Link Prediction in Protein Networks

KEYWORDS Link Prediction, Local Random Walk, Protein Protein Interactions

Sminu Izudheen
Department of Computer Science & Engineering,
Rajagiri School of Engineering & Technology, Kerala

Sheena Mathew
Division of Computer Science, School of Engineering,
Cochin University of Science & Technology, Kerala

ABSTRACT *Protein-protein interactions play a vital role in identifying the outcome of a vast majority of cellular mechanisms. Complexity of living systems arises as a result of these interactions. Predicting protein interactions has attracted much attention in recent years. The main difficulty for link prediction in protein network is the huge size of the network. This paper proposes an efficient method for link prediction in a weighted protein network using local random walk. The result shows that the method give better prediction than other random-walk based methods.*

INTRODUCTION

Protein interactions are important for numerous biological functions. For example, signal transduction, the process by which signals from exterior of a cell is mediated to interior of the cell is controlled by protein-protein interaction(PPI) of the signaling molecules. This plays a fundamental role in many biological processes and in many diseases like cancers. Several efforts have been made to identify protein interactions, so that biological systems can be understood better. The cost for experimentally detecting physically interaction between proteins in laboratory is very high and hence our current knowledge about protein networks is substantially incomplete [1,2]. Instead of blindly checking all the possible interactions, predictions based on the observed interactions and focusing on links most likely to exist can sharply reduce the experimental costs [3]. This motivated us towards link prediction which is one of major computational problem in this area.

Protein network is a complex network with proteins as nodes and their interactions as links. Protein networks are very dynamic objects, since new edges and vertices are added to the graph over the time. Understanding the dynamics that drives the evolution of protein network is a complex problem due to a large number of variable parameters. But, understanding the association between two specific nodes is comparatively an easier problem. One such problem is addressed in this paper, the Link Prediction problem. Given a protein network, link prediction problem is predicting the protein interactions that probably appear in future. It is one of the main approaches to understand the dynamics of a protein network. But designing an efficient and effective algorithm is a main challenge in link prediction. In this paper we propose an efficient method for link prediction using local random walk. We tried to improve the existing method on link prediction using local random walk by including protein interaction rate also while calculating the similarity index. The results assert that our method can be used as an effective method for link prediction in protein network.

RELATED WORK

Protein-protein interactions (PPIs) are one of the most intensely analyzed networks in biology and there are a multitude of biochemical and biophysical methods to detect them [4,5]. Since molecular biology techniques used are very expensive and time-consuming, researchers depend on graph theory techniques to study them.

Nantia Iakovidou, Panagiotis Symeonidis and Yannis Manolopoulos[6] uses a multiway spectral clustering analysis, a technique that uses information obtained from the top few eigen vectors and eigenvalues of the normalized laplacian matrix as a method to predict links in PPI network. W. Pentney and M. Meila[7] prove that their algorithm applying spectral clustering offers competitive performance on sequence data. A simple and unified derivation of spectral clustering of biological data is presented in [8]. A tool for the identification of PPIs, which can be used to detect interactions across the entire proteome of an organism is given in [9]. Local Protein Community Finder is a tool developed by authors on [10] to find community close to a queried protein in any network specified by the user. To predict protein interactions in yeast network Y. Yamanishi, J.P. Vert and M.Kanehisa[11] introduced a method based on variant of kernel canonical correlation analysis.

Link prediction has also attracted researchers from the area of social networking. Commonly, two nodes are more likely to be connected if they are more similar. A Comparison between similarity indices is presented in [12], where node-dependent indices like Common Neighbors[13], Jaccard coefficient [14], Adamic-Adar Index [15], Preferential Attachment [16] and path-dependent indices like Katz Index [17],Hitting Time [18], Commute Time [19], Rooted PageRank [20], SimRank [21] and Blondel Index [22] were considered. T. Zhou, L. L¨u and Y.-C. Zhang[23] proposed Resource Allocation index and Local Path index as a measure to compare two nodes. Results shows that the local path index provides much accurate prediction compared with the global index[24]. On a weighted network, weak links play an important role than strong links[25]. Likelihood for the existence of a link between two nodes was estimated through local path index in [26].

Weiping Liu and Linyuan LU[27] present a method to find node similarity based on local random walk. They illustrate that the method has lower computational complexity compared with other random-walk-based similarity indices, such as average commute time (ACT) and random walk with restart (RWR). In this paper we propose an improvement over the local random walk by considering protein interaction rate to calculate the similarity index. We defined a weighted protein network with protein interaction rate as the weight of the link. We calculated local random walk on this protein network. We compare our result with the existing local random walk, and the result shows that link prediction in protein networks can be improved by including interaction rate.

METHOD AND DATA

- **Data**

PPI data for the work consisting of 187455 interactions among 12119 proteins was extracted from MINT database. Since the number of interactions was very huge, we sample the data. We randomly selected k interactions and identified the proteins involved in those k interactions. From the main set of 187455 interactions, we generate a subset containing all interactions of the above identified proteins.

- **Link Prediction Based on Local Random Walk**

Consider an undirected weighted protein network G(V,E), where V is the set of proteins and E is set of links which represents the interaction between proteins. Here protein interaction rate is represented as the weight of the link. Two nodes are more likely to be connected if they are similar. In the proposed method a similarity score is generated based on Local Random Walk(LRW). Our method is an improvement over existing Link Prediction algorithm using Local Random Walk[27], in that here we generate the score based on Protein Interaction rate. To predict the missing links, sort the non existing links in the descending order based on the score generated. The links which are in the top of the list are likely to exist.

Random walk is a path consisting of sequence of random steps. The probability that a random walker starting at node x will move to y in the next step is represented by transition probability matrix , P, with Pxy = a_{xy}/k_x, where a_{xy} equals 1 if node x and node y are connected, 0 otherwise, and k_x denotes the degree of node x. The probability that a random walker located at node x will be located at node y after t steps is given by

$$\pi_x(t) = P^T\ \pi_x(t-1) \qquad (1)$$

where $\pi_x(0)$ is an Nx1 matrix with x = 1 and all other values are 0's and T is the transpose matrix. The similarity between node x and node y is given by

$$s_{xy}{}^{LRW}(t) = \frac{kx \quad kx}{2|E|2|E|}\ .\ \pi_{xy}(t) + \frac{ky \quad ky}{2|E|2|E|}\ .\pi_{yx}(t) \qquad (2)$$

where |E| represents the number of links in the network.

Random walk based similarity measures are more sensitive to nodes far away from target nodes[28]. Hence, the probability for the random walker to go farther from x and y, even though they are close to each other is high. Since proteins have a tendency to connect with ones nearby rather than far way, this may lead to low prediction accuracy. To solve this problem we can continuously release the walkers at the starting point. By superposing the contribution of each walker, we get the next similarity index, Superposed Random Walk (SRW)

$$s_{xy}{}^{SRW}(t) = \sum_{l=1}^{t} \sum_{l=1}^{t}\ s_{xy}{}^{LRW}(l) \qquad (3)$$

- **Metrics**

To validate the result we use a supervised training method by which the selected protein interactions are divided into mutually exclusive train set and test set. We have done experiments with different ratio of train set and test set. The overall ranking precision of the algorithm is measured by plotting Area Under Curve (AUC). To measure the prediction accuracy of the top n predictions another standard metric, precision curve is also plotted. For this we rank all the non existing links in the decreasing order of their scores, and the top n links are used for prediction. We have verified the algorithm for different values of n.

RESULT AND DISCUSSION

From the protein-protein interaction data a protein network is created and represented as an adjacency matrix. In this paper we propose an improvement over the existing method by including protein interaction rate while calculating the LRW and SRW indices. From the results obtained the following observations are noted.

Fig.1 shows the AUC calculated for different similarity indices. AUC is used to evaluate the overall ranking of the algorithm. It may be noted that Local Random Walk with Interaction Rate (LRW with IR) gives better result that Local Random Walk without Interaction Rate (LRW without IR). Similarly Superposed Random Walk with Interaction Rate (SRW with IR) outperforms Superposed Random Walk without Interaction Rate (SRW without IR). Or we can say that the probability for a randomly chosen missing link to have higher score than a randomly chosen non existing link is high when we include protein interaction rate to calculate index.

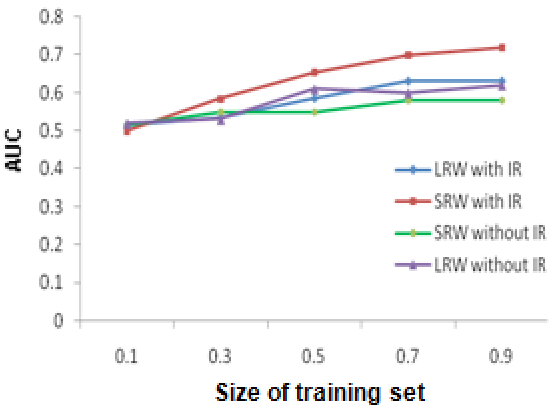

Fig. 1: Dependence of AUC on size of training set.

In order to evaluate the accuracy in predicting missing links we have calculated precision. Precision is the ratio of number of relevant items to the number of selected items. Hence we ranked all non existing links in descending order of their score and selected top n links. Fig. 2 shows the precision curve for various indices when the value of n is 100. In is clear from the figure that for both LRW and SRW, our method outperform the traditional method of index without protein interaction rate.

To demonstrate the tradeoff between sensitivity and specificity Receiver Operating Characteristic (ROC) curve was plotted. Figure 3 shows the ROC for various indices with training set containing 70% of the known links. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. From the figure it is clear that for both SRW and LRW indices our method do better than the traditional method.

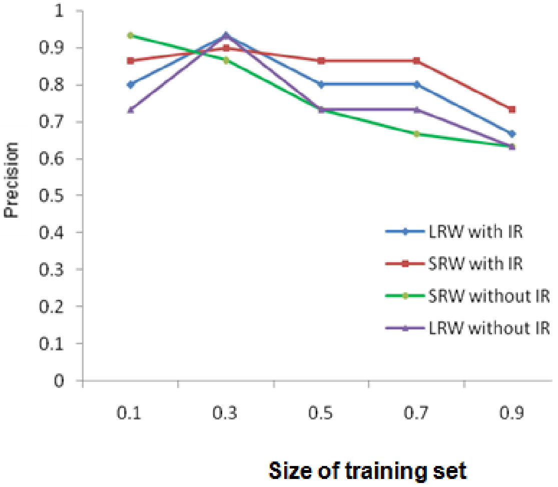

Fig. 2: Dependence of Precision on size of training set.

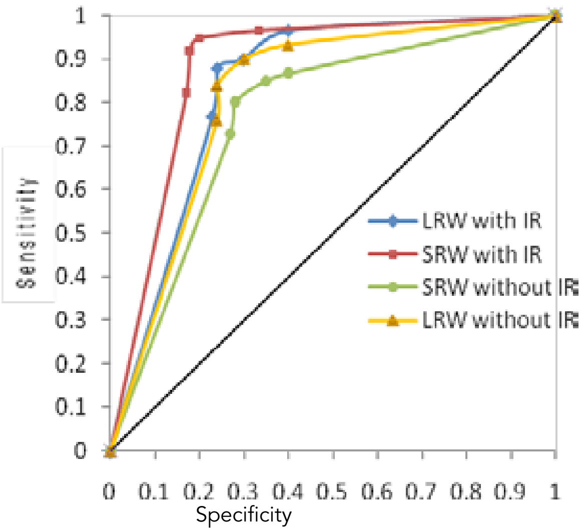

Fig. 3: ROC Curve.

CONCLUSIONS

Link prediction in protein network is an important problem and it is very helpful in analyzing and understanding cellular mechanisms. Such understanding can lead to efficient implementation of tools to identify hidden communities or to find missing members within communities. Through this work, we had shown that, link prediction in protein network can be solved with a very high accuracy by including protein interaction rate while calculating Local Random Walk (LRW) and Superposed Random Walk (SRW). We compared our method with traditional way of calculating LRW and SRW. Results show that our method provides more accurate and competitive result than traditional method.

REFERENCE

[1] N. D. Martinez, B. A. Hawkins, H. A. Dawah, B. P. Feifarek, Ecology 80, 1044 (1999). | [2] E. Sprinzak, S. Sattath, H. Margalit, J. Mol. Biol. 327, 919(2003). | [3] A. Clauset, C. Moore, M. E. J. Newman, Nature 453, 98(2008). | [4] T. Kocher and G. Superti-Furga, "Mass spectrometry based functional proteomics: from molecular machines to protein networks", Nature Methods, vol. 4, 2007, pp 807-815. | [5] L. Liua, Y. Caic, W. Lua, K. Fenge, C. Penga and B. Niu, "Pre-diction of protein-protein interactions based on PseAA composition and hybrid feature selection", Biochemical and Biophysical Research Communications, vol. 380, 2009, pp 318-322. | [6] Nantia Iakovidou, Panagiotis Symeonidis and Yannis Manolopoulos," Multiway Spectral Clustering Link Prediction in Protein-Protein Interaction Networks". | [7] W. Pentney and M. Meila, "Spectral clustering of biological sequence data", in the 12th National Conference on Artificial Intelligence, Pittsburgh, Pennsylvania, 2005, pp. 845-850. | [8] D. J. Higham, G. Kalna and M. Kibble, "Spectral clustering and its use in bioinformatics", Journal of Computational and Applied Mathematics, vol. 204, 2007, pp 25-37. | [9] U. Stelzl, U. Worm, M. Lalowski, et al. "A human protein-protein interaction network: a resource for annotating the proteome", Elsevier, vol. 122, 2005, pp 957-968. | [10] K. Voevodski, S. Teng and Y. Xia, "Finding local communities in protein networks", BMC Bioinformatics, vol. 10, 2009, pp 297-310. | [11] Y. Yamanishi, J.P. Vert and M.Kanehisa, "Protein network inference from multiple genomic data: a supervised approach" Bioinformatics, vol. 20, 2004, pp i363-i370. | [12] D. Liben-Nowell, J. Kleinberg, J. Am. Soc. Inf. Sci. &.Technol. 58, 1019 (2007). | [13] F. Lorrain, H. C. White, J. Math. Sociol. 1, 49 (1971). | [14] P. Jaccard, Bulletin de la Societe Vaudoise des ScienceNaturelles 37, 547 (1901). | [15] L. A. Adamic, E. Adar, Soc. Netw. 25, 211 (2003). | [16] A.-L. Barab´asi, R. Albert, Science 286, 509 (1999). | [17] L. Katz, Psychmetrika 18, 39 (1953). | [18] F. Gobel, A. Jagers, Stochastic Processes and Their Applications 2, 311 (1974). | [19] F. Fouss, A. Pirotte, J.-M. Renders, M. Saerens, IEEE Trans. Knowl. Data Eng. 19, 355 (2007). | [20] S. Brin, L. Page, Comput. Netw. ISDN Syst. 30, 107(1998). | [21] G. Jeh, J. Widom, SimRank: A Measure of Structural-Context Similarity, in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM Press, New York, 2002). | [22] V. D. Blondel, A. Gajardo, M. Heymans, P. Senellart, P.V. Dooren, SIAM Rev. 46, 647 (2004). | [23] T. Zhou, L. L¨u, Y.-C. Zhang, Eur. Phys. J. B 71, 623(2009). | [24] L. L¨u, C.-H. Jin, T. Zhou, Phys. Rev. E 80, 046122 (2009). | [25] L. L¨u, T. Zhou, Europhys. Lett. 89, 18001 (2010). | [26] L. Lu, C. Jin and T. Zhou, "Similarity index based on local paths for link prediction of complex networks" Physical Review E, vol. 80, 2009, pp 1-9. | [27] Weiping Liu and Linyuan LU, Link Prediction Using Local Random Walk. | [28] D. Liben-Nowell, J. Kleinberg, J. Am. Soc. Inf. Sci. &.Technol. 58, 1019 (2007). |