Pamukkale Univ Muh Bilim Derg, 30(3), 333-342, 2024

Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi

Pamukkale University Journal of Engineering Sciences

Protein complex detection from protein-protein interaction networks with machine learning methods

Protein-Protein etkileşim ağlarından makine öğrenmesi yöntemleriyle protein kompleksi tespiti

Yasin KARAKUŞ[1 *], *Volkan ALTUNTAŞ*[2]

[1]Department of Computer Engineering, Graduate School, Bursa Technical University, Bursa, Turkey.
22435004005@ogrenci.btu.edu.tr

[2]Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Bursa Technical University, Bursa, Turkey.
volkan.altuntas@btu.edu.tr

Received/Geliş Tarihi: 08.02.2023 Revision/Düzeltme Tarihi: 25.07.2023 doi: 10.5505/pajes.2023.56887
Accepted/Kabul Tarihi: 28.07.2023 Research Article/Araştırma Makalesi

Abstract

Understanding Protein - Protein interaction networks, which show the interactions between proteins involved in tasks that are very important for our organisms such as structural support, storage, signal transduction and defence, provides a better understanding of cellular processes. One of the important studies carried out for this purpose is to try to detect protein complexes from protein - protein interaction networks. Supervised and unsupervised machine learning methods were used to detect protein complexes. It is known that the machine learning methods used produce better performance when more than one method is used together. Based on this knowledge, a method that detects protein complexes from protein-protein interaction networks is proposed in this study. The method first weights protein-protein interaction networks using biological and topological properties of proteins. Then it estimates local and global protein complex core. Then it builds a protein complex detection model using the structural modularity of proteins and the voting regression model. We predict that XGB regression, gaussian process regression, catboost regression and histogram-based gradient boosting regression supervised learning methods can achieve more successful results when used together in the voting regression model. When we compare the success of the model with other models, it has shown the best performance many times among the compared models.

Keywords: Protein-Protein interaction networks, Protein complex detection, Machine learning, Voting regression, Bioinformatics, Network embedding.

Öz

Yapısal destek, depolama, sinyal iletimi, savunma gibi organizmalarımız için çok önemli olan görevlerde yer alan proteinlerin birbirleriyle olan ilişkilerinin gösterildiği Protein-Protein etkileşim ağlarını anlayabilmek hücresel süreçleri daha iyi anlayabilmeyi sağlamaktadır. Bu amaçla yapılan önemli çalışmalardan birisi protein-protein etkileşim ağlarından protein komplekslerini tespit etmeye çalışmaktır. Protein komplekslerini tespit etmek için denetimli ve denetimsiz makine öğrenmesi yöntemleri kullanılmıştır. Kullanılan makine öğrenmesi yöntemlerinin birden fazla yöntem bir arada kullanıldığında daha iyi performans ürettiği bilinmektedir. Buna benzer bilgilere dayanarak bu çalışmada protein-protein etkileşim ağlarından protein komplekslerini tespit eden bir yöntem önerilmiştir. Yöntem, ilk olarak protein-protein etkileşim ağlarını proteinlerin biyolojik ve topolojik özelliklerini kullanarak ağırlıklandırır. Ardından yerel ve global protein kompleksi çekirdeklerini tahmin eder. Sonra proteinlerin yapısal modülerliğini ve oylama regresyon modelini kullanarak protein kompleksi tespit eden model oluşturur. XGB regresyonu, gauss süreci regresyonu, catboost regresyonu ve histogram tabanlı gradyan artırma regresyonu denetimli öğrenme yöntemlerinin oylamalı regresyon modelinde birlikte kullanıldığında daha başarılı sonuçlar elde edebileceğini öngörüyoruz. Modelin başarısını diğer modellerle kıyasladığımızda kıyaslanan modeller arasında birçok kez en iyi performansı göstermiştir.

Anahtar kelimeler: Protein-Protein etkileşim ağları, Protein kompleksi tespiti, Makine öğrenmesi, Oylama regresyon, Biyoenformatik, Ağ gömme.

1 Introduction

Proteins are a complex of molecules consisting of one or more chains of amino acids. Proteins, which make up more than half of the dry weight of many cells, take part in almost every metabolic work that takes place in organisms, such as structural support, storage, transport, signal transmission, and defense. While proteins can rarely perform these tasks alone, they are usually performed by protein complexes, which we can call biological machines composed of more than one protein. Networks used to represent the physical relationship of proteins to each other are called Protein-Protein Interaction (PPI) networks. In the last few years, it has become popular to conduct community identity studies in complex networks such as PPI networks. The crucial issue in bioinformatics is investigating protein complexes in PPI networks. Researching protein complexes helps to better understand cellular systems [1]. It is also useful for predicting protein functions, disease genes and drug-disease associations [2]. In PPI networks, it can be said that densely connected subgraphs are potential protein complexes or functional modules. Computational methods are preferred because experimental methods take a lot of time and take up a lot of memory space in the calculation of protein complexes. PPI networks can be modeled as graphs. When PPI networks are modeled as graphs, proteins correspond to nodes and interprotein interactions to edges. More detailed information about the current study will be given in the related work section.

*Corresponding author/Yazışılan Yazar

1.1 Related work

In the last 15 years, various computational methods have been presented for the detection of protein complexes from protein-protein interaction (PPI) networks. The methods presented can be considered as methods based on unsupervised learning, model optimization, and supervised learning.

It has been seen that the local heuristic search strategy is used among many developed methods. Altaf-Ul-Amin et al. [3] identified the protein with the highest number of connections in the PPI network, that is, the highest density, and considered this protein as the core of the protein complex. Then, they expanded the core of this protein complex by using local heuristic search until the protein complex could no longer expand. Thus, the protein complex is obtained. Once the resulting protein complex is recorded, it is removed from the network, facilitating the detection of other protein complexes. Wang et al. [4] started by obtaining a weighted dynamic and static PPI network, using topological and biological information to detect protein complexes. They then predict the protein complex cores and expand the predicted core proteins by greedy local heuristic search. Dilmaghani et al. [5] proposed a local community detection algorithm (lcda-go) that uniquely utilises functionality information from gene ontology along with network topology for protein complex identification. Their algorithm identifies the community of each protein based solely on topological and functional information obtained from local neighbouring proteins within the ppi network. Yu and Kong [6] proposed a novel solution by combining node resource allocation and gene expression information into a weighted protein network (NRAGE-WPN) in which protein complexes are detected from PPI networks based on coreference and second-order neighbours.

Although it is quite easy to perform a local search for the methods developed using the local heuristic search strategy, it is very difficult to produce successful global solutions.

Methods using the global heuristic search strategy have been developed to detect protein complexes from PPI networks. Cho et al. [7] sent the information from an informative node to all other nodes in the PPI network that it could reach from every possible edge, and the information from which informative node the information came is kept with the information. Finally, proteins with strong connections are greedily combined into protein complexes. Omranian et al. [8] modeled PPI networks as bipartite diffuse subgraphs containing both sparse and dense subgraphs to detect protein complexes. Based on the fundamental properties of bipartite subgraphs, they designed a parameterless greedy approximation algorithm called Protein Complexes from Coherent Partition (PC2P). Wang et al. [9] started protein complex detection by weighting the PPI network. They then used gene expression and subcellular localisation data to identify local protein complex core. They then used Markov clustering algorithm to identify global protein complex core. Then, they defined a fitness function using multiple topological features to identify protein complexes. Then, they developed a novel protein complex formation strategy to expand the global and local protein complex kernels to form protein complexes. Finally, they used GO annotation data to filter candidate protein complexes and improve the accuracy of protein complex detection. The methods developed using the global heuristic search strategy are very successful in global searches. However, they are disadvantaged in terms of running times and memory space usage.

The fact that PPI networks contain false interactions is a problem that needs to be resolved in detecting protein complexes from PPI networks. Various studies have been carried out to overcome this problem. Xu et al. [10], Using functional annotations of proteins, Wang et al. [11] tried to overcome this problem by using topological and biological properties.

Some studies have suggested that the detection of protein complexes can be an optimization problem that can be solved using network topology and protein features. Zhang et al. [12] constructed the PPI network according to the Poisson distribution and smoothed the estimators of the trend parameters using the Laplace modifier.

In supervised learning methods, features are extracted from topological and biological features and the model is trained with these features. This trained model is used for protein complex prediction.

Several of the recent studies have proposed methods with supervised learning-based classification or regression to find protein complexes from PPI networks. Xu et al. [13] combined PPI information from six different sources and produced protein complex predictions from this PPI network using the SVM model.

In some of the recent studies, graph neural networks (GNN) and graph convolutional networks (GCN) methods have been used. As a result of these studies, graph-based networks have achieved acceptable success in detecting protein complexes from PPI networks. Zaki et al. [14] proposed the graph convolutional network approach to increase the success of detecting protein complexes from PPI networks.

1.2 Observations and contributions

The methods developed for protein complex detection from PPI networks are based on supervised learning or unsupervised learning approaches. The biggest advantage of unsupervised learning methods over supervised learning methods is that they do not need labeled training data. Although this has been a major advantage, unsupervised learning methods do not use labeled training data, resulting in fewer topological protein complexes to detect, which is a disadvantage. Although supervised learning methods are more successful than unsupervised learning methods in detecting protein complexes in PPI networks containing different topological information, they are generally weak in producing successful results on data outside the training set because they are trained with a single training model.

In this study, in order to overcome these difficulties, firstly, a weighted PPI network was constructed by using the similarity information of proteins in terms of common expression, function, intracellular localisation and topological structure. Secondly, the core of local protein complexes and the core of global protein complexes were determined using CPredictor2.0 [10]. Thirdly, a voting regression model is constructed, which includes various regression models that play an important role in the detection of protein complexes with different topological structures.

In this study, supervised learning, which has proven to be more successful in the literature, and voting regression model, which is based on the performance of multiple predictors so that

unsuccessful results in one predictor can be balanced with successful results in another, are used. In addition to the studies in the literature, we predict that the use of XGB regression, gaussian process regression, catboost regression and histogram-based gradient boosting regression models within the voting regression model may be more successful and experimental results show that our study gives more consistent and more successful results than most studies in the literature.

2 Materials and method

The flow diagram of the proposed method for protein complex detection is given in Figure 1.

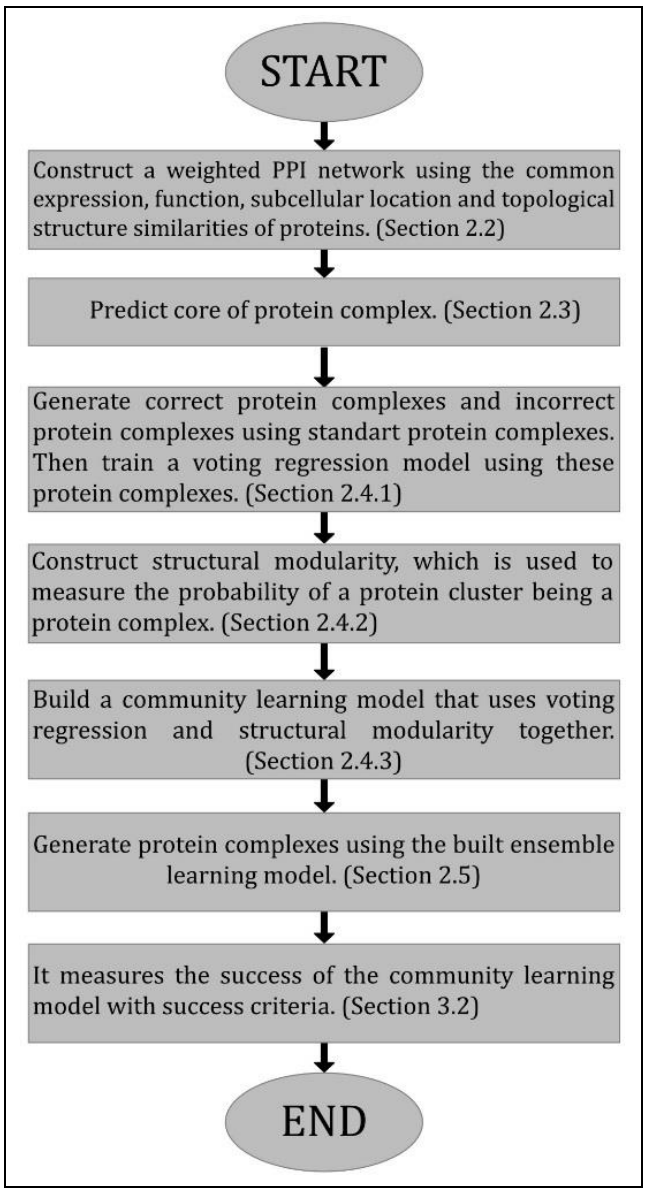

Figure 1. Flow chart of the developed method for protein complex detection from PPI networks.

2.1 Datasets

Although proteins are rarely alone to perform a task in metabolic processes, they usually perform tasks with protein complexes formed by more than one protein. PPI datasets are datasets that contain information about the physical relationships of proteins with each other. They can be used for protein complex detection. The standard protein complexes dataset is a dataset of protein complexes of known accuracy. The standard protein complex dataset can be used for purposes such as measuring the accuracy of predicted protein complexes.

In our study, four different PPI datasets were used, in which the repetitive interactions between two different proteins and the interactions of the proteins with themselves were inferred. These are DIP [15], Gavin [16], Krogan Core [17], and MIPS [18]. Detailed information about the datasets is given in Table 1.

Table 1. Detailed information about PPI datasets.

Datasets	Number of Nodes	Number of Edges
DIP	4930	17201
Gavin	1855	7669
Krogan Core	2674	7075
MIPS	4553	12318

In our study, two different standard protein complex datasets were used. The first standard protein complex dataset consists of protein complexes from MIPS [18], SGD [19], TAP06 [16], ALOY [20], CYC 2008 [21], and NEWMIPS [22]. The second standard protein complex dataset consists of Wodak [21], PINdb, and GO [23] protein complexes. Detailed information about protein complex datasets is given in Table 2.

Table 2. Detailed information about standard protein complex datasets.

Datasets	Number of Protein Complex	Protein Coverage	Average Size
Standart Protein Complex 1	812	2773	8.92
Standart Protein Complex 2	1045	2778	8.97

2.2 Weighting the PPI network

When the edges of the graphs are weighted, success in detecting protein complexes from PPI networks increases. Similarly, using more than one PPI network together increases the success results [4].

Protein complexes are made up of proteins and interactions between proteins. This indicates that proteins may have similar functions and positions. For this reason, it is beneficial to consider a large number of topological and biological data in weighting the PPI network. Considering this information, topological structure similarities, common expression similarities, functional similarities, and subcellular location similarities of proteins were used together to weight the PPI network.

2.2.1 Protein co-expression similarity

A protein complex consists of proteins interacting with each other at the same time and place [24]. In other words, proteins in a protein complex are co-localized, co-expressed, and functionally similar in biology [25]. Using this similarity increases the success rate of protein complex detection. The Pearson correlation coefficient (PCC) was used to measure the co-expression of two interacting proteins. PCC is calculated according to the formula in Equation 1, where given protein sequences $X = \{X_1, \ldots, X_i, \ldots, X_n\}$, $Y = \{Y_1, \ldots, Y_i, \ldots, Y_n\}$ and $\bar{X}$ is the average gene expression of X proteins at a given time point and $\bar{Y}$ is the average gene expression of Y proteins at a given time point.

$$PCC(X,Y) = \frac{\sum_{i=1}^{m}(x_i - \bar{X}) \times (y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{X})^2} \times \sqrt{\sum_{i=1}^{m}(y_i - \bar{Y})^2}} \quad (1)$$

2.2.2 Protein function similarity

The more common GO-slim descriptions a pair of proteins have, the more likely they are to have the same function. This means that the edges between them, namely the interactions, are stronger. Protein function similarity (FS) is calculated according to Equation 2, where $|FS(X)|$ is the number of GO-slim annotations of X proteins, $|FS(Y)|$ is the number of GO-slim annotations of Y proteins, and $|FS(X) \cap FS(Y)|$ is the number of GO-slim annotations common to X and Y proteins.

$$FS(X,Y) = \begin{cases} \dfrac{|FS(X) \cap FS(Y)|}{min\{|FS(X)|, |FS(Y)|\}}, & min\{|FS(X)|, |FS(Y)|\} \geq 1 \\ 0, & other \end{cases} \quad (2)$$

2.2.3 Protein subcellular location similarity

Subcellular locations determine the environments in which proteins operate. Therefore, it influences protein function by controlling the access and availability of all types of molecular interaction partners. Protein subcellular location similarity is calculated according to Equation 3, where $|SL(X)|$ is the number of subcellular localizations of X proteins, $|SL(X)|$ is the number of subcellular localizations of Y proteins, and $|SL(X) \cap SL(Y)|$ is the number of subcellular localizations common to X and Y proteins.

$$SL(X,Y) = \frac{2 \times |SL(X) \cap SL(Y)|}{|SL(X)| + |SL(Y)|} \quad (3)$$

2.2.4 Protein topological structure similarity

Network embedding is the process of vector representation of nodes corresponding to proteins in PPI networks in a multidimensional space without disturbing the network structure. Node2Vec [26] was used as the network embedding method in the study.

The calculating similarity between vector representations of two different proteins will give us topological structure similarity. Cosine similarity is used to calculate vector similarity in the study. Protein topological structure similarity is calculated according to Equation 4, where $F(X) = (x_1,..., x_i,..., x_n)$ is the n-dimensional vector of protein X returned by the Node2Vec algorithm and $F(Y) = (y_1, \ldots, y_i, \ldots, y_n)$ is the n-dimensional vector of protein Y returned by the Node2Vec algorithm.

$$TSS(X,Y) = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n} {x_i}^2} \times \sqrt{\sum_{i=1}^{n} {y_i}^2}} \quad (4)$$

The weight of an edge is calculated as in Equation 5 by averaging the similarity measures.

$$W(X,Y) = \frac{PCC(X,Y) + FS(X,Y) + SL(X,Y) + TSS(X,Y)}{4} \quad (5)$$

2.3 Prediction of protein complex core

Since we weight the edges of PPI networks using biological and topological features, a high edge weight means that the two terminal proteins are more likely to be in the same protein complex. Also, protein complex cores often correspond to dense subgraphs in PPI networks [27],[28].

To find the local protein complex core, the edges are first sorted by weight. The protein with the highest weighted edge is then expanded to become the core of the protein complex according to the proximity with its neighbouring proteins. The core of the protein complex formed is preserved if it contains two or more proteins. Meanwhile, the core edge cannot be used as the core edge of another cluster. Process of prediction a distinct protein complex core continues with the selection of the next highest weighted edge.

CPredictor2.0 [10] was also used to detect global protein complex cores. CPredictor2.0 first groups proteins with similar functions. Then, for each group, it discovers clusters within the group using the Markov Clustering Algorithm and merges the discovered clusters according to a certain concordance rate. Finally, the local protein complex cores found by us and the global protein complex cores found by CPredictor2.0 were merged.

2.4 Ensemble learning model

2.4.1 Voting regression model

Several steps were followed to train the regression model. Firstly, the protein complexes, whose accuracy has been proven by experiments, were collected from the standard protein complex dataset and the PPI network was weighted using the biological and topological similarities between the proteins. Secondly, the protein complexes collected from the standard protein complex dataset were matched with weighted and unweighted PPI networks. Thirdly, false protein complexes were produced in weighted and unweighted PPI networks. Then, topological feature extraction was performed for the protein complexes that were collected from the standard protein complexes dataset, that is, proven by experiments, and the false protein complexes produced. Fourthly, from the extracted topological features, the most important features that most influence protein complex prediction were selected. Fifthly, XGB regression, gaussian process regression, catboost regression and histogram-based gradient boosting regression, which we individually believe to produce the most successful results, were included in the voting regression model. The regression models were then trained.

The purpose of generating false protein complexes and extracting topological features is that the model needs topological properties to compare true and false protein complexes during the training phase to predict protein complexes correctly.

Since standard protein complexes 1 and 2 databases are determined experimentally in the laboratory, it is very important to obtain protein complexes with known accuracy from these two databases. They are used as experimentally proven protein complexes to train a model.

Because the voting regression model relies on the performance of more than one estimator, unsuccessful results in one estimator can be balanced with successful results in other estimators. For this reason, the voting regression model was chosen. In the voting regression model, XGB regression, Gaussian processes regression, Catboost regression, and Histogram-based gradient boosting regression models were used. Voting regression model estimators used with default parameters.

2.4.1.1 XGB regression model

XGB, which has features such as result generation speed, parallelization, and performance, is a gradient boost-based supervised machine learning model that is frequently used for regression prediction modeling.

2.4.1.2 Gauss process regression model

The Gaussian processes model is a probabilistic supervised machine learning framework widely used for regression and classification tasks. A Gaussian process regression (GPR) model can make predictions that contain preliminary information (kernels) and provide measures of uncertainty on the predictions [29].

2.4.1.3 Catboost regression model

Catboost is an open-source machine learning method based on gradient boost theory and decision trees, developed by Yandex company in 2017.

The main idea of the catboost regression model is to sequentially combine models that perform slightly better than randomness, thereby creating a greedy, robust competitive prediction model through search.

2.4.1.4 Histogram-based gradient boosting regression model

Gradient boosting algorithms use decision trees. They are quite popular for classification and regression. The main problem with gradient boosting algorithms is that model training takes quite a long time, especially on datasets containing tens of thousands of rows of data. The histogram-based gradient boosting model was developed to train models faster than gradient boosting approaches.

2.4.2 Structural modularity of protein complexes

$C = (V_c, E_c, W_c)$ is the set of proteins in a PPI network, where V_c is the vertices corresponding to the proteins belonging to set c, E_c is the edges corresponding to the interaction between proteins belonging to set c, and W_c is the weights of the edges belonging to set c. Structural modularity[30] is an effective quantitative measurement method used to estimate the probability that cluster C is a protein complex. Structural modularity is achieved by using the Cohesion and Coupling equations. Cohesion is calculated according to equation 6, where W_{in} is the total weight of edges that belong to cluster C and do not go outside the cluster, and $|C|$ is the number of nodes in cluster C. Coupling is calculated according to equation 7, where W_{out} is the weight of boundary edges connecting cluster C to other clusters and proteins. Structural modularity is calculated according to equation 8.

$$Cohesion(C) = \frac{2 \times W_{in}(C)}{sqrt(|C|) \times (|C| - 1)} \quad (6)$$

$$Coupling(C) = \frac{W_{out}(C)}{|C|} \quad (7)$$

$$SM(C) = \frac{Cohesion(C)}{Cohesion(C) + Coupling(C)} \quad (8)$$

2.4.3 Ensemble learning model building

At this stage, an ensemble learning model was constructed using the voting regression model and structural modularity to measure the probability of a $C = (V_c, E_c, W_c)$ cluster being a protein complex. The Ensemble learning model for a C set is as in Equation 9, where $VR(C)$ is the voting regression score, a measure of the probability that cluster C is a true protein complex, and $SM(C)$ is the structural modularity score of cluster C.

$$Fitness(C) = VR(C) \times SM(C) \quad (9)$$

2.5 Formation of protein complexes

Based on the knowledge that the protein, that is the core of the protein complex can pass through the edges and form a protein complex with junction proteins that are not included in the core of the protein complex, some cores of protein complexes were obtained. A set of outer boundary proteins based on interprotein neighbourhood was then constructed for each protein complex core. The outer boundary proteins are proteins that are not present in the protein complex core. When a new protein is added to the cluster, the score of the cluster is calculated according to equation 9. The proteins for which the cluster achieves the best score are provisionally recognised as protein complexes. Then, for the provisionally recognised protein complex, the proteins in the core of the protein complex are sequentially resmede from the provisional protein complex, and after each modification the score is calculated according to equation 9. The transient protein complex with the best score is recognised as the permanent protein complex. Finally, the protein complex is deleted if it has less than three proteins.

3 Results

The work was implemented in the PyCharm 2022.2.3 IDE using python 3.9. Successfully executed on a computer with Intel i7-6700HQ 2.60 GHz CPU and 16 GB RAM.

3.1 Parameter selection

We examined how effective the parameter ratio was on our study by increasing the parameter ratio by 5 from 1 to 20. According to the results we obtained, parameter ratio=5 obtained the best total score for standard protein complex 1 and standard protein complex 2. In addition, as we can see from the results obtained, the parameter ratio does not affect the total score much and the total score tends to be constant.

Figure 2 shows how much the change in parameter ratio in standard protein complex 1 changes the total score. Figure 3 shows how much the change in parameter ratio in standard protein complex 2 changes the total score.

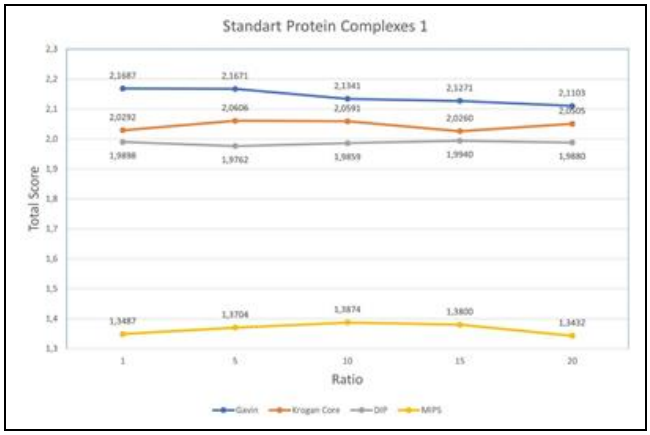

Figure 2. Total score change in standard protein complex 1 depending on the parameter ratio.

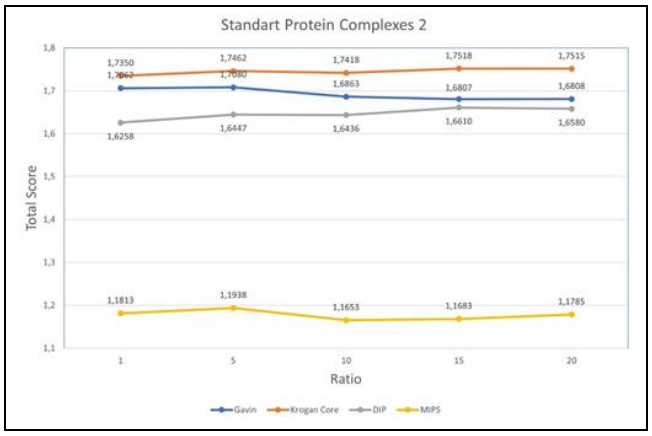

Figure 3. Total score change in standard protein complex 2 depending on the parameter ratio.

3.2 Performance evaluation

In the study, accuracy, f-score, highest matching rate, coverage ratio, and Jaccard were used to measure the success of the model developed to predict protein complexes from PPI networks and to compare with other models. S represents protein complexes obtained by experiments, that is, known to be accurate, and D represents protein complexes detected by the model developed to detect protein complexes.

3.2.1 Accuracy

T_{ij}, is the number of proteins included in both the standard protein complex S_i and the protein complex D_j detected by the model, and N_i is the number of proteins included in the standard protein complex S. To calculate the accuracy (ACC) value, the Sensivity (S_n) and Positive Predictive Value (PPV) values must be calculated. Equation 10 is used to calculate the S_n value, Equation 11 is used to calculate the PPV value, and Equation 12 is used to calculate the accuracy value.

$$S_n = \frac{\sum_{i=1}^{|S|} max_{j=1}^{|D|}\{T_{ij}\}}{\sum_{i=1}^{|S|} N_i} \tag{10}$$

$$PPV = \frac{\sum_{j=1}^{|D|} max_{i=1}^{|S|}\{T_{ij}\}}{\sum_{j=1}^{|D|}\sum_{i=1}^{|S|} T_{ij}} \tag{11}$$

$$ACC = \sqrt{S_n \times PPV} \tag{12}$$

3.2.2 F-score

N_{im}, is the number of protein complexes detected by the model that match at least one standard protein complex, N_{sm}, is the number of standard protein complexes detected by the model that match at least one protein complex. To calculate the F-score, recall (R) and precision (P) must be calculated first. Equation 13 is used for precision calculation, Equation 14 is used for precision calculation, and Equation 15 is used to calculate f-score.

$$R = \frac{N_{sm}}{|S|} \tag{13}$$

$$P = \frac{N_{im}}{|D|} \tag{14}$$

$$F = \frac{2 \times R \times P}{R + P} \tag{15}$$

3.2.3 Coverage rate

The coverage rate (CR) is used to measure how many proteins in standard protein complexes can be covered by the predicted complexes [31]. T matrix $|S| \times |D|$ generated from the process. T_{ij}, is the number of proteins shared between the i'th experimentally proven protein complex and the j'th protein complex predicted by model. The coverage rate is calculated by Equation 16.

$$CR = \frac{\sum_{i=1}^{|S|} max\{T_{ij}\}}{\sum_{i=1}^{|S|} N_i} \tag{16}$$

3.2.4 Maximum matching rate

For the highest matching rate, a binary graph is first created between S and D. The highest matching rate (MMR) is calculated by Equation 17.

$$MMR = \frac{\sum_{i=1}^{|S|} max_j NA(S_i, D_j)}{|S|} \tag{17}$$

3.2.5 Jaccard

A standard protein complex is $S_i \in S$ and a protein complex $D_j \in D$ detected by the model is calculated by Jaccard Equation 23. To calculate the Jaccard value, Equation 18, Equation 19, Equation 20, Equation 21, and Equation 22 must be calculated.

$$Jac(S_i, D_j) = \frac{|S_i \cap D_j|}{|S_i \cup D_j|} \tag{18}$$

$$Jac(S_i) = max_{D_j \in D} Jac(S_i, D_j) \tag{19}$$

$$Jac(D_j) = max_{S_i \in S} Jac(D_i, S_i) \tag{20}$$

$$JaccardS = \frac{\sum_{S_i \in S} |s_i| Jac(S_i)}{\sum_{S_i \in S} |s_i|} \tag{21}$$

$$JaccardD = \frac{\sum_{D_j \in D} |D_j| Jac(D_j)}{\sum_{D_j \in D} |D_j|} \tag{22}$$

$$Jaccard = \frac{2 \times JaccardD \times JaccardS}{JaccardD + JaccardS} \tag{23}$$

3.3 Comparison of the model with other models

The model developed in the study was tested in four different PPI networks, Gavin, Krogan Core, DIP and MIPS, using two different standard protein complexes, which are factual to help train and evaluate the model. Five different criteria mentioned in the title of success criteria were used to evaluate the performance of the model. The sum of these five different criteria is called the total score, and a different criterion has been produced to facilitate comparison. To compare the performance of the model, seven different unsupervised learning models including DPClus [3], ClusterONE [32], PEWCC [33], WPNCA [31], CPredictor2.0 [10], Zhang [34], PC2P [8] and ClusterEPs [35], ClusterSS [36] and ELF-DPC [30], three different supervised learning models were used.

The comparison of the model developed for standard protein complex 1 with other models is in Table 3, and the comparison of the model developed for standard protein complex 2 with other models is in Table 4. In Table 3 and Table 4, the highest values are written in bold numbers.

Table 2. Comparison of the model we created with other models for standard protein complex 1.

Model	Number	ACC	F-score	CR	MMR	Jaccard	Total Score
Gavin							
MCL	220	**0.3657**	0.5358	0.4891	0.1494	0.3610	1.9010
DPClus	285	0.3466	0.5972	0.4382	0.1736	0.4025	1.9581
ClusterONE	258	0.3458	0.5976	0.4514	0.1921	0.3974	1.9844
PEWCC	**664**	0.3146	0.6576	0.4316	**0.3538**	0.3969	2.1546
WPNCA	484	0.3114	0.6428	**0.4949**	0.2557	0.3554	2.0602
CPredictor2.0	266	0.3062	0.6286	0.3750	0.2144	0.4124	1.9365
Zhang	438	0.3156	0.6475	0.3976	0.3182	0.4084	2.0872
PC2P	219	0.3551	0.5769	0.4439	0.1825	0.3922	1.9505
ClusterEPs	271	0.2841	0.6014	0.3656	0.2166	0.4090	1.8766
ClusterSS	482	0.3218	0.5600	0.3941	0.2535	0.3685	1.8979
ELF-DPC	286	0.3391	0.6674	0.4792	0.2516	0,4330	**2.1702**
Our Model	327	0.3308	**0.6757**	0.4666	0.2422	**0.4519**	2.1671
Krogan Core							
MCL	370	**0.3192**	0.4004	0.3895	0.1361	0.2902	1.5354
DPClus	497	0.3071	0.4138	0.3672	0.1745	0.3235	1.5861
ClusterONE	240	0.2829	0.4694	0.3085	0.1523	0.3324	1.5454
PEWCC	383	0.2309	0.5289	0.3231	0.1471	0.3786	1.6085
WPNCA	369	0.2758	0.5446	0.3897	0.1912	0.3415	1.7428
CPredictor2.0	236	0.2725	0.5895	0.3037	0.1954	0.3688	1.7298
Zhang	326	0.2549	0.5563	0.2884	0.2182	0.3408	1.6585
PC2P	249	0.2970	0.4356	0.3458	0.1337	0.3190	1.5310
ClusterEPs	410	0.2621	0.5836	0.3352	0.2209	0.3448	1.7467
ClusterSS	**722**	0.3072	0.4377	0.3758	0.2402	0.3357	1.6966
ELF-DPC	304	0.2984	0.6287	**0.4239**	**0.2687**	0.4302	2.0499
Our Model	325	0.2939	**0.6397**	0.4186	0.2533	**0.4550**	**2.0606**
DIP							
MCL	628	0.2684	0.3106	0.3578	0.0932	0.2155	1.2455
DPClus	909	0.2720	0.3085	0.3792	0.1237	0.2645	1.3480
ClusterONE	904	**0.3270**	0.5118	**0.5062**	0.1752	0.3297	1.8499
PEWCC	648	0.2262	0.6004	0.3783	0.1573	0.3514	1.7136
WPNCA	623	0.2594	0.5888	0.4307	0.2070	0.3360	1.8219
CPredictor2.0	293	0.2287	0.5008	0.2302	0.1110	0.2825	1.3533
Zhang	502	0.2426	0.5622	0.3257	0.1811	0.3223	1.6339
PC2P	441	0.2542	0.3419	0.3401	0.0854	0.2324	1.2540
ClusterEPs	804	0.2147	0.5730	0.2954	0.2154	0.3087	1.6073
ClusterSS	**2375**	0.2577	0.3230	0.3335	**0.2331**	0.2573	1.4047
ELF-DPC	564	0.2768	0.6200	0.4922	0.2273	0.3454	1.9617
Our Model	570	0.2820	**0.6235**	0.4985	0.1968	**0.3754**	**1.9762**
MIPS							
MCL	594	0.1577	0.0681	0.1686	0.0214	0.1064	0.5221
DPClus	207	0.2133	0.3784	0.2031	0.0820	0.2264	1.1031
ClusterONE	690	**0.2489**	0.2925	0.2719	0.0989	0.2044	1.1167
PEWCC	382	0.1389	0.2802	0.1900	0.0566	0.1679	0.8335
WPNCA	527	0.1824	0.3301	0.2603	0.1017	0.1798	1.0543
CPredictor2.0	265	0.2288	0.4344	0.2212	0.1140	0.2545	1.2529
Zhang	406	0.2025	0.3702	0.2051	0.1077	0.2176	1.1031
PC2P	374	0.2137	0.2347	0.2371	0.0652	0.1662	0.9170
ClusterEPs	645	0.1943	0.4610	0.2426	0.1580	0.2543	1.3102
ClusterSS	**1266**	0.2320	0.2309	0.2400	0.1242	0.1942	1.0213
ELF-DPC	483	0.2237	**0.4811**	0.2914	**0.1678**	**0.2599**	**1.4239**
Our Model	458	0.2208	0.4725	**0.2940**	0.1352	0.2525	1.3750

Table 4. Comparison of the model we created with other models for standard protein complex 2.

Model	Number	ACC	F-score	CR	MMR	Jaccard	Total Score
			Gavin				
MCL	220	**0.3587**	0.3756	0.4091	0.1153	0.3126	1.5713
DPClus	285	0.3293	0.3854	0.3483	0.1405	0.3147	1.5182
ClusterONE	258	0.3359	0.4090	0.3633	0.1419	0.3200	1.5703
PEWCC	**664**	0.3137	0.4185	0.3483	**0.2152**	0.2999	1.5955
WPNCA	484	0.3305	0.4217	**0.4116**	0.1670	0.2962	1.6270
CPredictor2.0	266	0.2816	0.4820	0.3076	0.1564	0.3309	1.5584
Zhang	438	0.2942	0.4365	0.3209	0.2057	0.3186	1.5758
PC2P	219	0.3413	0.4025	0.3610	0.1295	0.3204	1.5547
ClusterEPs	271	0.2715	0.4331	0.2906	0.1670	0.3173	1.4795
ClusterSS	487	0.3170	0.3729	0.3279	0.1716	0.2924	1.4819
ELF-DPC	265	0.3259	0.4546	0.3838	0.1745	0.3619	1.7006
Our Model	324	0.3189	**0.4837**	0.3717	0.1704	**0.3632**	**1.7080**
			Krogan Core				
MCL	370	**0.3088**	0.3214	0.3534	0.0944	0.2559	1.3339
DPClus	**497**	0.2899	0.3577	0.3335	0.1200	0.2893	1.3904
ClusterONE	240	0.2756	0.3913	0.2729	0.1058	0.2826	1.3282
PEWCC	383	0.2125	0.4228	0.2913	0.0987	0.3247	1.3500
WPNCA	369	0.2614	0.4361	0.3572	0.1250	0.2960	1.4757
CPredictor2.0	236	0.2421	0.4932	0.2787	0.1258	0.3216	1.4614
Zhang	326	0.2373	0.4637	0.2634	0.1456	0.2957	1.4057
PC2P	249	0.2884	0.3636	0.3141	0.0951	0.2818	1.3429
ClusterEPs	410	0.2390	0.4658	0.3021	0.1444	0.2975	1.4488
ClusterSS	342	0.2705	0.4304	0.3201	0.1318	0.3140	1.4669
ELF-DPC	281	0.2827	0.5336	0.3768	**0.1750**	0.3785	1.7467
Our Model	313	0.2813	**0.5459**	**0.3771**	0.1629	**0.3887**	**1.7560**
			DIP				
MCL	628	0.2504	0.2409	0.3025	0.0613	0.1921	1.0473
DPClus	909	0.2493	0.2784	0.3424	0.0898	0.2445	1.2044
ClusterONE	904	**0.2937**	0.4232	**0.4358**	0.1184	0.2874	1.5585
PEWCC	648	0.2182	0.4812	0.3336	0.0950	0.2986	1.4266
WPNCA	623	0.2472	0.4603	0.3709	0.1226	0.2866	1.4876
CPredictor2.0	293	0.2077	0.4653	0.2265	0.0736	0.2635	1.2367
Zhang	502	0.2215	0.4929	0.2928	0.1223	0.2818	1.4113
PC2P	441	0.2337	0.2662	0.2967	0.0588	0.2083	1.0636
ClusterEPs	804	0.1929	0.4611	0.2646	0.1323	0.2652	1.3162
ClusterSS	**2179**	0.2360	0.3676	0.3168	**0.1588**	0.2340	1.3132
ELF-DPC	545	0.2607	0.5126	0.3998	0.1386	0.3020	1.6137
Our Model	565	0.2694	**0.5261**	0.4147	0.1243	**0.3103**	**1.6447**
			MIPS				
MCL	594	0.1475	0.0551	0.1640	0.0125	0.1031	0.4822
DPClus	207	0.1948	0.3307	0.1934	0.0547	0.2049	0.9785
ClusterONE	690	**0.2148**	0.2473	0.2384	0.0630	0.1801	0.9435
PEWCC	382	0.1166	0.2309	0.1700	0.0296	0.1301	0.6773
WPNCA	527	0.1549	0.2640	0.2383	0.0621	0.1522	0.8716
CPredictor2.0	265	0.1966	0.3843	0.2086	0.0672	**0.2264**	1.0831
Zhang	406	0.1857	0.3413	0.1944	0.0710	0.2002	0.9925
PC2P	374	0.1941	0.2078	0.2136	0.0432	0.1524	0.8112
ClusterEPs	645	0.1720	0.3582	0.2115	0.0884	0.2120	1.0421
ClusterSS	**1581**	0.2074	0.2539	0.2566	0.0894	0.1867	0.9940
ELF-DPC	469	0.1937	0.4026	0.2599	**0.1011**	0.2249	1.1822
Our Model	469	0.1927	**0.4163**	**0.2791**	0.0823	0.2235	**1.1938**

The developed model uses standard protein complex 1 as the known protein complex.

- Compared to other models for the Gavin dataset, it ranked first on f-score and Jaccard, second on total score, sixth on accuracy and MMR, and fourth on CR.
- Compared to other models for the Krogan Core dataset, it ranked first on f-score, Jaccard, and total score, second on CR and MMR, and sixth on accuracy.
- Compared to other models for the DIP dataset, it ranked first on f-score, Jaccard, and total score, second on accuracy and CR, and fifth on MMR.

- Compared to other models for the MIPS dataset, it ranked first on CR, second on f-score and total score, third on MMR, fourth on Jaccard, and fifth on accuracy.

The developed model uses standard protein complex 2 as the known protein complex.

- Compared to other models for the Gavin dataset, it ranked first on f-score, Jaccard, and total score, fourth on CR, fifth on MMR, and seventh on accuracy.
- Compared to other models for the Krogan Core dataset, it ranked first on f-score, CR, Jaccard and total score, second on MMR, and fifth on accuracy.
- Compared to other models for the DIP dataset, it ranked first on f-score, Jaccard, and total score, second on accuracy and CR, and fourth on MMR.
- Compared to other models for the MIPS data set, it ranked first on f-score, CR, and total score, third on Jaccard, fourth on MMR, and seventh on accuracy.

4 Conclusions

Although there are many models proposed to detect protein complexes from PPI networks, the lack of a model that can achieve perfect performance is still a problem for bioinformatics that needs to be solved. In this study, we developed a model as an alternative to the solutions produced for this problem. First, the common expression, function, subcellular location, and topological structure similarities of the proteins were used to create the weighted PPI network. Secondly, protein complex core has been predicted. Third, correct protein complexes and incorrect protein complexes were produced using standard protein complexes. A voting regression model was then trained using these protein complexes. Fourth, structural modularity was measured, which is used to measure the probability of a protein cluster being a protein complex. An ensemble learning model has been built in which voting regression and structural modularity are used together. Protein complexes were produced using the ensemble learning model constructed fifth. Finally, the success of the community learning model was measured with success criteria.

In our study, we used a voting regression model based on supervised learning and the performance of multiple predictors, which have proven to be more successful in the literature, so that unsuccessful results in one predictor can be balanced with successful results in the other, and in addition to the studies in the literature, we proved that the use of XGB regression, gaussian process regression, catboost regression and histogram-based gradient boosting regression models within the voting regression model may be more successful. When we compare our study with other studies, for standard protein complex 1, DIP and Krogan Core PPI networks showed the best performance and Gavin and MIPS PPI networks showed the second-best performance. For standard protein complex 2, it showed the best performance in Gavin, Krogan Core, DIP and MIPS PPI networks.

In future studies, it is planned to include different data sources to discover protein complexes where the number of proteins is low, which we believe will positively affect the success of the model we have developed, to treat the manual parameter selection as an optimization problem and to test optimization algorithms and to use deep learning methods.

5 Author contribution statement

Yasin KARAKUŞ has participated in the design of the study, wrote the code, performed the experiments and drafted the manuscript. Volkan ALTUNTAS has contributed to the manuscript and provided the initial idea, and to the design and supervision of the study. All authors have read and approved the final manuscript.

6 Ethics committee approval and conflict of interest statement

"There is no need to obtain an ethics committee approval for this manuscript".

"There is no conflict of interest with any person/institution in the manuscript".

7 References

[1] Sabzinezhad A, Jalili S. "DPCT: a dynamic method for detecting protein complexes from TAP-aware weighted ppi network". *Frontiers in Genetics*, 11, 1-15, 2020.

[2] Xu B, Li K, Zheng W, Liu X, Zhang Y, Zhao Z, He Z. "Protein complexes identification based on go attributed network embedding". *BMC Bioinformatics,* 19, 1-10, 2018.

[3] Altaf-Ul-Amin M, Shinbo Y, Mihara K, Kurokawa K, Kanaya S. "Development and implementation of an algorithm for detection of protein complexes in large interaction networks". *BMC Bioinformatics*, 7, 207-213, 2006.

[4] Wang R, Wang C, Liu G. "A novel graph clustering method with a greedy heuristic search algorithm for mining protein complexes from dynamic and static ppi networks". *Information Sciences*, 522, 275-298, 2020.

[5] Dilmaghani S, Brust MR, Ribeiro CHC, Kieffer E, Danoy G, Bouvry P. "From communities to protein complexes: a local community detection algorithm on PPI networks". *Plos One,* 17(1), 1-17, 2022.

[6] Yu Y, Kong D. "Protein complexes detection based on node local properties and gene expression in PPI weighted networks". *BMC Bioinformatics,* 23, 1-15, 2022.

[7] Cho Y, Hwang W, Zhang A. "Identification of overlapping functional modules in protein interaction networks: information flow-based approach". *Sixth IEEE International Conference on Data Mining-Workshops (ICDMW'06),* Hong Kong, China, 18-22 December 2006.

[8] Omranian S, Angeleska A, Nikoloski Z. "PC2P: parameter-free network-based prediction of protein complexes". *Bioinformatics*, 37(1), 73-81, 2021.

[9] Wang R, Wang C, Ma H. "Detecting protein complexes with multiple properties by an adaptive harmony search algorithm". *BMC Bioinformatics,* 23(1), 1-32, 2022.

[10] Xu B, Wang Y, Wang Z, Zhou J, Zhou S, Guan J. "An effective approach to detecting both small and large complexes from protein-protein interaction networks". *BMC Bioinformatics,* 18, 1-10, 2017.

[11] Wang X, Zhang N, Zhao Y, Wang J. "A new method for recognizing protein complexes based on protein interaction networks and GO terms". *Frontiers in Genetics,* 12, 1-7, 2021.

[12] Zhang XF, Dai DQ, Li XX. "Protein complexes discovery based on protein-protein interaction data via a regularized sparse generative network model". *IEEE-ACM Transactions on Computational Biology and Bioinformatics*, 9(3), 857-870, 2012.

[13] Xu B, Lin H, Chen Y, Yang Z, Liu H. "Protein complex identification by integrating protein-protein interaction evidence from multiple sources". *Plos One*, 8(12), 1-12, 2013.

[14] Zaki N, Singh H, Mohamed EA. "Identifying protein complexes in protein-protein interaction data using graph convolutional network". *IEEE Access*, 9, 123717-123726, 2021.

[15] Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D. "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions". *Nucleic Acids Research*, 30(1), 303-305, 2002.

[16] Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dümpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G. "Proteome survey reveals modularity of the yeast cell machinery". *Nature*, 440(7084), 631-636, 2006.

[17] Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrín-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, Onge PS, Ghanny S, Lam MHY, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF. "Global landscape of protein complexes in the yeast saccharomyces cerevisiae". *Nature*, 440(7084), 637-643, 2006.

[18] Güldener U, Münsterkötter M, Oesterheld M, Pagel P, Ruepp A, Mewes HW, Stümpflen V. "Mpact: the mips protein interaction resource on yeast". *Nucleic Acids Research,* 34, D436-441, 2006.

[19] Hong EL, Balakrishnan R, Dong Q, Christie KR, Park J, Binkley G, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Krieger CJ, Livstone MS, Miyasato SR, Nash RS, Oughtred R, Skrzypek MS, Weng S, Wong ED, Zhu KK, Dolinski K, Botstein D, Cherry JM. "Gene ontology annotations at sgd: new data sources and annotation methods". *Nucleic Acids Research*, 36, D577-D581, 2007.

[20] Aloy P, Böttcher B, Ceulemans H, Leutwein C, Mellwig C, Fischer S, Gavin AC, Bork P, Superti-Furga G, Serrano L, Russell RB. "Structure-based assembly of protein complexes in yeast". *Science*, 303(5666), 2026-2029, 2004.

[21] Pu S, Wong J, Turner B, Cho E, Wodak SJ. "Up-to-date catalogues of yeast protein complexes". *Nucleic Acids Research*, 37(3), 825-831, 2009.

[22] Friedel CC, Krumsiek J, Zimmer R. "Bootstrapping the interactome: unsupervised identification of protein complexes in yeast". *Journal of Computational Biology*, 16(8), 971-987, 2009.

[23] Ma CY, Chen YP, Berger B, Liao CS. "Identification of protein complexes by integrating multiple alignment of protein interaction networks". *Bioinformatics*, 33(11), 1681-1688, 2017.

[24] Spirin V, Mirny LA. "Protein complexes and functional modules in molecular networks". *Proceedings of the National Academy of Sciences of the United States of America*, 100(21), 12123-12128, 2003.

[25] Zhang J, Zhong C, Huang Y, Lin HX, Wang M. "A method for identifying protein complexes with the features of joint co-localization and joint co-expression in static PPI networks". *Computers in Biology and Medicine*, 111, 1-10, 2019.

[26] Grover A, Leskovec J. "node2vec: scalable feature learning for networks". *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, California, USA, 13-17 August 2016.

[27] Wu M, Li X, Kwoh CK, Ng SK. "A core-attachment based method to detect protein complexes in ppi networks". *BMC Bioinformatics*, 10, 1-16, 2009.

[28] Wang R, Wang C, Sun L, Liu G. "A seed-extended algorithm for detecting protein complexes based on density and modularity with topological structure and go annotations". *BMC Genomics*, 20, 1-28, 2019.

[29] Ghahramani Z. "A tutorial on gaussian processes (or why i don't use SVMs)". *Machine Learning Summer School (MLSS),* Bordeaux, Fransa, 4-17 September 2011.

[30] Wang R, Ma H, Wang C. "An ensemble learning framework for detecting protein complexes from PPI networks". *Frontiers in Genetics,* 13, 1-22, 2022.

[31] Peng W, Wang J, Zhao B, Wang L. "Identification of protein complexes using weighted pagerank-nibble algorithm and core-attachment structure". *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(1), 179-192, 2014.

[32] Nepusz T, Yu H, Paccanaro A. "Detecting overlapping protein complexes in protein-protein interaction networks". *Nature Methods*, 9(5), 471-472, 2012.

[33] Zaki N, Efimov D, Berengueres J. "Protein complex detection using interaction reliability assessment and weighted clustering coefficient". *BMC Bioinformatics,* 14, 1-9, 2013.

[34] Zhang Y, Lin H, Yang Z, Wang J, Liu Y, Sang S. "A method for predicting protein complex in dynamic ppi networks". *BMC Bioinformatics*, 17, 533-543, 2016.

[35] Liu Q, Song J, Li J. "Using contrast patterns between true complexes and random subgraphs in ppi networks to predict unknown protein complexes". *Scientific Reports*, 6, 1-15, 2016.

[36] Dong Y, Sun Y, Qin C. "Predicting protein complexes using a supervised learning method combined with local structural information". *Plos One*, 13(3), 1-23, 2018.