International Journal of
Molecular Sciences

MDPI

Article

cpxDeepMSA: A Deep Cascade Algorithm for Constructing Multiple Sequence Alignments of Protein–Protein Interactions

Zi Liu and Dong-Jun Yu *

School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; liuzi189836@163.com
* Correspondence: njyudj@njust.edu.cn

Abstract: Protein–protein interactions (PPIs) are fundamental to many biological processes. The coevolution-based prediction of interacting residues has made great strides in protein complexes that are known to interact. A multiple sequence alignment (MSA) is the basis of coevolution analysis. MSAs have recently made significant progress in the protein monomer sequence analysis. However, no standard or efficient pipelines are available for the sensitive protein complex MSA (cpxMSA) collection. How to generate cpxMSA is one of the most challenging problems of sequence coevolution analysis. Although several methods have been developed to address this problem, no standalone program exists. Furthermore, the number of built-in properties is limited; hence, it is often difficult for users to analyze sequence coevolution according to their desired cpxMSA. In this article, we developed a novel cpxMSA approach (cpxDeepMSA. We used different protein monomer databases and incorporated the three strategies (genomic distance, phylogeny information, and STRING interaction network) used to join the monomer MSA results of protein complexes, which can prevent using a single method fail to the joint two-monomer MSA causing the cpxMSA construction failure. We anticipate that the cpxDeepMSA algorithm will become a useful high-throughput tool in protein complex structure predictions, inter-protein residue-residue contacts, and the biological sequence coevolution analysis.

Keywords: protein–protein interactions; protein complex; multiple sequence alignment; genomic distance; phylogeny information; STRING interaction network; sequence coevolution analysis

Citation: Liu, Z.; Yu, D.-J. cpxDeepMSA: A Deep Cascade Algorithm for Constructing Multiple Sequence Alignments of Protein–Protein Interactions. *Int. J. Mol. Sci.* **2022**, *23*, 8459. https://doi.org/10.3390/ijms23158459

Academic Editors: Suren Rao Sooranna and Edmond Dik Lung Ma

Received: 11 June 2022
Accepted: 28 July 2022
Published: 30 July 2022

1. Introduction

Proteins play crucial roles in almost all biological processes in cells. These important biomolecules, particularly proteins, accomplish their roles by using intermolecular interactions, such as the identity, dynamics, and specificity of protein interactions [1,2]. Experimental screens have identified tens of thousands of protein–protein interactions (PPI) or protein complexes, and structural biology has provided detailed functional insight into select 3D protein complexes. However, the structures of many protein complexes are unknown, and there is still little, or no, 3D information for a significant percentage of currently known PPIs or protein complexes in bacteria, yeast, and humans [3,4]. The structures of many essential PPI complexes, including those bound with the cell membrane, are difficult, if not impossible, to solve using the current techniques. The computational approach has therefore become an increasingly important means to obtain protein complex structures, especially for large-scale protein complex structure modeling [5].

With the rapid growth in our knowledge of genetic variation at the sequence level, there is increased interest in linking sequences with the change in molecular interactions. However, the current experimental approaches cannot meet the demand for residue-level information on these interactions. Recent work has demonstrated the accuracy of coevolution-based contact prediction for monomeric proteins using global statistical models [6–8]. The chain's multiple sequence alignment (MSA) is the fundament of the quantified

coevolution. MSA provides more information by showing conserved regions and motifs of structural and functional importance within the protein family. Furthermore, the MSA is an essential part of protein structure prediction [9], protein contact map [10], second structure feature [11], ligand-binding site prediction [12], homologous templates [13], gene ontology [14], phylogenetic analysis [15], and many other valuable procedures in sequence research [16]. Therefore, many MSA construct methods have been developed, such as BLAST [17], HHblits [18] from the HH-suite [19], and Jackhammer and HMMsearch tools from the HMMER suite [20], MetaPSICOV2 [21], and DeepMSA [22].

In contrast to the extensive work on monomeric proteins, little is known about the utility of such statistical models for predicting protein–protein interactions or protein complexes. Coevolution is at the basis of many modern computational techniques for characterizing protein−protein interactions. Therefore, as shown in Figure 1, how to build the multiple sequence alignment (MSA) of the protein–protein interaction or protein complex is an important issue that needs to be addressed. EVcomplex [3], Gremlin-Complex [23], and ComplexContact [4] are based on the genomic distances to build protein complex multiple sequence alignments. ComplexContact [4] also creates protein complex multiple sequence alignment by using a phylogeny-based method.

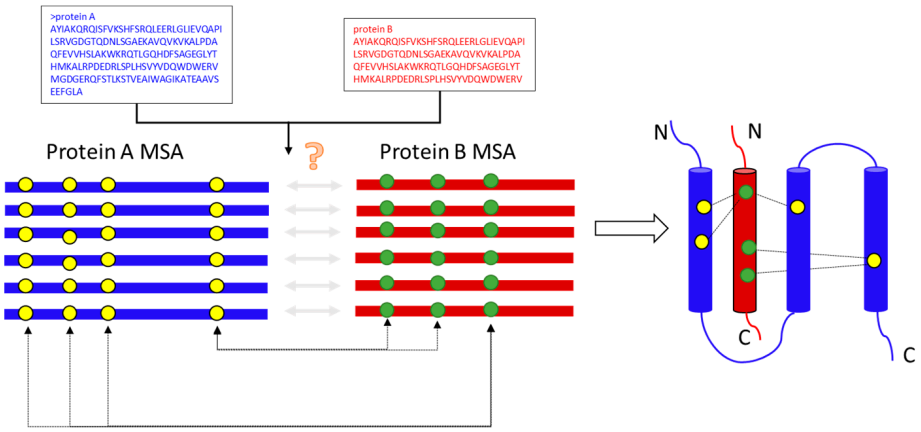

Figure 1. The flowchart of the building cpxMSA. The two circles (yellow and green) connected by double arrow lines indicate sites of coevolution (**left**) to identify evolutionary couplings between co-evolving inter-chain residue pairs (**right**).

Although the above methods developed the genomic-based and phylogeny-based methods to generate protein complex multiple sequence alignments, few standalone pipelines/programs exist that efficiently generate sensitive protein complex MSAs from the input protein complex sequences; hence, there was an urgent need to address this issue. Inspired by the protein monomer MSA algorithm DeepMSA, we developed and released cpxDeepMSA, a new open-source program to construct deep and sensitive protein complex MSAs by merging sequences from three different strategies through a hybrid homology–detection approach.

2. Results and Discussion

2.1. Evaluation

We evaluated our cpxMSA method for contact prediction using the state-of-the-art programs CCMpred [24] and trRosettaX [25,26]. We calculated the accuracy of the top 50, 20, 10, 5, and top L/k (k = 5, 10, 20, 50) predicted contacts where L is the total length of the two protein chains. The prediction is defined as the percentage of correctly predicted contacts among the top predictions.

2.2. cpxDeepMSA Increases Protein Complex Contact Prediction Accuracy

The genomic-, phylogeny-, and STRING-based methods for cpxMSA construction complement each other. Generally, for prokaryotic species, the genome-based method works better, and for eukaryotes, our phylogeny-based method works better, as shown in Table 1, which was tested on the PDB100 (of 100 heterodimers) database by using the predictor trRosettaX (with defaults: "predict.py -i input.a3m -o output.npz -mdir./model_res2net_202012"). The benchmark PDB100 was extracted from a Protein Data Bank (PDB) [27], and the sequence identity cutoff in the benchmark was 40%. The results indicate that for the cpxMSA construction method there is little difference between genomic-based (stage 1) and phylogeny-based (stage 2) and all of them are better than STRING-based (stage 3). The MSA from cpxDeepMSA outperforms the other three MSAs for contact prediction. For instance, when using the MSA from cpxDeepMSA, the precision for the top five contacts was 0.673; this was 58.7%, 16.6%, and 99.1% higher than that of the MSA from genomic-, phylogeny- and STRING-based, respectively.

Table 1. Inter-protein contact prediction precision on the PDB100 database by trRosettaX. Bold font indicates the highest value in each category.

MSA	L/5	L/10	L/20	L/50	50	20	10	5
Genomic-based	0.273	0.325	0.372	0.414	0.302	0.365	0.397	0.424
Phylogeny-based	0.353	0.430	0.499	0.564	0.394	0.485	0.538	0.577
STRING-based	0.210	0.253	0.295	0.333	0.228	0.282	0.316	0.338
cpxDeepMSA	0.398	0.491	0.572	0.645	0.449	0.560	0.629	0.673

To further investigate the effectiveness of cpxDeepMSA, we list in Table 2 the comparison of the contact map prediction results of cpxDeepMSA and RoseTTAFold (RF) MSA [28] on the Baker's dataset [23]. We used CCMpred with parameters "CCMpred input.aln output.mat -n 100 -e 0 -A" to detect the co-evolution on each alignment. Significant improvement of cpxDeepMSA was shown for the contact map prediction over RF MSA on the predictor CCMpred and trRosettaX. In comparison, the corresponding precision for the *top* 10 contacts of RF MSA by CCMpred and trRosettaX were 0.4% and 51.9%, respectively. cpxDeepMSA achieved precision for the *top* 10, with 38.5% and 55.2%, which were 9225.0% and 6.4% higher than RF MSA with CCMpred and trRosettaX, respectively.

Table 2. Inter-protein contact prediction precision (%) on Baker's data.

Predictor	MSA	L/5	L/10	L/20	L/50	50	20	10	5
CCMpred	RF MSA	0.012	0.007	0.006	0.009	0.010	0.006	0.004	0.007
	cpxDeepMSA	0.137	0.214	0.306	0.377	0.242	0.333	0.385	0.400
tRosettaX	RF MSA	0.340	0.416	0.462	0.535	0.406	0.461	0.519	0.556
	cpxDeepMSA	0.334	0.418	0.487	0.565	0.410	0.494	0.552	0.578

2.3. Web-Server and User Guide

To enhance the value of its practical applications, the web server for cpxDeepMSA was established. Below, we further give a step-by-step guide on how to use the web server to obtain the desired results.

2.3.1. Server Input

Opening the web server at https://zhanggroup.org/cpxDeepMSA/, you will see the top page of the cpxDeepMSA on your computer screen, as shown in Figure 2. The input to the cpxDeepMSA server involves two single-chain amino acid sequence files in FASTA format. After submitting a job, a URL link with a random job ID is generated, allowing the user to check the results and keep the data private. The user must provide an email

address when submitting a job, and the server will automatically send a notification email with a link to the results page upon the job completion.

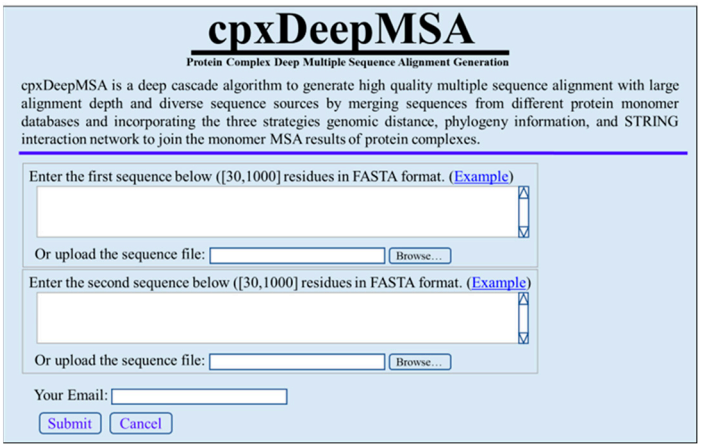

Figure 2. A semi-screenshot showing the top page of the cpxDeepMSA web server at https://zhanggroup.org/cpxDeepMSA/.

2.3.2. Server Output

The cpxDeepMSA results page consists of seven sections: (i) A summary of the multiple sequence alignments and sequence analysis compressed package files (Figure 3A), (ii) a submission including a query sequence (Figure 3B), (iii)–(vi) protein complex multiple sequence alignment-generated-based string, genomics, phylogeny, and cpxDeepMSA, respectively (Figure 3C–F), (G) the multiple sequence alignment file (Figure 3G). As an illustration, Figure 3 presents an example from the conformationally-strained, circular permutant of barnase (PDB ID: 3da7) to explain section vi of the results page.

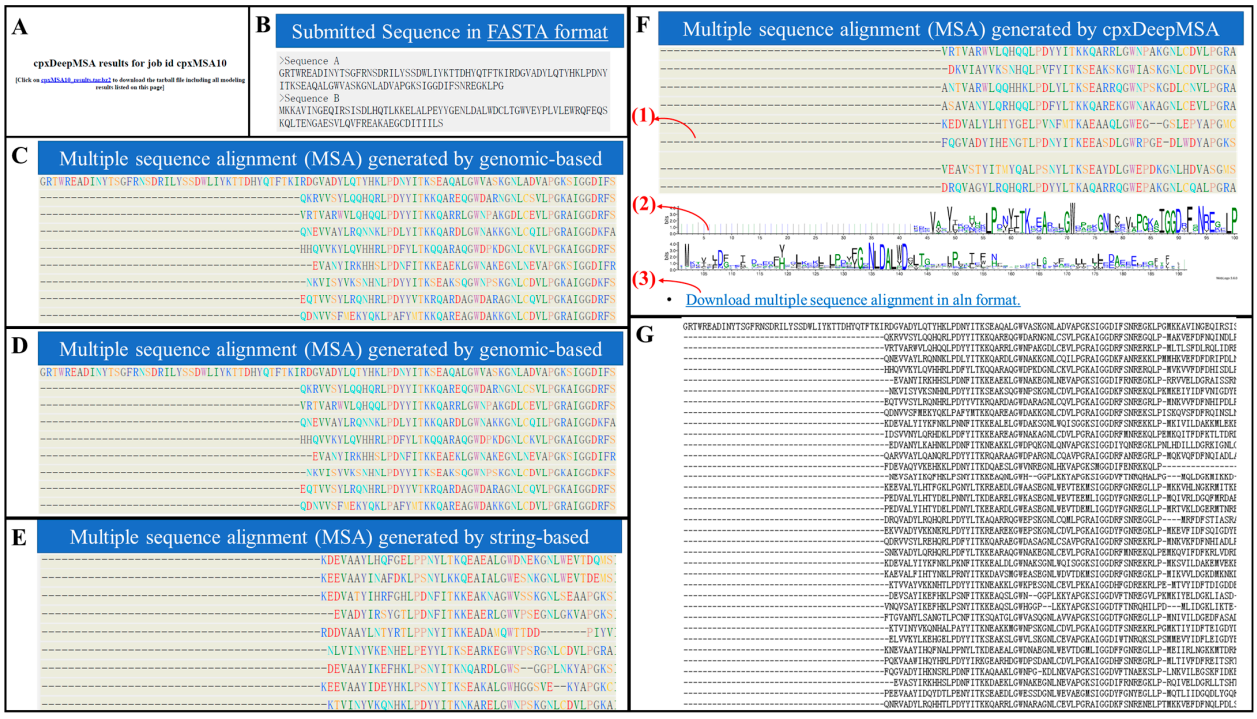

Figure 3. Illustration of the cpxDeepMSA server output, including (**A**) a summary of the multiple sequence alignment results; (**B**) summary of the user input; (**C**–**F**) protein complex multiple sequence alignment-generated-based string, genomics, phylogeny, and cpxDeepMSA, respectively; (**G**) multiple sequence alignment file.

Section vi (Figure 3F) shows the cpxMSA generated by cpxDeepMSA, which lists three parts, (1), (2), and (3). For (1), which shows the cpxMSA on the page, users can drag or zoom in on the table to check the cpxMSA. Additionally, (2) presents the sequence analysis of cpxMSA using the software WebLogo 3.6 [29]. In the last subsection, (3), an aln-formatted file can be downloaded by clicking on the link at the bottom table.

3. Materials and Methods

3.1. cpxDeepMSA Pipeline for MSA Construction

Figure 4 shows a complex pipeline that can be divided into three stages, which correspond to searching two protein sequence databases, Uniclust30 [30] and STRING [31], combining the HH-suite [19] program, and through three matching databases, ENA [32], Taxonomy [33], and STRING linker [31].

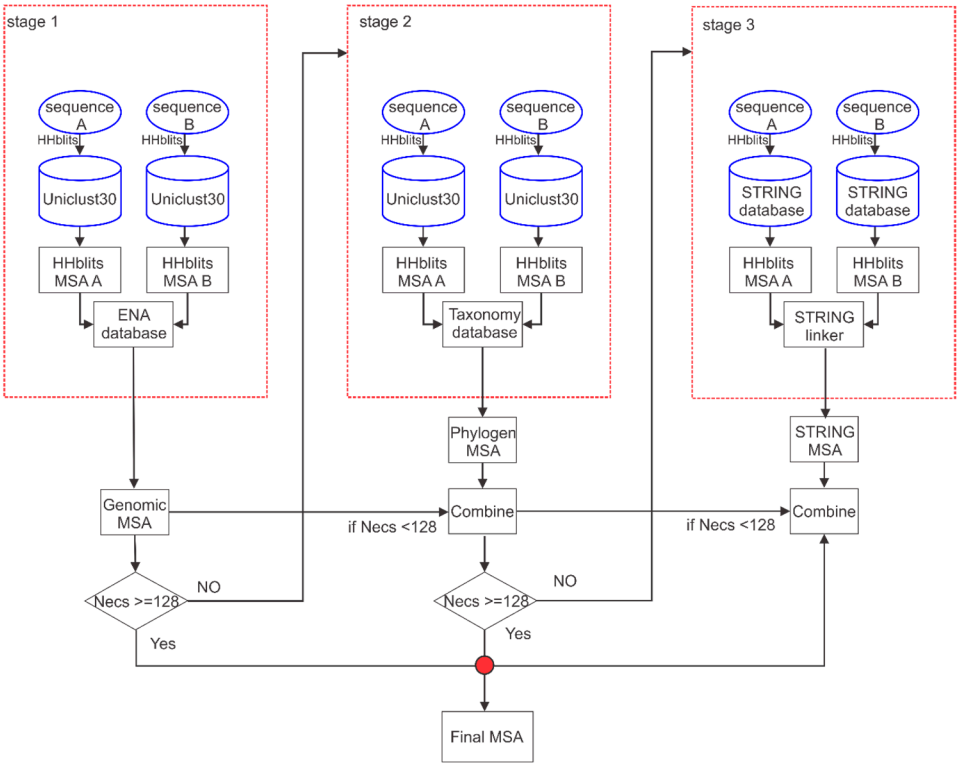

Figure 4. The flowchart of cpxDeepMSA. Three stages of MSA generations were performed consecutively using sequences from the HHblits search through Uniclust30 and pairing with genomic distance (first column), phylogeny information (second column), and the STRING interaction network (third column).

Stage 1. First, download the Uniclust30 (version: 2018_08) [30] protein monomer sequence database from the whole genome data of the protein monomer sequence. Secondly, use the multiple sequence alignment software HHblits (with the parameters "-diff inf -id 99 -cov 50 -n 3") from the HH-suite 2.0.16 program to search the protein sequence database Uniclust30 for query sequence A and sequence B, respectively. Additionally, obtain the multiple sequence alignment information MSA_A and MSA_B of the protein monomer sequence, respectively. Third, compare the results MSA_A and MSA_B in the genome database (ENA), and obtain the gene information MSA_A_gene and MSA_B_gene of the multiple sequence alignment results. Fourth, according to the gene distance $\Delta gene$ of the two protein sequences i and j with the same gene in the MSA_A_gene and MSA_B_gene, if $1 \leq \Delta gene \leq 20$, connect the protein sequence i and j. Finally, according to the above steps, construct a multiple sequence alignment (MSA) of the protein complex based on gene distance, as shown in Figure 5.

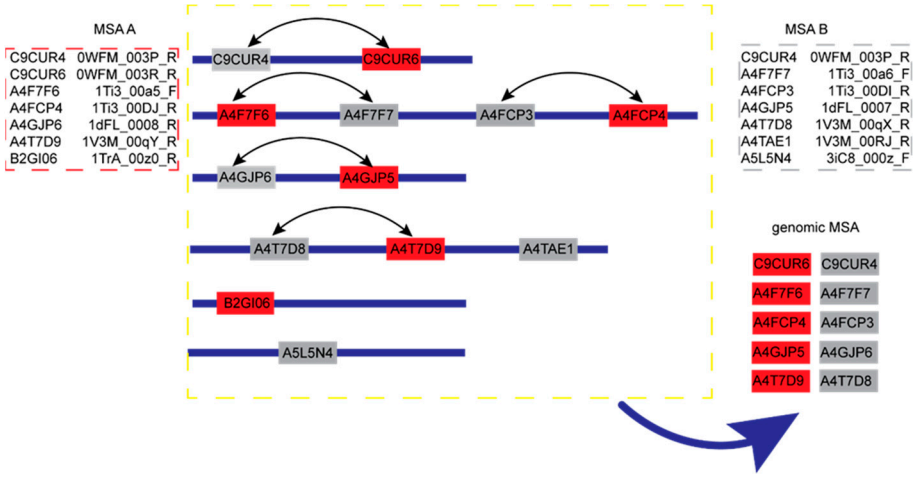

Figure 5. An example of the genomic-based MSA concatenation.

Stage 2. There are five steps in the cpxMSA construction based on the protein monomer sequence and species similarity search. First, download the taxonomy database from the National Center for Biotechnology Information (NCBI) public database. Secondly, compare the multiple sequence alignment information MSA_A and MSA_B of sequence A and sequence B in Stage 1 with the taxonomy database, respectively, to obtain the species information of the proteins in MSA_A_phy and MSA_B_phy, respectively. Third, rank the similarity of proteins and query sequences in each species in MSA_A_phy and MSA_B_phy from high to low. Fourth, let $P_1, P_2, \cdots, P_m$ be the species-specific proteins in MSA_A_phy sorted by sequence similarity, and $Q_1, Q_2, \cdots, Q_n$ be the species-specific proteins in MSA_B_phy ranked by sequence similarity. Then, connect P_i with Q_i, where $i \leq \min(m, n)$. Finally, according to the species comparison result, the two monomer multi-sequence comparisons are concatenated to obtain the species-based multi-sequence comparison result of the protein complex (see Figure 6).

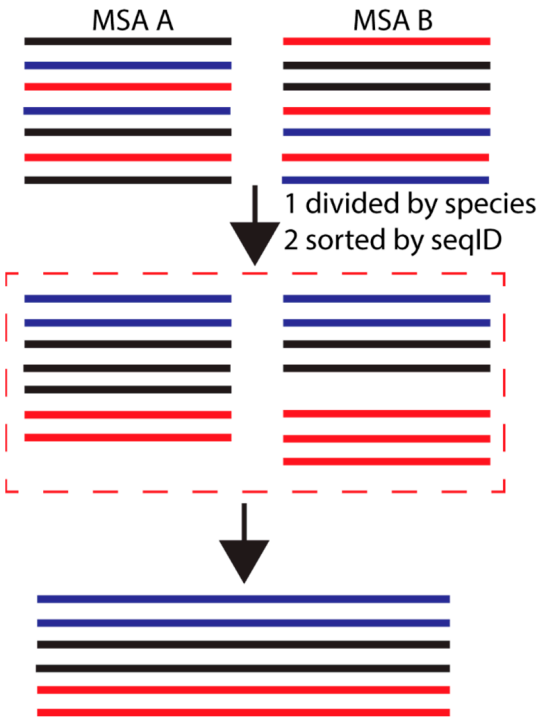

Figure 6. The flowchart of the phylogeny information-based method.

Stage 3. The main points of the process (according to the protein interaction network to build cpxMSA) are as follows: (i) Download the protein interaction information (STRING linker) and protein interaction sequence information (STRING database) from the protein interaction network database (STRING version 10.5v: https://cn.string-db.org/) of the public database. (ii) Use the multiple sequence alignment HHblits program to search for the protein interaction sequence information (STRING) of sequence A and sequence B, respectively, and obtain the multiple sequence alignment information MSA_stringA and MSA_stringB, respectively. (iii) According to the protein interaction information (STRING linker), determine whether any two proteins, protein i and j in the MSA_stringA and MSA_stringB, have interactions. If there is an interaction, connect the two. In summary, according to steps (i)–(iii), construct an interaction-based multiple sequence alignment (MSA) of the protein complexes.

3.2. Selection of the Protein Complex Multiple Sequence Alignment Method

The number of effective sequences of the protein complex multiple sequences alignment (*Necs*):

$$Necs = \frac{1}{\sqrt{L}} \sum_{i=1}^{N} \frac{1}{1 + \sum_{j=1, i \neq j}^{N} \delta\left(S_{i,j} \geq 0.8\right)} \tag{1}$$

$$S_{i,j} = \frac{2}{\frac{1}{S_{iA,jA}} + \frac{1}{S_{iB,jB}}} \tag{2}$$

$$\delta\left(S_{i,j}\right) = \begin{cases} 1, \; if \; S_{i,j} \geq 0.8 \\ 0, if \; S_{i,j} \leq 0.8 \end{cases} \tag{3}$$

where L is the length of the query protein complex and N is the number of sequences in the protein complex multiple sequence alignment (MSA). $S_{iA,jA}$ is the sequence identity between chain A in sequence i and chain A in sequence j. $S_{iB,jB}$ is the sequence identity between chain B in sequence i and chain B in sequence j.

Selection of protein complex multiple sequence alignment method: First, calculate the number of effective sequences in the multiple sequence alignment of the protein complex based on genomic distance in stage 1. Secondly, if the number of sequences in the multiple sequence alignment in stage 1 meets the requirements, the sequence alignment in stage 1 is used as the input in the step of removing redundant sequences. Otherwise, combine the multiple sequence alignment in step 1 with the multiple sequence alignment based on the species category in stage 2, and calculate the number of effective sequences. Thirdly, if the number of valid sequences after the merging of stage 1 and stage 2 meets the condition, the merging result is used as the input of the redundant sequence step. Otherwise, combine the multiple sequence alignments based on the protein interaction network in stage 1, stage 2, and stage 3 as the input to the step of removing the redundant sequences.

4. Conclusions

We developed an open-source pipeline, cpxDeepMSA, to provide a cpxMSA algorithm that is high-quality, large-depth, and provides a wide range of sequence sources and strong generalization abilities. cpxDeepMSA was proposed to solve the shortcomings of low-quality cpxMSA results due to a single database and low search depth. The advantages of cpxMSA by cpxDeepMSA are as follows: (i) It increases the depth of MSA. The depth of MSA, not just using the single search algorithm or database to align, can also judge according to the number of valid sequences in the MSA results from the previous layer. (ii) The proposed method enhances the generalization ability by using different protein monomer databases and three different monomer MSA strategies (genomic distance, phylogeny information, and STRING interaction network) to join the monomer MSA results in protein complexes. The online server and the standalone program of cpxDeepMSA are freely available at https://zhanggroup.org/cpxDeepMSA/ (accessed on 1 July 2022).

Author Contributions: Conceptualization, Z.L. and D.-J.Y.; methodology, Z.L.; software, Z.L.; validation, Z.L.; formal analysis, Z.L.; investigation, Z.L.; resources, Z.L.; data curation, Z.L.; writing—original draft preparation, Z.L.; writing—review and editing, Z.L. and D.-J.Y.; visualization, Z.L.; supervision, D.-J.Y.; project administration, D.-J.Y.; funding acquisition, D.-J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grants 62072243, 61772273, and 61872186) and the Natural Science Foundation of Jiangsu (grant no. BK20201304).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We are grateful to Chengxin Zhang and Yang Zhang for the suggestion of the pipeline and web design.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

PPI	protein–protein interaction
MSA	multiple sequence alignment
cpxMSA	protein complex multiple sequence alignment
Necs	effective sequences of the protein complex multiple sequences alignment

References

1. Kuzmanov, U.; Emili, A. Protein-protein interaction networks: Probing disease mechanisms using model systems. *Genome Med.* **2013**, *5*, 1–12. [CrossRef] [PubMed]
2. Shi, T.L.; Li, Y.X.; Cai, Y.D.; Chou, K.C. Computational methods for protein-protein interaction and their application. *Curr. Protein Pept. Sci.* **2005**, *6*, 443–449. [CrossRef] [PubMed]
3. Hopf, T.A.; Scharfe, C.P.; Rodrigues, J.P.; Green, A.G.; Kohlbacher, O.; Sander, C.; Bonvin, A.M.; Marks, D.S. Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* **2014**, *3*, e03430. [CrossRef] [PubMed]
4. Zeng, H.; Wang, S.; Zhou, T.; Zhao, F.; Li, X.; Wu, Q.; Xu, J. ComplexContact: A web server for inter-protein contact prediction using deep learning. *Nucleic Acids Res.* **2018**, *46*, W432–W437. [CrossRef]
5. Zhang, Q.C.; Petrey, D.; Deng, L.; Qiang, L.; Shi, Y.; Thu, C.A.; Bisikirska, B.; Lefebvre, C.; Accili, D.; Hunter, T.; et al. Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **2012**, *490*, 556–560. [CrossRef] [PubMed]
6. Li, Y.; Hu, J.; Zhang, C.; Yu, D.-J.; Zhang, Y. ResPRE: High-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. *Bioinformatics* **2019**, *35*, 4647–4655. [CrossRef]
7. Li, Y.; Zhang, C.; Bell, E.W.; Yu, D.J.; Zhang, Y. Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13. *Proteins Struct. Funct. Bioinform.* **2019**, *87*, 1082–1091. [CrossRef]
8. Ju, F.; Zhu, J.; Shao, B.; Kong, L.; Liu, T.-Y.; Zheng, W.-M.; Bu, D. CopulaNet: Learning residue co-evolution directly from multiple sequence alignment for protein structure prediction. *Nat. Commun.* **2021**, *12*, 1–9. [CrossRef]
9. Zhang, C.; Mortuza, S.; He, B.; Wang, Y.; Zhang, Y. Template-based and free modeling of I-TASSER and QUARK pipelines using predicted contact maps in CASP12. *Proteins Struct. Funct. Bioinform.* **2018**, *86*, 136–151. [CrossRef] [PubMed]
10. Hanson, J.; Paliwal, K.; Litfin, T.; Yang, Y.; Zhou, Y. Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. *Bioinformatics* **2018**, *34*, 4039–4045. [CrossRef] [PubMed]
11. Wu, S.T.; Zhang, Y. ANGLOR: A Composite Machine-Learning Algorithm for Protein Backbone Torsion Angle Prediction. *PLoS ONE* **2008**, *3*, e3400. [CrossRef] [PubMed]
12. Gil, N.; Fiser, A. The choice of sequence homologs included in multiple sequence alignments has a dramatic impact on evolutionary conservation analysis. *Bioinformatics* **2019**, *35*, 12–19. [CrossRef] [PubMed]
13. Zheng, W.; Wuyun, Q.Q.G.; Zhou, X.G.; Li, Y.; Freddolino, P.L.; Zhang, Y. LOMETS3: Integrating deep learning and profile alignment for advanced protein template recognition and function annotation. *Nucleic Acids Res.* **2022**, *50*, W454–W464. [CrossRef] [PubMed]
14. Zhang, C.; Freddolino, P.L.; Zhang, Y. COFACTOR: Improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic Acids Res.* **2017**, *45*, W291–W299. [CrossRef] [PubMed]
15. Liu, K.; Warnow, T. Large-Scale Multiple Sequence Alignment and Tree Estimation Using SATe. *Mult. Seq. Alignment Methods* **2014**, *1079*, 219–244.
16. Wang, Y.Y.; Wu, H.Y.; Cai, Y.P. A benchmark study of sequence alignment methods for protein clustering. *BMC Bioinform.* **2018**, *19*, 529. [CrossRef] [PubMed]

17. Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef] [PubMed]
18. Remmert, M.; Biegert, A.; Hauser, A.; Soding, J. HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **2012**, *9*, 173–175. [CrossRef] [PubMed]
19. Steinegger, M.; Meier, M.; Mirdita, M.; Vöhringer, H.; Haunsberger, S.J.; Söding, J. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinform.* **2019**, *20*, 473. [CrossRef]
20. Potter, S.C.; Luciani, A.; Eddy, S.R.; Park, Y.; Lopez, R.; Finn, R.D. HMMER web server: 2018 update. *Nucleic Acids Res.* **2018**, *46*, W200–W204. [CrossRef]
21. Buchan, D.W.A.; Jones, D.T. Improved protein contact predictions with the MetaPSICOV2 server in CASP12. *Proteins-Struct. Funct. Bioinform.* **2018**, *86*, 78–83. [CrossRef] [PubMed]
22. Zhang, C.; Zheng, W.; Mortuza, S.; Li, Y.; Zhang, Y. DeepMSA: Constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics* **2020**, *36*, 2105–2112. [CrossRef] [PubMed]
23. Ovchinnikov, S.; Kamisetty, H.; Baker, D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife* **2014**, *3*, e02030. [CrossRef] [PubMed]
24. Seemayer, S.; Gruber, M.; Soding, J. CCMpred-fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics* **2014**, *30*, 3128–3130. [CrossRef]
25. Du, Z.Y.; Su, H.; Wang, W.K.; Ye, L.S.; Wei, H.; Peng, Z.L.; Anishchenko, I.; Baker, D.; Yang, J.Y. The trRosetta server for fast and accurate protein structure prediction. *Nat. Protoc.* **2021**, *16*, 5634–5651. [CrossRef]
26. Su, H.; Wang, W.K.; Du, Z.Y.; Peng, Z.L.; Gao, S.H.; Cheng, M.M.; Yang, J.Y. Improved Protein Structure Prediction Using a New Multi-Scale Network and Homologous Templates. *Adv. Sci.* **2021**, *8*, e2102592. [CrossRef]
27. Burley, S.K.; Berman, H.M.; Kleywegt, G.J.; Markley, J.L.; Nakamura, H.; Velankar, S. Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. *Methods Mol. Biol.* **2017**, *1607*, 627–641.
28. aek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G.R.; Wang, J.; Cong, Q.; Kinch, L.N.; Schaeffer, R.D.; et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871–876.
29. Crooks, G.E.; Hon, G.; Chandonia, J.M.; Brenner, S.E. WebLogo: A sequence logo generator. *Genome Res.* **2004**, *14*, 1188–1190. [CrossRef] [PubMed]
30. Mirdita, M.; von den Driesch, L.; Galiez, C.; Martin, M.J.; Söding, J.; Steinegger, M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.* **2017**, *45*, D170–D176. [CrossRef]
31. Szklarczyk, D.; Morris, J.H.; Cook, H.; Kuhn, M.; Wyder, S.; Simonovic, M.; Santos, A.; Doncheva, N.T.; Roth, A.; Bork, P.; et al. The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **2017**, *45*, D362–D368. [CrossRef]
32. Harrison, P.W.; Alako, B.; Amid, C.; Cerdeno-Tarraga, A.; Cleland, I.; Holt, S.; Hussein, A.; Jayathilaka, S.; Kay, S.; Keane, T.; et al. The European Nucleotide Archive in 2018. *Nucleic Acids Res.* **2019**, *47*, D84–D88. [CrossRef] [PubMed]
33. Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res.* **2012**, *40*, D136–D143. [CrossRef]