Journal of Algorithms and Computation

journal homepage: http://jac.ut.ac.ir

An attributed network embedding method to predict missing links in protein-protein interaction networks

Ali Golzadeh Kermani*[1], Ali Kamandi†[2] and Hossein Rahami‡[3]

[1,2,3]School of Engineering Science, College of Engineering, University of Tehran, Tehran, Iran

ABSTRACT

Predicting missing links in noisy protein-protein interaction networks is an essential computational method. Recently, attributed network embedding methods have been shown to be significantly effective in generating low-dimensional representations of nodes to predict links; in these representations, both the nodes'features and the network's topological information are preserved. Recent research suggests that models based on paths of length 3 between two nodes are more accurate than models based on paths of length 2 for predicting missing links in a protein-protein interaction network. In the present study, an attributed network embedding method termed ANESITI is recommended to combine protein sequence information and network topological information.

Keyword: link prediction, protein-protein interaction networks, attributed network embedding, biased random walks.

ARTICLE INFO

Article history:
Research paper
Received 29, March 2023
Received in revised form 11, May 2023
Accepted 28 May 2023
Available online 01, June 2023

AMS subject Classification: 05C78.

*aligolzadeh@ut.ac.ir
†Corresponding author: A. Kamandi. Email: kamandi@ut.ac.ir
‡hrahami@ut.ac.ir

1 Abstract continued

In addition, to improve accuracy, network topological information also considers paths of length 3 between two proteins. The results of this experiment demonstrate that ANE-SITI outperforms the compared methods on various protein-protein interaction (PPI) networks.

2 Introduction

Networks are effective tools for describing and simulating complex systems that depict interactions between diverse real-world entities. Extracting knowledge from biological networks has recently become a popular but challenging topic [26]. Protein-protein interaction (PPI) networks are one of the most important biological networks in which proteins interact with each other in biological activities. In a PPI network represented as a graph, proteins are depicted as nodes and their interactions as edges. When PPI is considered a weighted network, the weight indicates the probable interaction of each protein pair concerning its reliability [33]. PPIs are involved in many signal transductions between cellular processes and metabolic pathways, and other molecules in the biological system [37]. Considering their mechanisms requires a thorough understanding of all the physical relations among proteins [14]. PPI data coverage is still relatively low, and there is a lot of noisy data in the PPI dataset [20], which has led to various computational approaches in recent years to predict potential interactions despite major efforts in high-efficiency techniques.

Discovering possible relations in a network is an important and challenging task. In a network, deducing missing relations based on a currently observed network or predicting possible future links is called link prediction [35,37]. The application domains for link prediction are biological, scientific collaboration, and social networks [35]. Many link prediction methods have been implemented for different types of networks and proposed to meet the needs of related applications [55]. Obviously, these methods can also be employed to predict interactions in PPI networks as a single complex network. Link prediction approaches based on network topology are grouped into three categories: 1) similarity-based, 2) probabilistic-based, and 3) embedding-based approaches [30]. Similarity-based approaches are the simplest methods in link prediction, in which a similarity score is initially calculated for each node pair. The score between each pair is based on topological information. In this method, unobserved links are assigned points based on their similarities, and any pair of nodes with a higher score is more likely to be related. Methods such as common neighbor [44], Adamic/Adar [2], Katz Index [25], SimRank [24], and Local Path Index [36] belong to this category.

The probabilistic-based approaches optimize an objective function to configure a model consisting of various parameters. Yu et al. [61], Clauset et al. [13], and Guimerà et al. [18] are examples of studies conducted on the probabilistic-based approach.

An embedding-based approach is recognized as a dimensional reduction model in which nodes in networks are mapped to a lower-dimensional space while maintaining the neighboring structures of the nodes. This category includes methods such as DeepWalk [48], Node2vec [17], graph factorization [4], and GraphGAN [53].

Most of the methods in the aforementioned link prediction approaches rely on network topology and disregard the valuable characteristics of proteins. These features include protein domains, protein structure information, phylogenetic profiles, gene neighborhood, and gene expression [54]. Thus, some methods are based on protein biological information, meaning that two proteins are more likely to interact if they have features close to one other. With the advancement of machine learning technology, machine learning-based methods have extensively been used to predict PPI [59]. For example, DeepPPI [15], EnsDNN [63], InterSPPI [59], and GcForest-PPI [60] are examples of these methods.

Recently, the L3 principle was introduced, a novel approach to link prediction based on biological information [29, 47, 42]. This principle is based on the assumption that two proteins linked by different paths of length 3 are more likely to interact with each other directly [29]. The L3 principle scores the relationship between the two proteins deduces new interactions, and retains candidates with the highest scores. It can be argued that if there are multiple paths of length 2 between two proteins, this may exert the opposite effect on their direct interaction, meaning that the L3 principle works better than many well-known link prediction approaches, such as the Common Neighbor [44].

Most of the methods above only consider the topological information of the network or the protein sequence information. This study presents an attributed network embedding (ANE) technique that combines sequence and topological information named ANE-SITI. Initially, these two types of information are combined to improve the efficiency of link prediction; the L3 principle and edge weights are considered in topological information. This combination of data results in the formation of an enriched network. Then, sequences of nodes are created using a biased random walk on the generated network. These sequences are moved to the skip-gram with negative sampling (SGNS) model [39] to generate low-dimensional vectors of each protein. A binary classification algorithm on these vectors is then used to predict missing interactions.

The remainder of this paper is structured as follows. Details of related works and background information are provided in Section 2. The link prediction approach of this article is discussed in Section 3. In Section 4, the performance of the approach is verified with real-world networks and other popular methods for comparison purposes. Finally, conclusions and potential future issues are offered in Section 5.

3 Related works

As stated previously, diverse methods have been proposed to predict protein interactions based on the properties of the proteins or the network structure. In recent years, researchers have become interested in network embedding applications. Network embedding methods have achieved favorable outcomes compared to the other methods in complex networks, especially downstream functions such as link prediction [40,8]. This method aims to learn a low-dimensional feature representation for every node. Low-dimensional representations are learned to preserve network data and can therefore be used as a component in machine learning model construction [62]. Network embedding methods include four models: matrix factorization, random walks, neural networks, and attributed networks.

The classic network embedding model is a matrix factorization-based model that has been utilized in numerous articles involving link prediction [38,1,49]. This model factorizes the matrix of input data into lower-dimension matrices. For example, GraRep [9] considers various powers of the adjacency matrix to capture higher-order graph proximity. The optimization problem is solved by a classical matrix factorization technique, singular value decomposition (SVD) [16]. HOPE [45] also investigates high-order proximity and applies several important similarity criteria, such as the Katz index [25]. Cho et al. [12] aimed to learn each protein's dense low-dimensional vector representation that best defines their patterns for the PPI networks of the input by employing a matrix factorization technique. Zhang et al. [64] proposed the drug feature-based adjustment and represented a new matrix factorization method for predicting possible drug-drug interactions.

The random walk-based model uses random walks to build network neighborhoods from each node in the network and focuses on embedding nodes in low-dimension vector spaces. This network embedding model has successfully been used for various bioinformatics applications such as protein function prediction [65], disease-pathway analysis [3], and PPI prediction. For example, DeepWalk is considered the first method in this model [48]. This method was initially proposed for embedding nodes in a social network using linguistic literature-based designs. Node2vec [17] has a flexible neighborhood sampling strategy, which provides an alternative between the breadth-first search (BFS) and the depth-first search (DFS). Liu et al. [34] combined several PPI datasets of dissimilar types into a sole network. The study used a network embedding method to encode protein nodes in continuous vector spaces and a seed-and-extend method to identify protein complexes. HerGePred [58], a framework for disease gene prediction, is another method. This approach proposes a random walk with a restart method on a reconstructed disease-gene network to predict disease genes efficiently.

Recent years have witnessed the success of neural network models in various fields. Several neural networks have also been introduced in the network embedding area, such as multilayer perceptron (MLP) [40], autoencoder [10,52], generative adversarial network (GAN) [53], and graph convolutional network (GCN) [11]. For example, VGAE [28] includes a dual-layer GCN and a straightforward interior product decoder to determine meaningful latent embedding based on the variational autoencoder [27]. SAGE [19] samples and collects features from a node's local neighborhood and learns embedding by long short-term memory and pooling. Lim et al. [32] suggest a new deep-learning method for predicting drug-target communication using neural networks and extracting network features of intermolecular interactions straight from three-dimensional structural information on proteins. Decagon [66] is a method for demonstrating the side effects of polypharmacy. This method generates a multidimensional graph of polypharmacy side effects, drug-protein target interactions, and PPIs, representing drug-drug interactions. Furthermore, it develops a convolutional neural network to predict multi-link interconnections in multidimensional networks.

The attributed network embedding (ANE) model was recently proposed, and the current article's approach is based on this model. This model states that, in addition to the observed network topological information, many nodes in the network are associated with node feature-rich information. Embedding a network is more challenging in attributed networks because it requires learning low-dimensional representations of nodes that preserve topological and feature information [54]. For example, ASNE [31] uses structural proximity and attribute proximity to learn representations for social nodes. TADW [56] is proposed to combine the text features of nodes in the DeepWalk algorithm and demonstrates better performance. S-VGAE [57] studies PPI prediction based on both sequence information and network structure. Pan et al. [46] employed the convolutional network as an encoder to embed the topological information and node content into a vector representation.

4 The proposed method

This section, describes the ANE-SITI method to predict missing links in PPI networks. ANE-SITI consists of three phases: 1) Protein sequence information is obtained, 2) The topological information is obtained based on the L3 principle with edges weight, and 3) A biased random walk on the generated network of a combination of these two types of information is considered to generate low- dimensional vectors of each protein. The main notations used are presented in Table 1, and the basic definitions are explained as follows:

Table 1
Summary of notations

Notations	**Descriptions**
$\boldsymbol{G}$	An undirected weighted network
$\boldsymbol{V}$	Proteins set
$\boldsymbol{E}$	Interactions between proteins set
$\boldsymbol{B}$	Sequence information similarity matrix
$\boldsymbol{T}$	Topological information similarity matrix
$\boldsymbol{W_{ij}}$	Weight of interaction between protein v_i and protein v_j
$\boldsymbol{\lambda}$	Balancing factor
$\boldsymbol{d}$	Length of vector representation
$\boldsymbol{\vert V\vert = n}$	Number of proteins
$\boldsymbol{k}$	Number of attributes of each protein

Definition 1: $G = (V,\ E,\ B,\ W)$ is an undirected weighted attributed network representing a PPI network, where V, E, B, and W are described in Table 1.
Definition 2: Attributed network embedding aims to find a mapping function f : $V \rightarrow \mathbb{R}^d$ ($d \ll |V|$), that creates an embedded vector for each node in the d dimension to preserve the topological information and protein sequence information.

4.1 Protein sequence information

In order to obtain the protein sequence information, evolutionary information is first extracted from the Position Specific Scoring Matrix (PSSM), and then Moran autocorrelation (MAC) [41] is used to extract features from PSSMs. PSSM is created by the PSI-BLAST program [6] to probe the NCBI's NR database through a cutoff E-value of 0.001 and three iterations for multiple sequence alignment to the protein sequence. PSSM is a matrix of size L×20 in which L denotes the length of the protein amino acid sequence, and the number 20 means 20 native amino acid types. Each element in the matrix is represented by $P_{i,j}$ that signifies the score of the residue of amino acid in the ith place of the protein sequence being changed to amino acid type j in the biology evolution procedure.
This paper uses MAC to transform PSSM vectors of different lengths into vectors of equal length. MAC can be calculated as follows:

$$MAC_{r,j} = \frac{\frac{1}{L-r}\ \sum_{i=1}^{L-r} (P_{i,j} - \overline{P_j})\ (P_{i+r,j} - \overline{P}_j)}{\frac{1}{L}\ \sum_{i=1}^{L}\ (P_{i,j} - \overline{P}_j)^2}\ , (j = 1, 2, \ldots, 20; \quad r = 1, 2, \ldots, 10) \tag{4.1}$$

$$\overline{P}_j = \frac{1}{L} \sum_{i=1}^{L} P_{i,j} \tag{4.2}$$

where L denotes the length of the protein sequence, r is the distance between a residue and its neighbors (its maximum value is assumed to be 10 [41]), $P_{i,j}$ and $P_{i+r,j}$ represent the score values in ith and $i+rth$ place of the sequence of protein being transformed to amino acid type j throughout the evolution procedure, and $\overline{P}_j$ is the average value of P_j. Therefore, $MAC_{r,j}$ consists of a total of 20*10 = 200 descriptor values for each protein. For each vector, $MAC_{r,j}$ is considered an equivalent vector $MACP_k(n)$, where n represents the protein number and k is the index ($1 \leq k \leq 200$), that is, a 200-dimensional vector to represent each protein sequence. Finally, the sequence information similarity matrix B is obtained with cosine similarity [50] on these vectors as follows:

$$B_{i,j} = \frac{\sum_{k=1}^{200}\ MACP_k(i)\ MACP_k(j)}{\sqrt{\sum_{k=1}^{200}\ (MACP_k(i))^2}\sqrt{\sum_{k=1}^{200}\ (MACP_k(j))^2}} \tag{4.3}$$

4.2 Topological information

L3 principle is considered to obtain topological information. As previously stated, according to the L3 principle, the greater the number of paths of length three between two proteins, the greater the probability of a link between them. This principle outperforms many well-known link prediction techniques [29, 47, 42] in predicting missing links in PPI networks. The calculation of the weight of the edges is the feature that ANE-SITI adds to the L3 principle.
The fundamental concept underlying topological link prediction methods is the development of a node similarity measurement that specifies the probability of a link between each pair of nodes. One of these methods is Jaccard [23], which has been widely utilized in link prediction, which is defined as follows in unweighted networks:

$$Jacard_{u,v} = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|} \tag{4.4}$$

where $\Gamma(j)$ is a series of node j neighbors in the PPI network. In weighted networks, the following can also be considered:

$$J_{u,v} = \frac{\sum_{k \in \Gamma(u) \cap \Gamma(v)} \; W_{u,k} + W_{k,v}}{\sum_{m \in \Gamma(u)} \; W_{u,m} \; + \; \sum_{n \in \Gamma(v)} \; W_{v,n}} \tag{4.5}$$

where W denotes the weighted adjacency matrix.
In this study, the weighted Jaccard similarity is not used directly between proteins i and j to predict their links, but rather between the neighbors of one protein and another (according to the L3 principle, paths of length three are considered). Thus, the high-weighted Jaccard similarity between the two proteins does not indicate a relation between them but has a large proportion of reciprocal partners, that is, the ratio of their common neighbors to all their neighbors. In other words, if the neighbors of node i and j have high weighted Jaccard similarity, then a relation between i and j can be predicted. Accordingly, the topological information similarity matrix is defined as follows:

$$T_{ij} = \sum_{m \in \Gamma(i)} J_{m,j} + \sum_{n \in \Gamma(j)} J_{n,i} \tag{4.6}$$

4.3 Biased Random Walks

In this section, the two types of obtained information are combined first. To this end, the two matrices (i.e., matrix B and matrix T) are normalized to obtain the transition matrix of each. In both matrices, nodes that have no relation are not considered because they do not influence the performance of the ANE-SITI. In matrix B, negative values are assumed zero because it is considered that either the two proteins are similar that have a value greater than zero or the two proteins have no resemblance to each other that value is assumed to be zero. In addition, to increase the efficiency of the

ANE-SITI, only the top-m most similar proteins are measured in matrix B. Thus, each row of the B has only an m non-zero value. Consequently, the transition matrix for the matrix B is $B^{(T)}$ as follows:

$$B_{i,j}^{(T)} = \frac{B_{i,j}}{\sum_{j=1}^{m} B_{i,j}}, (0 \leq B_{i,j}^{(T)} \leq 1;\ \sum_{j=1}^{m} B_{i,j}^{(T)} = 1) \tag{4.7}$$

There is no negative value in matrix T according to Eq. 4.6. In addition, the transition matrix T is $T^{(T)}$ as follows:

$$T_{i,j}^{(T)} = \frac{T_{i,j}}{\sum_{j=1}^{n} T_{i,j}}, (0 \leq T_{i,j}^{(T)} \leq 1;\ \sum_{j=1}^{n} T_{i,j}^{(T)} = 1) \tag{4.8}$$

The combination of these two types of information is as follows:

$$C_{i,j} = (1-\lambda)\ B_{i,j}^{(T)} + \lambda\ T_{i,j}^{(T)}\ (0 \leq \lambda \leq 1) \tag{4.9}$$

where λ is a factor that controls the trade-off between protein sequence information and topological information. Since matrix $B^{(T)}$ and $T^{(T)}$ are transitions, matrix C is also a transition matrix. Using the C matrix, a weighted network is reconstructed that combines the two types of information.

At this stage, a random walks method is employed to learn node embeddings on the reconstructed network. The alias sampling technique, as described in Node2Vec [17], is used to implement weighted random walks on the constructed network. These weighted random walks perform on each protein t times with length L and create a set of $t \times n$ protein sequences. A constant size window w is applied to slide beside every protein sequence, and a number of training pairs (v, u) are created for each window in which v is the central protein, and $u \in V$ are the neighboring proteins. Then, to train protein embeddings, the training pairs are moved to the SGNS model [30]. Finally, logistic regression is used as a binary classification algorithm to predict missing interactions on the concatenated embed vectors of each protein as an edge feature. An outline of the ANE-SITI method, which contains the above three steps, is presented in Fig. 1.

5 Experiments

The ANE-SITI method was compared to six cutting-edge network algorithms to determine its effectiveness and efficiency. Each algorithm's category is specified in Table 2, and each is described below:

L3 [29], if two proteins are similar, they tend not to be related, but if one of them is similar to the other's partners, this tendency is high.

DeepWalk [48] embeds network nodes using random walks in addition to the skip-gram model.

FSFDW [43] employs structural and non-structural properties of PPI networks in conjunction with a biased random-walk-based embedding method to predict PPIs.
AANE [22] is an accelerated attributed network embedding algorithm that considers network topology and node attributes.
GraphSage [19] is a method based on convolutional neural networks that aggregate the attributes of neighboring nodes.

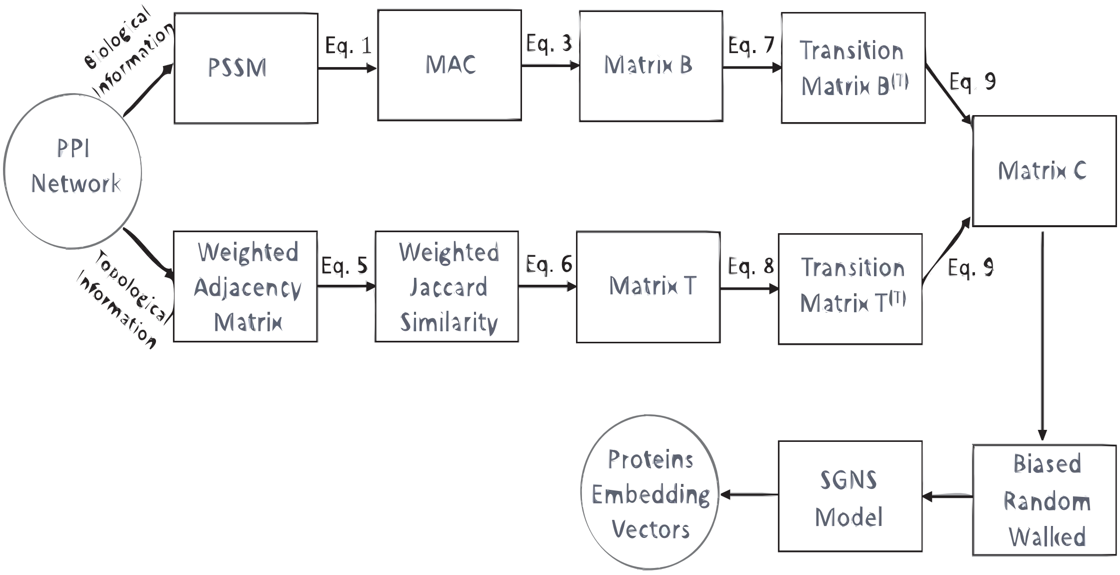

Fig. 1. The structure of the ANE-SITI method

Table 2
Compared algorithms

Algorithm name	**Category**
L3	Network similarity
DeepWalk	Random-walk-based network embedding
FSFDW	Attributed network embedding
AANE	Attributed network embedding
S-VGAE	Attributed network embedding
GraphSage	Neural network embedding

S-VGAE [57] uses sequence information and network structure to predict PPIs.
Most binary classification evaluation criteria are used in link prediction evaluation because the link prediction problem can be considered a binary classification task [5]. This approach is evaluated based on three criteria with the algorithms described above: The Area Under the Receiver Operating Characteristic curve (AUROC) demonstrates the method's ability to distinguish between positive and negative samples.

The F1-score calculates the harmonic mean between precision and recall and combines them into one metric. The F1-score can be computed using the following equations:

$$precision = \frac{TP}{TP + FP} \tag{5.1}$$

$$recall = \frac{TP}{TP + FN} \tag{5.2}$$

$$F1 - score = \frac{precision \times recall}{precision + recall} \tag{5.3}$$

The Matthew correlation coefficient (MCC) is a common measure of binary classification quality. MCC is not only capable of displaying the correlation coefficient between predicted protein pairs and the original truth, but it can also handle cases in which the number of interacting and non-interacting proteins are of very different sizes.

$$MCC \ = \frac{TP \times TN - FP \times FN}{\sqrt{(TN + FN)(TN + FP)(TP + FP)(TP + FN)}} \tag{5.4}$$

The Precision-Recall (PR) curve and the corresponding area under the PR curve (AUPRC) are commonly used to evaluate the classification performance when negative samples are greater than positive ones. Therefore, it is a useful criterion to evaluate link prediction models in PPI networks where the number of identified interactions between proteins is much less than the number of unidentified interactions.

5.1 Datasets

Homo sapiens (*H. sapiens*), Mus musculus (*M. musculus*), and Saccharomyces cerevisiae (*S. cerevisiae*), three real-world PPI networks, are selected from the STRINGDB [51] dataset to assess the function of the ANE-SITI on real PPI network. In the field of bioinformatics, STRINGDB is the most frequently used database that gathers a lot of PPI data from different species [51]. The links with weights <0.7 are removed to normalize and prevent false positives. The information on these three networks is presented in Table 3.

Table 3 Characteristics of the PPI Networks used in the Experiment

Network	**|V|**	**|E|**
H. sapiens	10216	164122
M. musculus	9507	139998
S. cerevisiae	3355	92846

5.2 Experimental results

All identified links are established as positive instances and are separated into a training set (80%) and test set (20%) to compare the ANE-SITI method with other methods, and these instances are then eliminated from the network, confirming that the network remains connected. Since unidentified links are much more prevalent than identified cases, unidentified links are randomly selected as negative instances in training alongside an equal number of positive instances. Because the use of the same number of positive and negative instances during the testing provides a biased picture of the actual performance of the method, the ratio of negative to positive instances is chosen to be 10:1. For each data set, the training set and the test set are randomly selected ten times, and then the average of these ten repetitions is considered for each metric.

Parameter settings for the six compared methods are identical to their original papers. For the ANE-SITI, walk length, denoted by h, is 80, the dimensions of embedding vectors, denoted by d, is 64, the number of walks per protein, denoted by t, is 10, and the window size, denoted by w, is 10.

In Fig. 2, the ANE-SITI is compared to the other methods through the F1-score on the three networks. As indicated in Fig. 2, the ANE-SITI obtains the largest F1-score on *H. sapiens* with a value of 0.87, *M. musculus* with a value of 0.89, and *S. cerevisiae* with a value of 0.87.

A comparison of the methods based on the AUROC is shown in Fig. 3. As seen in the figure, the proposed method yields better results than other methods. This may be due to the application of the composition of the L3 principle and sequence information. As is well-known, the GraphSage method has also achieved excellent results concerning the AUROC criterion. In principle, GraphSage has shown to be highly efficient on large networks.

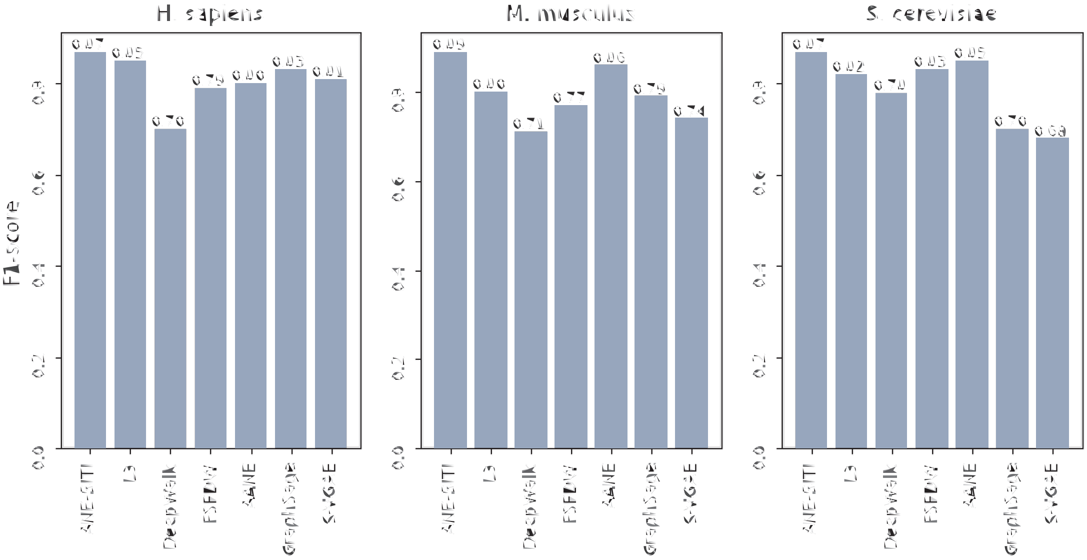

Fig. 2. F1-score of methods on the three networks

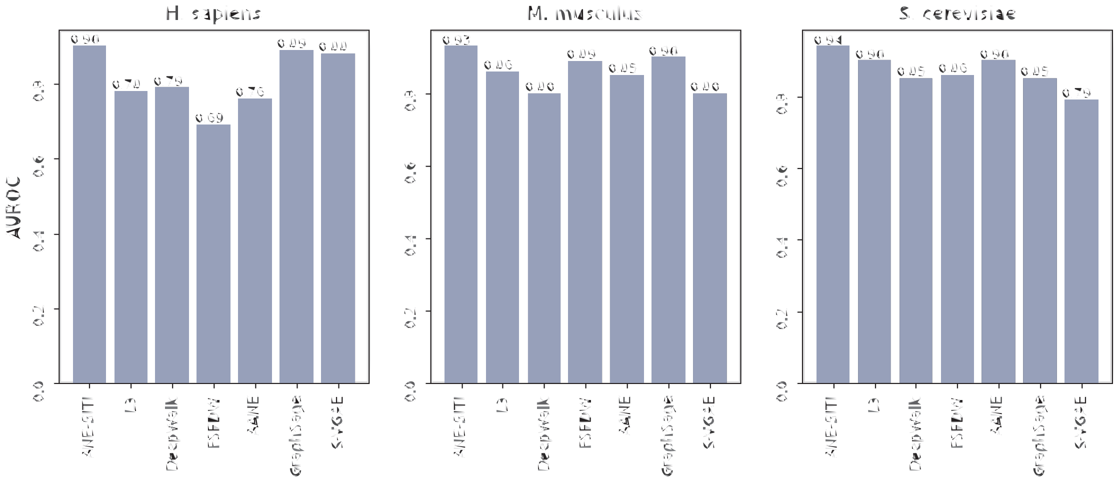

Fig. 3. AUROC of the compared methods

According to Table 4, ANE-SITI receives the highest MCC among the three networks. As shown in Table 4, the L3 method yields superior results to the other methods. This indicates that the probability of a link between two proteins increases as the number of paths of length three between them increases. In addition, ANE-SITI is superior to L3 because it uses sequence information in its link prediction process.

Fig. 4 compares the ANE-SITI method and other methods along the Precision-Recall curve. The ANE-SITI method outperforms the other methods, as shown in Fig. 4. As can be seen, the area under the Precision-Recall (PR) curve, or AUPRC, is greater than the compared methods at nearly all thresholds across the three networks.

In the last part of the evaluation, the methods are compared based on their running times, and all methods are simulated on a workstation. Fig. 5 presents the method's running times on the three networks. The ANE-SITI's running time is more than that of some methods, which is normal, given that it obtains both sequence and topological information. However, this time is not significantly increased, and it makes sense to add a small amount of computational time to improve performance. On the other hand, this time can also be reduced if obtaining sequence and topological information is done in parallel. As seen, the DeepWalk and L3 methods have the minimum running time due to calculating only the structural information, and the GraphSage method has the maximum running time due to using a neural network.

Table 4. MCC of methods on the three networks

	H. sapiens	***M. musculus***	***S. cerevisiae***
ANE-SITI	**0.78**	**0.73**	**0.75**
L3	0.74	0.72	0.74
DeepWalk	0.66	0.71	0.71
FSFDW	0.70	0.68	0.66
AANE	0.67	0.70	0.69
GraphSage	0.75	0.70	0.70
S-VGAE	0.72	0.68	0.74

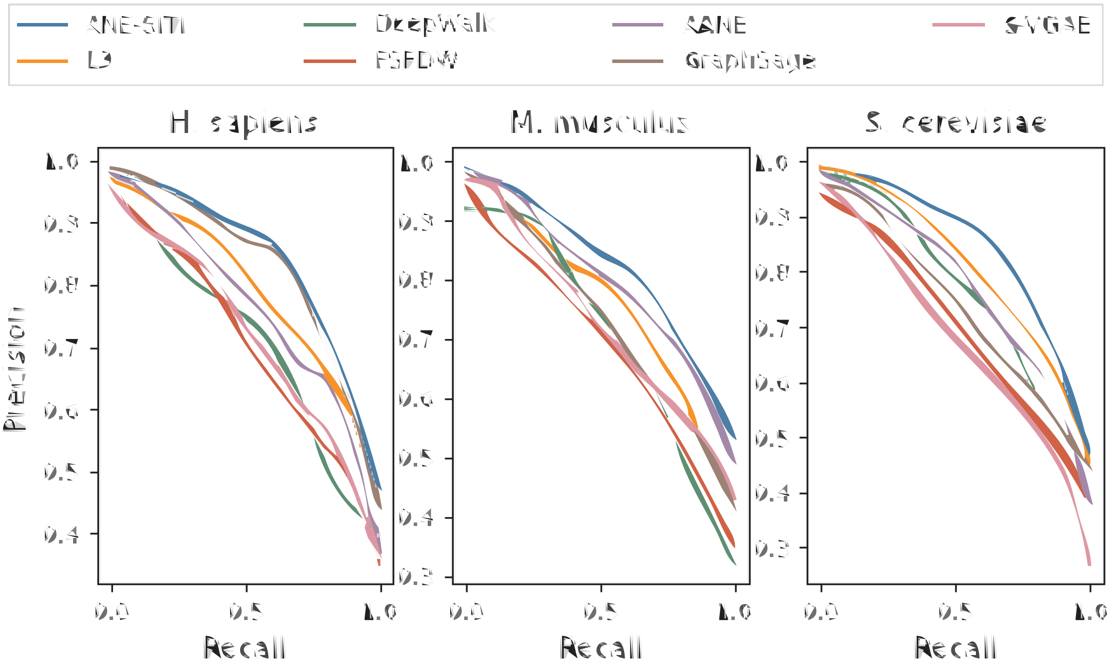

Fig. 4. Precision-Recall curves of the compared methods

As demonstrated in this section, ANE-SITI is more effective than other methods. In addition to considering the specific information of proteins in the sequence information combined with the topological information that creates a new enriched graph, the weight of paths of length 3 between two proteins is also used in the topological information, which has increased the efficiency of ANE-SITI.

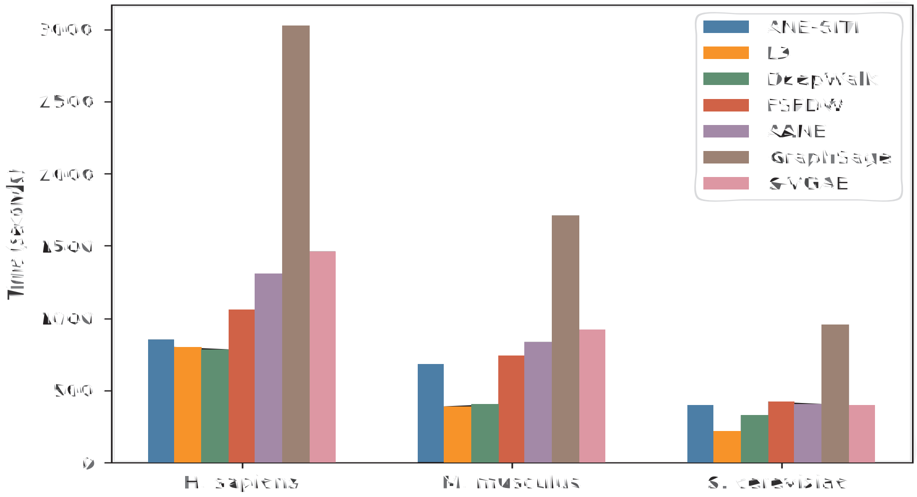

Fig. 5. Running time of the compared methods

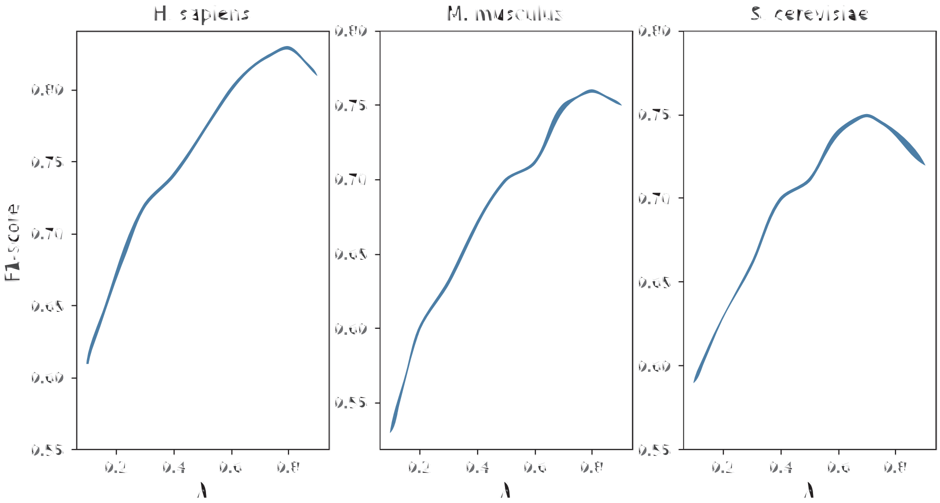

Fig. 6. Influence of the λ parameter on ANE-SITI

5.3 Parameter sensitivity

In this section, the ANE-SITI's sensitivity is examined concerning two factors on the F1-score criterion, that is, the harmonic value λ and the length of vector representation d. The values of the other variables are permanent when examining these two factors, according to Section 4.2.

The harmonic factor λ balances the contribution of protein sequence information and topological information. This ranges from 0.1 to 0.9 for determining its effectiveness. When low, protein sequence information has a significant impact on ANE-SITI performance. By increasing it, topological information exerts a meaningful effect. Examining the λ parameter on the three networks is presented in Fig. 6. As can be seen, as λ increases, so does the F1-score. This means that the effect of topological information on the efficiency of ANE-SITI is more than that of the protein sequence information. The best λ value is 0.8 on the *H. sapiens* and *M. musculus* networks and 0.7 on the *S. cerevisiae* network, and the value of the F1-score decreases somewhat for higher λ values.

To examine the parameter d, it varies from 2^4 to 2^9 on the three networks. Fig. 7 displays that as the value of d increases, so the value of the F1-score does, and from one place onwards, by increasing the value of d, the value of the F1-score remains almost constant. Although the best results on different networks are obtained with different dimensional parameters, 2^7 is a relatively good choice in practice.

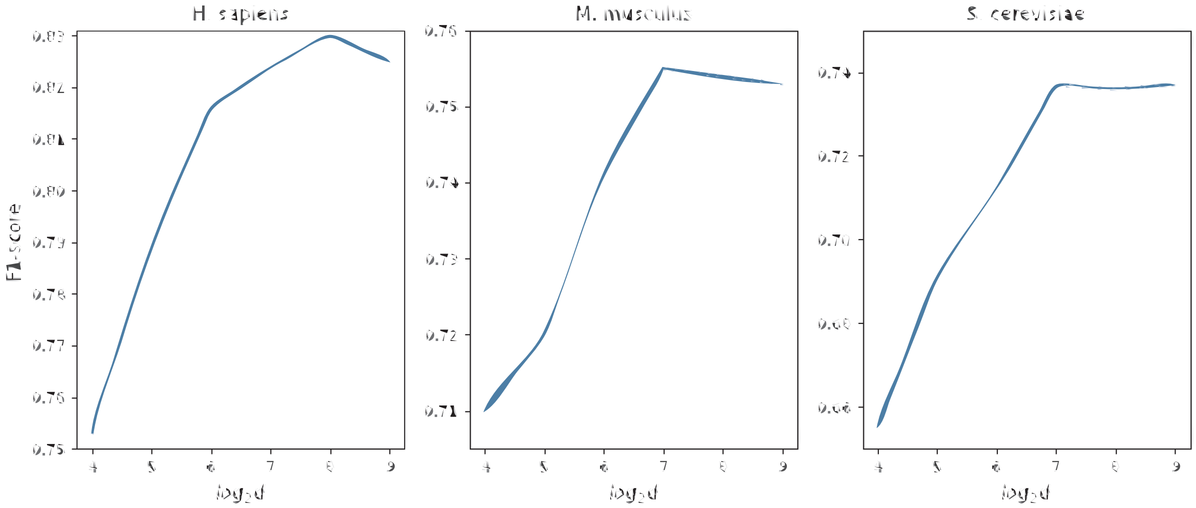

Fig. 7. Influence of the *d* parameter on the ANE-SITI

6 Conclusion

Predicting missing links in PPI networks is important for understanding cell activity. This paper presents an attributed network embedding method termed ANE-SITI that can efficiently predict missing links in incomplete and noisy PPI networks. To this end, the topological information of the network and the protein sequence information are combined, and the L3 principle with weights of edges is considered in topological information to increase the link prediction efficiency. L3 Principle states that the greater the number of paths of length 3 between two proteins, the greater the likelihood of a link between the two proteins. This information incorporation generates an enriched network. Then, sequences of nodes are created using a biased random walk on the generated network. These sequences are moved to the SGNS model to generate low-dimensional vectors of each protein. Finally, these vectors are used to predict the missing links between proteins. The experiments conducted in this study prove that ANE-SITI outperforms the compared link prediction methods in three real-world PPI networks. Future research may investigate the ANE-SITI method for linking prediction for additional biological interaction networks.

References

[1] Acar, E., Dunlavy, D.M. and Kolda, T.G., 2009, December. Link prediction on evolving data using matrix and tensor factorizations. In 2009 IEEE International conference on data mining workshops (pp. 262-269). IEEE.

[2] Adamic, L.A. and Adar, E., 2003. Friends and neighbors on the web. Social networks, 25(4.3), pp.211-230.

[3] Agrawal, M., Zitnik, M. and Leskovec, J., 2018. Large-scale analysis of disease pathways in the human interactome. In PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018: Proceedings of the Pacific Symposium (pp. 111-122).

[4] Ahmed, A., Shervashidze, N., Narayanamurthy, S., Josifovski, V. and Smola, A.J., 2013, May. Distributed large-scale natural graph factorization. In Proceedings of the 22nd international conference on World Wide Web (pp. 37-48).

[5] Al Hasan, M., Chaoji, V., Salem, S. and Zaki, M., 2006, April. Link prediction using supervised learning. In SDM06: workshop on link analysis, counter-terrorism and security (Vol. 30, pp. 798-805).

[6] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 25(**??**), pp.3389-3402.

[7] Barabasi, A.L. and Oltvai, Z.N., 2004. Network biology: understanding the cell's functional organization. Nature reviews genetics, 5(4.2), pp.101-113.

[8] Berg, R.V.D., Kipf, T.N. and Welling, M., 2017. Graph convolutional matrix completion. arXiv preprint arXiv:1706.02263.

[9] Cao, S., Lu, W. and Xu, Q., 2015, October. Grarep: Learning graph representations with global structural information. In Proceedings of the 24th ACM international conference on information and knowledge management (pp. 891-900).

[10] Cao, S., Lu, W. and Xu, Q., 2016, February. Deep neural networks for learning graph representations. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 30, No. 1).

[11] Chen, H., Yin, H., Sun, X., Chen, T., Gabrys, B. and Musial, K., 2020, August. Multi-level graph convolutional networks for cross-platform anchor link prediction. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 1503-1511).

[12] Cho, H., Berger, B. and Peng, J., 2016. Compact integration of multi-network topology for functional analysis of genes. Cell systems, 3(4.6), pp.540-548.

[13] Clauset, A., Moore, C. and Newman, M.E., 2008. Hierarchical structure and the prediction of missing links in networks. Nature, 453(7191), pp.98-101.

[14] Collins, F.S., Green, E.D., Guttmacher, A.E. and Guyer, M.S., 2003. A vision for the future of genomics research. nature, 422(6934), pp.835-847.

[15] Du, X., Sun, S., Hu, C., Yao, Y., Yan, Y. and Zhang, Y., 2017. DeepPPI: boosting prediction of protein-protein interactions with deep neural networks. Journal of chemical information and modeling, 57(4.6), pp.1499-1510.

[16] Eckart, C. and Young, G., 1936. The approximation of one matrix by another of lower rank. Psychometrika, 1(4.3), pp.211-218.

[17] Grover, A. and Leskovec, J., 2016, August. node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 855-864).

[18] Guimerà, R. and Sales-Pardo, M., 2009. Missing and spurious interactions and the reconstruction of complex networks. Proceedings of the National Academy of Sciences, 106(**??**), pp.22073-22078.

[19] Hamilton, W., Ying, Z. and Leskovec, J., 2017. Inductive representation learning on large graphs. Advances in neural information processing systems, 30.

[20] Hashemifar, S., Neyshabur, B., Khan, A.A. and Xu, J., 2018. Predicting protein-protein interactions through sequence-based deep learning. Bioinformatics, 34(**??**), pp.i802-i810.

[21] Hong, R., He, Y., Wu, L., Ge, Y. and Wu, X., 2019. Deep attributed network embedding by preserving structure and attribute information. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 51(4.3), pp.1434-1445.

[22] Huang, X., Li, J. and Hu, X., 2017, June. Accelerated attributed network embedding. In Proceedings of the 2017 SIAM international conference on data mining (pp. 633-641). Society for Industrial and Applied Mathematics.

[23] Jaccard, P., 1912. The distribution of the flora in the alpine zone. 1. New phytologist, 11(4.2), pp.37-50.

[24] Jeh, G. and Widom, J., 2002, July. Simrank: a measure of structural-context similarity. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 538-543).

[25] Katz, L., 1953. A new status index derived from sociometric analysis. Psychometrika, 18(4.1), pp.39-43.

[26] Kermani, A.G., Kamandi, A. and Moeini, A., 2022. Integrating graph structure information and node attributes to predict protein-protein interactions. Journal of Computational Science, 64, p.101837.

[27] Kingma, D.P. and Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

[28] Kipf, T.N. and Welling, M., 2016. Variational graph auto-encoders. arXiv preprint arXiv:1611.07308.

[29] Kovács, I.A., Luck, K., Spirohn, K., Wang, Y., Pollis, C., Schlabach, S., Bian, W., Kim, D.K., Kishore, N., Hao, T. and Calderwood, M.A., 2019. Network-based prediction of protein interactions. Nature communications, 10(4.1), pp.1-8.

[30] Kumar, A., Singh, S.S., Singh, K. and Biswas, B., 2020. Link prediction techniques, applications, and performance: A survey. Physica A: Statistical Mechanics and its Applications, 553, p.124289.

[31] Liao, L., He, X., Zhang, H. and Chua, T.S., 2018. Attributed social network embedding. IEEE Transactions on Knowledge and Data Engineering, 30(5.4), pp.2257-2270.

[32] Lim, J., Ryu, S., Park, K., Choe, Y.J., Ham, J. and Kim, W.Y., 2019. Predicting drug-target interaction using a novel graph neural network with 3D structure-embedded graph representation. Journal of chemical information and modeling, 59(5.1), pp.3981-3988.

[33] Lin, C., Cho, Y.R., Hwang, W.C., Pei, P. and Zhang, A., 2007. Clustering methods in protein-protein interaction network. Knowledge Discovery in Bioinformatics: techniques, methods, and application, pp.1-35.

[34] Liu, X., Yang, Z., Sang, S., Lin, H., Wang, J. and Xu, B., 2019. Detection of protein complexes from multiple protein interaction networks using graph embedding. Artificial Intelligence in Medicine, 96, pp.107-115.

[35] Lü, L. and Zhou, T., 2011. Link prediction in complex networks: A survey. Physica A: statistical mechanics and its applications, 390(4.6), pp.1150-1170.

[36] Lü, L., Jin, C.H. and Zhou, T., 2009. Similarity index based on local paths for link prediction of complex networks. Physical Review E, 80(4.4), p.046122.

[37] Martínez, V., Berzal, F. and Cubero, J.C., 2016. A survey of link prediction in complex networks. ACM computing surveys (CSUR), 49(4.4), pp.1-33.

[38] Menon, A.K. and Elkan, C., 2011, September. Link prediction via matrix factorization. In Joint european conference on machine learning and knowledge discovery in databases (pp. 437-452). Springer, Berlin, Heidelberg.

[39] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems 2013 (pp. 3111-3119). Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J. and Mei, Q., 2015, May. Line: Large-scale information network embedding. In Proceedings of the 24th international conference on world wide web (pp. 1067-1077).

[40] Moran, P.A., 1950. Notes on continuous stochastic phenomena. Biometrika, 37(1/2), pp.17-23.

[41] Muscoloni, A., Abdelhamid, I. and Cannistraci, C.V., 2018. Local-community network automata modeling based on length-three-paths for prediction of complex network structures in protein interactomes, food webs and more. bioRxiv, p.346916.

[42] Nasiri, E., Berahmand, K., Rostami, M. and Dabiri, M., 2021. A novel link prediction algorithm for protein-protein interaction networks by attributed graph embedding. Computers in Biology and Medicine, 137, p.104772.

[43] Newman, M.E., 2001. Clustering and preferential attachment in growing networks. Physical review E, 64(4.2), p.025102.

[44] Ou, M., Cui, P., Pei, J., Zhang, Z. and Zhu, W., 2016, August. Asymmetric transitivity preserving graph embedding. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1105-1114).

[45] Pan, S., Hu, R., Fung, S.F., Long, G., Jiang, J. and Zhang, C., 2019. Learning graph embedding with adversarial training methods. IEEE transactions on cybernetics, 50(4.6), pp.2475-2487.

[46] Pech, R., Hao, D., Lee, Y.L., Yuan, Y. and Zhou, T., 2019. Link prediction via linear optimization. Physica A: Statistical Mechanics and its Applications, 528, p.121319.

[47] Perozzi, B., Al-Rfou, R. and Skiena, S., 2014, August. Deepwalk: Online learning of social representations. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 701-710).

[48] Sharan, U. and Neville, J., 2008, December. Temporal-relational classifiers for prediction in evolving domains. In 2008 Eighth IEEE International Conference on Data Mining (pp. 540-549). IEEE.

[49] Strehl, A., Ghosh, J. and Mooney, R., 2000, July. Impact of similarity measures on web-page clustering. In Workshop on artificial intelligence for web search (AAAI 2000) (Vol. 58, p. 64).

[50] Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P. and Jensen, L.J., 2019. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic acids research, 47(D1), pp.D607-D613.

[51] Wang, D., Cui, P. and Zhu, W., 2016, August. Structural deep network embedding. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1225-1234).

[52] Wang, H., Wang, J., Wang, J., Zhao, M., Zhang, W., Zhang, F., Xie, X. and Guo, M., 2018, April. Graphgan: Graph representation learning with generative adversarial nets. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1). You, Z.H., Chan, K.C. and Hu, P., 2015. Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest. PloS one, 10(4.5), p.e0125811.

[53] Wang, P., Xu, B., Wu, Y. and Zhou, X., 2015. Link prediction in social networks: the state-of-the-art. Science China Information Sciences, 58(4.1), pp.1-38.

[54] Yang, C., Liu, Z., Zhao, D., Sun, M. and Chang, E., 2015, June. Network representation learning with rich text information. In Twenty-fourth international joint conference on artificial intelligence.

[55] Yang, F., Fan, K., Song, D. and Lin, H., 2020. Graph-based prediction of Protein-protein interactions with attributed signed graph embedding. BMC bioinformatics, 21(4.1), pp.1-16.

[56] Yang, K., Wang, R., Liu, G., Shu, Z., Wang, N., Zhang, R., Yu, J., Chen, J., Li, X. and Zhou, X., 2018. HerGePred: heterogeneous network embedding representation for disease gene prediction. IEEE journal of biomedical and health informatics, 23(4.4), pp.1805-1815.

[57] Yang, X., Yang, S., Li, Q., Wuchty, S. and Zhang, Z., 2020. Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. Computational and structural biotechnology journal, 18, pp.153-161.

[58] Yu, B., Chen, C., Wang, X., Yu, Z., Ma, A. and Liu, B., 2021. Prediction of protein-protein interactions based on elastic net and deep forest. Expert Systems with Applications, 176, p.114876.

[59] Yu, K., Chu, W., Yu, S., Tresp, V. and Xu, Z., 2006, December. Stochastic relational models for discriminative link prediction. In NIPS (Vol. 6, pp. 1553-1560).

[60] Yue, X., Wang, Z., Huang, J., Parthasarathy, S., Moosavinasab, S., Huang, Y., Lin, S.M., Zhang, W., Zhang, P. and Sun, H., 2020. Graph embedding on biomedical networks: methods, applications and evaluations. Bioinformatics, 36(4.4), pp.1241-1251.

[61] Zhang, L., Yu, G., Xia, D. and Wang, J., 2019. Protein-protein interactions prediction based on ensemble deep neural networks. Neurocomputing, 324, pp.10-19.

[62] Zhang, W., Chen, Y., Li, D. and Yue, X., 2018. Manifold regularized matrix factorization for drug-drug interaction prediction. Journal of biomedical informatics, 88, pp.90-97.

[63] Zitnik, M. and Leskovec, J., 2017. Predicting multicellular function through multilayer tissue networks. Bioinformatics, 33(**??**), pp.i190-i198.

[64] Zitnik, M., Agrawal, M. and Leskovec, J., 2018. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics, 34(**??**), pp.i457-i466.