Intelligent Exploration of Negative Interaction from Protein-Protein Interaction Network and its Application in Healthcare

Sminu Izudheen1[1]*, Sheena Mathew[2]

[1]Department of Computer Science & Engineering, Rajagiri School of Engineering & Technology, Kakkanad, India
[2]Division of Computer Science, School of Engineering, Cochin University of Science & Engineering, Kerala, India
*sminu_i@rajagiritech.edu.in

ABSTRACT

Predicting protein-protein interactions (PPIs) has attracted much attention in recent years. Complexity of living systems depends on these interactions, as it controls healthy and diseased states in any organism. Even though recent advances in high throughput technologies have amplified PPI data repository; high level of noise, sparseness and skewed degree distribution of data has been a hindrance in making any useful findings from these data. Most of the works in this area concentrated on missing link prediction, and only very few explored the possibility of predicting negative links, or links that might disappear from the network. This paper proposes a method to predict these negative links from PPI network using an adaptive genetic algorithm, which is further optimized using Minimum Weak Edge-Edge Domination (WEED) set. The promising result obtained on MINT dataset asserts that the method can improve the quality of PPI data.

Keywords
Link Prediction: Genetic Algorithm: Protein-Protein Interaction: Domination Set: Weak Edge-Edge Domination Set.

Introduction

Protein interactions are vital for proper functioning of an organism. Even though they are vital, aberrant interactions may lead to disease like cancer. Hence research community is interested on these protein interactions; as more insight on these interactions may help of uncover the mystery behind many complex biological processes. But our knowledge about these interactions is incomplete, as identifying them through wet lab experiments are expensive [1], [2]. But experimental cost can be reduced by performing prediction on observed interactions, and focusing only on those predicted interactions [3]. This motivated towards link prediction which is one of major computational problem in this area.

Protein network is a very complex network as the link change dynamically. Due to this dynamicity these networks raises lot of challenges. Like, how long a pair of proteins will remain connected? What is the probability for two unconnected proteins to get connected in future? The challenge behind understanding these dynamics made link prediction an interesting problem for research community. While most of the existing works in these area concentrates on predicting links that might be added to the network; only very few addresses shrinking problem of predicting links that might drop from the network. This paper proposes an efficient method for predicting links that may get dropped from the network. The algorithm works in two phases. First phase consists of an adaptive genetic algorithm which predicts weak links. In the second phase, the result obtained is further optimized using Minimum Weak Edge-Edge Domination set [4] of the protein network. The results obtained assert that the method can be used as an effective method for link prediction in protein network.

Literature Review

Protein-protein interaction networks are considered to be one of the most intensively analyzed networks in biology. A large number of biochemical and biophysical methods exist to detect these interactions [5],[6]. Nowadays, there is a paradigm shift to graph theory techniques to study these interactions, as molecular biology techniques used are very expensive and time-consuming. A method to identify negative links from positive links predicted was presented by Wadhah Almansoori et. al.[7], which authors applied to health care and stock market models. Spurious and missing interactions were identified through a generative network model by Yuan Zhu et. al.[8]. It assumes that link exists between two proteins if they have higher propensities on one or more dominant latent factors in the generative network model. Olesksii et.al.[9] proposes a method to predict interactions using geometric graph model. It uses spectral decomposition to identify indirect neighbors in the network, which make it computationally expensive and intractable for large and incomplete PPI networks. Another drawback of this method is that, in addition to the confidence score generated by the algorithm, it uses GO annotation also to identify negative interactions. All the above methods, one way or other uses biological information also to predict the links. Since biological experimental methods are very time consuming and expensive, there is need for computational methods to predict protein-protein interactions with minimum prior knowledge from biological experiments. Yi et.al.[10] suggests a global geometric affinity method based on diffusion process to avoid the spectral decomposition problem in the method suggested by Olesksii et.al.[9]. Here, dot product between a pair of high-dimensional vectors represents the affinity (similarity).But, the main drawback of the method is that the probability based algorithm proposed to select the optimal propagation step which plays a critical role in fitting the PPI network into the geometric space is not well defined. A combination of common-neighborhood and distance-based method is used to improve the quality of a PPI network was proposed by Chengwei et.al.[11]. Yu Chen et.al.[12] presents “Sim”, a link prediction approach based on proteins‘ complementary interfaces and gene duplication. Huiyan Sun et.al. proposed Extended Local Path Gain (ELPG) method, which improved local path method by integrating neighbors’ relationship of the target protein pairs to identify target protein pairs in PPI network[13].

In this paper the spurious links in a protein network is predicted using only topological properties of the network. Here, the interactions which have a chance to get missed from a protein network were first predicted using Adaptive Genetic Algorithm and later optimized through all possible Minimum Weak Edge-Edge Domination (WEED) set of the network.

Methods and Data

Data

Protein-protein interaction data downloaded from MINT[14] database consisting of 187455 protein interactions among 12119 proteins was used for the present study. To learn more about the dataset degree distribution of all proteins were plotted. From Figure1(a) the skewness value of the dataset is 4.64, with degree of the proteins ranging from 1 to 600. Most of the proteins has small degree, with around 50 percent of them within degree 4. Average degree is about 20, but 75 percent of the nodes have a degree 12 or less. Although most nodes are with small degree, there are a few nodes with degree above 500, forming long tail in the distribution, probably form hubs in the network. These hub proteins with degree greater than majority of the proteins makes the degree distribution follows power law; and hence the scale free property of the PPI network.

As the interactions are very huge, random sampling was done without disturbing the degree distribution of the network. Performance of the algorithm was then verified using test data set generated from sampled data. All existing method prepares the data for simulation by randomly inserting interactions. This may disturb the statistics of the dataset. But here, care is taken to generate statistically significant data without disturbing the global structure of data. Assuming that the network follows a Gaussian distribution, a link was inserted into the network based on a Gaussian probability value. This is done by selecting p_{ref} random proteins from the sample data set. Let the set of all proteins within a given circumference from p_{ref} be represented by p_{cur}. Calculate the Mahalanobis distance d, from p_{ref} to p_{cur}. Then p = e^{-d} represents the probability of protein p_{cur} with respect to the reference protein p_{ref}. If there is no connection between p_{ref} and p_{cur} and the probabilty p is greater than a randon value, then a connections is established between two proteins. Degree distribution of the generated data set was compared with sampled data, to ensure that it follows the same pattern. If it shows different degree distribution than the original dataset, it is discarded and the process is repeated until both the dataset follows similar distribution. The degree distribution of the sampled dataset and the test data set is given in Figure 1(b). and Figure 1(c). respectively.

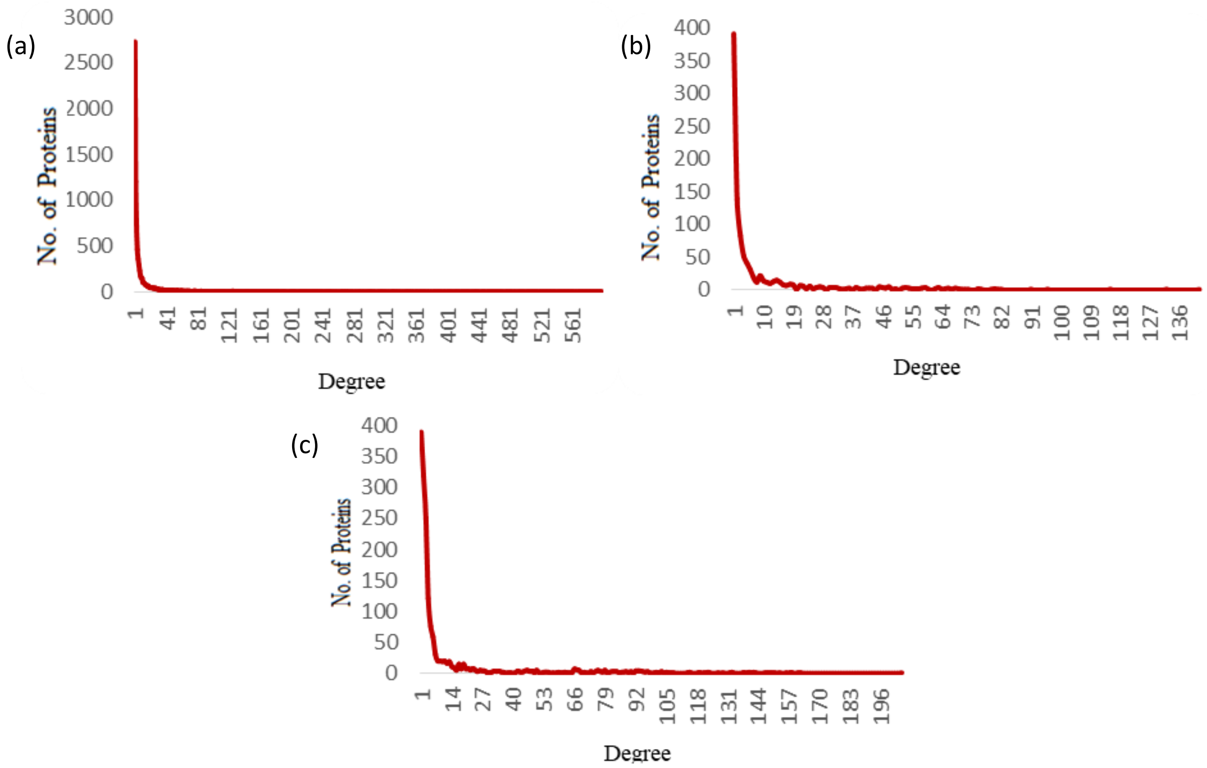


Figure 1(a). Degree distribution of MINT dataset (b). Degree distribution of sample dataset (c). Degree distribution of test dataset

Adaptive Genetic Algorithm

Consider an undirected graph G(V,E), with V and E as set of vertices and edges respectively. More similar two nodes are, more likely they are connected. In the proposed method missing links in a protein network were predicted using Adaptive Genetic Algorithm [15], with population size and chromosome length as the only parameters passed. As the algorithm automatically selectst the parameters selection, mutation, and crossover are all parameter-free.

Consider a population of N_P chromosomes, encoded by L locus, arranged as a population matrix A with N_P rows and L columns. Thus each row in matrix A represents one chromosome with a probability p ($0 < p < 1$) to undergo mutation. From matrix A, a mutation matrix M is formed with M_{ij}($i = 1, \dots, N_P$; $j = 1, \dots, L$) representing the mutation probability of A_{ij}. Arrange the chromosomes in the non-increasing order of their fitness. i.e., for $i < j$, $f_i \geq f_j$, where f_i and f_j represents the fitness value of a chromosomes i and j respectively. In mutation matrix, loci are arranged according to the standard deviation σ(j) of the allele distribution as given in Eq. (1), with V represening the number of possible alleles for a locus.

$$\sigma(j) = \sqrt{\frac{\sum_{i=1}^{N_P}(A_{ij} - \overline{h(j)})^2 \times C(f(i))}{\sum_{i=1}^{N_P} C(f(i))}} \qquad (1)$$

$$\overline{h(j)} = \frac{1}{N_P}\sum_{i=1}^{N_P} A_{ij}.$$

where (2)

Here, C(f(i)) represents fitness cumulative probability of chromosome i, which is directly proportional to the information gain as given in Eq. (3).

$$C(f(i)) = \frac{1}{N_P}\sum_{g \leq f} N(g). \qquad (3)$$

where N(g) is the number of chromosomes with fitness value g. In any chromosome, for any two loci i and j, information content in locus j will be greater than locus i, if σ(j) < σ(i). Hence standard deviation is used here to evaluate information content of a locus. Arrange the loci in the non-increasing order of standard deviation. i.e., for i < j, σ(j) < σ(i). And hence in mutation matrix more information is with loci located in higher rows and those closer to the right side. In each generation, based on the fitness cumulative probability chromosomes were selected for mutation. The probability for ith row to get selected for mutation is given in Eq. (4).

$$\alpha(i) = 1 - C(f(i)). \quad (4)$$

Therefore, more fitter a chromosome is, less chance to get mutatuted and hence high proabability to survive. The number of loci that undergo mutation N_{mg} will be as represented in Eq. (5).

$$N_{mg} = \alpha(i) \ x \ L . \quad (5)$$

Thus, in the selected chromosome, mutate Nmg leftmost loci which are actually the least informative loci of the chromosome. When the algorithm starts, the chromosomes are randomly generated and alleles are randomly distributed, hence the standard deviation will not be so informative. But as the algorithm progresses, chromosomes acquire good structure. The algorithm gives better performance than traditional genetic algorithm in terms of both speed and quality of solution because, using the above described fitness ranking and loci statistics it can determine M_{ij} dynamically.

Crossover is performed by calculating hamming distance between two chromosomes. Calculate an N_P x N_P distance matrix H with an element $H_{ii'}$ equal to the number of different alleles in ith and i'-th chromosome, which actually represents the distance between two chromosomes. Hence smaller difference between i and i', more similar they are. After defining the matrix H, chromosomes are selected for crossover. Probability to select the first chromosome is given in Eq. (6), which is the fitness cumulative probability.

$$P_{CI}(i) = C(f(i)) . \quad (6)$$

The probability to select the second chromosome depends upon the first chromosome already selected, and it is as in Eq. (7).

$$P_{CII}(i') = \frac{H_{ii'}}{\sum_{k=1}^{N_P} H_{ik}}. \quad (7)$$

If both the chromosomes are the same, the second one is chosen again from remaining chromosomes until both are different. After selecting two chromosomes, selective crossover [16] is preformed to maintain balance between exploration and exploitation. Selective crossover is based on the fact that even though chromosomes may only have a life span of one generation, its genetic unit lasts for many generations. It uses an extra dominance vector with every chromosome to accumulate knowledge about previous generations; also to promote prosperous genes during crossover to next generation. In selective crossover, the probability for performing crossover at a position depends on previous generation, which gives selective crossover an extra edge over uniform crossover where the probability is fixed. There is no limit in number of times a chromosome can participate in crossover during one generation.

One of the main concerns of any genetic algorithm is to maintain a proper balance between exploration and exploitation. To address this problem, a history table is created which stores all the chromosomes generated during various generations. It also maintains activation frequencies with respect to all individuals evaluated during evolution. Chromosomes for next generation are selected based on its novelty value, which is calculated as the reciprocal of its activation frequency in the history table. It helps to maintain the diversity among chromosomes in the population. In each generation, first carry out crossover and then perform mutation. After crossover and mutation fitness and cumulative fitness probability of all selected chromosomes is updated.

Link Prediction using Adaptive Genetic Algorithm and Minimum WEED set

In a PPI network, less similar the proteins are, the more chance that the interaction between them may get dropped in future. Here, these negative links can be predicted using Adaptive Genetic Algorithm. Algorithm uses a population matrix P where each row represents a chromosome. A single row can be considered as a consolidated form of an adjacency matrix. Given a graph G(V,E), an adjacency matrix, A is a binary matrix were for every pair of nodes i,j ∈ V, A_{ij} equal to 1 if node i is connected with node j, 0 otherwise. A single row in the population matrix is formed by concatenating n rows of the adjacency matrix A, where n represents cardinality of vertices in G. Hence xth entry in a row corresponds to (x div n) row and (x mod n) column in the adjacency matrix A, i.e., it represents the interaction between proteins (x div

n) and (x mod n). Algorithm begins with an initial population size of 24 chromosomes. First 5 chromosomes of the population matrix represents the link predicted using five standard algorithms, viz., Common Neighbors[17], Jaccard coefficient[18], Adamic Adar[19], Preferential Attachment [20] and Local Random Walk [21], while remaining 19 chromosomes are generated randomly. System is then trained to predict links by applying Adaptive Genetic Algorithm on the population matrix. Once the algorithm converges, better chromosome emerges with fittest chromosome in the topmost row. In any chromosome, loci are arranged in non-increasing order of standard deviation, or loci towards left represent least informative ones and corresponding links more likely to get dropped in future. For every loci x in the selected chromosome, a similarity score is calculated as

$$S_x = 1 - \sigma(x)\ . \qquad (8)$$

This gives the similarity score for the interaction between proteins (x div n) and (x mod n) in the selected chromosome. The interactions representing loci towards left will have low similarity score or those interactions are more likely to get dropped in future. As edges in the minimum WEED set represent weak connection in any network[22][23], the predicted result is further optimized by finding minimum WEED set of the graph. Schematic overview of the method is given in Figure 2.

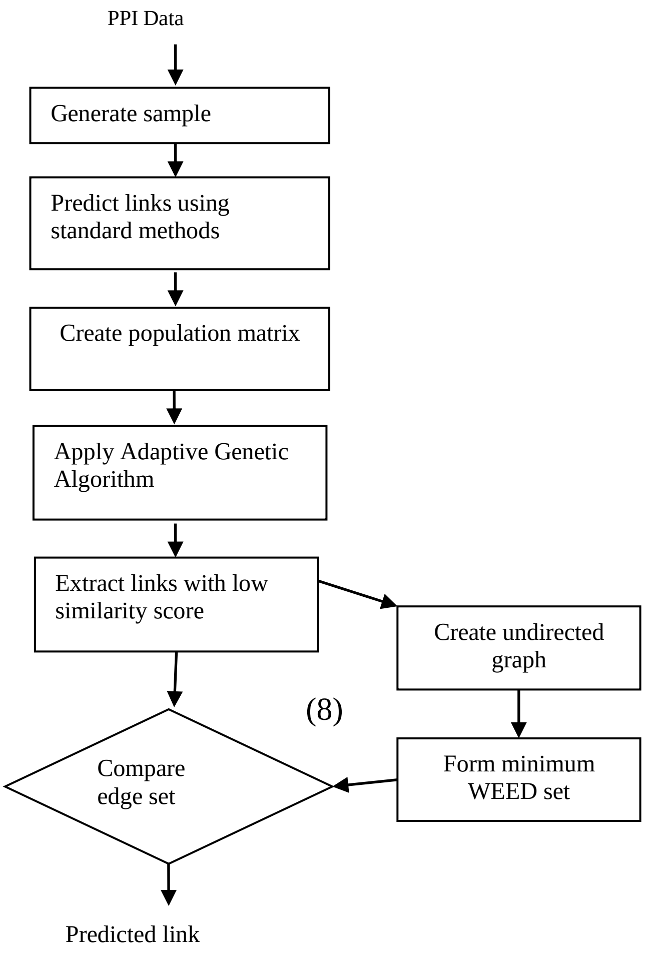


Figure 2. Schematic overview of Negative Link Prediction in Protein Networks

Results and Discussion

Probable links to get dropped in future is predicted using five standard methods viz., Common Neighbors, Jaccard coefficient, Adamic Adar, Preferential Attachment and Local Random Walk. Results of these methods represent five different chromosomes in the population matrix, while remaining 19 chromosomes are generated randomly. Adaptive Genetic Algorithm is applied on a population matrix of size 24 including chromosomes generated using standard methods. The links predicted from Adaptive Genetic Algorithm is further optimized by calculating all possible minimum WEED-set.

Reconstructed Network with WEED improves accuracy

To evaluate the effectiveness of WEED algorithm, experiments were conducted with different ratio of edges added according to Gaussian probability value. A comparison of result before and after applying WEED algorithm is given in Figure 3. From the figure it is clear that the WEED algorithm is able to reduce the false positive rate by 9.5 percent.

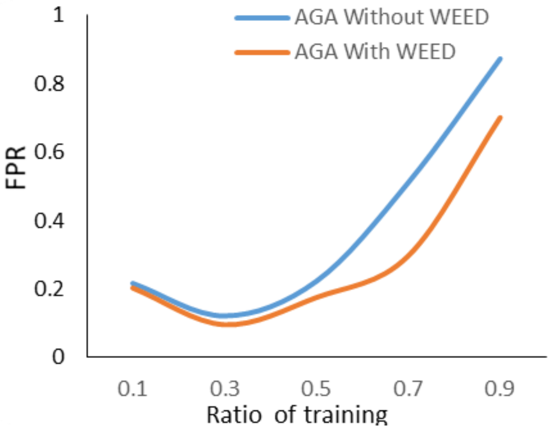


Figure 3. False Positive Rate on MINT dataset

Reconstructed Network has better functional relevance

The edges in the PPI network are divided into different groups. Edges which were present in the input network and the reconstructed network are called “before” and “after” groups respectively. The newly added edges in the reconstructed network are called “new” group and those which were present in the original dataset but removed from reconstructed network are called “removed” group. The edges which were present in both original and the reconstructed network are called “confirmed” group.

The functional relevance of the reconstructed PPI network is evaluated by comparing data obtained from essentiality of a gene in a protein complex. Since proteins interact with each other for proper biological activity, there is more chance for two interacting proteins to be in same protein complex. If they belong to same protein complex, there essentiality will also be same. i.e., if one gene is essential the other is also expected to be essential and if one is not essential, the other is also expected to be not essential. Similarly, essentiality of two non interacting proteins will be different. This essentiality difference is used as a measure to identify weak interactions from a PPI network.

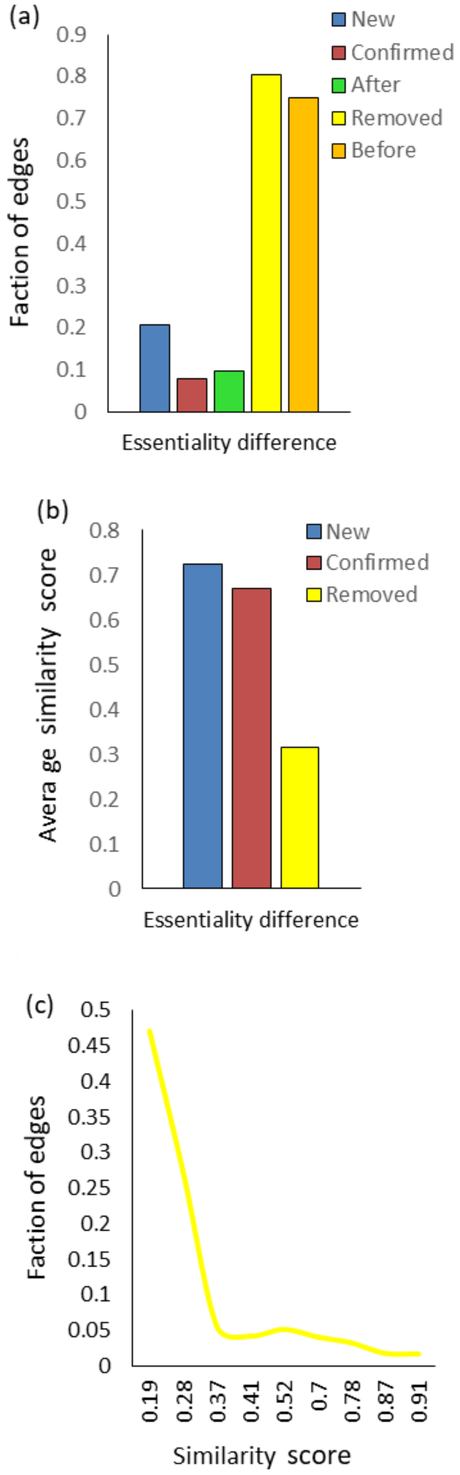


Figure 4. Result on MINT dataset. (a) Fraction of edges in different groups. (b) Average Similarity score of edges in different groups. (c) Distribution of edges with various similarity score in removed group.

To analyze the functional relevance of various groups, distribution based on essentiality difference was calculated. Figure. 4(a) shows that fraction of essentiality difference on confirmed, new, after, removed and before groups were 7.9%, 20.6%, 9.8%, 80.4% and 74.6% respectively. It may be noted that, highest difference is for removed group and it contribute the major portion of essentiality difference to the before group. Hence by removing interactions in removed group the essentiality difference in before group can be reduced to a great extent.

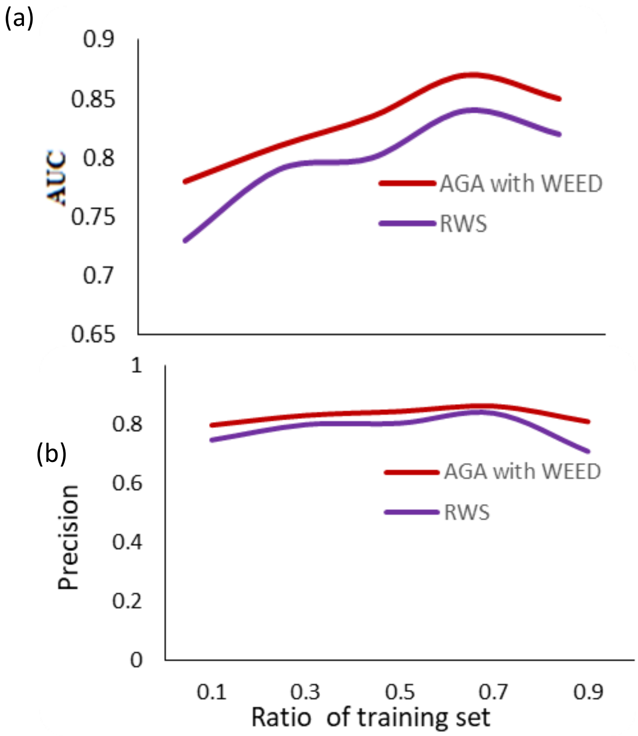
To evaluate the performance of prediction, average similarity of various groups based on essentiality difference is presented in Figure 4(b). While the average similarity score based in essentiality difference of confirmed and new edges were 0.72 and 0.67 respectively, similarity score of removed group is only 0.31. Hence weak interactions can be eliminated by targeting interactions with low similarity score.

Further investigation on similarity distribution shows that removed group has larger ratio of edges for smaller similarity values and smaller ratio for high similarity scores. From Figure 4(c) it is clear that, similarity score for removed group is less than 0.3 for 74% of edges. From all the above facts it can be concluded that the edges which were removed are less functionally relevant.

Comparison with Existing Methods

AUC and precision were calculated to quantify the accurarcy of the proposed method. Area Under Curve (AUC) is used to evaluate the general performance of the algorithm, while consistency in prediction is evaluated using precision, which is the ratio of number of relevant items to the number of selected items. A comparison of AUC and Precision values of AGA with Random]Walk with Resistance (RWS) [11] is given in Figure 5(a) and Figure 5(b) respectively. It may be noted from Figure 5(a) that, the probability for a randomly chosen spurious link to have lower score than a randomly chosen non existing link can be improved using Adaptive Genetic Algorithm. Similarly, Figure 5(b) shows that Adaptive Genetic Algorithm was able to predict an average of 82.9 % links while precision for the spurious links predicted using RWS method is only 78%.

Figure 5. (a) Dependence of Precision on ratio of dataset (b) Dependence of AUC on ratio of



dataset

Conclusion

Link prediction in protein network is an important problem as it is very helpful in analyzing and understanding cellular mechanisms. Most of the works in this area were limited to predicting links that may get added to the network during an interval of time. Very few works have explored the possibility of the links getting dropped in future. This paper proposes a method to predict negative links in protein network using Adaptive Genetic Algorithm. It also suggests a method to improve the performance through minimum Weak Edge-Edge Domination set of the network. The promising result obtained on MINT dataset asserts that proposed method can be an answer to the link prediction problem in protein networks.

References

[1] Neo D Martinez, Bradford A Hawkins, Hassan Ali Dawah and Brian P Feifarek (1993) Effects of sampling effort on

characterization of food-web structure. Ecology 80(3):1044-55.

[2] Sprinzak E., Sattath S. and Margalit H. (2003) How reliable are experimental protein-protein interaction data?. J. Molecular Biology,327(5):919-23.

[3] Aaron Clauset, Cristopher Moore and M. E. J. Newman (2008) Hierarchical structure and the prediction of missing links in networks. Nature, 453: 98-101.

[4] R. S. Bhat, S. S. Kamath, Surekha R. Bhat (2012) Strong (Weak) Edge-Edge Domination Number of a Graph. Applied Mathematical Sciences, 111(6), 2012: 5525 – 5531.

[5] T. Kocher and G. Superti-Furga (2007) Mass spectrometry based functional proteomics: from molecular machines to protein networks. Nature Methods, 4: 807-815.

[6] L. Liua, Y. Caic, W. Lua, K. Fenge, C. Penga and B. Niu (2009) Pre-diction of protein-protein interactions based on PseAA composition and hybrid feature selection, Biochemical and Biophysical Research Communications. 380: 318-322.

[7] Wadhah Almansoor, Shang Gao, Tamer M. Jarada, Reda Alhaj and Jon Rokne (2011) Link Prediction and Classification in Social Networks and its Application in Healthcare. IEEE IRI: 422-428.

[8] H Yuan Zhu, Xiao-Fei Zhang, Dao-Qing Dai, and Meng-Yun Wu (2013) Identifying Spurious Interactions and Predicting Missing Interactions in the Protein-Protein Interaction Networks via a Generative Network Model. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1(10).

[9] Oleksii Kuchaiev, Marija Rasˇajski, Desmond J. Higham and Natasˇa Przˇulj (2009) Geometric De-noising of PPI Networks. PLoS Computational Biology 5(8).

[10] Yi Fang, William Benjamin, Mengtian Sun and Karthik Ramani (2011) Global Geometric Affinity for Revealing High Fidelity Protein Interaction Network. PLoS 6(5).

[11] Chengwei Lei and Jianhua (2013) A novel link prediction algorithm for reconstructing protein-protein interaction networks by topological similarity. Bioinformatics, 29(3): 355-64.

[12] Yu Chen, Wei Wang, Jiale Liu, Jinping Feng and Xinqi Gong, "Protein Interface Complementarity and Gene Duplication Improve Link Prediction of Protein-Protein Interaction Network", Frontiers in Genetics, 02 April 2020.

[13] Huiyan Sun, Yanchun Liang, Yan Wang, Liang Chen, Wei Du, Yuexu Jiang, Xiaohu Shi, "Link Prediction Based on Extended Local Path Gain in Protein-Protein Interaction Network", Tehnički vjesnik 26, 1(2019), 177-182.

[14] MINT: the Molecular INTeraction database, http://mint.bio.uniroma2.it/mint/Welcome.do

[15] Nga Lam Law and K. Y. Szeto, Adaptive Genetic Algorithm with Mutation and Crossover Matrices, Proceedings of the 20th International Joint Conference on Artifical intelligence IJCAI'07, pp 2330-2333, 2007.

[16] Kanta Vekaria and Chris Clack (1998) Selective Crossover in Genetic Algorithms: An empirical study, Proceding of 5th International Conference Amsterdam, 1498: 438-447.

[17] F. Lorrain, H. C. White (1971) Structural equivalence of individuals in social networks. Journal of Mathematical Sociology,1: 49-80.

[18] P. Jaccard (1901) Etude de la distribution florale dans une portion des Alpes et du Jura. Bulletin de la Societe Vaudoise des Science Naturelles, 37(142): 547-79.

[19] L. A. Adamic, E. Adar (2003) Friends and neighbors on the Web. Social Networks, 25: 211-30.

[20] Albert Laszlo Barabasi, Rekha Albert (1999) Emergence of Scaling in Random Networks. Science, 286: 509 - 512.

[21] Weiping Liu and Linyuan LU (2010) Link Prediction Using Local Random Walk, Europhys. Letter 89,5.

[22] Sminu Izdheen and Sheena Mathew (2016) Identifying Negative Interactions in Protein-Protein Interaction Network Using Weak Edge-edge

Domination Set, Procedia Technology, 24: 1423-1430.

[23] Razika Boutrig and Mustapha Chellali (2012) A note on relation between the Weak and Strong Domination Numbers of a graph. Opuscula Mathematica, 2 (32): 235-238.