Front. Comput. Sci., 2020, 14(4): 144901
https://doi.org/10.1007/s11704-019-8232-z

REVIEW ARTICLE

A survey of current trends in computational predictions of protein-protein interactions

Yanbin WANG[1,2], Zhuhong YOU (✉)[1], Liping LI[1], Zhanheng CHEN[1,2]

1 Xinjiang Technical Institutes of Physics and Chemistry, Chinese Academy of Science, Urumqi 830011, China
2 University of Chinese Academy of Sciences, Beijing 100049, China

Abstract Proteomics become an important research area of interests in life science after the completion of the human genome project. This scientific is to study the characteristics of proteins at the large-scale data level, and then gain a holistic and comprehensive understanding of the process of disease occurrence and cell metabolism at the protein level. A key issue in proteomics is how to efficiently analyze the massive amounts of protein data produced by high-throughput technologies. Computational technologies with low-cost and short-cycle are becoming the preferred methods for solving some important problems in post-genome era, such as protein-protein interactions (PPIs). In this review, we focus on computational methods for PPIs detection and show recent advancements in this critical area from multiple aspects. First, we analyze in detail the several challenges for computational methods for predicting PPIs and summarize the available PPIs data sources. Second, we describe the state-of-the-art computational methods recently proposed on this topic. Finally, we discuss some important technologies that can promote the prediction of PPI and the development of computational proteomics.

Keywords proteomics, protein-protein interactions, protein feature extraction, computational proteomics

1 Introduction

Since the official implementation of the human genome project (HGP) in the 1990s, the study of computational analysis methods for genome sequence information has become one of the most concentrated research topics in bioinformatics. With the announcement of the completion of HGP in 2003 and the increased awareness of researchers on the role of proteins in living systems, comprehensive research on proteomics has gradually become a crucial task in the post-genome era. Proteomics is a new discipline that studies the level of protein expression, post-translational modification, protein-protein interaction and so on. Related research not only provides molecular-level insights into the law of life activities, but also provides theoretical grounds and solutions for the elucidation and capture of numerous disease mechanisms.

For example, abnormalities in protein interactions will affect the activity of cells and their function, leading to many diseases such as neurodegenerative diseases and cancers. Computational proteomics is the study of how to use computational techniques to solve key biological problems in proteomics, such as protein identification, structure prediction, functional classification, subcellular localization, post-translational modification analysis, protein interactions, quantitative analysis, disease diagnosis and drug design. It has become a major branch of computational biology and bioinformatics. In particular, the rapid development of computer hardware, information processing technology, and network technology in recent years has provided mature conditions for the extensive development of computational proteomics research. Computational proteomics is playing an increasingly important role in the research of contemporary life

Received July 1, 2018; accepted July 3, 2019

E-mail: zhuhongyou@ms.xjb.ac.cn

sciences [1,2].

Protein is the material basis of life and the main bearer of life activities and functions. The vast majority of intracellular biochemical functions and biological processes involve protein-protein interactions (PPIs) [3]. Therefore, a comprehensive identification of PPIs is beneficial to deciphering the molecular mechanisms of specific biological functions, and providing a global picture of biological processes and cellular functions. In addition, although significant progress has been made in the fields of genomics and molecular biology, the function of most proteins is still unknown [4]. Comprehensively revealing the function of a protein is a complex and long-term work involving many proteins that potential functions have not yet been discovered and some proteins that may be involved in performing multiple functions. Jansen et al. [5] studies show that the interaction between unknown function proteins and known functional proteins can contribute greatly to the determination of protein function. Therefore, the prediction of protein-protein interactions (PPIs) is an important challenge currently facing bioinformatics and proteomics. Several physiochemical experimental techniques have assisted in identifying PPIs within the interactome. However, these technologies are expensive, time-consuming, and the data generated in the lab covers only a small part of the whole PPI network. Moreover, these data usually contain high false-positive, false-negative data, and few overlapping data observed between experimentally generated datasets. This may indicate that a portion of this data is not trusted. According to statistics, only about 10% of human PPI networks and 50% of yeast PPI networks have been characterized. Due to the limitations of experimental methods and the demand of determining PPIs, additional reliable computational methods need to be developed to accelerate the discovery of PPIs [6–9].

In this paper, we focus on reviewing current computational methods for inferring PPIs based on sequence information. We describe the progress from three aspects: PPIs data source, protein feature representation and prediction strategy. At the end of the paper, we also discuss the prospect of deep learning technology in computational proteomics and the opportunities of quantum mechanics in this field.

2 Technical challenges and available data

2.1 Problems confronting

The computational approach to PPI prediction presents several technical challenges:

1) A large amount of protein data was generated in the past genomic era. In the context of big data, how to succinctly and delicately handle these data and maintain good scalability are issues that must be considered.
2) How to establish a comprehensive, accurate and reliable gold standard dataset to train and evaluate prediction methods?
3) Proteins have a variety of physical and chemical properties and different structural characteristics, so the efficient and accurate extraction of features is a common problem faced by PPIs.
4) The original features are usually rough. Effectively reducing feature size and noise, eliminating similar information can contribute to improve the accuracy of the model, reduce the computational complexity of the model, and improve the interpretability of the model. However, the noise removal technology used to process biological data has not yet been officially launched.
5) How to effectively code protein pairs? or How to combine the features of two sequences to form the feature vectors of protein pairs?
6) How to choose an efficient and accurate prediction algorithm, which can make full use of the current information and establish an effective model to reduce the error of PPI prediction.
7) Some machine learning algorithms are easy to be overfitting or trapped in local optimization when they are applied.
8) Most of the existing PPI prediction models are based on balanced data sets. But realistic PPIs datasets are often unbalanced, which leads to training a predictor with "preference".

2.2 PPIs data sources

Several previous international collaborative genome projects have accumulated a great deal of protein interaction information, which is contained in several large databases in different formats. With the deepening of research, these databases have gradually become an important resource for global biological researchers. A list of popular PPIs databases is presented in Table 1.

3 Protein feature representation

The protein feature extraction methods appearing in the PPI prediction methods proposed in recent years are mainly based

Table 1 A list of popular PPIs databases

Database	Database description
DIP [10]	A simple, highly trusted PPI public database, including multiple species.
BioGRID [11]	1728498 protein were found in 70208 publications.
BIND [12]	Including Human, Fruit Fly, Yeast, Nematode.
SGD [13]	The Saccharomyces Genome Database provides comprehensive integrated biological information for the budding yeast Saccharomyces cerevisiae.
HPRD [14]	The Human Protein Reference Database integrate interaction networks information for each protein in the human proteome.
MIPS [15]	The MIPS Database is a collection of manually curated high-quality PPI data collected from the scientific literature by expert curators.
IntACT [16]	Data sources come from literature collation or direct user submission.

on the sequence [17–19]. Amino acids constitute the basic unit of protein sequences. Protein sequence determines the structural information and function of protein, and is the most basic and easily accessible feature information of protein. Previously, a sequence-based approach studies and analyzes the basic properties of amino acids and compositional information of protein sequences to find out their intrinsic rules and the relationship between them and to predict the interactions between proteins [20]. Although the continuous development of classification algorithms has greatly improved the accuracy of PPI prediction, as the study continues, researchers have gradually found that the sequence information obtained by the method of amino acid composition is very limited, which makes the accuracy of PPI prediction greatly restricted.with the further research, several feature extraction algorithms including more protein information are proposed, including pseudo amino acid composition, auto covariance, resonance recognition model etc. These methods improve prediction performance with varying degrees, but they are also based on the physicochemical properties of amino acids. Recently, some successful research cases based on the evolutionary information of proteins have reminded researchers that methods based on the evolutionary information of proteins deserve to be reconsidered [21]. In the past two years, some methods based on image processing technology or signal processing technology have been reported for obtaining protein evolution information. These methods have achieved excellent results on various PPIs data sets. In addition, the fusion of protein features is also a subject of continuous attention by researchers recently. In this section, we briefly review the progress made in the feature extraction of protein sequences.

3.1 Auto covariance

Auto covariance (AC) is a concept in statistics that refers to the covariance between a signal and their neighbor signals. Guo et al. [22] first introduced this concept to bioinformatics and proposed auto-covariance coding scheme for proteins. The method is based on the assumption that there are certain signal-like fluctuations in the protein sequence. This method treats the protein sequence as a set of signals. The protein sequence is then replaced with digital sequence using appropriate physicochemical properties. Finally, the digital sequence signal is further analyzed for obtaining protein features. The AC method takes into account both the physicochemical properties and the positional information of amino acids indirectly.

3.2 Resonant recognition model

The resonance recognition model (RRM) [23] was originally proposed in the field of mathematical physics. This method converts amino acid sequences into numerical sequences by assigning a set of specific physicochemical parameters. There was a certain degree of correlation between the physicochemical parameters and protein activity of the protein. Previous studies have found that the relationship between the local ionization parameters of amino acids and the activity of proteins is very close. This indicates that the ionization energy of amino acids has a great effect on the distribution of protein sequences. Therefore, such a numerical sequence actually reflects the distribution of ionization energy over the entire protein sequence. Based on this, the signal processing technology of the RRM can be used for protein sequence analysis and obtaining the energy features of protein sequences.

3.3 Conjoint triad and local descriptor

Shen et al. believe that sufficient consideration should be given to the local environment of the amino acid sequence to obtain reliable predictive performance [21]. Therefore, the conjoint triad (CT) encoding method was proposed. CT is one of the k-mer sequence assembly algorithms, and the default value of K is 3. CT considers the correlation between each amino acid and its neighboring amino acids, divides the adjacent three consecutive amino acids into one combined unit, and calculates the frequency of occurrence of each combination in the entire sequence.

Protein-protein interactions occur not only in the contiguous amino acid region of the sequence, but also in the discontinuous amino acid region. Some residues are far apart in

one-dimensional space, but they may be adjacent or close in high-dimensional space through the rotation and folding of proteins. In order to capture the interaction between discontinuous sequence fragments, Yang et al. proposed segmented local descriptors. Afterwards, improved versions of some local descriptors were successively proposed [24].

3.4 Feature extraction based on protein evolution information

Hu suggests using a new co-evolutionary feature extraction method called CoFex to jointly consider the features of two sequences in a protein pair during the process of feature extraction of protein evolution. The key to CoFex's co-evolutionary features is the discovery of co-variations in co-evolutionary locations. Based on the presence and absence of these co-evolutionary features in the two protein sequences, the feature vector can consist of a pair of proteins rather than a single protein [25].

There is also a method for obtaining protein evolution information by calculating the position-specific scoring matrix of protein sequences. The algorithm obtains conservative scoring of amino acid residues through homologous protein multiple sequence alignment. Higher scores indicate that the evolutionary histories of the two protein sequences are more similar and the interactions are more likely to occur. The PSSM scoring matrix for a protein sequence is obtained by iterative comparison of the sequence with the data in the non-redundant database SWISS-Prot using the PSI-Blast tool. A PSSM contains 20*L elements, where L is the length of the protein sequence [26–28].

The recently reported method for feature extraction of protein evolution information does not directly use PSSM, and most of them often use a combination of PSSM and image processing techniques to obtain the evolutionary information of proteins. Wang et al. have proposed several new methods to extract the evolution information of proteins in PSSM using Zernike moment and Legendre moment algorithm [29]. Li et al. used alow-rank approximation method to obtain the evolutionary information of the proteins contained in the PSSM [30]. Song et al. use an integrated strategy to extract protein evolution information, which incorporates discrete cosine transform, fast Fourier transform, singular value decomposition [31]. An et al. proposed a method for obtaining protein evolution information combining local phase quantization with PSSM. These successful studies are reminding researchers to examine the problem of protein feature extraction from the perspective of image feature extraction [32]. Figure 1 shows the basic flow of constructing PPIs prediction mode using image processing technology to extract protein evolution information. Next we briefly introduce several available image processing techniques.

3.4.1 Scale-invariant feature transform

The scale-invariant feature transform (SIFT) [33] features are widely used in local feature descriptor for extracting local scale-invariant features of a target, and using these features to match two images. At present, SIFT is relatively popular among the whole local feature descriptors of images, and SIFT features are robust to complex clutter and occlusion. It can be used to match objects in large databases. For a very small object, SIFT algorithm also can produce a large number of features. In addition to these properties, SIFT has the advantage of fast calculation speed and high efficiency. It is widely used in image processing.

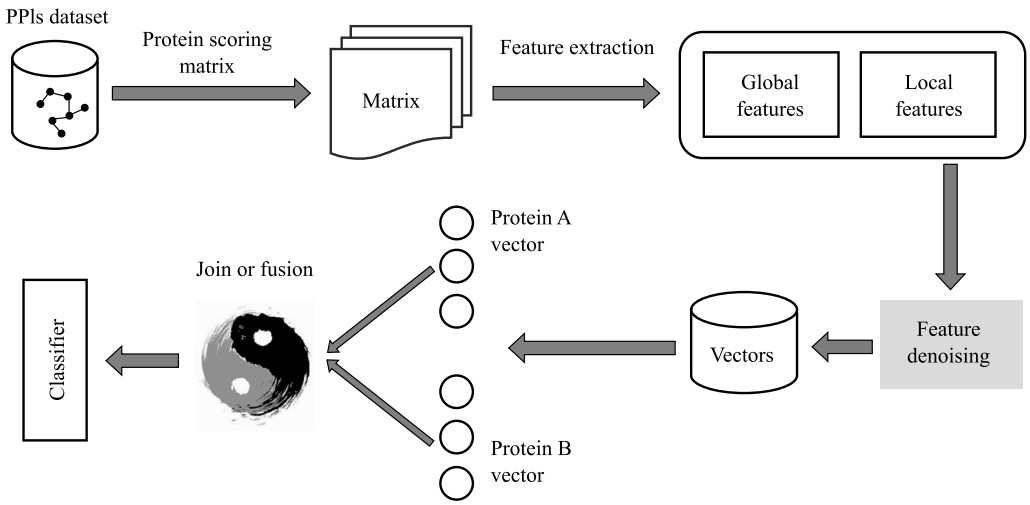


Fig. 1 The basic flow of constructing PPIs prediction mode using image processing technology to extract protein evolution information

3.4.2 Speeded up robust features

The speeded up robust features (SURF) [34] is an interest point detection and description sub-algorithm similar to SIFT. It determines the position of the interest point through the determinant of the Hessian matrix, and then determines the descriptor according to the Haar wavelet response of the neighbor of the interest point. The dimensions of descriptor size are only 64 (it can also be expanded to 128 dimensions, the effect is better). In a word, it is a very good interest point detection algorithm. The algorithm not only maintains the good performance of the SIFT algorithm, but also improves the performance result relative to the SIFT algorithm. It consumes less time and has lower computational complexity, and it also overcomes some shortcomings in the aspect of the extraction at the interest point and feature vector description. The algorithm has improved the speed of calculation.

3.4.3 Moment feature

The moment feature mainly represents the geometric feature of the image region, which also known as the geometric moment. Because it has the invariant features of rotation,translation, scale and other characteristics, it is also called invariant moment. In image processing, geometric invariant moment can be used as an important feature to represent objects, and images can be classified and manipulated based on this feature. At present, the moment features used for image recognition mainly include Hu moment, Zernike moment, wavelet moment and so on. 1) Hu's seven invariant moments are constructed using second-order and third-order normalized center moments. The feature quantities composed of Hu's moments are very fast when identifying images. 2) Zernike moment is a kind of orthogonal moment and is a formal function based on Zernike polynomial. Compared with Hu's moment, although it is computationally complex, it can construct arbitrary higher-order moments, and has better feature expression ability, and has less noise sensitivity, which mainly reflects the size of the image and the stability of the feature. 3) Wavelet moment combines the advantages of wavelet multi-scale analysis and invariant moments, that is, it has the ability of wavelet multi-scale analysis to reflect the local information of signals, and it also has the characteristics of invariant moments, which improves the degree of grasp of the precise structure of the image by moments [35–37].

3.4.4 Histogram of oriented gradient

Histogram of oriented gradient (HOG) [38]features are descriptors which can be used for object detection in computer vision and image processing. HOG features are formed by computing and counting the histogram of gradient directions in local regions of the image. First, the protein sequence is divided into small cell units. Then, the gradient or edge orientation histogram of each pixel is collected in cell units. Finally, we can obtain the feature descriptors by combining these histograms. HOG features combined with SVM classifier have been widely used in image recognition.

3.4.5 Haar-like feature

Haar feature, also known as Haar-Like feature, was first proposed by Whitehill and Omlin [39]. Haar features can be divided into three types: edge feature, liner feature, central and diagonal feature. They are a simple and inexpensive image features based on intensity differences between rectangle-based regions that share similar shapes to the Haar wavelets. In other words, the value of a rectangular feature refers to the sum *D*-value between the gray value of all pixels in the two or more rectangles of the same shape and size, that is, the sum of the gray values of all pixels in the white rectangle region is used to minus the sum of the gray values of all pixels in the black rectangle region.

3.4.6 Local binary pattern

Ojala et al. proposed local binary pattern (LBP) features [40]. It is an operator used to describe the local features of an image. LBP features have significant advantages such as gray invariance and rotation invariance and so on. Because of the simple calculation and better effect, LBP features have been widely applied in a great deal of fields in computer vision. The LBP operator is defined in the neighborhood of the pixel 3*3. To compared the gray value of the 8 adjacent pixels with the neighborhood central pixel values. If the surrounding pixels are larger than the central pixel values, the location of the pixel points is marked as 1, otherwise, it is recorded as 0. In this way, 8 points in the neighborhood of the pixel 3*3 can be generated as 8 bit binary numbers, and the 8 bit binary numbers are arranged in turn to form a binary numeral, which is the LBP value of the central pixel.

3.4.7 Texture spectrum

Texture spectrum [41] is a description method based on the local structural features of images. The specific expression is a change in the gray level or color of pixels in a pixel neighborhood of a pixel, and this change is related to space statistics. It is composed of two elements of the arrangement of

texture elements and basic elements. The texture spectrum is invariant when the image is rotated arbitrarily. This feature not only makes the texture spectrum features symmetric and invariant, but also is robust to texture rotation. Texture features analysis methods include statistical methods, structural methods and model-based methods.

3.4.8 Gabor transform

The Gabor feature [42] can be used to describe the texture information of the image. The frequency and direction of the Gabor filter are similar to the human visual system, which is especially suitable for texture representation and discrimination. It also can be employed to extract the features of protein sequence. It mainly relies on the Gabor kernel to windowed the sequence information in the frequency domain, so that the local information of the protein sequence can be described. In essence, it is the Fourier transform. In the original space domain, a Gabor kernel is actually a result of the Gauss kernel and the sine-wave modulation, which can be regarded as the Gauss kernel application in the frequency domain of the sine wave. As a result, we can use Gabor transform to obtain the main features for the protein sequence information.

3.5 Coding protein pairs

One important section in PPIs prediction is to integrate the features of two sequences to form the feature vectors of protein pairs. For a protein pair, assume that the representation of protein A is $x_1, x_2, \ldots, x_N$ and that of protein B is $y_1, y_2, \ldots, y_N$.

A common approach for coding protein pairsis to directly link the feature of two protein sequences, which can be represented as Cod 1

$$\text{Cod 1: } \mathrm{A} \oplus \mathrm{B} = [x_1, x_2, \ldots, x_N, y_1, y_2, \ldots, y_N], \quad (1)$$

There is a potential concern in this method, some conflicts may arise when the positions of two sequences are reversed. Feature fusion methods based on numerical operation can avoid such conflicts. Yang et al. tested three feature fusion schemes, in which protein pairs were coded as follows:

$$\text{Cod 2: } \mathrm{abs}(\mathrm{A} - \mathrm{B}) = [|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_N - y_N|], \quad (2)$$

$$\text{Cod 3: } \mathrm{A} + \mathrm{B} = [x_1 + y_1, x_2 + y_2, \ldots, x_N + y_N], \quad (3)$$

$$\text{Cod 4: } (\mathrm{abs}(\mathrm{A} - \mathrm{B}) \oplus (\mathrm{A} + \mathrm{B}) = [|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_N - y_N|, (x_1 + y_1), (x_2 + y_2), \ldots, (x_N + y_N)], \quad (4)$$

However, the experimental results reported by Yang's work [24] shows that the prediction quality of Cod 2–4 is significantly lower than that of Cod 1. The reason is that this kind of feature fusion method may lead to information loss, because the features of the protein A and protein B become undetermined. How to overcome this loss? Theoretically, merging two difference feature fusion methods without abs operation is helpful to ensure the integrity of information. Zeng et al. [43] proposed a protein pairs coding scheme by combing three feature fusion methods, involving average operation, 2-norm operation and relevance operation, which achieved a considerable prediction performance.

The study of coding protein pairs in PPIs prediction has not been taken much concern. Its importance was ignored for some time. But that does not mean it has no research value, it is an area to be explored. Introducing or developing efficient and practical feature fusion technology for the representation of protein pairs is beneficial for classifiers to recognize protein-protein interaction patterns and provide more stable and reliable prediction results.

4 Prediction algorithm

4.1 Support vector machine

Support vector machine (SVM) is one of the most advanced algorithms at present, which has the advantages of good classification performance and strong generalization ability. The basic idea of support vector machine is to map the training data set nonlinearly to a high-dimensional feature space. The purpose of this non-linear mapping is to make the linearly inseparable data set in the original space linearly separable. An optimal separation hyperplane with the largest isolation distance is then established in the feature space, which means that an optimal nonlinear decision boundary is created in the input space. Optimal separation hyperplane of SVM not only minimizes the empirical risk, but also reduces the generalization error. Shen et al. [21] proposed a PPI prediction model based on the SVM algorithm using amino acid sequence information. This method overcomes the limitations of most prediction methods at that time, but this method must understand the homology of proteins before perform prediction. In order to overcome this limitation, they proposed the conjoint triad feature for describing amino acids and selected the SVM with a kernel function as a predictor for protein interaction prediction. Guo et al. [22] proposed a PPI prediction combined Auto Covariance code method with the SVM with radial basis function.

4.2 Relevance vector machine

Relevance vector machine (RVM) is a supervised learning algorithm in Bayesian framework proposed by Tipping in 2000, which removes irrelevant points based on the automatic relevance determination (ARD) to obtain a sparse model. During the iterative learning of sample data, the posterior distribution of most parameters tends to zero. The points corresponding to those non-zero parameters are called relevance vectors, which represent the core features of the data. Compared with the support vector machine, the biggest advantage of the correlation vector machine is that it greatly reduces the computational complexity of the kernel function, and also overcomes the shortcomings that the selected kernel function must satisfy the Mercer condition [44–46]. RVM also has the following advantages: 1) RVM not only obtains binary output but also obtains probability output; 2) RVM does not need to set the penalty factor. The penalty factor in the SVM is a constant that balances empirical risk and confidence intervals. The experimental result is very sensitive to this value. Improper setting will cause some problems such as over-fitting. However, parameters are automatically assigned in RVM; 3) The RVM model is sparser than SVM, and thus takes less time in the testing phase and is more suitable for online detection. As we all know, the number of support vectors for SVM grows linearly with the increase of training samples. When training samples are large, it is obviously not appropriate. Although the RVM correlation vector also increases with training samples, the growth rate is much slower than SVM; 4) The results of practice show that RVM has stronger generalization ability.

4.3 Recurrent neural network

With the continuous deepening of the research work on artificial neural networks [47], it has successfully solved many practical problems that are difficult to solve in the fields of pattern recognition, intelligent robots, automatic control, biology, medicine, and economics. The recurrent neuron network [48] is a neural network that models sequence data. In recent years, RNN has achieved extraordinary performance in natural language processing, image recognition, and speech recognition. In the traditional neural network model, the structure between layers is fully connected, and the neurons within the layer are disconnected. This type of neural network is powerless for certain problems. The current output of a sequence in RNNs is affected by the previous output. Specifically, the network will memorize the previous information and apply it to the calculation of the current output, i.e., the nodes between the hidden layers are connected. The input of the hidden layer not only comes from the output of the input layer but also includes the output of the hidden layer at the previous moment. In other words, the neurons in the hidden layer of the RNN are sequential from left to right. Figure 2 shows the structure of a recurrent neural network. Applying a recurrent neural network to process biological sequence data can take into account the positional relationship of amino acids that cannot be considered by other algorithms. The potential of this technology in the field of biological information has not yet been released, but its unique ability deserves the attention of biologists. Because this strong front-to-back positional relationship often exists in biological sequence data [49,50].

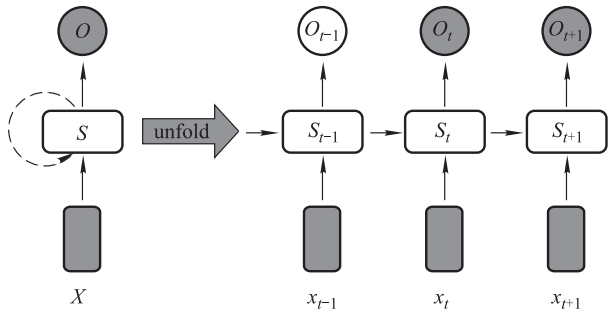


Fig. 2 A recurrent neural network structure

4.4 Long short-term memory

When the interval between the related information and the predicted position is small, the RNNs can learn to use the previous information, but as the time interval grows, the ordinary RNNs cannot learn the long-distance information. With more effort in this area being paid, long short-term memory (LSTM) neural network [51,52], which can learn long-term dependence, is proposed to overcome this problem. The main difference between LSTM network and other networks is its use of complex memory block instead of the neurons of general network. The memory block contains three multiplicative "gate" units (the input, forget, and output gates.) along with some memory cells (one or more). The gate unit is used to control the information flow, and the memory cell is used to store the historical information. The gate removes or restore information to the cell state by controlling the information flow. More specific, the input and output of the information flow are respectively handled by the input and output gates. The forget gate determines how much of the previous unit's information is retained to the current unit [53–55].

5 Evaluation of several existing methods

In recent years, many computational PPIs prediction methods have been proposed based on various data sources, such as genomic information, evolutionary knowledge, structural information, protein domain and protein sequence information. In this survey, we primarily focus on several state-of-the-art sequence-based approaches.

The method proposed by Sprinzak and Margalit [56] predicted interactions between proteins using known sequence-signatures that can be found from the InterPro. This method assumes that interactions occur in these well-defined domain-domain interactions. Bock and Gough [57] develop a more general computational tool by employing protein substructure and physiochemical descriptors based on the amino acid, and then training an SVM model to detect PPIs from these descriptors. Another approach come from Martin's work [58] called signature product combines advantages from Sprinzak et al. [8], Bock and Gough [57] which extended the concept of sequence signatures via subsequence pairing. Benhur and Noble [59] proposed a new pair-wise kernel to integrate the information from homologous interactions, protein sequences, GO annotations, and then they also make use of SVM to perform prediction on yeast PPIs data. Chou and Cai [60] have noted that sequence order effects to affect the quality of classification models, so they proposed pseudo-amino acid composition for representing protein, that incorporate the amino acid position effect. Shen et al. [21] come up with a theory that the local environments of amino acids affect reliability and stability of the prediction model on the basis of amino acid composition method, so they proposed conjoint triad method considers the effects of the two most adjacent amino acids. Guo et al. [22] reported a method that combined the auto-covariance and SVM,which can account for the interactions between residues long-distance apart in the sequence. There are also some other methods based on fusion feature and ensemble learning to predict PPIs. We recount the characters of various PPIs prediction methods in Table 2.

Table 2 Comparison of several state-of-the-art PPIs prediction approaches

Methods	Type of information	Characters
Sprinzak and Margali [56]	Experimental binding data structural domains	Good interpretability but poor generalization ability
Bock and Gough [57]	Sequence information physicochemical properties	Good interpretability but poor accuracy
Martin et al. [58]	Experimental data sequence information	Does not require physio-chemical information prior knowledge of domains
Benhur and Noble [59]	Various information including homologous interactions, GO annotations, protein sequence et.	Can integrate kinds of knowledge, but not very efficient.
Shen et al. [21]	Only protein primary sequence	Good generalization ability does not require any knowledge
Guo et al. [22]	Protein primary sequence physicochemical properties	Good performance but sensitive to data sources
Chou and Cai [60]	Protein primary sequence physicochemical properties	Does not require prior knowledge but sensitive to physicochemical properties
Wang et al. [61]	Only protein primary sequence	Excellent performance but not very efficient.

6 Future trends

Several advanced neural network technologies have been introduced previously, and their potential in dealing with biological sequence data problems has been briefly explored. Recently, the ability of deep learning technology in the field of bioinformatics has been initially verified, and its application in proteomics has also increased year by year. Sun et al. use a stacked auto-encoder to detect protein-protein interactions [62]. Almagro Armenteros et al. developed a predictor based on deep learning technology named DeepLoc to infer protein subcellular localization [63]. Yi et al. proposed a deep learning framework to accurately predict ncRNA-protein interactions [64]. Wang et al. proposed adeep coding-decoding network for protein-protein interaction prediction [65]. From the current popularity of deep learning technology in the field of biological information, it can be inferred that deep neural networks will dominate the field of computational biology in the next five years.

Application of natural language processing technology in Bioinformatics promised to be a hot spot in recent years, especially in sequence-driven biological problems. Biological sequences can be seen as meaningful genetic languages, which have strong similarities with human language. In languages, words can be arranged into meaningful sentences; in biology, amino acid arrangements determine the structure and function of proteins, which can be viewed as meaningful words to analyze the structure and function of proteins. Rescanning some fundamental theory problems with biological

language viewpoint can make us gain much new revelation and find a new solution to bio information problems. There is a very important problem here, how to find "bio words" from biological sequences? And how to develop a reasonable word segmentation system to divide a biological sequence into sentences with biological words as the basic unit? Unlike natural language, the "word" in biological sequence is unknown. Applying mechanically the natural language processing technology to biological problems will only greatly reduce its value in the field of biology, and can hardly produce interpretable results and make a breakthrough. Therefore, the discovery of words in biological sequences is a crucial step. Inspired by natural language, amino acids or bases that are often together or have a high probability of being together can be considered as potential bio-words. For example, in in massive protein sequence database, if the probability of the amino acid fragment "GADE" is high, we can treat it as a word. Two natural language processing techniques, byte-pair-encoding (BPE) [66] and unigram language model (ULM) [67], can be used to find these fragments. There is also a tool called SentencePiece [68] that can be used directly to generate bio words, which integrates BPE and ULM algorithms.

In addition, the development of quantum mechanics may greatly affect the future computational proteomics research. As the chemical structure of a large number of biomolecules becomes clearer, the use of quantum mechanics and quantum machine learning techniques [69–71] to simulate the interaction of molecules in biology maybe likely set off a new technological revolution.

7 Conclusion

There is no doubt that computational proteomics, which is rapidly developing and receiving sufficient attention, is having an unparalleled impact on science and the business revolution. Large scale biological data available will promote the development of all aspects of life sciences. As more and more computing methods infer laboratory results prior to actual physical experiments, narrowing the scope of experimental testing and increasing laboratory productivity, such acceleration will continue over the next decade. Computational methods for infer protein interactions will accelerate the accumulation of this knowledge, thereby facilitating the discovery of molecular mechanisms of complex diseases, promoting the development of disease diagnosis, and improving the efficiency of new drug development. This paper analyzes in detail the several challenges for computational methods for predicting protein-protein interactions, surveys the recent advances in computational methods for predicting protein-protein interactions, briefly reviews several classic PPI prediction methods and focuses on exploring some techniques that can be used to enrich proteomics research and improve computational performance in the future. We have found that proteomics research will enter the multidisciplinary integration phase. With the development of research, image processing technology and signal recognition technology have the potential to be widely used to acquire protein knowledge. In addition, we recommend that deep learning techniques should be strongly introduced in the application of computational proteomics, and it is also essential to increase the investment in quantum information technology in proteomics.

Acknowledgements This work was supported in part by Awardee of the NSFC Excellent Young Scholars Program in 2017, in part by the National Natural Science Foundation of China (Grant Nos. 61902342, 61722212 and 61572506). The authors would like to thank the editors and anonymous reviewers for their constructive advices.

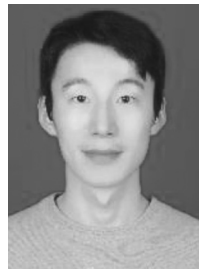
References

1. Colinge J, Bennett K L. Introduction to computational proteomics. PLoS Computational Biology, 2007, 3(7): e114
2. Matthiesen R. Methods, algorithms and tools in computational proteomics: a practical point of view. Proteomics, 2010, 7(16): 2815–2832
3. Jones S, Thornton J M. Principles of protein-protein interactions. Proceedings of the National Academy of Sciences of the United States of America, 1996, 93(1): 13–20
4. Phizicky E M, Fields S. Protein-protein interactions: methods for detection and analysis. Microbiological Reviews, 1995, 59(1): 94–123
5. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan N J, Chung S, Emili A, Snyder M, Greenblatt J F, Gerstein M. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science, 2003, 302(5644): 449–453
6. Rhodes D R, Tomlins S A, Varambally S, Mahavisno V, Barrette T, Kalyanasundaram S, Ghosh D, Pandey A, Chinnaiyan A M. Probabilistic model of the human protein-protein interaction network. Nature Biotechnology, 2005, 23(8): 951–959
7. Oti M, Snel B, Huynen M A, Brunner H G. Predicting disease genes using protein-protein interactions. Journal of Medical Genetics, 2006, 43(8): 691–698
8. Sprinzak E, Sattath S, Margalit H. How reliable are experimental protein-protein interaction data? Journal of Molecular Biology, 2003, 327(5): 919–923
9. Letovsky S, Kasif S. Predicting protein function from protein-protein interaction data: a probabilistic approach. Intelligent Systems in Molecular Biology, 2003, 19: 197–204
10. Xenarios I, Salwinski L, Duan X J, Higney P, Kim S, Eisenberg D. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Research, 2002, 30(1): 303–305

11. Chatr-Aryamontri A, Breitkreutz B J, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, Odonnell L. The BioGRID interaction database. Nucleic Acids Research, 2013, 41: D816–D823
12. Bader G D, Betel D, Hogue C W V. BIND: the biomolecular interaction network database. Nucleic Acids Research, 2001, 31(1): 248–250
13. Cherry J M, Adler C, Ball C A, Chervitz S A, Dwight S S, Hester E T, Jia Y, Juvik G, Roe T, Schroeder M. SGD: saccharomyces genome database. Nucleic Acids Research, 1998, 26(1): 73–79
14. Peri S, Navarro J D, Amanchy R, Kristiansen T Z, Jonnalagadda C K, Surendranath V, Niranjan V, Muthusamy B, Gandhi T K, Gronborg M. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Research, 2003, 13(10): 2363–2371
15. Pagel P, Kovac S, Oesterheld M, Brauner B, Dungerkaltenbach I, Frishman G, Montrone C, Mark P, Stumpflen V, Mewes H. The MIPS mammalian protein–protein interaction database. Bioinformatics, 2005, 21(6): 832–834
16. Samuel K, Bruno A, Lionel B, Alan B, Fiona B C, Carol C, Margaret D, Marine D, Marc F, Ursula H. The IntAct molecular interaction database in 2012. Nucleic Acids Research, 2012, 40(Database issue): 841–846
17. Wei L, Xing P, Zeng J, Chen J, Su R, Guo F. Improved prediction of protein–protein interactions using novel negative samples, features, and an ensemble classifier. Artificial Intelligence in Medicine, 2017, 83: 67–74
18. Ding Y, Tang J, Guo F. Predicting protein-protein interactions via multivariate mutual information of protein sequences. BMC Bioinformatics, 2016, 17(1): 398
19. Wang T, Li L, Huang Y, Zhang H, Ma Y, Zhou X. Prediction of protein-protein interactions from amino acid sequences based on continuous and discrete wavelet transform features. Molecules, 2018, 23(4): 823
20. Wang Y, You Z, Li L, Huang Y, Yi H. Detection of interactions between proteins by using legendre moments descriptor to extract discriminatory information embedded in PSSM. Molecules, 2017, 22(8): 1366
21. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H. Predicting protein–protein interactions based only on sequences information. Proceedings of the National Academy of Sciences of the United States of America, 2007, 104(11): 4337–4341
22. Guo Y, Yu L, Wen Z, Li M. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. Nucleic Acids Research, 2008, 36(9): 3025–3030
23. Cosic I, Hearn M T. Studies on protein-DNA interactions using the resonant recognition model: application to repressors and transforming proteins. FEBS Journal, 2010, 205(2): 613–619
24. Yang L, Xia J F, Gui J. Prediction of protein-protein interactions from protein sequence using local descriptors. Protein & Peptide Letters, 2010, 17(9): 1085–1090
25. Hu L, Chan K C C. Extracting coevolutionary features from protein sequences for predicting protein-protein interactions. IEEE/ACM Transactions Computational Biology and Bioinformatics, 2017, 14(1): 155–166
26. Wei Z S, Yang J Y, Yu D J. Predicting protein-protein interactions with weighted PSSM histogram and random forests. In: Proceedings of International Conference on Intelligent Science and Big Data Engineering. 2015, 326–335
27. Zahiri J, Yaghoubi O, Mohammad-Noori M, Ebrahimpour R, Masoudi-Nejad A. PPIevo: protein-protein interaction prediction from PSSM based evolutionary information. Genomics, 2013, 102(4): 237–242
28. Lin C Y, Chen Y C, Lo Y S, Yang J M. Inferring homologous protein-protein interactions through pair position specific scoring matrix. BMC Bioinformatics, 2013, 14(S2): S11
29. Wang Y, You Z, Li X, Chen X, Jiang T, Zhang J. PCVMZM: using the probabilistic classification vector machines model combined with a zernike moments descriptor to predict protein-protein interactions from protein sequences. International Journal of Molecular Sciences, 2017, 18(5): 1029
30. Li L P, Wang Y B, You Z H, Li Y, An J Y. PCLPred: a bioinformatics method for predicting protein-protein interactions by combining relevance vector machine model with low-rank matrix approximation. International Journal of Molecular Sciences, 2018, 19(4): 1029
31. Song X Y, Chen Z H, Sun X Y, You Z H, Li L P, Zhao Y. An ensemble classifier with random projection for predicting protein–protein interactions using sequence and evolutionary information. Applied Sciences, 2018, 8(1): 89
32. An J Y, Meng F R, You Z H, Fang Y H, Zhao Y J, Zhang M. Using the relevance vector machine model combined with local phase quantization to predict protein-protein interactions from protein sequences. BioMed Research International, 2016, 2016: 1–9
33. Cheung W, Hamarneh G. n-SIFT: n-dimensional scale invariant feature transform. IEEE Transactions on Image Processing, 2009, 18(9): 2012
34. Bay H, Ess A, Tuytelaars T, Van Gool L. Speeded-up robust features. Computer Vision & Image Understanding, 2008, 110(3): 404–417
35. Žunić J, Hirota K, Rosin P L. A Hu moment invariant as a shape circularity measure. Pattern Recognition, 2010, 43(1): 47–57
36. Khotanzad A, Hong Y H. Invariant image recognition by Zernike moments. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1990, 12(5): 489–497
37. Zhang F, Liu S Q, Wang D B, Guan W. Aircraft recognition in infrared image using wavelet moment invariants. Image & Vision Computing, 2009, 27(4): 313–318
38. Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of International Conference on Computer Vision and Pattern Recognition. 2005, 886–893
39. Whitehill J, Omlin C W. Haar features for FACS AU recognition. In: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition. 2006, 97–101
40. Ojala T, Pietikainen M, Harwood D. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In: Proceedings of the 12th International Conference on Pattern Recognition. 1994, 582–585
41. He D C, Wang L. Texture features based on texture spectrum. Pattern Recognition, 1991, 24(5): 391–399
42. Qian S, Chen D. Discrete gabor transform. IEEE Transactions on Signal Processing, 1993, 41(7): 2429–2438
43. Zeng J, Li D, Wu Y, Zou Q, Liu X. An empirical study of features fusion techniques for protein-protein interaction prediction. Current Bioinformatics, 2016, 11(1): 4–12
44. Tipping M E. Sparse Bayesian learning and the relevance vector machine. Journal of Machine Learning Research, 2001, 1(3): 211–244
45. Tipping M E. The relevance vector machine. In: Proceedings of the 12th International Conference on Neural Information Processing Systems. 2000, 652–658
46. Wei L, Yang Y, Nishikawa R M, Wernick M N, Edwards A. Relevance vector machine for automatic detection of clustered microcalci-

fications. IEEE Transactions on Medical Imaging, 2005, 24(10): 1278

47. Zhou Z. Learnware: on the future of machine learning. Frontiers of Computer Science, 2016, 10(4): 589–590
48. Rong W, Peng B, Ouyang Y, Li C, Xiong Z. Structural information aware deep semi-supervised recurrent neural network for sentiment analysis. Frontiers of Computer Science, 2015, 9(2): 171–184
49. Mikolov T, Karafiat M, Burget L, Cernocký J, Khudanpur S. Recurrent neural network based language model. In: Proceedings of the 11th Annual Conference of the International Speech Communication Association. 2010, 1045–1048
50. Gregor K, Danihelka I, Graves A, Rezende D J, Wierstra D. DRAW: a recurrent neural network for image generation. In: Proceedings of International Conference of Machine Learning. 2015, 1462–1471
51. Sainath T N, Vinyals O, Senior A, Sak H. Convolutional, long short-term memory, fully connected deep neural networks. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. 2015, 4580–4584
52. Dyer C, Ballesteros M, Ling W, Matthews A, Smith N A. Transition-based dependency parsing with stack long short-term memory. Computer Science, 2015, 37(2): 321–332
53. Sak H, Senior A, Beaufays F. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. In: Proceedings of the 15 Annual Conference of the International Speech Communication Association. 2014
54. Li Z, Wang Y, Zhi T, Chen T. A survey of neural network accelerators. Frontiers of Computer Science, 2017, 11(5): 746–761
55. Lazib L, Qin B, Zhao Y, Zhang W, Liu T. A syntactic path-based hybrid neural network for negation scope detection. Frontiers of Computer Science, 2020, 14(1): 84–94
56. Sprinzak E, Margalit H. Correlated sequence-signatures as markers of protein-protein interaction. Journal of Molecular Biology, 2001, 311(4): 681–692
57. Bock J R, Gough D A. Predicting protein–protein interactions from primary structure. Bioinformatics, 2001, 17(5): 455–460
58. Martin S, Roe D, Faulon J L. Predicting protein–protein interactions using signature products. Bioinformatics, 2004, 21(2): 218–226
59. Benhur A, Noble W S. Kernel methods for predicting protein–protein interactions. Intelligent Systems in Molecular Biology, 2005, 21(1): 38–46
60. Chou K, Cai Y. Predicting protein-protein interactions from sequences in a hybridization space. Journal of Proteome Research, 2006, 5(2): 316–322
61. Wang Y, You Z, Li L, Cheng L, Zhou X, Zhang L, Li X, Jiang T. Predicting protein interactions using a deep learning method-stacked sparse autoencoder combined with a probabilistic classification vector machine. Complexity, 2018, 2018: 1–12
62. Sun T, Zhou B, Lai L, Pei J. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC Bioinformatics, 2017, 18(1): 277
63. Almagro Armenteros J J, Sønderby C K, Sønderby S K, Nielsen H, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics, 2017, 33(21): 3387–3395
64. Yi H C, You Z H, Huang D S, Li X, Jiang T H, Li L P. A deep learning framework for robust and accurate prediction of ncRNA-protein interactions using evolutionary information. Molecular Therapy Nucleic Acids, 2018, 11: 337–344
65. Wang Y B, You Z H, Li X, Jiang T H, Chen X, Zhou X, Wang L. Predicting protein-protein interactions from protein sequences by a stacked sparse autoencoder deep neural network. Molecular Biosystems, 2017, 13(7): 1336–1344
66. Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016, 1715–1725
67. Kudo T. Subword regularization: improving neural network translation models with multiple subword candidates. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018, 66–75
68. Kudo T, Richardson J. SentencePiece: a simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2018, 66–71
69. Rebentrost P, Mohseni M, Lloyd S. Quantum support vector machine for big data classification. Physical Review Letters, 2013, 113(13): 130503
70. Crawford D, Levit A, Ghadermarzy N, Oberoi J S, Ronagh P. Reinforcement learning using quantum boltzmann machines. 2016, arXiv preprint arXiv:1612.05695
71. Qiu D, Li L. An overview of quantum computation models: quantum automata. Frontiers of Computer Science, 2008, 2(2): 193–207



Yanbin Wang received his BE degree in Computer Science and Technology from Zhengzhou University, China in 2015. He obtained his MS degree in Computer Science from University of Chinese Academy of Sciences (UCAS), China in 2018. He is currently a research assistant with the Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, China. His current research interests include deep neural networks, big data, signal processing, and its applications in bioinformatics.

Zhuhong You received his BE degree in Electronic Information Science and Engineering from Hunan Normal University, China in 2005. He obtained his PhD degree in control science and engineering from University of Science & Technology of China (USTC), China in 2010. From June 2008 to November 2009, he was a visiting research fellow at the Center of Biotechnology and Information, Cornell University, USA. He is currently a professor with the Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, China. His current research interests include neural networks, big data, intelligent information processing, sparse representation, and their applications in bioinformatics.

Liping Li received his BE degree in architectural engineering from Gansu Agricultural University, China in 2006. She received his Master degree in School of Computer Science from Shenzhen University, China in 2016. She is currently an association professor with the Xijing University, China. Her current research interests include data mining algorithms, neural networks, pattern recognition, and its applications in bioinformatics.

Zhanheng Chen is currently pursuing the PhD degree with the University of Chinese Academy of Sciences, China. His current research interests include data mining, natural language processing, and pattern identification. He has several publications in journals (Published in Molecular Therapy-Nucleic Acids, BMC Genomics, BMC Systems Biology, Frontiers in Genetics, International Journal of Molecular Sciences, and so on), and international conferences (such as RECOMB, ISBRA, ICIC, ICIBM, and so on).