ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)

https://doi.org/10.17559/TV-20180829122435
Original scientific paper

# Link Prediction Based on Extended Local Path Gain in Protein-Protein Interaction Network

Huiyan SUN, Yanchun LIANG, Yan WANG, Liang CHEN, Wei DU, Yuexu JIANG, Xiaohu SHI

**Abstract:** Protein–protein interaction (PPI) plays key role in each cellular process of any living cell, however, almost all organisms' PPIs are still incomplete. In this study, we firstly proposed a computational method Extended Local Path (ELP), which estimated links' existence likelihood by integrating all their neighbours' local paths in the network. In addition, on this basis, we extended it to Extended Local Path Gain (ELPG), which estimated gain effect when adding or deleting one potential link to the network. Applying both ELPG and ELP methods and other four recognized outstanding methods on four public PPI data of Yeast, *E. coli*, Fruit fly and Mouse, we demonstrated that ELPG and ELP obtained better performance under two standard measures: area under curve (AUC) and Precision. Besides, ELP and ELPG were identified as the best features for classifying existing and unknown links by using support vector machine-recursive feature elimination (SVM-RFE) for feature selection.

**Keywords:** Extended Local Path (ELP); Extended Local Path Gain (ELPG); Link prediction; Protein–Protein Interaction (PPI)

## 1 INTRODUCTION

Link prediction has increasingly become one of the most important problems to explore associations between nodes based on their attributes and observed links. Protein as the functional product of gene executes relevant biological function in a cell [1]. However, most cellular processes and biochemical events are ultimately achieved by interactions between proteins rather than single protein itself [2], and one certain protein would perform different functions through interaction with different proteins. Therefore, reliable identification of proteins interaction can contribute to drug design, disease diagnosis and medical treatments [3]. PPI data are organized into networks named PPI network, one of the most widely studied networks currently, in which nodes represent proteins and links between the nodes represent physical interactions between proteins.

Over the years, numerous PPIs have been identified in a variety of organisms using high-throughput techniques such as the yeast two-hybrid technique [4]. Other biology methods such as affinity chromatography, co-purification, co-immuno precipitation, and cross-linking are common to identify or predict PPIs [5]. Among all the organisms, Saccharomyces cerevisiae PPI has become the most widely studied in the past years [6], and as of now, a number of PPI networks have been constructed [4] as well as *E. coli* and *C. elegans* [7], which have been organized into several online PPI databases for the public, such as DIP [8], BioGRID [9], STRING [10] and MIPS [11]. However, for most other organisms, PPIs data are still far from complete, hence PPIs discovery in alternative approaches is an urgent need. Compared with biological experiments that are both time-consuming and laborious, computational prediction measures are superior in many aspects, especially could provide a candidate guideline and reduce workload if the prediction is accurate enough.

Some computational methods integrating biology data and mathematics models have been developed to predict PPIs. Aloy et al. [12] predicted PPI through proteins' tertiary structure, and McDowall et al. [13] predicted human PPI by combining information of biology expression, ortholog, domain co-occurrence, post-translational modifications and sub-cellular location. Gomez et al. [14] developed probabilistic models and Bock and Gough in 2001 [15] applied the support vector machine learning method for PPI prediction. Some other link prediction methods, which are widely used in social networks, like similarity index, have been gradually applied into PPI prediction [16, 17]. Compared with afore mentioned methods, these social networks based link prediction methods mainly focus on the topology characters of PPI network and rarely biology information.

Similarity index is the intuitionistic, easiest, and effective measure, and widely used in various methods. Among those, Common Neighbors (CN) method [18] could achieve high prediction accuracy, but it only emphasizes the number of common neighbors and ignores individual contributions. Adamic-Adar Index (AA) method [19] and Resource Allocation Index (RA) method [20] make up such a drawback, in which low-degree common neighbors are advocated by assigning more weight to them. Local Path Index (LP) method [21] takes consideration of local paths, with wider horizon than CN and makes a good tradeoff of accuracy and computational complexity.

In this paper, we firstly proposed a model named Extended Local Path (ELP), an improved LP method by adding the information of their neighbors' closeness for the target protein pairs in PPI network. On this basis, we added gain effect to ELP model, named ELPG. For each potential protein pairs, it scored the difference between original network status and the status after adding the target link. ELPG score could validly reflect the link power to the entire network. To assess the performance of our algorithm, we applied it on Saccharomyces cerevisiae (Yeast) PPI networks, Escherichia Coli (*E. coli*) PPI networks, Fruit fly PPI networks and Mouse PPI networks which were obtained from DIP database. Comparing with aforementioned methods CN, AA, RA and LP, both ELPG and ELP showed better performance under Precision and AUC measures.

In addition, we treated the score of each method as a feature and assessed their classification ability on existing links and unknown ones in the network. It is naturally expected that methods with good link prediction performance could have good classification ability as well. SVM-RFE, a feature selection method was applied, which obtained the weight of each feature and removed the one with the smallest weight iteratively, and finally got a ranking feature list. Based on the ranking list, we built a

classifier to classify exiting links and unknown ones, and it demonstrated that *ELP* and ELPG could achieve better classify performance than others.

## 2 MATERIALS AND METHOD

### 2.1 Date Source

All the PPI networks in this paper were downloaded from DIP database (release of Jan. 1st, 2015) which collects the report and experiment confirmed two-hybrid of protein interactions, as well as protein complexes from PDB (protein data bank) database. Now it is admittedly the simple and highly reliable PPI public database. DIP database contains two parts: DIP CORE and DIP FULL. The DIP CORE includes interactions confirmed at least by two high-throughput experimental methods. In this work, we chose Yeast, *E. coli*, Fruit fly and Mouse to estimate the performance of our method.

#### 2.1.1 Yeast and *E. coli*

Among all the species, Yeast is the most widely studied model organism at present and its PPI network is admittedly relatively complete and credible. And *E. coli* is another well-studied organism. After the self-interactions, repeated interactions and isolated sub-networks are removed, there are total 4467 interactions among 2036 proteins in Yeast core PPI network, 22060 interactions among 4978 proteins in Yeast full PPI network, total 1113 interactions among 683 proteins in *E. coli* core PPI network and 11516 interactions among 2537 proteins in *E. coli* full PPI network.

#### 2.1.2 Fruit Fly and Mouse

Further, we choose another two organisms, Fruit fly and Mouse, after tradeoff of available data size and complexity. Due to the core PPI network structures of Fruit fly is too sparse and scattered, and the core PPI and full PPI networks of Mouse are nearly identical, we just choose their full PPI network to assess the performance of our method. There are 1020 interactions among 690 proteins in Fruit fly and 741 interactions among 581 proteins in Mouse.

### 2.2 Similarity Index Method

Considering an undirected and unweight network $G<N, L>$, where $N$ is the set of nodes and $L$ is the set of links in the network, multiple links between two nodes and self-connections are removed. For a node $x$, let $N(x)$ denote the set of neighbors of $x$ in $G$ but not include node $x$ itself. Considering a pair of nodes, $x$ and $y$, which are not directly connected, similarity index methods are defined as:

**Common Neighbors** (*CN*).Two nodes are more likely to have a link if they have many common neighbors. The simplest measure of this neighborhood overlap is shown as

$$S_{xy}^{CB} = \left|N(x) \cap N(y)\right| \tag{1}$$

**Adamic-Adar Index** (*AA*).This index assigns the less-connected neighbors more weight and improves the common neighbors model, and it is defined as

$$S_{xy}^{AA} = \sum_{z \in N(x) \cap N(y)} \frac{1}{\log k_z} \tag{2}$$

where $k_x$ is the degree of node $x$.

**Resource Allocation Index** (*RA*). Common neighbors could transmit the information from node $x$ to target node $y$. In the process, each transmitter equally distributes its resource to all neighbors and the similarity between $x$ and $y$ would be defined as the amount of resource $y$ received from $x$, which is:

$$S_{xy}^{RA} = \sum_{z \in N(x) \cap N(y)} \frac{1}{k_z} \tag{3}$$

**Local Path Index** (*LP*). To provide a good tradeoff of accuracy and computational complexity, local path index was introduced with wider horizon than CN. It is defined as

$$S^{LP} = A^2 + \in A^3 \tag{4}$$

In Eq. (4), $A$ is the adjacent matrix of $G$, $\in$ is a free parameter, $LP(x, y)$ value is the element of the $x^{th}$ row and the $y^{th}$ column in matrix $S^{LP}$ and presents the possibility of potential interaction between node $x$ and node $y$. It has been proved that *LP* method would get the best performance when $\in$was assigned as 0.01 [21].

### 2.3 Extended Local Path (*ELP*)

In this paper, we proposed Extended Local Path (*ELP*) method which improved the *LP* similarity by adding more information to each node pair (x, y). *ELP* included three terms with the same weight, they were: *LP* similarity between node $x$ to each neighbor of node $y$, *LP* similarity between node $y$ to each neighbor of node $x$, and *LP* similarity between each neighbor of node y to each neighbor of node $x$ respectively, it was defined as:

$$ELP(x, y) = \sum_{\substack{u \in N(x), \\ v \in N(y)}} LP(x, v) + LP(y, u) + LP(u, v) \tag{5}$$

According to this definition, it is obvious that *ELP* additionally takes neighbors into consideration, not only the two nodes themselves. Compared with *LP* method, *ELP* increased calculation scope under the assumption that one interaction could be influenced by more factors.

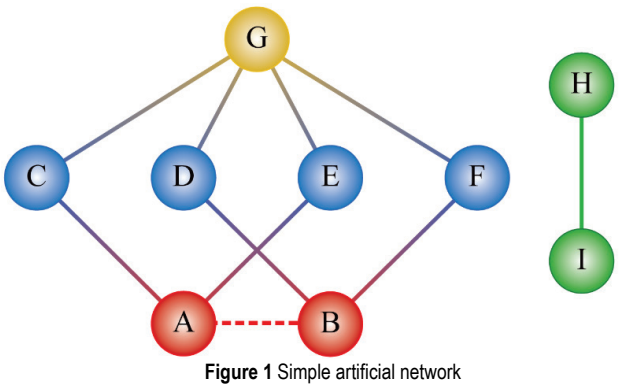


**Figure 1** Simple artificial network

Taking Fig. 1 as an example, when calculating the link existing probability between indirectly connected nodes $A$ and $B$ which are joined with the imaginary line, original $LP$ method (path=3) would score $LP(A, B) = 0$. However, as their neighbors connect frequently through the intermediary node $G$, it should exist a potential link intuitively.

### 2.4 Extended Local Path Gain ($ELPG$)

Suppose one link deletion would reduce or one link addition could increase much entire energy of the network, the link or potential link must play a significant role in the network, and the decrement (negative) and the increment (positive) are all defined as gain. In this paper, we used gain to measure the impact of a non-observed link when adding it to the network, defined as:

$$ELPG(x, y) = \\ = ELP(x, y|\text{status} = 1) - ELP(x, y|\text{status} = 0) \quad (6)$$

$ELP(x, y|\text{status} = 1)$ referred to the $ELP$ value of node pair $(x, y)$ after adding a link between them, $ELP(x, y|\text{status} = 0)$ represented the $ELP$ value of node pair $(x, y)$ which kept original indirectly connected status, and $ELPG$ was defined to measure the difference between $ELP(x, y|\text{status} = 1)$ and $ELP(x, y|\text{status} = 0)$. As ELP was obtained through their neighbors' relation and adding one link between two indirectly connected nodes would change neighbors' path characters, thus the change in different status was actually the neighbors' relation transformation under the effect of link. All above, Eq. (6) could reflect significance of the target nodes pair. Those nodes pairs with higher gain would be chosen, that is to say, they tended to connect, and their further connection would be more significant for the network.

## 3 RESULTS AND DISCUSSION

### 3.1 Data Preparation

Considering an undirected and unweight network $G<N, L>$, $U$ denoted the upper bound of links set containing all $|N|\cdot(|N| - 1)/2$ possible links in the complete connected network, where $|N|$ denoted the number of elements in set $N$, and the set of non-observed links was $U - L$.

#### 3.1.1 Train and Test Datasets

In order to evaluate our method, we separated data into two sets: train datasets for constructing method model and test datasets for assessing the prediction performance of the model.

In yeast and *E. coli* PPI network, we used their core PPI networks as train dataset, and ignored the links that belonged to full PPI networks but both of their vertexes do not appear in core PPI nodes set. The rest links in network are used for test dataset, that is, links in test dataset are the ones which belonged to full PPI networks and their vertexes included in core PPI, but the links themselves were not in core PPI network. The full PPI network has been defined as $G<N, L>$, in a similar way, the core full PPI network was represented as $G_{core}<N_{core}, L_{core}>$, where $N_{core}$ was the set of nodes and $L_{core}$ was the set of links in the core network. $L'$ was the set of links whose two nodes were all in $N_{core}$ and defined it as $L' = L\ [N_{core}]$. The test link set was represented as $L_{test} = L' - L_{core}$, as shown in the following Fig. 2.

Regarding Mouse and Fruit fly data sets, as mentioned in 2.1.2, we applied our method on their full PPI networks. To ensure all observations links were used for both training and testing, we performed 5-fold cross validation.

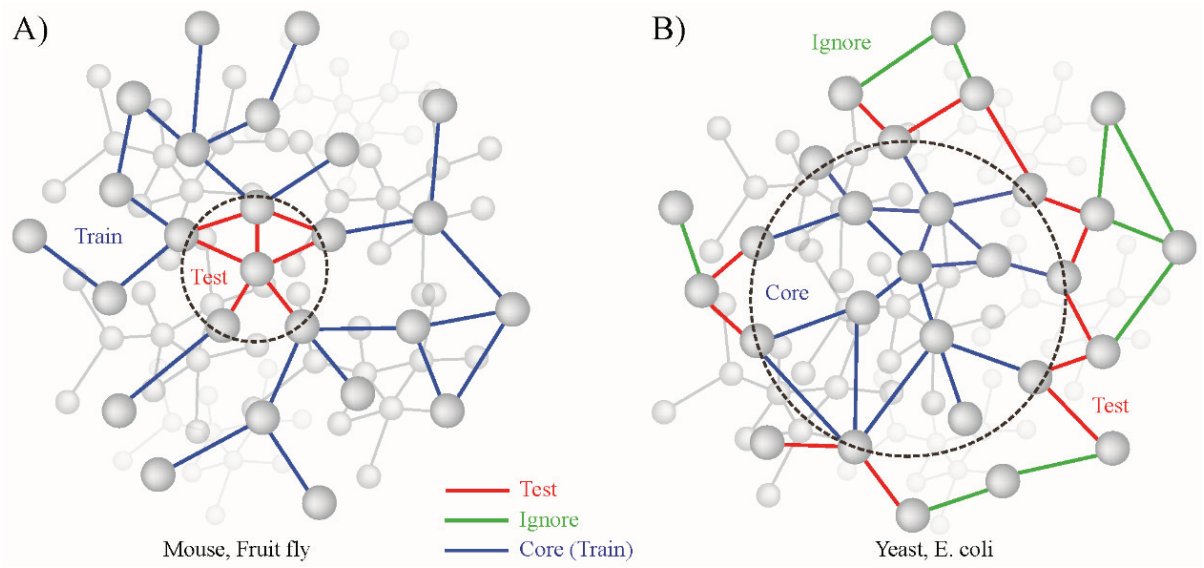


**Figure 2** Train and test dataset for four organisms

#### 3.1.2 Unknown Dataset

To evaluate our method, we analyzed the values of non-observed links and the links in test dataset which were known as truly existed. In Yeast and *E. coli* PPI network, the unknown dataset was $[U_{core} - L_{core} - L_{test}]$. $U_{core}$ was the links set containing all $|N_{core}|\cdot(|N_{core}| - 1)/2$ possible links in the core PPI network. For Mouse and Fruit fly, unknown datasets were $(U - L)$. Detail of each data set is shown in Tab. 1.

**Table 1** Data size of nodes, links, train links, test links, unknown links for four organisms

| | #Node | #Link | #Train_Link Set | #Test_Link Set | #Unknown_Link Set |
|---|---|---|---|---|---|
| Fruit fly | 690 | 1020 | 816 | 204 | 236685 |
| Mouse | 581 | 741 | 593 | 148 | 167749 |
| *E. coli* Core | 683 | 1113 | 1113 | 2964 | 228826 |
| Yeast_Core | 2036 | 4467 | 4467 | 6869 | 2060294 |

### 3.2 ELP Performance

ELPG was proposed based on the ELP model, so performance of ELP was the precondition for further evaluating ELPG. We applied precision curves to evaluate the overall performance of ELP and other similarity index methods. After sorting the value obtained by each method in descending order respectively, the top n links were selected as prospective ones, and then we examined the precision curves of each method in their own candidates set.

$$Precision(n) = TP(n) / (TP(n) + FP(n)) \tag{7}$$

where $TP(n)$ was the true positive which indicates the number of top n observed ordered links correctly predicted, and $FP(n)$ was the false positive which was the indirectly connected links incorrectly predicted as true links in the top n ordered links. For mouse and Fruit fly applying 5-fold cross validation, precision was the average value. As it would produce many zero values in CN, RA, AA methods and in order to avoid zero values disturbing evaluation, we selected top 1000 links which were almost with non-zero value in each method. The result was shown as Fig. 3.

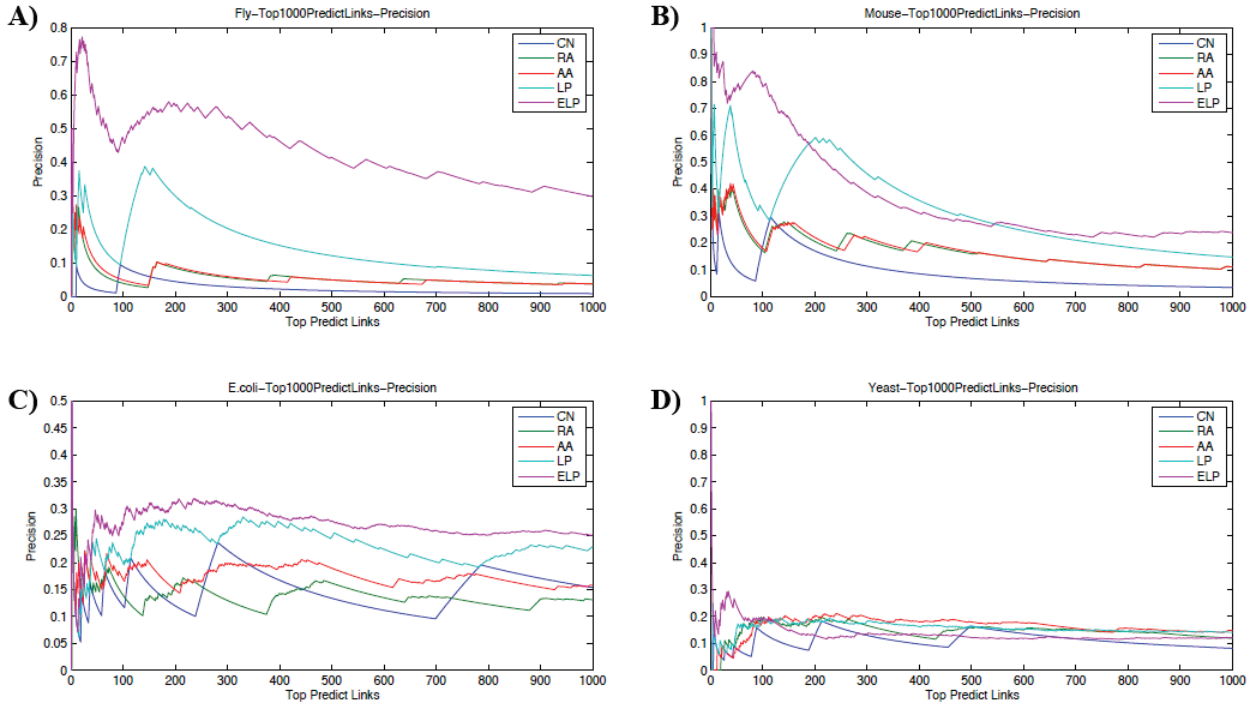


**Figure 3** Precision curve of each prediction method in Fruit fly, Mouse, Yeast and *E. coli*

Fig. 3, A), B), C) and D) described the precision curve of Fruit fly, Mouse, *E. coli* and Yeast respectively. In these sub-figures, x-axis presented the top n predicted links and y-axis was the precision. When compared with *CN*, *RA*, *AA*, and *LP* methods, *ELP* possessed an overall better performance in four organism PPI networks.

When applied into *E. coli* and Yeast, one reason for relative lower precision may be that the proportion between positive (test) and negative (unknown) links was extremely unbalanced. The other reason may be that the PPI network was still far from incomplete as we mentioned, which would minify test datasets and treat true positive links as false positive links.

### 3.3 ELPG Performance

Due to weakness of *CN*, *AA*, *RA* methods and unbalanced data as aforementioned, we used *AUC* [22] to compare and measure the performance of the following methods. The *AUC* evaluates the algorithm's performance according to all the links while the precision only focused on the n links with the top ranks or the highest scores. *AUC* value was approximate to Eq.(8): each time, we randomly picked a link in test set and another link in unknown set to compare their scores; if among n independent comparisons, there were $n'$ comparisons that the test link having a higher score and $n''$ they had the same score, the *AUC* value was

$$AUC = \frac{n' + 0.5n''}{n} \tag{8}$$

*AUC* value exceeding 0.5 indicated how much better the algorithm performed than random event. Here we assigned $n$ as $|test\ set| * |unknown\ set|$ and compared thoroughly. When applying 6 different methods on Fruit fly, Mouse, Yeast and *E. coli*, the performance was shown as follows:

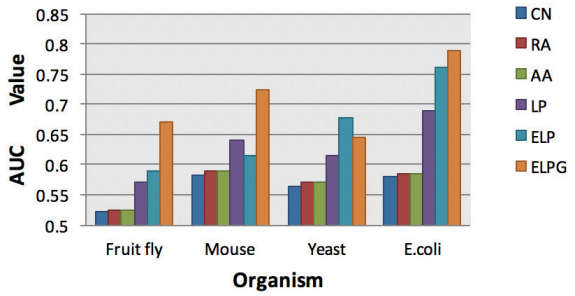


**Figure 4** AUC performance of each prediction method in Fruit fly, Mouse, Yeast and *E. coli*

In Fig. 4, it is obvious that when comparing with others, the overall performance of both *ELP* and *ELPG* was better. *ELPG* was more effective and had higher score than *ELP* in Fruit fly, Mouse, *E. coli* dataset. *ELP* under *AUC* measure was not good enough in Mouse dataset, but it ranked in the first place in Yeast dataset. It could be found that the *AUC* value of *CN*, *AA*, *RA* in all four organisms tended to be similar because of too many zero and repeating values which restricted evaluation ability, and *LP* was the relatively better method among *CN*, *AA*, *RA* and *LP*.

In addition, the superiority of *ELPG* does not only limit in *AUC* performance, but also significance of its wider calculation range. When analyzing PPI network, there were lots of isolate connect components which were difficult for other methods to integrate them into the network due to unusable common neighbor or infinity path, but *ELPG* could overcome such imperfection to some extent. Taking Fig. 1 as an example, by other methods, values of certain nodes pair between *H* or *I* and $\{A-G\}$ would be zero value while their *ELPG* value was non-zero, and (*G*, *H*) and (*G*, *I*) had the highest *ELPG* score. For Fruit fly and Mouse, we speculated that the network was probably spited into several small isolate parts with 5-fold cross validation as we have observed that these ELPG values for isolate parts tend to be much higher than the nodes connecting with hub node, which is consistent with previous viewpoints [23].

### 3.4 Weigh Each Measure Using SVM-RFE Method

When evaluating each method, its intrinsic ability for classifying existing links and unknown links is important. We used link scores obtained by each method as the features for classification and evaluated the classification performance of each feature (i.e. method) by SVM-RFE.

SVM-RFE is one of the most successful wrapper method based algorithms in the feature ranking. RFE is an iterative procedure for SVM classifier. Through calculating cost function *DJ*(*i*) as shown below, we iteratively removed the features with the smallest score. When there was only one feature left, the iterative process stoped, and finally it formed a ranking list with weights from high to low.

$$DJ(i) = \frac{1}{2}\frac{\partial^2 J}{\partial w_i^2}(Dw_i)^2 \tag{9}$$

This is explained by the Optimal Brain Damage (OBD) algorithm [24] in which a cost function *J* computed on training samples is used as an objective function.

As the links numbers were not the same in different organisms, we randomly selected different amount of links for different organisms to execute SVM-RFE process. In order to use the experimental data reasonably, we chose balanced links (samples), which contain an equal amount of existing links (positive samples) and unknown links (negative samples). The details are shown in table below.

We selected 5000 positive examples and 5000 negative examples in yeast, and 2000 positive examples and 2000 negative examples in *E. coli*. In Mouse, we selected 200 existing links and 200 unknown links, and in Fruit fly, we selected 500 existing links and 500 unknown ones for each cross validation. It was demonstrated that the rank lists were highly conserved as *ELP*>*ELPG*>*LP*>*CN*>*AA*>*RA* in 4 organisms.

**Table 2** Rank list obtained by SVM-RFE for four organisms

| | Data set size | Rank list-rank the weight of each method from high to low |
|---|---|---|
| Yeast | 10000 | *ELP ELPG LP CN AA RA* |
| *E. coli* | 4000 | *ELP ELPG LP CN AA RA* |
| Fruit fly | 1000 | *ELP ELPG LP CN AA RA* |
| Mouse | 400 | *ELP ELPG LP CN AA RA* |

## 4 CONCLUSIONS

In this paper, we proposed an *ELPG* method, which improved *LP* method by integrating neighbors' relationship of the target protein pairs, and focused on a link's impact on its surrounding proteins even the whole network. *ELP* possesses an overall better performance in precision, and it provides a good foundation for *ELPG* model. Both *ELP* and *ELPG* methods achieved better performance under Precision and *AUC* measure than the state-of-the-art methods when applied to Yeast, *E. coli*, Fruit fly and Mouse. Besides, *ELP* and *ELPG* were the best two features for classifying existing links and unknown links in these four organisms. We expect that the *ELPG* could serve as a useful tool not just for predicting PPIs of any organisms, but also relation recommendation in social network.

### Acknowledgement

The authors thank funding support from National Natural Science Foundation of China (61572228, 61572227, 61772227, 61872418), the Science-Technology Development Project from Jilin Province (20170520063JH, 20170101006JC, 20180101050JC, 20190201293JC), Development Project of Jilin Province of China (20180414012GH), Jilin Provincial Key Laboratory of Big Date Intelligent Computing (20180622002JC), the Zhuhai Premier Discipline Enhancement Scheme, and the Guangdong Premier Key-Discipline Enhancement Scheme.

## 5 REFERENCE

[1] Eisenberg, D., Marcotte, E. M., Xenarios, I., & Yeates, T. O. (2000). Protein function in the post-genomic era. *Nature, 405*(6788), 823-826. https://doi.org/10.1038/35015694

[2] von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., & Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature, 417*(6887), 399-403. https://doi.org/10.1038/nature750

[3] Archakov, A. I., Govorun, V. M., Dubanov, A. V., Ivanov, Y. D., Veselovsky, A. V., Lewi, P., & Janssen, P. (2003). Protein-

protein interactions as a target for drugs in proteomics. *Proteomics, 3*(4), 380-391. https://doi.org/10.1002/pmic.200390053

[4] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., & Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature, 403*(6770), 623-627. https://doi.org/10.1038/35001009

[5] Deng, M., Mehta, S., Sun, F., & Chen, T. (2002). Inferring domain-domain interactions from protein-protein interactions. *Genome Res, 12*(10), 1540-1548. https://doi.org/10.1101/gr.153002

[6] Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S., & Sakaki, Y. (2000). Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc Natl AcadSci U S A, 97*(3), 1143-1147. https://doi.org/10.1073/pnas.97.3.1143

[7] Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidal, P.-O. et al. (2004). A map of the interactome network of the metazoan C. elegans. *Science, 303*(5657), 540-543. https://doi.org/10.1126/science.1091403

[8] Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S. M., & Eisenberg, D. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res, 30*(1), 303-305. https://doi.org/10.1093/nar/30.1.303

[9] Stark, C., Breitkreutz, B. J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M. S., …, Tyers, M. (2011). The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res, 39*(Database issue), D698-704. https://doi.org/10.1093/nar/gkq1116

[10] Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., …, von Mering, C. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res, 39(Database issue), D561-568. https://doi.org/10.1093/nar/gkq973

[11] Mewes, H. W., Frishman, D., Mayer, K. F., Munsterkotter, M., Noubibou, O., Pagel, P., …, Stumpflen, V. (2006). MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res, 34*(Database issue), D169-172. https://doi.org/10.1093/nar/gkj148

[12] Aloy, P. & Russell, R. B. (2003). InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics, 19*(1), 161-162. https://doi.org/10.1093/bioinformatics/19.1.161

[13] McDowall, M. D., Scott, M. S., & Barton, G. J. (2009). PIPs: human protein-protein interaction prediction database. *Nucleic Acids Res, 37*(Database issue), D651-656. https://doi.org/10.1093/nar/gkn870

[14] Morimura, T., Fujita, K., Akita, M., Nagashima, M., & Satomi, A. (2008). The proton pump inhibitor inhibits cell growth and induces apoptosis in human hepatoblastoma. *PediatrSurgInt, 24*(10), 1087-1094. https://doi.org/10.1007/s00383-008-2229-2

[15] Bock, J. R. & Gough, D. A. (2001). Predicting protein--protein interactions from primary structure. *Bioinformatics, 17*(5), 455-460. https://doi.org/10.1093/bioinformatics/17.5.455

[16] Lu, L., Jin, C. H., & Zhou, T. (2009). Similarity index based on local paths for link prediction of complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys, 80*(4 Pt 2), 046122. https://doi.org/10.1103/PhysRevE.80.046122

[17] Lei, C. & Ruan, J. (2013). A novel link prediction algorithm for reconstructing protein-protein interaction networks by topological similarity. *Bioinformatics, 29*(3), 355-364. https://doi.org/10.1093/bioinformatics/bts688

[18] Newman, M. E. (2001). Clustering and preferential attachment in growing networks. *Phys Rev E Stat Nonlin Soft Matter Phys, 64*(2 Pt 2), 025102. https://doi.org/10.1103/PhysRevE.64.025102

[19] Adamic, L. A. & Adar, E. (2001). Friends and Neighbors on the Web. *Social Networks, 25*(3), 211-230. https://doi.org/10.1016/S0378-8733(03)00009-1

[20] Zhou, T., Lü, L., & Zhang, Y. C. (2009). Predicting missing links via local information. European Physical Journal B 71(4), 623-630. https://doi.org/10.1140/epjb/e2009-00335-8

[21] Lü, L. & Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications, 390*(6), 1150-1170. https://doi.org/10.1016/j.physa.2010.11.027

[22] Hanley, J. A. & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29-36. https://doi.org/10.1148/radiology.143.1.7063747

[23] Liu, Y. Y., Slotine, J. J., & Barabasi, A. L. (2011). Controllability of complex networks. *Nature, 473*(7346), 167-173. https://doi.org/10.1038/nature10011

[24] Lecun, Y., Denker, J., & Solla, S. (1989). Optimal Brain Damage. Advances in Neural Information Processing Systems. 2. 598-605.

**Contact information:**

**Huiyan SUN,** PhD
Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, #2699 Qianjin St, Changchun, 130012, China
sunhuiyan111@foxmail.com

**Yanchun LIANG,** PhD
Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, #2699 Qianjin St, Changchun, 130012, China

Zhuhai Laboratory of Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, Department of Computer Science and Technology, Zhuhai College of Jilin University,
Zhuhai, 519041, China
ycliang@jlu.edu.cn

**Yan WANG,** PhD
(Corresponding author)
Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, #2699 Qianjin St, Changchun, 130012, China
wy6868@jlu.edu.cn

**Liang CHEN,** PhD
University of Macau
Taipa, Macau S.A.R., 999078, China
liangliangabc123@163.com

**Wei DU,** PhD
Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, #2699 Qianjin St, Changchun, 130012, China
weidu@jlu.edu.cn

**Yuexu JIANG,** PhD
University of Missouri
Columbia, MO 65211, USA
bcsuperjiang@163.com

**Xiaohu SHI,** PhD
(Corresponding author)
Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, #2699 Qianjin St, Changchun, 130012, China
shixh@jlu.edu.cn