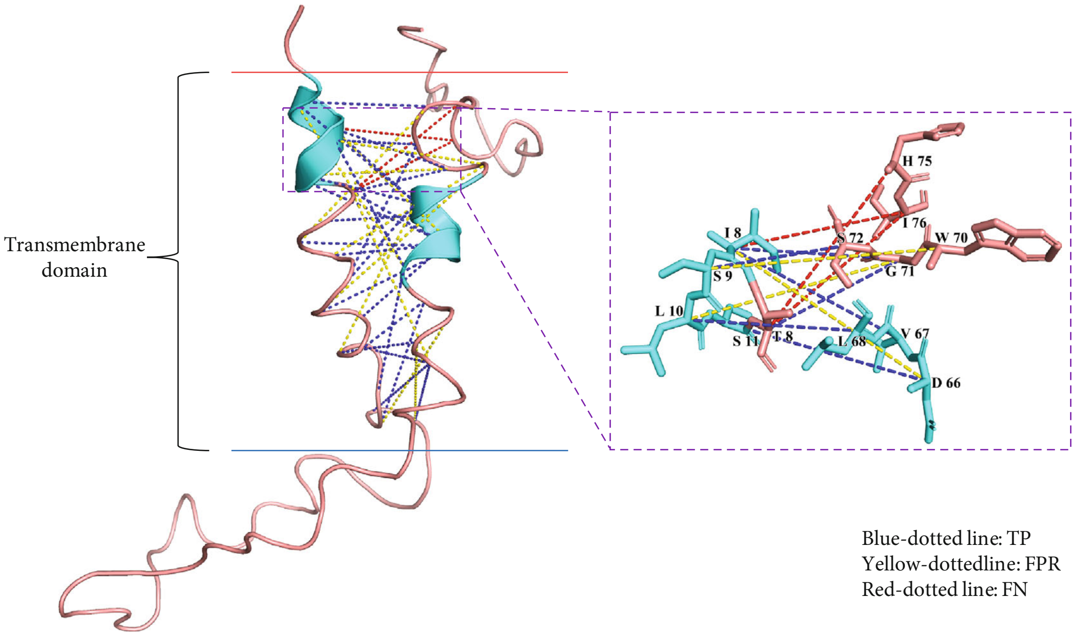


Figure 5: Prediction case of 2WSC_G.

classifier. After the experiments, the ELSC method proved to be better at discovering the evolutionary covariation in transmembrane segments, but it cannot identify the IHRCs by itself. The deep learning method showed that it was efficient in predicting IHRC based on the TMP-specific features. Compared to the CMA-based methods, our method achieved better performance on a testing dataset. In the case studies, IHRCs in the regular transmembrane helixes were better predicted than in the irregular ones. It is still a challenge to accurately predict IHRCs for all the TMPs.

Data Availability

All the Training and Testing datasets are available at https://github.com/NENUBioCompute/IMPContact.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Funds of China (Nos. 81671328 and 61802057), the Jilin Scientific and Technological Development Program (Nos. 20180414006GH, 20180520028JH, and 20170520058JH), the Science and Technology Research Project of the Education Department of Jilin Province (Nos. JJKH20190290KJ and JJKH20191309KJ), and the Fundamental Research Funds for the Central Universities (Nos. 2412019FZ052, 2412019FZ048).

References

[1] H. Wang, J. Wang, L. Zhang, P. Sun, N. Du, and Y. Li, "A sequential segment based alpha-helical transmembrane protein alignment method," *International Journal of Biological Sciences*, vol. 14, no. 8, pp. 901–906, 2018.

[2] A. Engel and H. E. Gaub, "Structure and mechanics of membrane proteins," *Annual Review of Biochemistry*, vol. 77, pp. 127–148, 2008.

[3] M. A. Yildirim, K. I. Goh, M. E. Cusick, A. L. Barabasi, and M. Vidal, "Drug-target network," *Nature Biotechnology*, vol. 25, no. 10, pp. 1119–1126, 2007.

[4] X. F. Wang, Z. Chen, C. Wang, R. X. Yan, Z. Zhang, and J. Song, "Predicting residue-residue contacts and helix-helix interactions in transmembrane proteins using an integrative feature-based random forest approach," *PLoS One*, vol. 6, no. 10, article e26767, 2011.

[5] D. T. Jones, D. W. A. Buchan, D. Cozzetto, and M. Pontil, "PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments," *Bioinformatics*, vol. 28, no. 2, pp. 184–190, 2012.

[6] S. Seemayer, M. Gruber, and J. Soding, "CCMpred—fast and precise prediction of protein residue-residue contacts from correlated mutations," *Bioinformatics*, vol. 30, no. 21, pp. 3128–3130, 2014.

[7] P. Di Lena, K. Nagata, and P. Baldi, "Deep architectures for protein contact map prediction," *Bioinformatics*, vol. 28, no. 19, pp. 2449–2457, 2012.

[8] J. Yang, Q. Y. Jin, B. Zhang, and H. B. Shen, "R2C: improving ab initio residue contact map prediction using dynamic fusion strategy and Gaussian noise filter," *Bioinformatics*, vol. 32, no. 16, pp. 2435–2443, 2016.

[9] A. Fuchs, A. J. Martin-Galiano, M. Kalman, S. Fleishman, N. Ben-Tal, and D. Frishman, "Co-evolving residues in membrane proteins," *Bioinformatics*, vol. 23, no. 24, pp. 3312–3319, 2007.

[10] F. Morcos, A. Pagnani, B. Lunt et al., "Direct-coupling analysis of residue coevolution captures native contacts across many protein families," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 49, pp. E1293–E1301, 2011.

[11] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell, "Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models," *Physical Review E*, vol. 87, no. 1, article 012707, 2013.

[12] H. Kamisetty, S. Ovchinnikov, and D. Baker, "Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 39, pp. 15674–15679, 2013.

[13] A. Fuchs, A. Kirschner, and D. Frishman, "Prediction of helix-helix contacts and interacting helices in polytopic membrane proteins using neural networks," *Proteins: Structure, Function, and Bioinformatics*, vol. 74, no. 4, pp. 857–871, 2009.

[14] A. Lo, Y. Y. Chiu, E. A. Rodland, P. C. Lyu, T. Y. Sung, and W. L. Hsu, "Predicting helix-helix interactions from residue contacts in membrane proteins," *Bioinformatics*, vol. 25, no. 8, pp. 996–1003, 2009.

[15] T. Nugent and D. T. Jones, "Predicting transmembrane helix packing arrangements using residue contacts and a force-directed algorithm," *PLoS Computational Biology*, vol. 6, no. 3, article e1000714, 2010.

[16] H. Zhang, Q. Huang, Z. Bei, Y. Wei, and C. A. Floudas, "COMSAT: residue contact prediction of transmembrane proteins based on support vector machines and mixed integer linear programming," *Proteins: Structure, Function, and Bioinformatics*, vol. 84, no. 3, pp. 332–348, 2016.

[17] P. Honigschmid and D. Frishman, "Accurate prediction of helix interactions and residue contacts in membrane proteins," *Journal of Structural Biology*, vol. 194, no. 1, pp. 112–123, 2016.

[18] L. Zhang, H. Wang, L. Yan, L. Su, and D. Xu, "OMPcontact: an outer membrane protein inter-barrel residue contact prediction method," *Journal of Computational Biology*, vol. 24, no. 3, pp. 217–228, 2017.

[19] M. Gao, H. Zhou, and J. Skolnick, "DESTINI: a deep-learning approach to contact-driven protein structure prediction," *Scientific Reports*, vol. 9, no. 1, article 3514, 2019.

[20] J. Hou, T. Wu, R. Cao, and J. Cheng, "Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13," *Proteins*, vol. 87, no. 12, pp. 1165–1178, 2019.

[21] B. Adhikari, J. Hou, and J. Cheng, "DNCON2: improved protein contact prediction using two-level deep convolutional neural networks," *Bioinformatics*, vol. 34, no. 9, pp. 1466–1472, 2018.

[22] S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu, "Accurate de novo prediction of protein contact map by ultra-deep learning model," *PLoS Computational Biology*, vol. 13, no. 1, article e1005324, 2017.

[23] D. Kozma, I. Simon, and G. E. Tusnady, "PDBTM: Protein Data Bank of transmembrane proteins after 8 years," *Nucleic Acids Research*, vol. 41, Database issue, pp. D524–D529, 2013.

[24] A. J. Li and R. Nussinov, "A set of van der Waals and coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking," *Proteins*, vol. 32, no. 1, pp. 111–127, 1998.

[25] Q. Zou, Q. Hu, M. Guo, and G. Wang, "HAlign: fast multiple similar DNA/RNA sequence alignment based on the centre star strategy," *Bioinformatics*, vol. 31, no. 15, pp. 2475–2481, 2015.

[26] L. H. Greene, T. E. Lewis, S. Addou et al., "The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution," *Nucleic Acids Research*, vol. 35, Database issue, pp. D291–D297, 2007.

[27] L. Wei, J. Tang, and Q. Zou, "Local-DPP: an improved DNA-binding protein prediction method by exploring local

evolutionary information," *Information Sciences*, vol. 384, pp. 135–144, 2017.

[28] S. F. Altschul, T. L. Madden, A. A. Schaffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[29] M. M. Gromiha and P. K. Ponnuswamy, "Prediction of protein secondary structures from their hydrophobic characteristics," *International Journal of Peptide and Protein Research*, vol. 45, no. 3, pp. 225–240, 1995.

[30] D. T. Jones, W. R. Taylor, and J. M. Thornton, "A model recognition approach to the prediction of all-helical membrane protein structure and topology," *Biochemistry*, vol. 33, no. 10, pp. 3038–3049, 1994.

[31] B. Rost, P. Fariselli, and R. Casadio, "Topology prediction for helical transmembrane proteins at 86% accuracy," *Protein Science*, vol. 5, no. 8, pp. 1704–1718, 1996.

[32] R. Casadio, P. Fariselli, C. Taroni, and M. Compiani, "A predictor of transmembrane alpha-helix domains of proteins based on neural networks," *European Biophysics Journal*, vol. 24, no. 3, pp. 165–178, 1996.

[33] H. Wang, Y. Yang, J. Yu et al., "DMCTOP: topology prediction of alpha-helical transmembrane protein based on deep multiscale convolutional neural network," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 36–43, San Diego, CA, USA, November 2019.

[34] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[35] Q. Zou, P. Xing, L. Wei, and B. Liu, "Gene2vec: gene subsequence embedding for prediction of mammalian N^6-methyladenosine sites from mRNA," *RNA*, vol. 25, no. 2, pp. 205–218, 2019.

[36] Z. Lv, C. Ao, and Q. Zou, "Protein function prediction: from traditional classifier to deep learning," *Proteomics*, vol. 19, no. 14, article 1900119, 2019.

[37] L. Wei, Y. Ding, R. Su, J. Tang, and Q. Zou, "Prediction of human protein subcellular localization using deep learning," *Journal of Parallel and Distributed Computing*, vol. 117, pp. 212–217, 2018.

[38] L. Yu, X. Sun, S. Tian, X. Shi, and Y. Yan, "Drug and nondrug classification based on deep learning with various feature selection strategies," *Current Bioinformatics*, vol. 13, no. 3, pp. 253–259, 2018.

[39] L. Peng, M. Peng, B. Liao, G. Huang, W. Li, and D. Xie, "The advances and challenges of deep learning application in biological big data processing," *Current Bioinformatics*, vol. 13, no. 4, pp. 352–359, 2018.

[40] J. P. Dekker, A. Fodor, R. W. Aldrich, and G. Yellen, "A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments," *Bioinformatics*, vol. 20, no. 10, pp. 1565–1572, 2004.

[41] W. R. Atchley, K. R. Wollenberg, W. M. Fitch, W. Terhalle, and A. W. Dress, "Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis," *Molecular Biology and Evolution*, vol. 17, no. 1, pp. 164–178, 2000.

[42] A. A. Fodor and R. W. Aldrich, "Influence of conservation on calculations of amino acid covariance in multiple sequence alignments," *Proteins*, vol. 56, no. 2, pp. 211–221, 2004.

[43] T. A. Hopf, A. G. Green, B. Schubert et al., "The EVcouplings Python framework for coevolutionary sequence analysis," *Bioinformatics*, vol. 35, no. 9, pp. 1582–1584, 2019.

[44] C. A. Brown and K. S. Brown, "Validation of coevolving residue algorithms via pipeline sensitivity analysis: ELSC and OMES and ZNMI, oh my!," *PLoS One*, vol. 5, no. 6, article e10779, 2010.

[45] D. Y. Little and L. Chen, "Identification of coevolving residues and coevolution potentials emphasizing structure, bond formation and catalytic coordination in protein evolution," *PLoS One*, vol. 4, no. 3, article e4762, 2009.

[46] I. Kass and A. Horovitz, "Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations," *Proteins*, vol. 48, no. 4, pp. 611–617, 2002.

[47] W. S. Noble, "What is a support vector machine?," *Nature Biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[48] C. Meng, S. Jin, L. Wang, F. Guo, and Q. Zou, "AOPs-SVM: a sequence-based classifier of antioxidant proteins using a support vector machine," *Frontiers in Bioengineering and Biotechnology*, vol. 7, p. 224, 2019.

[49] Y. Wang, F. Shi, L. Cao et al., "Morphological segmentation analysis and texture-based support vector machines classification on mice liver fibrosis microscopic images," *Current Bioinformatics*, vol. 14, no. 4, pp. 282–294, 2019.

[50] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[51] X. Ru, L. Li, and Q. Zou, "Incorporating distance-based top-n-gram and random forest to identify electron transport proteins," *Journal of Proteome Research*, vol. 18, no. 7, pp. 2931–2939, 2019.

[52] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 79, no. 8, pp. 2554–2558, 1982.

[53] X. Zeng, W. Wang, G. Deng, J. Bing, and Q. Zou, "Prediction of potential disease-associated microRNAs by using neural networks," *Molecular Therapy–Nucleic Acids*, vol. 16, pp. 566–575, 2019.

[54] K. M. Chen, E. M. Cofer, J. Zhou, and O. G. Troyanskaya, "Selene: a PyTorch-based deep learning library for sequence data," *Nature Methods*, vol. 16, no. 4, pp. 315–318, 2019.

[55] F. Pedregosa, G. Varoquaux, A. Gramfort et al., "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[56] X. Zhang, J. Wang, Y. Feng et al., "Structure and molecular mechanism of an anion-selective mechanosensitive channel of small conductance," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 44, pp. 18180–18185, 2012.

[57] A. Amunts, H. Toporik, A. Borovikova, and N. Nelson, "Structure determination and improved model of plant photosystem I," *Journal of Biological Chemistry*, vol. 285, no. 5, pp. 3478–3486, 2010.