TOPICAL REVIEW — Modeling and simulations for the structures and functions of proteins and nucleic acids

Review of multimer protein–protein interaction complex topology and structure prediction[*]

Daiwen Sun(孙黛雯)[1], Shijie Liu(刘世婕)[1], and Xinqi Gong(龚新奇)[1,2,†]

[1] *Mathematics Intelligence Application Laboratory, Institute for Mathematical Sciences, Renmin University of China, Beijing* 100872, *China*
[2] *Beijing Advanced Innovation Center for Structural Biology, Tshinghua University, Beijing* 100094, *China*

(Received 29 June 2020; revised manuscript received 31 August 2020; accepted manuscript online 9 September 2020)

Protein–protein interactions (PPI) are important for many biological processes. Theoretical understanding of the structurally determining factors of interaction sites will help to understand the underlying mechanism of protein–protein interactions. At the same time, understanding the complex structure of proteins helps to explore their function. And accurately predicting protein complexes from PPI networks helps us understand the relationship between proteins. In the past few decades, scholars have proposed many methods for predicting protein interactions and protein complex structures. In this review, we first briefly introduce the methods and servers for predicting protein interaction sites and interface residue pairs, and then introduce the protein complex structure prediction methods including template-based prediction and template-free prediction. Subsequently, this paper introduces the methods of predicting protein complexes from the PPI network and the method of predicting missing links in the PPI network. Finally, it briefly summarizes the application of machine/deep learning models in protein structure prediction and action site prediction.

Keywords: protein complex prediction, protein–protein interaction

PACS: 87.15.km, 01.50.hv **DOI:** 10.1088/1674-1056/abb659

1. Introduction

Protein is an important executor of biological functions. With the completion of the Human Genome Project and the continuous development of protein research technology, protein science has received more and more attention. Among them, protein structure-function relationship, protein–protein interaction, and protein recognition are important contents of protein scientific research. Due to many difficulties in experimentally determining the structure of protein complexes, there is an urgent need to develop effective computer simulation methods to explore the interaction and recognition process between protein molecules, and then predict the three-dimensional structure of the complex formed by protein–protein binding. The development of theoretical simulation methods not only helps us understand the mechanism of specific recognition between protein molecules, but also provides theoretical guidance for rational drug development and new protein molecule design. In the past few decades, many high-throughput methods for protein structure determination based on x-ray and high-resolution NMR have appeared.[1] However, most proteins, especially protein–protein complexes, do not have corresponding structures in the Protein Data Bank (PDB), which increases the demand for predicting protein interactions and complex structures.

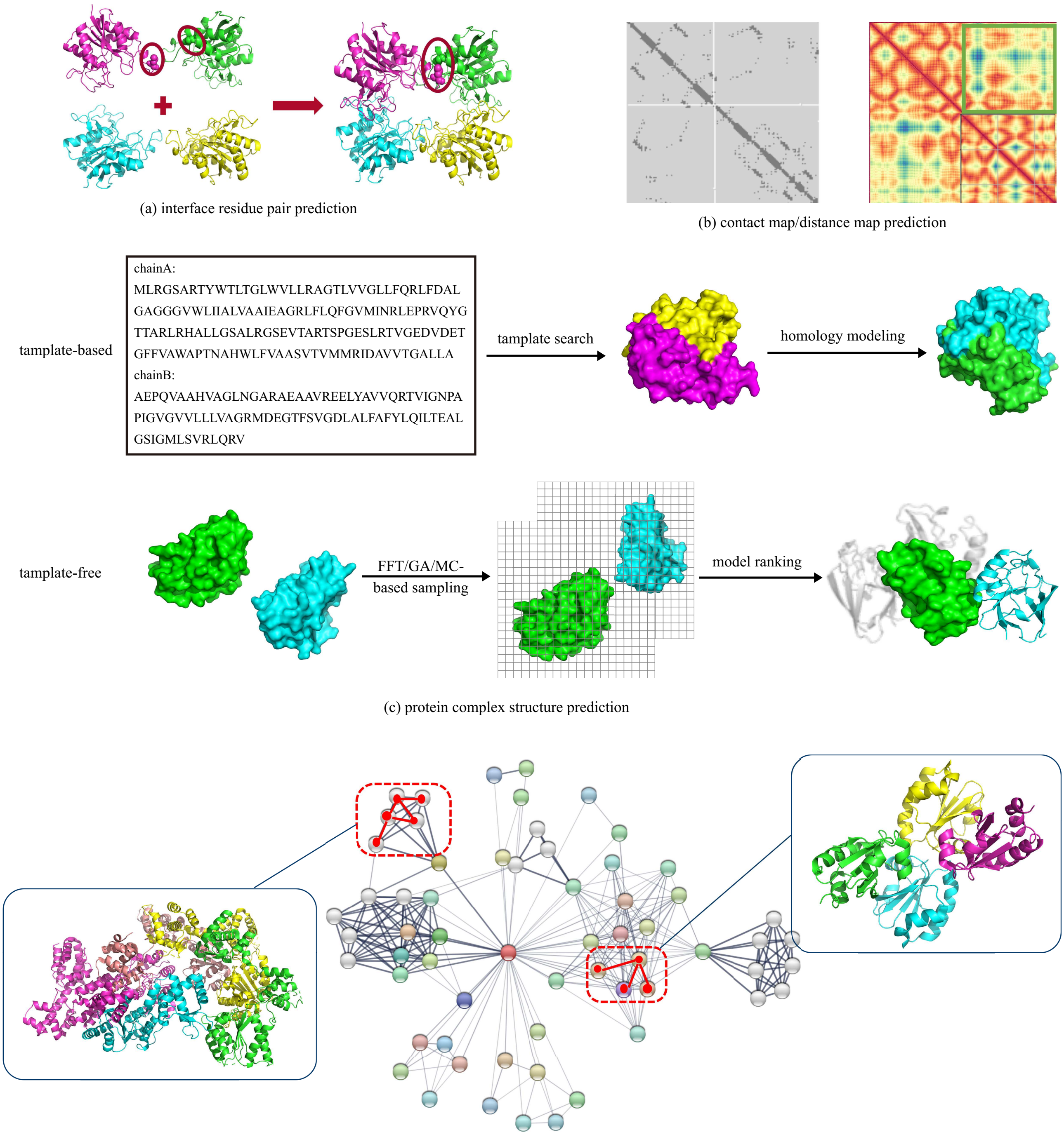
This review aims to introduce the current status of protein complex interactions and structure prediction. It starts with a review of existing methods and servers for predicting protein interaction residues. It is crucial to know protein–protein interaction interface binding sites (interface residue pairs and contact map) (see Figs. 1(a) and 1(b)) for comprehensively understanding the molecular mechanism and confirming potential drug targets.[2] Besides, the prediction results of protein–protein interaction interface residue pairs can assist to predict protein 3D structure. Then it continues with protein structure prediction methods, including template-based methods and template-free prediction methods (see Fig. 1(c)). After that, this paper introduces the prediction of protein complexes from protein–protein interaction (PPI) network from two aspects, including complex prediction based on PPI network clustering and complex interaction link prediction from PPI network (see Fig. 1(d)). Finally, the application of machine/deep learning in protein structure prediction is briefly summarized. The main content of protein interaction calculation is shown in Fig. 1.

2. Protein databases

Two databases are commonly used in the study of protein structure, they are Protein Data Bank (PDB) and the Electron Microscopy Data Bank (EMDB) at PDBe.

[*] Project supported by the National Natural Science Foundation of China (Grant No. 31670725).

[†] Corresponding author. E-mail: xinqigong@ruc.edu.cn


tamplate-based
chainA:
MLRGSARTYWTLTGLWVLLRAGTLVVGLLFQRLFDAL
GAGGGVWLIIALVAAIEAGRLFLQFGVMINRLEPRVQYG
TTARLRHALLGSALRGSEVTARTSPGESLRTVGEDVDET
GFFVAWAPTNAHWLFVAASVTVMMRIDAVVTGALLA
chainB:
AEPQVAAHVAGLNGARAEAAVREELYAVVQRTVIGNPA
PIGVGVVLLLVAGRMDEGTFSVGDLALFAFYLQILTEAL
GSIGMLSVRLQRV
tamplate search
homology modeling
tamplate-free
FFT/GA/MC-
based sampling
model ranking

(a) interface residue pair prediction

(b) contact map/distance map prediction

(c) protein complex structure prediction

(d) protein complex prediction from PPI network

Fig. 1. The main content of protein interaction calculation.

The PDB database is a macromolecular structure database established by the Brookhaven laboratory in 1971. The Research Collaboration for Structural Bioinformatics (RCSB) is responsible for the maintenance of the PDB database. Its data source mainly determines the three-dimensional structure of biological macromolecules through experiments (x-ray crystal diffraction, nuclear magnetic resonance, electron microscope methods, etc.). It is mainly the three-dimensional structure of proteins, and also includes the three-dimensional structure of nucleic acid, carbohydrate, protein, and nucleic acid complexes. The information of the PDB file includes the spatial coordinates of the atoms, the cited literature, the amino acid sequence forming -helix and -sheet, the disulfide bond linking mode, the ligand bound to the protein, the residue involved in the biochemical function, etc. The protein recorded in the PDB consists of a unique PDB-ID, including 4 characters, which can be composed of uppercase letters A to Z and numbers 0 to 9, such as 1A4S. PDB provides a query for each PDB record, which can be advanced searched according to some special query items (such as structure keywords, structure author, gene name). We searched the proteins in the PDB database according to the following conditions: Number of Protein Instances (Chains) ≥ 3, Entry Polymer Types = "Protein (only)", Polymer Entity Sequence Length ≥ 100, and Resolution (Å) ≤ 3. The number of retrieved protein multimers was counted according to the number of monomers included, as shown in Fig. 2. See supplementary data for specific protein details.

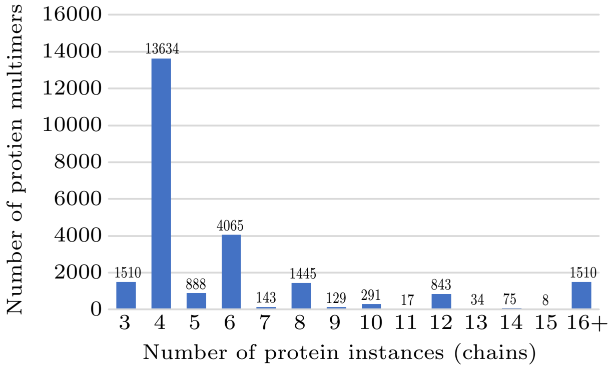


Fig. 2. Statistics of protein multimers in PDB database.

The EMDB database was established by Kim Henrick at EBI in 2002. Since 2007, it is jointly operated by PDBe and RCSB. The EMDB database contains images of different types of biological macromolecules obtained by cryo-electron microscopy, including proteins, nucleic acid, prokaryotic ribosome, eukaryotic ribosome, ligand, and cell component. Currently, there are 9355 protein cryo-EM images in EMDB, of which 3283 images have a resolution of less than 4 Å.

3. Interface residue pair prediction

Protein residues are an important part of proteins. The contact and structure of residue pair prediction has also become an important part of protein structure prediction. As the experiment progressed, there have been more and more algorithms for protein structure and protein interaction prediction, and then more servers appeared to be applied. In this section, we will review some of the servers for predicting protein interaction residues and residue pairs.

We know that proteins rarely perform their functions alone, and most of them form complexes or protein–protein interactions (PPI) networks with other proteins. However, structural analysis of PPI may require its three-dimensional structure. An important step in predicting the three-dimensional structure of protein complexes is to predict the interaction residue pairs between proteins. Direct evolutionary coupling analysis (DCA) is widely used to identifying coevolutionary residue pairs. DCA is very effective in intra-protein contact prediction,[3] but it does not perform well in inter-protein contact prediction, which is a great challenge because this kind of prediction requires too many interacting homologs to perform better, which is not easy to achieve. However, deep learning (DL) methods are better than DCA for predicting contacts between proteins. Wu *et al.* proposed ComplexContact[4] (http://raptorx.uchicago.edu/ComplexContact/), which is a web server of protein complex interface residue pair prediction. This server can predict the contact between residues from different proteins without using any structural template. For the interaction between a pair of protein sequences A and B, they first search for homologs of protein sequences A and B through HHbits and establish MSA separately. Then two methods (genome and phylogeny-based) are used to construct the paired MSA. The construction methods of genome and phylogeny-based MSA are complementary. Generally, for prokaryotes, the genome method is better. And for eukaryotes, biophylogeny-based methods are better. Through the DL model, the contact map between the two proteins is predicted, and the prediction result is calculated. The DL model is mainly composed of two ResNets. One processes sequence features; the sequence features are first subjected to a one-dimensional convolution transformation, and the output is converted into a two-dimensional matrix while being fed into the second ResNet together with the paired features. The second ResNet is used to process the paired features, and the inputs are subjected to two-dimensional convolution transformation, and its outputs will be input to the logistic regression, which can predict the contact probability of any two residues.

DCA is widely used in protein contact prediction, but DCA is only effective for proteins with a large number of sequence homologues.[5] Current contact prediction methods are roughly divided into two categories: evolutionary coupling analysis (ECA) and supervised machine learning. ECA predicts by recognizing co-evolutionary residues of proteins. Supervised machine learning requires various kinds of information to make predictions. The Jinbo Xu team also proposed a new server to accurately predict the protein contact map from the ultra-deep learning model. The DL model is still used in this server. RaptorX-Contact[6] is a web server (http://raptorx.uchicago.edu/ContactMap/), and it also can predict a contact map and distance matrix without any templates. It uses an ultra-deep convolutional residual neural network to predict contact and distance, and can perform well on proteins without many sequence homologs. This method predicts the contact and distance matrix with a whole protein, rather than predicting the interaction of a residue pair by each residue, thus greatly improving the accuracy of prediction. Contact prediction can help predict protein structure. In order to better model the 3D structure, the top contact prediction is used here as a constraint for de novo folding to construct a protein structure model. Similarly, RaptorX-Property[7] (http://raptorx.uchicago.edu/StructurePropertyPred/predict/) is a web server that can predict the structure of protein sequences without using any template. The model of deep convolutional neural fields (DeepCNF) is applied to this server.

Gremlin[8–10] (http://gremlin.bakerlab.org/complexes.php) is a statistical model method for learning protein families proposed by Baker's team. This model can capture the conservative and co-evolutionary patterns of the family, and predict the contact of residue pairs in the 3D structure of proteins through the strength of co-evolution. In this study, a statistical method based on pseudo-likelihoods was used to study the covariance of residue pairs across the protein–protein interface. They further found that significant residual pair covariance often occurs between physically interacting protein pairs, but rarely between non-interacting protein pairs, which is very useful for predicting whether two proteins interact.

Of course, there are also some other teams using different methods to build the servers, for example, another improved protein contact graph predictor DNCON2[11] (http://sysbio.rnet.missouri.edu/dncon2/) with a two-layer convolutional neural network. We know that the EVfold method opens up a new world with the aid of mean field direct coupling analysis (EVfold-mfDCA). The PSICOV[12] method applies the concept of estimating sparse inverse covariance matrix. But these two methods belong to interactive applications, which require too much CPU time. FreeContact[13] (https://rostlab.org/owiki/index.php/FreeContact) can accelerate EVfold-mfDCA. In the 140 proteins of the test set, FreeContact is nearly 8 times faster than PSICOV without deterioration of results. The last is the CoinFold Web server[14] proposed by Wang *et al.* CoinFold (http://raptorx.uchicago.edu/ContactMap/) is used for protein contact prediction and contact-assisted Web server structure prediction from scratch.[15] CoinFold predicts contact using integrating evolutionary coupling (EC) and machine learning.

In addition to the methods introduced above, our team also proposed some methods for this problem. Zhao *et al.*[16] proposed a multi-layered deep learning approach with long-short term memory (LSTM) networks to predict interface residue pairs. In this method, every surface residue pair was described using eighteen features, including geometric and physicochemical properties. Sun *et al.*[17] proposed a deep learning method for tetramer protein complex interface residue pair prediction with long-short term memory networks combined with graph representations. In this method, we considered that every residue of a protein monomer can be affected by residues around it. So we represented each residue with a graph, and the neighbor nodes in the graph were the k residues closest to the residue. Liu *et al.*[18] improved the LSTM with attention mechanism and residual architecture, and used the model to predict interface residue pairs, which achieved good performance.

4. Protein complex structure prediction

Protein complexes play an important role in life activities, participating in the regulation of gene expression, electron transmission, nerve transmission in cells, and even learning and memory. At present, there are many computational algorithms for predicting the structures of protein complexes. These methods can be roughly divided into two classes — template-free (docking) and template-based methods.[19] The docking methods derive the structure of the protein complex from the unbound structures of several proteins, which are obtained by x-ray or nuclear magnetic resonance (NMR). The template-based approaches use the similarities with known complex structures to predict. The following sections will introduce some template-free and template-based methods respectively.

4.1. Template-free prediction (multimer docking)

Usually, the protein complex structures are obtained by 'docking'. Docking is to find the best matched 3D structure of protein complex formed by several component proteins.[20] A docking algorithm includes a fast search program to search all possible spatial conformations and a scoring function to sort the searched conformations. The commonly used spatial search methods are fast Fourier transform (FFT), genetic algorithm (GA), and Monte Carlo (MC). In addition, there are some other methods for conformation search, such as spherical polar Fourier correlation,[21] conformational space annealing,[22] and molecular interaction field.[23] The commonly used scoring items currently include: geometric complementary, interface contact area, van der Waals forces and electrostatic, and statistical pairing preference.

Most docking procedures are for two proteins, receptor and ligand. First, in order to reduce the influence of the high degree of freedom of the protein backbone and side chains, the protein is treated as a rigid body. And the relative binding orientations of the ligand and the receptor are decided by translating and rotating the ligand in six degrees. Then the structure obtained in the first step is optimized by a local search, where some of the side chains and interface residues can move freely.

However, the large number of possibilities for the position and angle of protein residues and side chains makes the computational cost of search algorithms the most difficult problem in protein–protein docking. The search algorithms used for rigid protein docking can be roughly divided into 3 categories (Table 1).

There are currently many docking procedures between two proteins, but there are relatively few docking approaches applicable to three or more proteins. One of the representative methods is M-ZDOCK developed by Zhiping Weng *et* *al.*[24] M-ZDOCK can use the structure of monomeric proteins to predict the structure of circular symmetric (C_n) complexes. It uses a grid-based FFT method to search for the best structure in a fully symmetric space of multimers. The scoring function used by M-ZDOCK is similar to that used in ZDOCK.[25] It scores the searched structure through surface complementarity, electrostatics, and desolvation. Experiments showed that this method has improved accuracy and running time. Similar to the function of M-ZDOCK, there are some other methods for prediction of cyclically symmetric (C_n) multimers structure, such as RosettaDock,[26] ClusPro,[27] and HDOCK.[28,29] Besides methods for C_n complexes, there are also some docking methods for homo-oligomers with dihedral symmetric (D_n) complexes, such as HSYMDOCK[30] and SAM.[31]

Table 1. Classification of optimization algorithms applied by protein protein docking approached.

Optimization algorithms	Programs
Fast Fourier transformation	ZDOCK; GRAMM; DOT; SmoothDock; ClusPro; MolFit; FTDock; 3D-Dock; PIPER; pyDock; HDOCK; SDOCK; HEX; FRODOCK; InterEvDock; MDockPP; CoDockPP; HSYMDOCK; SAM
Monte Carlo	RosettaDock; ICM-DOCK; HADDOCK; ATTRACT
Genetic algorithm	DARWIN; Multi-LZerD; AutoDock

Another docking procedure for multimers is Multi-LZerD, which can dock proteins other than symmeric multimers and does rely on the prior availability of biological information (e.g., interation sites). It is a multi-stage method, where the first stage uses the docking program LZerD to dock the pairwise proteins to obtain a large number of docking conformations. The conformational space of entire complex is constructed using pairwise decoys as building blocks. Then it uses genetic algorithms to combine the docking results. After each iteration, the qualities of complex decoy structures are assessed by a physics-based scoring function. Finally, after many iterations, the final decoy is obtained and the optimal structure is refined.

4.2. Template-based prediction

With the increase in the amount of protein structure data, template-based protein structure prediction methods have been widely used, that is, using sequence or structure similarity to model protein complexes with known structures. For template-based methods, the process is simple and clear: selecting high-quality template data consisting of protein complexes of known structure, inferring known data by using sequence or structural similarities to identify unknown interactions, and according to the prediction, the results are ranked by a score function, that is, statistical potential or energy function. Template-based methods mainly reduce the possible structure by restricting the direction of protein binding. This method is more efficient than docking and can be applied to larger-scale protein complex prediction.

This kind of methods is based on the fact that if the protein sequence identity is as high as 30%–40%, then they are combined in a similar way. However, in order to reliably map the internal dialogue between protein–protein interactions between different species, it was found that the combined sequence identity is at least 80%.[32] Aloy and Russell's[33] preliminary analysis of the complex with known 3D structures showed that protein interactions occur through various main chain and side chian contacts. They first determined the interface residues at the interface between two proteins. Then, they used electrical potential to score the compatibility of homologous species sequences. If the score is high enough, the homologue will bind to the template in a similar manner. Later, they designed a web server to predict protein interactions using this method.[34]

Skolnick *et al.* have also achieved certain results in predicting protein interactions based on homology. They developed a multimeric threading approach, which is comprised of two phases.[35] In the first phase, a protein structure corresponding to each target sequence in the template library is found using a threading algorithm, PROSPECTOR developed by them before. In the second phase, the same method was applied to multiple chains. This algorithm has certain flaws in considering conformational changes and determining the relative position of proteins. M-Tasser[36] explicitly conbines the flexibility of a backbone with threadings for prediction. Simonson *et al.*[37] considered structural homology and complexes formed between single-domain proteins in the template dataset. In addition to homology, Shoemaker *et al.* integrated structural similarity of the overall protein structure to predict protein complexes. In addition to sequence similarity, Aloy *et al.*[38] also considered the overall structural folding similarity to make the structural interaction network more complete. Nye *et al.*[39] used statistical analysis to predict domain pairs that mediate protein–protein interactions. PISA[40] and ProtCID[41] were developed to model complexes using chemical thermodynamics and interface residue pair dynamics, respectively.

5. Protein complex prediction from PPI networks

With the development of high-throughput technology, we can get more and more protein–protein interaction (PPI) data.[42–45] These PPI data can be represented by graphs or networks, where vertices in the network represent proteins and edges represent protein–protein interactions. Biologists can use these large amounts of PPI data and PPI networks to gain insights into cells and their internal proteins.[46] For example, an overall analysis of the PPI network can understand the relationship between protein interactions and functioning.[47–49] We can also use the PPI network to predict protein pathways, gene functions, or protein complexes through machine learning methods.[50–58] In addition, many efforts have recently been made to integrate PPI networks to build a complete diagnosis and treatment system for complex diseases.[59–63]

However, the rapid increase in the number of protein tests and the expansion of the PPI network have brought great challenges to biologists. First, there are many unlabeled or mislabeled real interactions in the existing PPI network.[64] Second, most PPI networks are sparse, which brings misery to neighbor-based algorithms.[52,56] Third, PPI networks are known to have skewed degree distributions, which means that the number of their hub genes exceeds expectations.[65] Such central nodes often degrade the performance of existing graph theory algorithms. These algorithms are usually effective for networks with uniform degree distribution. In the following sections, we briefly review some methods for predicting protein complexes and link prediction through PPI networks.

5.1. Complex prediction based on PPI network clustering

In general, most methods assume that the protein complex is part of a known PPI network, that is, the graph composed of protein complexes and their interactions are a subgraph of the PPI network. Some of these methods only use the PPI network for clustering, and some use additional biological information, including structure, function, organization, co-evolution information, etc.[66]

MCODE (Molecular COmplex DEtection), proposed by Bader and Hogue[67] is one of the earliest computational methods to predict protein complex from PPI networks. The MCODE algorithm is implemented in three stages: vertex weighting, complex prediction, and a subsequent optional step to reduce or add some proteins to the resulting complex through the connectivity criterion. In the first stage, MCODE weights all vertices using the network density based on the highest k-core of the vertex neighborhood. In the second stage, the protein with the highest weight will be selected into the complex.

Van Dongen proposed a graph clustering algorithm, MCL (Markov clustering), which uses random walks to extract dense subnetworks on graphs for clustering. Subsequently, the MCL algorithm is very effective in clustering protein complexes and functional modules from PPI networks.[68–70] MCL adjusts the adjacency matrix of the network through expansion and inflation to increase the probability of intra-cluster walk and reduce the probability of inter-cluster walk. In this way, the network is divided into multiple non-overlapping regions through multiple iterations.

Blatt *et al.*[71] and Getz *et al.*[72,73] proposed SPC (superparamagnetic clustering), which was improved from the Potts model. The Potts model is a model of ferromagnetism, where each data point in the metric space is clustered according to the “spins” assigned to them. In 2003, Spirin and Mirny[74] applied SPC to protein complex prediction in PPI networks. In 2005, Li *et al.*[75] proposed LCMA (local clique merging algorithm), which can identify some small cliques and merge them into high-density subgraphs. In 2008, Qi *et al.*[76] proposed the SuperComplex (supervised protein complex prediction) method, which uses Bayesian network models to learn the features of real protein complexes to cluster PPI networks. The supervised Bayesian network (BN) method is a machine learning method. The positive samples in the training set come from real protein complexes, and the negative samples are randomly selected proteins from the PPI network into groups. Yong *et al.*[77] and Srihari *et al.*[78] proposed ensemble methods, which used several methods to cluster PPI networks and used majority voting to decide the final predicting results. In addition, there are some other methods such as CFinder[79] (complex finder), DPClus[80] (density-periphery-based clustering), IPCA[81] (interaction probability-based clustering algorithm), CMC[82] (clustering based on merging maximal cliques), ClusterONE[83] (clustering with overlapping neighborhood expansion), and HACO[84] (hierachical agglomerative clustering with overlaps).

5.2. Complex interaction link prediction from PPI network

In recent years, scientists have developed many methods to predict the actual links in the network,[85–88] and reviewed by Lü and Zhou.[89] There are some methods applied to the PPI network, which can be roughly divided into two categories, public neighbors and distance.

One of the simple ideas is there may be two nodes in the same module sharing many common neighbors.[86,87,90] These methods have limited effect on sparse networks. Fang *et al.*[91] considered neighbors with greater distances and proved that the measurement of global geometric affinity (GGA) can predict new PPI. Considering that the matrix model can be used to model the network structure, Xu Qian *et al.*[92] introduced a matrix factorization-based method to predict exact links in sparse networks based on relatively dense networks. In addition, Park and Bader have developed hierarchical agglomerative clustering (HAC) algorithm,[93] for rapid clustering of heterogeneous interactive networks. The algorithm uses maximum likelihood to cluster the network structure.

The second method measures the distance between all pairs of nodes in the network. Typical methods include random walk-based methods, such as Euclidean commute time (ECT)[88] and random walk with restart (RWR).[85] Kuchaiev *et al.*[94] attempted to use multiple dimensional scaling (MDS) to map the PPI network to a low-dimensional space, and to allocate edges between pairs of nodes in the embedded space that have short distances. Lei and Ruan[65,95] proposed an algorithm that uses random walk to determine that certain compounds belong to similar compounds by measuring the topological similarity of the two networks. In additional, Wang Liang *et al.*[96] used link-weighted PPI networks to enhance the robustness of the method.

6. Machine learning and deep learning applications

In recent years, artificial intelligence has been greatly developed. At the same time, methods based on machine learning and deep learning have been successfully applied to various prediction problems in biology, such as protein structure prediction and protein interaction prediction.

Since the 1990s, researchers have used contact maps to reconstruct the three-dimensional structure of proteins.[97–99] Although reconstructing the three-dimensional structure of proteins through contact maps is a NP-hard problem,[100] there are many methods to approximate this problem[99,101,102] and optimize the computational efficiency.[98] Using the distance map and the multi-class contact map can obtain more accurate protein structures and make the prediction results more robust.[103,104]

Machine/deep learning methods such as FFNN, BRNN, and multi-satge approaches have been used to predict whether any two residues in a protein are in contact since the late 1990s. Similarly, disulphide can be predicted by Monte Carlo simulation annealing,[105] or hybrid methods like hidden Markov models and FFNN,[106] multi-stage FFNN, SVM, and BRNN,[107] and machine learning methods like SVM,[108] deep learning methods like BRNN[109] and FFNN.[110]

Recently, with the increase of co-evolution information and computing resources, the prediction of contact graphs has broken through.[111] PSICOV,[12] FreeContact,[13] and CCMpred[112] are notable achievements of GPU development, they allow the development of a growing database and have triggered a new round of deep learning methods. RaptorX-Contact is a contact predictor based on residual CNN architecture.[6] DNCON2[11] is a two-stage CNN model with features similar to MetaPSICOV. DeepCov uses CNN and some co-evolution information to predict contact maps.[113] SPOT-Contact[114] was inspired by RaptorX-Contact and added the residual two-dimensional bidirectional LSTM layer behind the original CNN. AlphaFold[115] is currently the best performing protein structure predictor, which uses a very deep residual neural network with 220 residual blocks.

In addition, in recent years, there have also been some machine/deep learning methods that predict protein–protein interaction sites and interface residue pairs. Among the many machine learning methods, the following methods are the most successful: support vector machine (SVM)[116–121] and fuzzy SVM,[122] neural networks (NN),[123–128] Bayesian networks (BN),[129,130] naive Bayes classifier (NBC),[131,132] random forests (RF),[133–136] cascade random forests (CRF),[137] conditional random fields (CRF),[138] extreme learning machine (ELM),[139] L1-logreg classifier,[140] and the ensemble method.[141–144] The most commonly used algorithms in deep learning, convolutional neural network (CNN)[145] can be used to extract features from input. Besides, the long short-term memory neural network (LSTM) is also used to predict protein–protein interaction interface residue pairs.[16]

7. Conclusion and perspectives

In the past few decades, scientists have conducted in-depth studies on protein–protein interactions, and have proposed many interaction interface residues, protein docking, and complex prediction algorithms to predict the structure of protein complexes. We have organized the methods, descriptions, advantages, and limits of these types of problems, as shown in Table 2. However, there is still a lot of work on these issues that can be further improved. The future development trend of protein complex structure prediction should improve the complex prediction system shown in Fig. 3.

Two topics need to be considered: efficiency and accuracy. For example, in protein docking, how to search the conformation space more efficiently is a question that can be considered. When designing a deep learning model, a deep and complex network does not mean that there will be good prediction results. While ensuring accuracy, the model solves the problem more efficiently is what we need. When it comes to accuracy, many deep learning models are not very interpretable. This means that deep learning models still need to be optimized and more specific to the problem. Especially in the problem of protein–protein interaction interface residue pairs prediction with extreme imbalance of such positive and negative samples, the accuracy needs to be improved. The current methods can not actually provide practical help for biological experiments.

Table 2. The summary of protein complex calculations.

		Methods	Description	Advantages	Limits
Interface residue pair prediction		ComplexContact; RaptorX-Contact; RaptorX-Property; Gremlin; DNCON2; PSICOV; FreeContact; LSTM; LSTM with Graph Representation	Direct evolutionary coupling analysis (DCA), machine learning and deep learning methods	Interfacial residue pair prediction can help subsequent protein complex structure predictions, such as docking.Protein contact map prediction can help reconstruct the three-dimensional structure of protein complexes.	The accuracy of interface residues for prediction needs to be improved.
Protein structure prediction	Template-free	ZDOCK; GRAMM; DOT; SmoothDock; ClusPro; MolFit; FTDock; 3D-Dock; PIPER; pyDock; HDOCK; SDOCK; HEX; FRODOCK; InterEvDock; MDockPP; CoDockPP; HSYMDOCK; SAM; RosettaDock; ICM-DOCK; HADDOCK; ATTRACT; DARWIN; Multi-LZerD; AutoDock	The search strategies of these methods are mainly FFT, GA and MC.	Protein docking can give all possible complex structures, some of which can also dock Cn and Dn complexes.	Designing an effective scoring function to sort the docking structure remains to be further explored.
	Template-based	InterPreTS; Multimeric threading approach; M-Tasser; PISA; ProtCID	Using sequence or structure similarity to model protein complexes with known structures.	Template-based methods mainly reduce the possible structure by restricting the direction of protein binding. This method is more efficient than docking and can be applied to larger-scale protein complex prediction.	For proteins without a template, the structure of the complex cannot be predicted.
Protein complex prediction from PPI networks	Complex prediction based on PPI network clustering	MCODE; MCL; SPC; LCMA; SuperComplex; BN; CFinder; DPClus; IPCA; CMC; ClusterONE; HACO	The protein complex is part of a known PPI network, that is, the graph composed of protein complexes and their interactions is a subgraph of the PPI network.	Some of these methods only use the PPI network for clustering, and some use additional biological information, including structure, function, organization and coevolution infornation, etc.	The proteins that may form complexes can only be picked out from the existing PPI network.
	Complex interaction link prediction from PPI network	GGA; HAC; ECT; RWR; MDS; Link-weighted PPI	Methods to predict actual links in the network include public neighbors-based methods and distance-based methods.	This type of method predicts possible protein–protein interactions based on existing network information.	Public neighbors-based methods have limited effect on sparse networks.

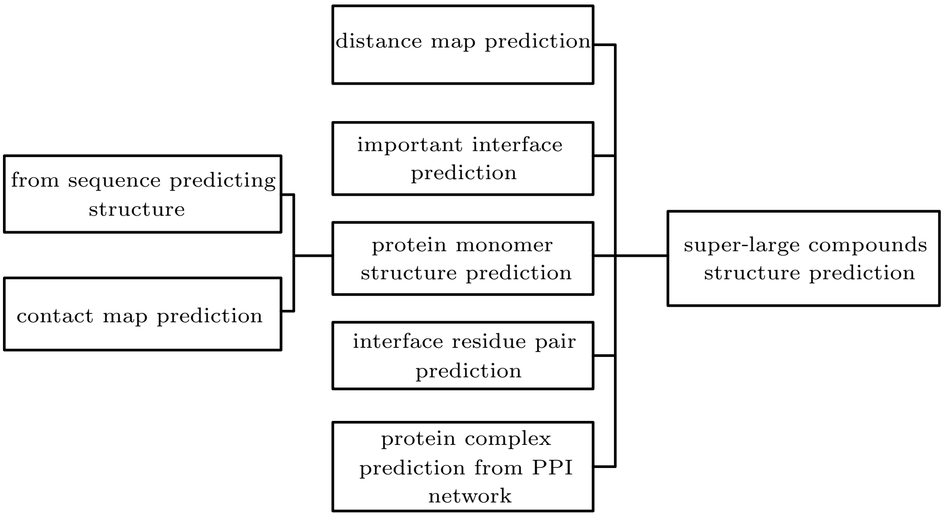

distance map prediction
important interface prediction
from sequence predicting structure
protein monomer structure prediction
super-large compounds structure prediction
contact map prediction
interface residue pair prediction
protein complex prediction from PPI network

Fig. 3. The system of protein complex prediction.

References

[1] Janin J 2010 *Molecular bioSystems* **6** 2351

[2] Sudha G, Nussinov R and Srinivasan N 2014 *Prog. Biophys. Mol. Biol.* **116** 141

[3] Zhou T M, Wang S and Xu J 2018 *bioRxiv*

[4] Zeng H, Wang S, Zhou T, Zhao F, Li X, Wu Q and Xu J 2018 *Nucleic Acids Research* **46** W432

[5] Ching T, Himmelstein D S, Beaulieu-Jones B K, *et al.* 2018 *J. R. Soc. Interface* **15** 20170387

[6] Wang S, Sun S, Li Z, Zhang R and Xu J 2017 *PLoS Comput. Biol.* **13** e1005324

[7] Wang S, Li W, Liu S and Xu J 2016 *Nucleic Acids Research* **44** W430

[8] Kamisetty H, Ovchinnikov S and Baker D 2013 *Proc. Natl. Acad. Sci. USA* **110** 15674

[9] Balakrishnan S, Kamisetty H, Carbonell J G, Lee S I and Langmead C J 2011 *Proteins* **79** 1061

[10] Ovchinnikov S, Kamisetty H and Baker D 2014 *Elife* **3** e02030

[11] Adhikari B, Hou J and Cheng J 2018 *Bioinformatics* **34** 1466

[12] Jones D T, Buchan D W, Cozzetto D and Pontil M 2012 *Bioinformatics* **28** 184

[13] Kaján L, Hopf T A, Kalaš M, Marks D S and Rost B 2014 *BMC Bioinformatics* **15** 85

[14] Wang S, Li W, Zhang R, Liu S and Xu J 2016 *Nucleic Acids Res.* **44** W361

[15] Xu G, Wang Q and Ma J 2020 *Journal of Chemical Theory and Computation* **16** 3970

[16] Zhao Z and Gong X 2019 *IEEEACM Transactions on Computational Biology and Bioinformatics* **16** 1753

[17] Sun D and Gong X 2020 *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* 140504

[18] Liu J and Gong X 2019 *BMC Bioinformatics* **20** 609

[19] Vreven T, Hwang H, Pierce B G and Weng Z 2014 *Brief Bioinform* **15** 169

[20] Zhang Q, Feng T, Xu L, Sun H, Pan P, Li Y, Li D and Hou T 2016 *Curr. Drug Targets* **17** 1586

[21] Ritchie D W and Kemp G J 2000 *Proteins: Structure, Function, and Bioinformatics* **39** 178

[22] Lee K, Czaplewski C, Kim S Y and Lee J 2005 *Journal of Computational Chemistry* **26** 78

[23] Ritchie D W 2003 *Proteins: Structure, Function, and Bioinformatics* **52** 98

[24] Pierce B, Tong W and Weng Z 2005 *Bioinformatics* **21** 1472

[25] Chen R, Li L and Weng Z 2003 *Proteins* **52** 80

[26] André I, Bradley P, Wang C and Baker D 2007 *Proc. Natl. Acad. Sci. USA* **104** 17656

[27] Comeau S R, Gatchell D W, Vajda S and Camacho C J 2004 *Bioinformatics* **20** 45

[28] Yan Y, Tao H, He J and Huang S Y 2020 *Nature Protocols* **15** 1829

[29] Yan Y, Zhang D, Zhou P, Li B and Huang S Y 2017 *Nucleic Acids Res.* **45** W365

[30] Yan Y, Tao H and Huang S Y 2018 *Nucleic Acids Res.* **46** W423

[31] Ritchie D W and Grudinin S 2016 *Journal of Applied Crystallography* **49** 158

[32] Yu H, Luscombe N M, Lu H X, Zhu X, Xia Y, Han J D J, Bertin N, Chung S, Vidal M and Gerstein M 2004 *Genome Research* **14** 1107

[33] Aloy P and Russell R B 2002 *Proc. Natl. Acad. Sci. USA* **99** 5896

[34] Aloy P and Russell R B 2003 *Bioinformatics* **19** 161

[35] Lu L, Lu H and Skolnick J 2002 *Proteins* **49** 350

[36] Chen H and Skolnick J 2008 *Biophys J.* **94** 918

[37] Launay G and Simonson T 2008 *BMC Bioinformatics* **9** 427

[38] Aloy P, Böttcher B, Ceulemans H, Leutwein C, Mellwig C, Fischer S, Gavin A C, Bork P, Superti-Furga G and Serrano L 2004 *Science* **303** 2026

[39] Nye T M W, Berzuini C, Gilks W R, Babu M M and Teichmann S A 2004 *Bioinformatics* **21** 993

[40] Krissinel E and Henrick K 2007 *Journal of Molecular Biology* **372** 774

[41] Xu Q and Dunbrack R L 2010 *Nucleic Acids Res.* **39** D761

[42] Yu H, Braun P, Yildirim M A, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N, Hao T, Rual J F, Dricot A, Vazquez A, Murray R R, Simon C, Tardivo L, Tam S, Svrzikapa N, Fan C, de Smet A S, Motyl A, Hudson M E, Park J, Xin X, Cusick M E, Moore T, Boone C, Snyder M, Roth F P, Barabási A L, Tavernier J, Hill D E and Vidal M 2008 *Science* **322** 104

[43] Tarassov K, Messier V, Landry C R, Radinovic S, Serna Molina M M, Shames I, Malitskaya Y, Vogel J, Bussey H and Michnick S W 2008 *Science* **320** 1465

[44] Krogan N J, Cagney G, Yu H, *et al.* 2006 *Nature* **440** 637

[45] Gavin A C, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen L J, Bastuck S, Dümpelfeld B, Edelmann A, Heurtier M A, Hoffman V, Hoefert C, Klein K, Hudak M, Michon A M, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick J M, Kuster B, Bork P, Russell R B and Superti-Furga G 2006 *Nature* **440** 631

[46] Pržulj N 2011 *Bioessays* **33** 115

[47] Yu H, Kim P M, Sprecher E, Trifonov V and Gerstein M 2007 *PLoS Comput. Biol.* **3** e59

[48] Jeong H, Mason S P, Barabási A L and Oltvai Z N 2001 *Nature* **411** 41

[49] Han J D J, Bertin N, Hao T, Goldberg D S, Berriz G F, Zhang L V, Dupuy D, Walhout A J M, Cusick M E, Roth F P and Vidal M 2004 *Nature* **430** 88

[50] Wang J, Li M, Deng Y and Pan Y 2010 *BMC Genomics* **11** Suppl 3 S10

[51] Ulitsky I and Shamir R 2009 *Bioinformatics* **25** 1158

[52] Sharan R, Ulitsky I and Shamir R 2007 *Mol. Syst. Biol.* **3** 88

[53] Lee K, Chuang H Y, Beyer A, Sung M K, Huh W K, Lee B and Ideker T 2008 *Nucleic Acids Res.* **36** e136

[54] King A D, Przulj N and Jurisica I 2004 *Bioinformatics* **20** 3013

[55] Friedel C C, Krumsiek J and Zimmer R 2009 *J. Comput. Biol.* **16** 971

[56] Chua H N, Sung W K and Wong L 2006 *Bioinformatics* **22** 1623

[57] Bader G D and Hogue C W V 2002 *Nature Biotechnology* **20** 991

[58] Asthana S, King O D, Gibbons F D and Roth F P 2004 *Genome Research* **14** 1170

[59] Kim Y A, Wuchty S and Przytycka T M 2011 *PLoS Comput. Biol.* **7** e1001095

[60] Ideker T and Sharan R 2008 *Genome Research* **18** 644

[61] Hidalgo C A, Blumm N, Barabási A L and Christakis N A 2009 *PLoS Comput. Biol.* **5** e1000353

[62] Hannum G, Srivas R, Guénolé A, van Attikum H, Krogan N J, Karp R M and Ideker T 2009 *PLoS Genet.* **5** e1000782

[63] Chuang H Y, Lee E, Liu Y T, Lee D and Ideker T 2007 *Molecular Systems Biology* **3** 140

[64] Huang H, Jedynak B M and Bader J S 2007 *PLoS Comput. Biol.* **3** e214

[65] Lei C and Ruan J *2012 IEEE International Conference on Bioinformatics and Biomedicine*, 4–7 October, 2012, pp. 1–6

[66] Srihari S, Yong C H and Wong L 2017 *Computational prediction of protein complexes from protein interaction networks* (Association for Computing Machinery)

[67] Bader G D and Hogue C W 2003 *BMC bioinformatics* **4** 2

[68] Pereira Leal J B, Enright A J and Ouzounis C A 2004 *PROTEINS: Structure, Function, and Bioinformatics* **54** 49

[69] Brohee S and Van Helden J 2006 *BMC Bioinformatics* **7** 488

[70] Pu S, Vlasblom J, Emili A, Greenblatt J and Wodak S J 2007 *Proteomics* **7** 944

[71] Blatt M, Wiseman S and Domany E 1996 *Phys. Rev. Lett.* **76** 3251

[72] Getz G, Vendruscolo M, Sachs D and Domany E 2002 *Proteins: Structure, Function, and Bioinformatics* **46** 405

[73] Getz G, Levine E and Domany E 2000 *Proc. Natl. Acad. Sci. USA* **97** 12079

[74] Spirin V and Mirny L A 2003 *Proc. Natl. Acad. Sci. USA* **100** 12123

[75] Li X L, Foo C S, Tan S H and Ng S K 2005 *Genome Informatics* **16** 260

[76] Qi Y, Balem F, Faloutsos C, Klein-Seetharaman J and Bar-Joseph Z 2008 *Bioinformatics* **24** i250

[77] Yong C H, Liu G, Chua H N and Wong L *BMC Systems Biology*, p. S13

[78] Srihari S and Leong H W 2012 *International Journal of Bioinformatics Research and Applications* **8** 286

[79] Adamcsek B, Palla G, Farkas I J, Derenyi I and Vicsek T 2006 *Bioinformatics* **22** 1021

[80] Altaf-Ul-Amin M, Shinbo Y, Mihara K, Kurokawa K and Kanaya S 2006 *BMC Bioinformatics* **7** 207

[81] Li M, Chen J E, Wang J X, Hu B and Chen G 2008 *BMC Bioinformatics* **9** 398

[82] Liu G, Wong L and Chua H N 2009 *Bioinformatics* **25** 1891

[83] Nepusz T, Yu H and Paccanaro A 2012 *Nat. Methods* **9** 471

[84] Wang H, Kakaradov B, Collins S R, Karotki L, Fiedler D, Shales M, Shokat K M, Walther T C, Krogan N J and Koller D 2009 *Mol. Cell Proteomics* **8** 1361

[85] Tong H, Faloutsos C and Pan J Y *Sixth international conference on data mining (ICDM'06)*, pp. 613–622

[86] Radicchi F, Castellano C, Cecconi F, Loreto V and Parisi D 2004 *Proc. Natl. Acad. Sci. USA* **101** 2658

[87] Li A and Horvath S 2007 *Bioinformatics* **23** 222

[88] Fouss F, Pirotte A, Renders J and Saerens M 2007 *IEEE Transactions on Knowledge and Data Engineering* **19** 355

[89] Lü L and Zhou T 2011 *Physica A* **390** 1150

[90] Wang C, Ding C, Yang Q and Holbrook S R 2007 *Genome Biol.* **8** R271

[91] Fang Y, Benjamin W, Sun M and Ramani K 2011 *PLoS One* **6** e19349

[92] Xu Q, Xiang E W and Yang Q 2011 *Proteomics* **11** 3818

[93] Park Y and Bader J S 2011 *BMC Bioinformatics* **12** S44

[94] Kuchaiev O, Rasajski M, Higham D J and Przulj N 2009 *PLoS Comput. Biol.* **5** e1000454

[95] Lei C and Ruan J 2012 *Bioinformatics* **29** 355

[96] Wang L, Hu K and Tang Y 2014 *Current Bioinformatics* **9** 246

[97] Bartoli L, Capriotti E, Fariselli P, Martelli P L and Casadio R 2008 *Methods Mol. Biol.* **413** 199

[98] Vassura M, Margara L, Di Lena P, Medri F, Fariselli P and Casadio R 2008 *IEEEACM Trans. Comput. Biol. Bioinform.* **5** 357

[99] Vendruscolo M, Kussell E and Domany E 1997 *Fold Des.* **2** 295

[100] Breu H and Kirkpatrick D G 1998 *Computational Geometry* **9** 3

[101] Zhang C, Mortuza S M, He B, Wang Y and Zhang Y 2018 *Proteins* **86** Suppl 1 136

[102] Baú D, Martin A J M, Mooney C, Vullo A, Walsh I and Pollastri G 2006 *BMC Bioinformatics* **7** 402

[103] Kukic P, Mirabello C, Tradigo G, Walsh I, Veltri P and Pollastri G 2014 *BMC Bioinformatics* **15** 6

[104] Walsh I, Baù D, Martin A J M, Mooney C, Vullo A and Pollastri G 2009 *BMC Structural Biology* **9** 5

[105] Fariselli P and Casadio R 2001 *Bioinformatics* **17** 957

[106] Martelli P L, Fariselli P, Malaguti L and Casadio R 2002 *Protein Engineering, Design and Selection* **15** 951

[107] Ceroni A, Passerini A, Vullo A and Frasconi P 2006 *Nucleic Acids Res.* **34** W177

[108] Tsai C H, Chen B J, Chan C H, Liu H L and Kao C Y 2005 *Bioinformatics* **21** 4416

[109] Vullo A and Frasconi P 2004 *Bioinformatics* **20** 653

[110] Ferrè F and Clote P 2005 *Nucleic Acids Res* **33** W230

[111] Schaarschmidt J, Monastyrskyy B, Kryshtafovych A and Bonvin A 2018 *Proteins* **86** Suppl 1 51

[112] Seemayer S, Gruber M and Söding J 2014 *Bioinformatics* **30** 3128

[113] Jones D T and Kandathil S M 2018 *Bioinformatics* **34** 3308

[114] Hanson J, Paliwal K, Litfin T, Yang Y and Zhou Y 2018 *Bioinformatics* **34** 4039

[115] Senior A W, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson A W and Bridgland A 2019 *Proteins: Structure, Function, and Bioinformatics* **87** 1141

[116] Dong Q, Wang X, Lin L and Guan Y 2007 *BMC Bioinformatics* **8** 147

[117] Minhas F u A A, Geiss B J and Ben-Hur A 2014 *Proteins* **82** 1142

[118] Zellner H, Staudigel M, Trenner T, Bittkowski M, Wolowski V, Icking C and Merkl R 2012 *Proteins* **80** 154

[119] Wang B, Chen P, Huang D S, Li J J, Lok T M and Lyu M R 2006 *FEBS Lett.* **580** 380

[120] Bradford J R and Westhead D R 2005 *Bioinformatics* **21** 1487

[121] Koike A and Takagi T 2004 *Protein Eng. Des. Sel.* **17** 165

[122] Sriwastava B K, Basu S and Maulik U 2015 *J. Biosci.* **40** 809

[123] Singh G, Dhole K, Pai P P and Mondal S 2014 *SPRINGS: prediction of protein–protein interaction sites using artificial neural networks*, Report No. 2167–9843

[124] Ofran Y and Rost B 2007 *Bioinformatics* **23** e13

[125] Chen H and Zhou H X 2005 *Proteins* **61** 21

[126] Ofran Y and Rost B 2003 *FEBS Lett.* **544** 236

[127] Fariselli P, Pazos F, Valencia A and Casadio R 2002 *Eur. J. Biochem.* **269** 1356

[128] Zhou H X and Shan Y 2001 *Proteins: Structure, Function, and Bioinformatics* **44** 336

[129] Bradford J R, Needham C J, Bulpitt A J and Westhead D R 2006 *J. Mol. Biol.* **362** 365

[130] Neuvirth H, Raz R and Schreiber G 2004 *J. Mol. Biol.* **338** 181

[131] Geng H, Lu T, Lin X, Liu Y and Yan F 2015 *Biochemistry Research International* **2015** 978193

[132] Murakami Y and Mizuguchi K 2010 *Bioinformatics* **26** 1841

[133] Chen X W and Jeong J C 2009 *Bioinformatics* **25** 585

[134] Northey T C, Barešić A and Martin A C R 2018 *Bioinformatics* **34** 223

[135] Li B Q, Feng K Y, Chen L, Huang T and Cai Y D 2012 *PLoS One* **7** e43927

[136] Sikić M, Tomić S and Vlahovicek K 2009 *PLoS Comput. Biol.* **5** e1000278

[137] Wei Z S, Yang J Y, Shen H B and Yu D J 2015 *IEEE Trans. Nanobioscience* **14** 746

[138] Li M H, Lin L, Wang X L and Liu T 2007 *Bioinformatics* **23** 597

[139] Wang D D, Wang R and Yan H 2014 *Neurocomputing* **128** 258

[140] Dhole K, Singh G, Pai P P and Mondal S 2014 *J. Theor. Biol.* **348** 47

[141] Jia J, Liu Z, Xiao X, Liu B and Chou K C 2016 *J. Biomol. Struct. Dyn.* **34** 1946

[142] Deng L, Guan J, Dong Q and Zhou S 2009 *BMC Bioinformatics* **10** 426

[143] Chen P and Li J 2010 *BMC Bioinformatics* **11** 402

[144] Du X, Sun S, Hu C, Li X and Xia J 2016 *J. Biol. Res. (Thessalon)* **23** 10

[145] Krizhevsky A, Sutskever I and Hinton G E *Advances in Neural Information Processing Systems*, pp. 1097–1105