Computers in Biology and Medicine 166 (2023) 107529

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Integrating unsupervised language model with multi-view multiple sequence alignments for high-accuracy inter-chain contact prediction

Zi Liu [a,b,1], Yi-Heng Zhu [c,1], Long-Chen Shen [a], Xuan Xiao [b], Wang-Ren Qiu [b,**], Dong-Jun Yu [a,*]

[a] School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei 200, Nanjing, 210094, China
[b] Computer Department, Jingdezhen Ceramic University, Jingdezhen, 333403 , China
[c] College of Artificial Intelligence, Nanjing Agricultural University, Nanjing, 210095 , China

ARTICLE INFO

Keywords:
Inter-chain contact prediction
Multiple sequence alignment
Pre-trained language models
Co-evolution diversity
Deep residual networks

ABSTRACT

Accurate identification of inter-chain contacts in the protein complex is critical to determine the corresponding 3D structures and understand the biological functions. We proposed a new deep learning method, ICCPred, to deduce the inter-chain contacts from the amino acid sequences of the protein complex. This pipeline was built on the designed deep residual network architecture, integrating the pre-trained language model with three multiple sequence alignments (MSAs) from different biological views. Experimental results on 709 non-redundant benchmarking protein complexes showed that the proposed ICCPred significantly increased inter-chain contact prediction accuracy compared to the state-of-the-art approaches. Detailed data analyses showed that the significant advantage of ICCPred lies in the utilization of pre-trained transformer language models which can effectively extract the complementary co-evolution diversity from three MSAs. Meanwhile, the designed deep residual network enhances the correlation between the co-evolution diversity and the patterns of inter-chain contacts. These results demonstrated a new avenue for high-accuracy deep-learning inter-chain contact prediction that is applicable to large-scale protein-protein interaction annotations from sequence alone.

1. Introduction

In living cells, proteins perform functions by interacting with other proteins. The accurate identification of protein complex structure is critical to understand the biological functions and design new drugs [1–8]. Direct determination of the 3D structures of protein complexes through biochemical methods, such as X-ray crystallography [9], nuclear magnetic resonance spectroscopy [10], and cryo-electron microscopy [11], is time-consuming, laborious, and often incomplete. Many computational methods have emerged to accelerate the deposition of protein complex structures. However, the accuracy of complex structure prediction methods is not satisfactory. To improve the complex structure prediction, a promising approach is to determine the inter-chain contacts, as the important constraint to predict the atom coordinates. In light of this, inter-chain contact prediction has been a hot topic.

The existing inter-chain contact prediction methods can be divided into direct coupling analysis-based and machine learning-based methods. In the early stage, direct coupling analysis-based methods lead the trend of inter-chain contact prediction, such as CCMPred [12], Germlin [13], Evcomplex [14,15], and EVfold [16]. Specifically, those methods identify inter-chain contacts by analyzing coevolution residues from a multiple sequence alignment (MSA) to distinguish between direct and indirect correlation effects. However, there is a common drawback: the accuracy of direct coupling analysis-based methods is contingent upon the number of homology sequences. To eliminate this dependence, machine learning-based methods have emerged to extract hand-crafted features from sequences and structures, which can then be used by machine learning approaches to implement inter-chain contact prediction, with typical examples including PAIRpred [17], I-Patch [18] and BIPSPI [19].

Despite the potential advantage, the prediction accuracy of many early machine learning methods was not satisfactory. One of the major reasons is due to the lack of informative feature representation methods, such as position-specific scoring matrices and physiochemical properties, as most of the approaches are based on simple feature representations, which cannot fully extract the complex pattern of inter-chain

* Corresponding author.
** Corresponding author.
E-mail addresses: qiuwangren@jci.edu.cn (W.-R. Qiu), njyudj@njust.edu.cn (D.-J. Yu).
[1] These authors contributed equally.

https://doi.org/10.1016/j.compbiomed.2023.107529
Received 1 August 2023; Received in revised form 30 August 2023; Accepted 19 September 2023
Available online 20 September 2023

contacts. To partly overcome this barrier, several methods, e.g., ComplexContact [20], DeepHomo [21], utilized deep learning technology to predict inter-chain contacts. Compared to traditional machine learning approaches, one advantage of deep learning-based methods is that they could extract more discriminative feature embeddings from preliminary sequences of the protein complex through designing complex neural networks. Nevertheless, the performance of deep learning methods is often hampered by the limitation of protein complex data. The insufficient experimental data significantly limits the effectiveness of training the high-accuracy deep neural network models for inter-chain contact prediction.

To alleviate the issue caused by the lack of protein complex data, a promising approach is to utilize protein language models pre-trained through deep-learning networks on large-scale MSAs. Due to the extensive sequence training and learning, important co-evolution patterns between residues, which are critical for inter-chain contact prediction, can be extracted through the language models and utilized for feature embedding. Recently, a new language model, ESM-MSA transformer [22], has achieved great success in protein monomer structure prediction. Meanwhile, a few deep learning-based methods, e.g., GLINTER [23] and HDIContact [24], utilized ESM-MSA to extract discriminate feature embeddings from MSAs to improve inter-chain contact prediction accuracy for protein complexes. Despite the promising results, there is still room for further improvements due to the following reasons. Specifically, the above-mentioned method extracts the feature embedding from a single type of MSA, in which the co-evolution patterns are very limited. Therefore, multiple types of MSA can provide complementary co-evolution knowledge for further improving the prediction accuracy of inter-chain contacts.

In this work, we proposed a new deep learning method, ICCPred, for high-accuracy inter-chain contact prediction through integrating unsupervised language models with multiple types of MSAs from different biological views. Specifically, the recently proposed ESM-MSA transformer is utilized to extract the complementary co-evolution diversity highly associated with inter-chain contacts from three designed MSAs, which are generated from the views of genomic distance, phylogeny information, and protein-protein interactions. Then, a deep residual network architecture is proposed to enhance the correlation between co-evolution diversity and the patterns of inter-chain contacts. ICCPred has been systematically tested on a large set of non-redundant protein complexes, where the results demonstrated a significant advantage in accurate inter-chain contact prediction over the current state-of-the-art of the field. The standalone package of ICCPred is made freely available through the URL https://github.com/yiheng-zhu/ICCPred.

2. Methods

2.1. Benchmark dataset

A set of 4204 non-redundant protein complexes with medium-to-long sequence lengths is collected from the PDB library and used for training and testing the ICCPred pipeline. The complexes in this set have a pair-wise sequence similarity of <40%. The protein set was randomly split into 3574 for training and validation, and the remaining 630 were independent testing complexes (i.e., TS630). In addition, the training and validation sets were divided at proportions of 90% and 10%, respectively. The average length in the ICCPred benchmark contains 424 residues, with the smallest complex containing 103 residues and the largest holding 700 residues.

In addition to TS630, the models were tested on benchmark datasets compiled in other studies, which include the Baker dataset (i.e., TS032) [13] used to evaluate the ComplexContact [20] and the E.coli dataset (i.e., TS047) [14] used to evaluate HDIContact [24].

2.2. The workflow of ICCPred

ICCPred is a deep learning-based inter-chain contact prediction method. The input and output are two chains (i.e., a receptor sequence and a ligand sequence) and an inter-chain contact map, respectively. As shown in Fig. 1, ICCPred consists of three procedures of multi-view MSA generation, feature embedding using the ESM-MSA transformer, and inter-chain contact prediction using the deep residual network.

2.2.1. Procedure I: multi-view MSA generation

For the input two sequences, we use HHblits software (with the parameters "-diff inf -id 99 -cov 50 -n 3" for HH-suite 2.0.16 program) [32] to search the corresponding MSAs from Uniclust30 [33] and STRING databases [25], denoted as (receptor MSA I, ligand MSA I) and (receptor MSA II, ligand MSA II), respectively. In the first pair of MSAs, we concentrate single sequences to generate two joint MSAs (i.e., joint MSAs I and II) from the views of genomic distance and phylogeny information, respectively; In the second MSA pair, the single sequences are concentrated as another joint MSA (i.e., joint MSA III) from the view of protein-protein interaction (see details in section 2.3).

2.2.2. Procedure II: Feature embedding using the ESM-MSA transformer

Each joint MSA is fed to the ESM-MSA transformer with 12 blocks to generate the corresponding feature embedding, represented as a $(L_1+L_2)\times(L_1+L_2)\times 144$ attention map, where L_1 and L_2 are the lengths of receptor and ligand sequences, respectively. 144 is a preset hyper-parameter in the ESM-MSA transformer (see details in section 2.4). In each attention map, the element in the i-th row and j-th column can be viewed as a 144-D correlation coefficient vector between the i-th position of the receptor sequence and the j-th position of the ligand sequence in the evolution process from a specific view, where $i \le L_1$ and $j \le L_2$. Then, three attention maps (i.e., attention map I, II, and III) are concentrated in the channels to generate a multi-view attention map (i.e., attention map IV) with the scale of $(L_1+L_2)\times(L_1+L_2)\times 432$. Finally, we extract the first L_1 rows and L_2 columns from attention map IV to generate a smaller attention map (i.e., attention map V) with the scale of $L_1\times L_2\times 432$, as the final feature embedding.

2.2.3. Procedure III: Inter-chain contact prediction using the deep residual network

The attention map V is fed to a deep residual network with 5 basic blocks to generate a $L_1\times L_2$ confidence score matrix of inter-chain contact, where the element in the i-th row and j-th column is the confidence score of contact between the i-th position of the receptor sequence and the j-th position of the ligand sequence. The output of the t-th residual basic block admits a representation of the form

$$x_t = dp(f(x_{t-1} + cb(f(cb(x_t))))) \quad (1)$$

where x_t is the output in the t-th block, dp (.) is the dropout layer with a ratio of 0.2 to prevent over-fitting in the training stage, $f(.)$ is the RELU activate function, and $cb(.)$ is the combination of the 2D convolution layer with the filter size of 3×3 and the batch normalization layer.

2.3. Multi-view MSA generation strategies

In ICCPred, we proposed three strategies from different views, including genomic distance, phylogeny information, and protein-protein interaction, separately concentrating two monomer MSAs as a joint MSA.

2.3.1. Genomic distance-based strategy (GDS)

We use HHblits software [32] to search against the Uniclust30 database [33] to generate the corresponding monomer MSAs for a receptor sequence and a ligand sequence. Then, two monomer MSAs are concentrated based on the proteins in close proximity on the genome

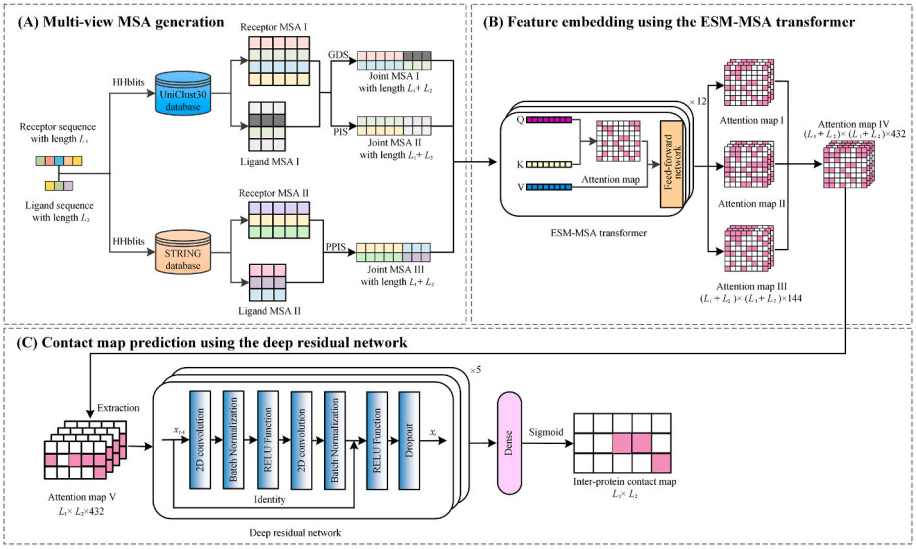

Fig. 1. The overview of the ICCPred architecture. (A) We generate MSAs from the views of genomic distance, phylogeny information, and protein-protein interaction. (B) The feature embeddings of MSAs are extracted by the ESM-MSA transformer. (C) We capture contact maps from MSA embeddings on inter-chain using the deep residual network.

[13,20,26]; In other words, proteins co-located on the chromosome into operons on the same operon are more likely to interact [13,20,26,27]. Therefore, we retrieved the coding DNA sequence (CDS) of the protein and fetched the genomic contig of CDS from European Nucleotide Archive (ENA) [28]. To summarize, the concatenated pairs of proteins are based on two criteria: first, the CDS of each concatenated protein pair must be located on the same genomic contig; second, the intergenic distance between the two concatenated proteins must be less than a certain threshold of 20 [13,14].

2.3.2. Phylogeny information-based strategy (PIS)

There is a challenging problem that two genes may interact even if their genomic distances are not close [27]. We use phylogeny information to concatenate two monomer MSAs to overcome this problem. First, according to the phylogeny tree in the Taxonomy database [29], we grouped the proteins in each MSA. Secondly, we ranked the similarity of query sequences and the proteins in each species of each MSA from high to low. In the two MSAs, the highest-ranked hit from one species was paired with the highest-ranked hit of the interacting chain from the same organism species. The paired sequences contain the inter-chain coevolutionary information in the concentrated MSA [30].

2.3.3. Protein-protein interaction-based strategy (PPIS)

The protein-protein interaction network database (i.e., STRING [25]) (https://cn.string-db.org/) records the confidence score of interaction between two proteins. Therefore, we determine whether two sequences in each monomer MSA are concentrated according to the corresponding confidence score of interaction, where each monomer MSA is generated using HHblits to search the STRING database.

2.4. The details of the ESM-MSA transformer

The architecture of the ESM-MSA transformer is illustrated in Fig. S1. The masking strategy is performed on the corresponding tokens (i.e., amino acids) for an input MSA. Then, the masked MSA is encoded as three one-hot encoding matrices from different views, respectively, which are then fused to generate an initial embedding matrix. Next, this embedding matrix is fed to a self-attention network with 12 blocks, each consisting of a row attention layer, a column attention layer, a feed-forward network, a batch normalization layer, and a dropout layer. The output of the self-attention network is a probability matrix, indicating the probability of belonging to types of amino acids for each position in masked MSA. Finally, the loss function is designed as a negative log-likelihood function between the masked MSA and the probability matrix, to ensure that the prediction model correctly predicts the true amino acids in the masked position as much as possible.

The Adam optimization algorithm optimizes the ESM-MSA transformer by minimizing the loss function. Then, the attention maps of 12 attention heads in all 12 blocks are concentrated to generate the final attention map with the scale of $L \times L \times 144$, as the feature embedding of ESM-MSA, where L is the length of a single sequence in the inputted MSA. A detailed description of the ESM-MSA transformer is given in **Text S1**.

2.5. Evaluation metrics

The performance of inter-chain contact map prediction was evaluated by the following criteria: precision, area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPR). Precision is the ratio of correctly predicted contacts in the top N ($N = 1$, 5, 10, 20, 50, and 100) number of predicted contacts. The overall performance in the benchmark dataset was represented by the average precision of all the targets. In addition, we utilized the precision of top L/K predicted contacts to illustrate the L-dependent accuracy commonly employed in intra- and inter-chain contact prediction, where L is the total length of the two protein chains and $K = 30$, 20, 10, 5, and 2 [20,21,24,31]. AUROC quantifies the balance between sensitivity and specificity, with higher scores indicating better performance. Similarly, AUPR measures the equilibrium between precision and recall, with elevated scores suggesting superior performance.

3. Results

3.1. Comparison with competing inter-chain contact predictors

To evaluate the performance of our proposed methodology for inter-chain contact map prediction, we compared ICCPred with the deep learning-based methods, including GLINTER and HDIContact on three different datasets.

3.1.1. Performance on the TS630 dataset

Fig. 2 shows that ICCPred significantly outperformed GLINTER and

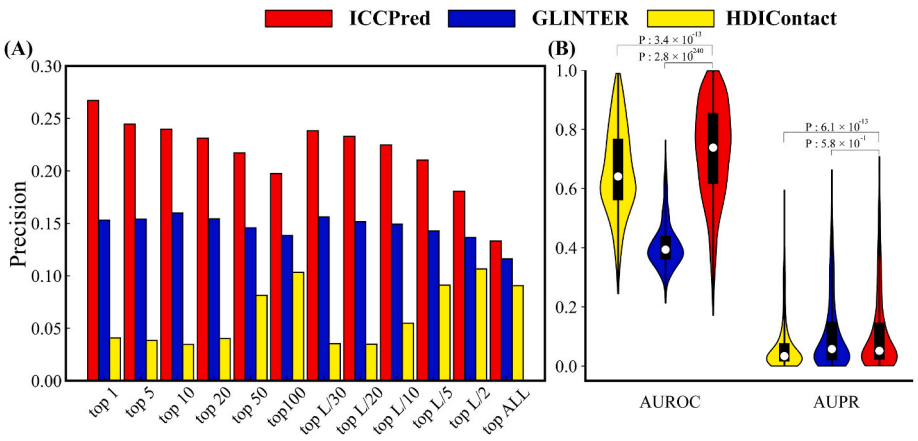

Fig. 2. ICCPred yielded robust performance and outperformed the GLINTER and HDIContact on the TS630 test dataset. (A) and (B) show the evaluation metrics with top N and top L/N, AUROC, and AUPR, respectively.

HDIContact across all the evaluated metrics on the TS630 dataset. When combining all targets, the overall top 10/20 precision by ICCPred (0.240/0.232) is 33.3%/33.4% and 84.0%/83.0% higher than the GLINTER (0.160/0.154) and HDIContact (0.034/0.040), respectively. Moreover, the performance of ICCPred is significantly superior to that of GLINTER and HDIContact in terms of AUROC values, showing an increase of up to 43.3% and 9.0%, respectively (as depicted in Fig. 2B). These differences correspond to remarkably low p-values of 2.8E-240 and 1.3E-13 in the student's t-test. Additionally, the AUPR value of ICCPred (0.15) surpasses that of GLINTER and HDIContact by 3.5% and 39.1%, respectively. Furthermore, we also evaluated top L/K predicted inter-chain contacts to show the L-dependent precision. ICCPred achieves 0.234 top $L/20$ precision, while GLINTER and HDIContact have 15.2% and 3.4% top $L/20$ precision, respectively. See detailed results in Supplementary Tables S1 and S2.

3.1.2. Performance on the TS047 dataset

We further benchmark three deep learning methods on the TS047 dataset, as shown in Tables 1 and 2. It could be found that ICCPred obtains 0.636/0.609/0.566 precision in the top 5/10/20 predicted contacts, which is 57.1%/56.0%/54.3% and 8.3%/8.1%/4.9% higher than the GLINTER and HDIContact, respectively. Moreover, regarding the precision in the top L/K (K = 20, 10, and 5) predicted contacts, ICCPred gains 111.5% and 2.1% average improvements in the above-mentioned five evaluation metrics compared to GLINTER and HDIContact, respectively. Additionally, ICCPred obtained the highest top ALL precision of 0.310, which gains 55.9% and 8.1% improvements compared to GLINTER and HDIContact, respectively.

3.1.3. Performance on the TS032 dataset

As shown in Table 3, our ICCPred model outperformed both direct coupling analysis-based (i.e., EVcomplex, GremlinComplex, EVfold, and CCMPred) and machine learning-based methods (i.e., ComplexContact and GLINTER) on the TS032 test dataset, in terms of precision in the top $L/5$, 50, 20, 10, and 5 predicted contacts. Specifically, ICCPred achieves 69.0%, 42.2%, 39.2%, 36.9%, 13.6%, and 70.4% average improvements in the five evaluation metrics as mentioned earlier in comparison to EVcomplex, GremlinComplex, EVfold, CCMPred, ComplexContact, and GLINTER, respectively.

Table 1
Average precision in top N predicted contacts on TS047 dataset for three deep learning methods.

Methods	Top 1	Top 5	Top 10	Top 20	Top 50	Top 100
ICCPred	**0.669**	**0.636**	**0.609**	**0.566**	**0.509**	**0.457**
GLINTER	0.255	0.272	0.268	0.259	0.234	0.206
HDIContact	0.660	0.583	0.560	0.538	0.509	0.461

3.1.4. Case study

Fig. 3 presents an example of a complex form of Mycobacterium tuberculosis VapBC11 toxin-antitoxin (PDB ID: 6a7v), where ICCPred shows significantly better performance than GLINTER and HDIContact in the inter-chain contact prediction. Taking the top 50 predicted residue pairs as an example, ICCPred, GLINTER, and HDIContact achieve 45, 12, and 29 correct residue pairs with precision rates of 90%, 58%, and 24%, respectively. Furthermore, the false positive contacts predicted by ICCPred are predominantly proximal to the native contacts, as illustrated in the bottom panel of Fig. 3C. In Contrasty, the false positives of HDIcontact and GLINTER are far away from the native contacts, displayed in the bottom panels of Fig. 3A and B.

3.2. Contribution analysis for different MSAs

We designed the following test to analyze the contributions of three types of MSAs (i.e., GDS, PIS, and PPIS) in inter-chain contact prediction. First, we generate seven feature embeddings by feeding different MSAs to the ESM-MSA transformer, including three individual embeddings from GDS, PIS, and PPIS, and four combination embeddings from GDS + PIS (GP), GDS + PPIS (GI), and PIS + PPIS (PP), GDS + PIS + PPIS (GSP), respectively, where "+" means that the individual feature embeddings from different MSAs are concentrated as a combination embedding. Moreover, to demonstrate the strongness of the MSA feature embedding, we select the feature embedding of the single sequence from the ESM2 transformer [32] as a comparison baseline. Additionally, we use the recently proposed cascade MSA construction algorithm, i.e., cpxDeepMSA [30], to concentrate three individual MSAs as a combination MSA (denoted as CPX), which is further fed to the ESM-MSA to extract the corresponding feature embedding, as another comparison baseline. Finally, each of the above-mentioned nine feature embeddings is fed to the designed deep residual network as an inter-chain contact prediction model, which is further benchmarked in our TS630 test dataset.

Fig. 4 illustrates the performance of nine feature embeddings on the TS630 test dataset, where the detailed results are listed in Table S3. In comparison with the feature embedding of the single sequence from the ESM2 model, three individual MSA feature embeddings (i.e., GDS, PIS, and PPIS) from the ESM-MSA transformer separately achieve 309.3%, 271.0%, and 397.8% average improvements in all of 12 evaluation metrics. This indicates that the MSA contains much more knowledge than the single sequence in the inter-chain contact prediction. Among seven in-house MSA feature embeddings, GSP consistently shows the best performance in all metrics, demonstrating that each of the three

Table 2
Average precision in top *L/K* predicted contacts, AUPR, and AUROC on the TS047 dataset. ALL represents the number of native contacts on the target.

Methods	AUPR	AUROC	Top				
			L/20	*L*/10	*L*/5	*L*/2	ALL
ICCPred	**0.272**	**0.832**	**0.575**	**0.531**	**0.491**	**0.424**	**0.310**
GLINTER	0.121	0.765	0.259	0.249	0.232	0.197	0.137
HDIContact	0.250	0.831	0.556	0.519	0.492	0.413	0.285

Table 3
Performance comparison between ICCPred and six competing methods on TS032.

Methods	Top L/5	Top 50	Top 20	Top 10	Top 5
Evcomplex[a]	0.096	0.144	0.216	0.266	0.096
GremlinComplex[a]	0.147	0.260	0.412	0.528	0.147
EVfold[a]	0.161	0.276	0.421	0.548	0.161
CCMPred[a]	0.176	0.299	0.460	0.555	0.176
ComplexContact[a]	0.385	0.504	0.605	0.659	0.385
GLINTER	0.172	0.187	0.192	0.216	0.172
ICCPred	**0.564**	**0.626**	**0.668**	**0.684**	**0.564**

Note: [a] Experimental results excerpt from the reference [20], which only lists five evaluation metrics, including top *L*/5, 50, 20, 10, and 5 precision.

MSAs helps improve contact prediction. Moreover, after removing PIS from GSP, the combination of the remaining two feature embeddings (i.e., GI) shows the worst performance in all four combination embeddings, indicating that PIS makes the most contribution among the three MSAs in contact prediction. Additionally, GPS achieves an 8.6% average increase over CPX in all 12 metrics. Fig. S2 demonstrates that the GSP features outperform other MSA features in terms of AUROC, showing improvements of 25.2%, 1.7%, 6.6%, 6.6%, 2.9%, 1.1%, 0.8%, and 2.8% in comparison with ESM2, CPX, GDS, PPIS, PIS, PP, GP, and GI, respectively. This observation further means that the combination of MSAs shows better performance at the feature embedding level than the sequence level.

Fig. 5 displays a representative protein complex (PDB ID: 1u5t), showcasing the superior performance of GSP compared to GDS, PIS, and PPIS. In the top 50 predicted contacts, GSP achieves a precision of 78%, while GDS, PIS, and PPIS separately exhibit a precision of 64%, 70%, and 56%. The best performance of GSP indicates that the knowledge among the three MSAs could be complementary to improve contact prediction accuracy. These data are consistent with the experiment results in Fig. 4.

4. Conclusions

We have presented ICCPred, a novel deep learning model for inter-chain interface residue-residue contact predictions, which integrates multi-view MSA generation methodology with unsupervised protein language transformers. This model generates MSAs from three different biology views and transfers the co-evolutionary patterns learned by pre-training a protein language model to extract MSA 2D embeddings. The deep residual network is utilized to capture the environment context of residue pairs from the 2D MSA embedding feature for high-accuracy contact prediction. ICCPred was systematically benchmarked on three independent test datasets, showing superior performance over existing methods. The improvement of ICCPred can be attributed to two advancements. First and most importantly, three types of MSA generation from different views provide complementary knowledge, extracted as the discriminative feature embeddings using the ESM-MSA transformer. Secondly, the designed deep residual network effectively captures the relationship between the MSA pattern and residue-residue contact.

Despite the promising results, there is still room for further improvement. First, the feature embedding fusion strategy, currently used in ICCPred, simply concentrates all embeddings in serial mode, easily leading to information redundancy. Therefore, a more advanced feature fusion method may alleviate the negative impact caused by information redundancy in our future work. Second, the recently proposed protein complex structure prediction models (e.g., AlphaFold-Multimer [33]) provide a promising way to further enhance the prediction of inter-chain contact. Studies along these lines are in progress.

Funding

This work was supported by grants from the National Natural Science Foundation of China (62372234, 62072243, 61772273, 62162032, and 32260154), the Natural Science Foundation of Jiangsu (BK20201304), and the Foundation of National Defense Key Laboratory of Science and Technology (JZX7Y202001SY000901), and the Scientific Research Plan of the Department of Education of Jiangxi Province, China (GJJ2201004).

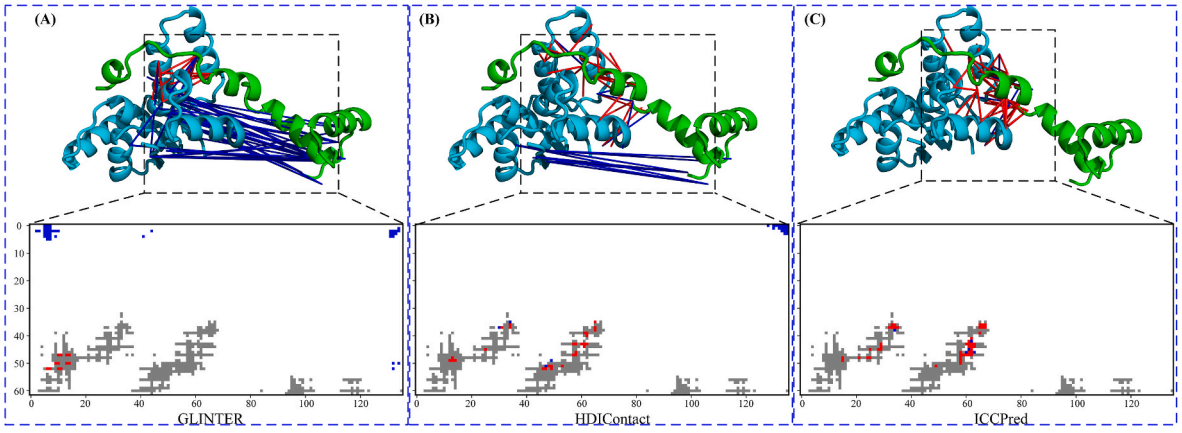

Fig. 3. Illustrative examples for GLINTER (A), HDIContact (B), and ICCPred (C) on protein complex 6A7V at the top 50 predicted contacts. Native structures of two monomers are shown in green and cyan, respectively. True positives are depicted in red, while solid blue lines represent false positives. On the bottom of each panel, grey dots indicate naive contacts, red dots represent the true positives in the top 50 predicted contacts, and blue dots are false positives.

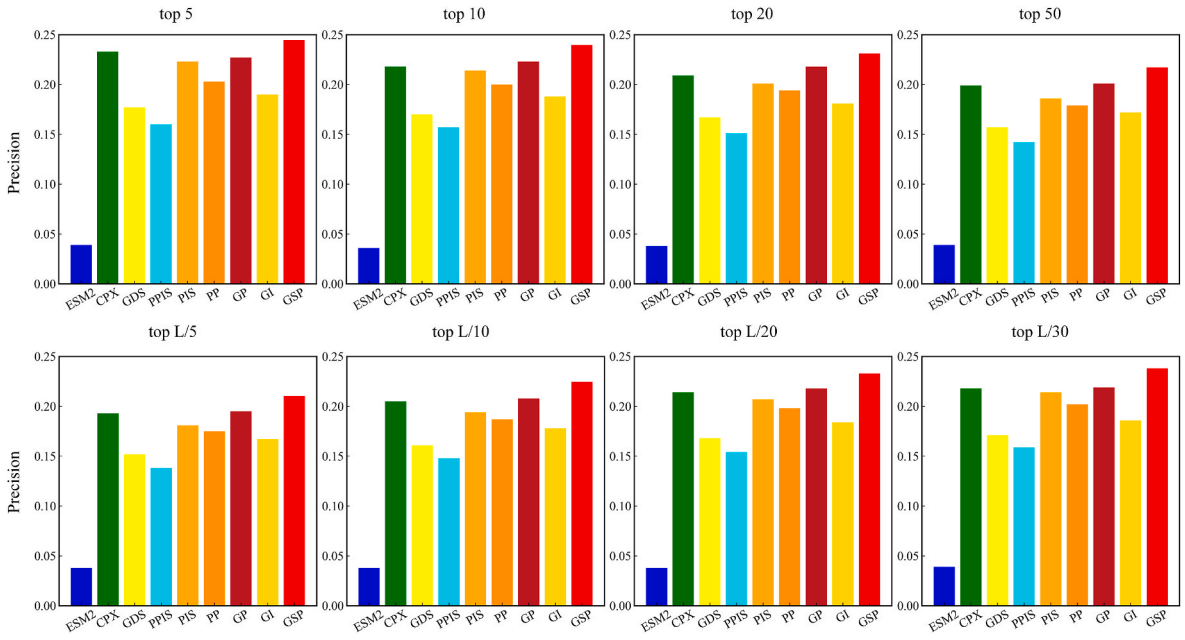

Fig. 4. Performance comparison between nine feature embeddings on the TS630 dataset.

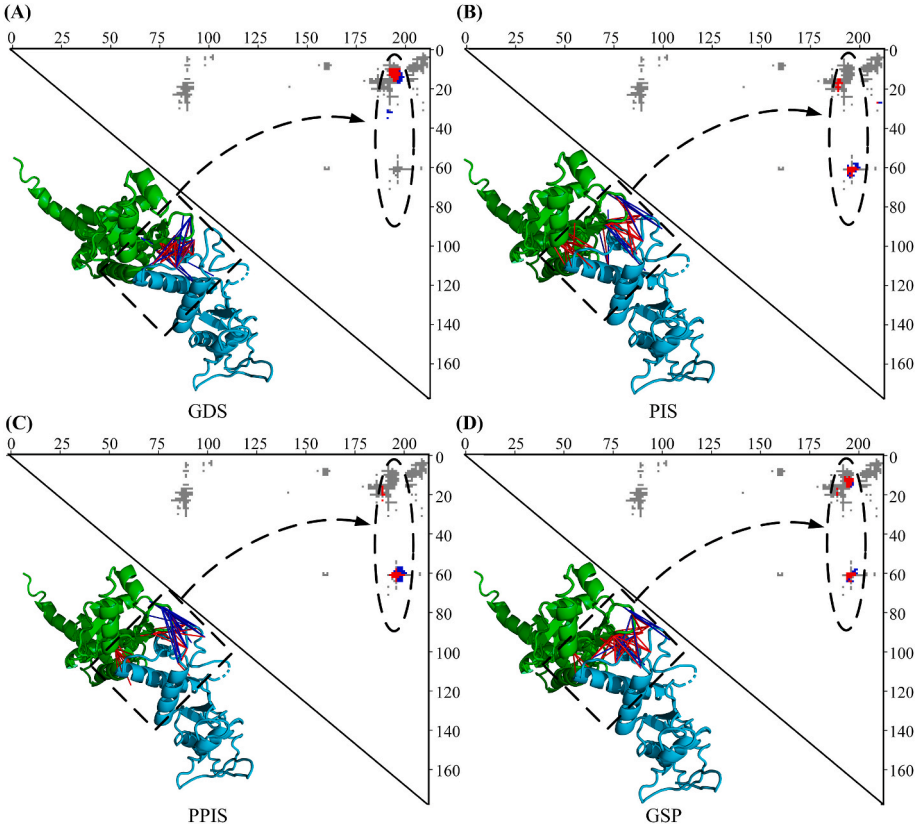

Fig. 5. An illustrative example of the protein complex (PDB ID: 1u5t) in the inter-chain contact prediction using four different feature embeddings, including GDS, PIS, PPIS, and GSP, as shown in panels (A), (B), (C), and (D), respectively. In each panel, the top 50 predicted contacts between two chains are displayed. Native structures of two monomers are shown in green and cyan, respectively. True positives are depicted in red, while false positives are represented by solid blue lines. On the upper-right side of each panel, grey dots indicate naive contacts, red dots represent the true positives in the top 50 predicted contacts, and blue dots are false positives.

Declaration of competing interest

The authors declare no conflict of interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2023.107529.

References

[1] P. Kong, G. Huang, W. Liu, Identification of protein complexes and functional modules in E. coli PPI networks, BMC Microbiol. 20 (1) (2020).
[2] A. Mafi, S.-K. Kim, W.A. Goddard III, The mechanism for ligand activation of the GPCR–G protein complex, Proc. Natl. Acad. Sci. USA 119 (18) (2022), e2110085119.
[3] C.S. Olver, Erythrocyte structure and function, Schalm's Veterin. Hematol. (2022) 158–165.
[4] J. Hu, Y. Li, J.-Y. Yang, H.-B. Shen, D.-J. Yu, GPCR–drug interactions prediction using random forest with drug-association-matrix-based post-processing procedure, Comput. Biol. Chem. 60 (2016) 59–71.
[5] D.-J. Yu, J. Hu, Q.-M. Li, Z.-M. Tang, J.-Y. Yang, H.-B. Shen, Constructing query-driven dynamic machine learning model with application to protein-ligand binding sites prediction, IEEE Trans. NanoBioscience 14 (1) (2015) 45–58.
[6] Y.-H. Zhu, C. Zhang, Y. Liu, G.S. Omenn, P.L. Freddolino, D.-J. Yu, Y. Zhang, TripletGO: integrating transcript expression profiles with protein homology inferences for gene function prediction, Dev. Reprod. Biol. 20 (5) (2022) 1013–1027.
[7] M. Arif, S. Ahmad, F. Ali, G. Fang, M. Li, D.-J. Yu, TargetCPP: accurate prediction of cell-penetrating peptides from optimized multi-scale features using gradient boost decision tree, J. Comput. Aided Mol. Des. 34 (2020) 841–856.
[8] Y.-H. Zhu, J. Hu, X.-N. Song, D.-J. Yu, DNAPred: accurate identification of DNA-binding sites from protein sequence by ensembled hyperplane-distance-based support vector machines, J. Chem. Inf. Model. 59 (6) (2019) 3057–3071.
[9] Y. Shi, A glimpse of structural biology through X-ray crystallography, Cell 159 (5) (2014) 995–1014.
[10] M.R. O'Connell, R. Gamsjaeger, J.P. Mackay, The structural analysis of protein–protein interactions by NMR spectroscopy, Proteomics 9 (23) (2009) 5224–5232.
[11] M. Serna, Hands on methods for high resolution cryo-electron microscopy structures of heterogeneous macromolecular complexes, Front. Mol. Biosci. 6 (2019) 33.
[12] S. Seemayer, M. Gruber, J. Soding, CCMpred-fast and precise prediction of protein residue-residue contacts from correlated mutations, Bioinformatics 30 (21) (2014) 3128–3130.
[13] S. Ovchinnikov, H. Kamisetty, D. Baker, Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information, Elife 3 (2014), e02030.
[14] T.A. Hopf, C.P.I. Scharfe, J.P.G.L.M. Rodrigues, A.G. Green, O. Kohlbacher, C. Sander, A.M.J.J. Bonvin, D.S. Marks, Sequence co-evolution gives 3D contacts and structures of protein complexes, Elife 3 (2014).
[15] H. Szurmant, M. Weigt, Inter-residue, inter-protein and inter-family coevolution: bridging the scales, Curr. Opin. Struct. Biol. 50 (2018) 26–32.
[16] D.S. Marks, L.J. Colwell, R. Sheridan, T.A. Hopf, A. Pagnani, R. Zecchina, C. Sander, Protein 3D structure computed from evolutionary sequence variation, PLoS One 6 (12) (2011), e28766.
[17] F.u.A. Afsar Minhas, B.J. Geiss, A. Ben-Hur, PAIRpred: partner-specific prediction of interacting residues from sequence and structure, Proteins: Struct., Funct., Bioinf. 82 (7) (2014) 1142–1155.
[18] R. Hamer, Q. Luo, J.P. Armitage, G. Reinert, C.M. Deane, I-Patch: interprotein contact prediction using local network information, Proteins: Struct., Funct., Bioinf. 78 (13) (2010) 2781–2797.
[19] R. Sanchez-Garcia, C.O.S. Sorzano, J.M. Carazo, J. Segura, BIPSPI: a method for the prediction of partner-specific protein–protein interfaces, Bioinformatics 35 (3) (2019) 470–477.
[20] H. Zeng, S. Wang, T. Zhou, F. Zhao, X. Li, Q. Wu, J. Xu, ComplexContact: a web server for inter-protein contact prediction using deep learning, Nucleic Acids Res. 46 (W1) (2018) W432–W437.
[21] Y. Yan, S.-Y. Huang, Accurate prediction of inter-protein residue–residue contacts for homo-oligomeric protein complexes, Briefings Bioinf. 22 (5) (2021) bbab038.
[22] R.M. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, A. Rives, In MSA transformer, Int. Conf. Mach. Learn. (2021) 8844–8856, 2021; PMLR.
[23] Z.W. Xie, J.B. Xu, Deep graph learning of inter-protein contacts, Bioinformatics 38 (4) (2022) 947–953.
[24] W. Zhang, Q. Meng, J. Wang, F. Guo, HDIContact: a novel predictor of residue–residue contacts on hetero-dimer interfaces via sequential information and transfer learning strategy, Briefings Bioinf. 23 (4) (2022) bbac169.
[25] D. Szklarczyk, J.H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N.T. Doncheva, A. Roth, P. Bork, L.J. Jensen, C. von Mering, The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible, Nucleic Acids Res. 45 (D1) (2017) D362–D368.
[26] T.A. Hopf, C.P. Scharfe, J.P. Rodrigues, A.G. Green, O. Kohlbacher, C. Sander, A. M. Bonvin, D.S. Marks, Sequence co-evolution gives 3D contacts and structures of protein complexes, Elife 3 (2014).
[27] C. Feinauer, H. Szurmant, M. Weigt, A. Pagnani, Inter-protein sequence Co-evolution predicts known physical interactions in bacterial ribosomes and the trp operon, PLoS One 11 (2) (2016), e0149166.
[28] P.W. Harrison, B. Alako, C. Amid, A. Cerdeno-Tarraga, I. Cleland, S. Holt, A. Hussein, S. Jayathilaka, S. Kay, T. Keane, R. Leinonen, X. Liu, J. Martinez-Villacorta, A. Milano, N. Pakseresht, J. Rajan, K. Reddy, E. Richards, M. Rosello, N. Silvester, D. Smirnov, A.L. Toribio, S. Vijayaraja, G. Cochrane, The European Nucleotide archive in 2018, Nucleic Acids Res. 47 (D1) (2019) D84–D88.
[29] S. Federhen, The NCBI Taxonomy database, Nucleic Acids Res. 40 (D1) (2012) D136–D143.
[30] Z. Liu, D.-J. Yu, cpxDeepMSA: a deep cascade algorithm for constructing multiple sequence alignments of protein–protein interactions, Int. J. Mol. Sci. 23 (15) (2022), 8459.
[31] Y. Li, C. Zhang, E.W. Bell, W. Zheng, X. Zhou, D.-J. Yu, Y. Zhang, Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks, PLoS Comput. Biol. 17 (3) (2021), e1008865.
[32] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, Evolutionary-scale prediction of atomic-level protein structure with a language model, Science 379 (6637) (2023) 1123–1130.
[33] R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, Protein complex prediction with AlphaFold-Multimer, bioRxiv (2021), 463034, 2021.10. 04.