Link Prediction in Protein-Protein Networks: Survey

Manu Kurakar[1], Sminu Izudheen[2]
[1](Department of computer science, Rajagiri School of Science and Technology-Kochi, India)
[2](Department of computer science, Rajagiri School of Science and Technology-Kochi, India)

ABSTRACT : *Protein networks have a great importance in biological activities. Protein-Protein interaction occurs when two or more proteins interact together to carry out some biological activities. For example signals from the exterior of a cell are mediated to the interior through these interactions. Identification of these interaction have a great significance in understanding complex diseases and also for designing drugs. With the availability of huge biological data, computational biology is at position such that, it can predict missing protein protein interactions. Here, this article summarizes technologies for missing link prediction.*

Keywords *- Link Prediction, Protein Networks, sequence similarity, clustering, interactions*

I. INTRODUCTION

Many real world information can be better represented asnetworks, where nodes represent entities and edges represents the relationship between the entities. The study of complex networks is there for a common interest of various branches of science. Consider the case of biological networks, large amount of data is available about protein protein interactions. These protein protein interactions have great importance in understanding biological activities, analyzing complex disease and also for designing new drugs for diseases. Protein protein interaction occurs when two or more protein bind together to perform certain biological functions. Interactions between proteins are important for biological functions. For example, signals from the exterior of the cells are mediated to the interior by protein protein interaction of the signaling molecules. Identification of these interactions through clinical study includes very complex procedures. Yeast two-hybrid screening and affinity capture mass spectrometry are two important methods for determining these interactions.

With the availability of large amount of biological data, various computational models are there to predict missing protein protein interactions. Important challenge with human protein protein interaction is that, our knowledge on these data is very limited. For example, 99.7 per cent of human molecular interactions are still unknown [1]. Blindly checking all possible interaction is very expensive and not possible. So prediction techniques are used to predict missing interaction based on the known interaction. This approach will reduce the cost effectively, provided the prediction technique must be accurate enough. Here, this article summarizes various techniques and algorithms for link prediction on protein protein interactions networks.

This article is organized as follows. The article start with an introduction to link prediction then describes the representation of genetic data and its notations. Section III describes important link prediction algorithms developed so far and finally the conclusion and references.

II. REPRESENTING GENETIC DATA

Protein protein interaction data can be represented by using graph data structures, where nodes represent the proteins and edges represent their interaction. So we can represent the PPI data using an undirected network *G(V,E)*, where *V* indicates the set of edges and *E* indicates the set of edges. Protein protein interaction network is created after removing multiple links and self-loops. The universal set *U* will contain all possible $\frac{|V|*(|V|-1)}{2}$ links, where $|V|$ represent number of elements in the set *V*. The set of non-existent links would be *U* - *E*. Link prediction is based on the assumption that, among the non-existent links, some links were missing and the task is to predict the links accurately.

III. LINK PREDICTION ALGORITHMS

There are different approaches for link prediction in protein protein networks. Important ones are link prediction based on sequence similarity between proteins and other one is by considering the network topological similarity. Network topological similarity methods can be classified into two local similarity based methods and global similarity based method. The local similarity based algorithms focus mainly on node's local structure for link prediction, whereas global approach considers the overall path structure of the network. Both the approaches have their own pros and cons. The advantages of local similarity based algorithms are that, they are easy to understand implement on real data. But the performance of those algorithm on real protein protein interaction network is not appreciable[2],whereas global approaches shows much better performance.

3.1 Prediction based on sequence similarity

Proteins are build up with amino acid and protein sequencing is the technique to determine amino acid sequence. It is determined that proteins with similar sequence identity (50%) shows similar structural and functional similarities[3]. Many works are there to develop a method to predict the protein interaction using the sequence similarity. Some works[4] showed that, sequence similarity information alone is sufficent to predict the protein interactions. Authors of the paper [5] transformed the complex sequence information into a numerical value representing certain physiochemical properties including hydrophobicity, polarizability, polarity. After this transformation, machine learning techniques are used to classify and predict links. One advantage of sequence similarity based link prediction is that prior knowledge about protein protein interaction network is not required for prediction. Authors of the work [6] presents an enhanced approach aims at improving efficiency and effectiveness of the prediction accuracy. Sequence-based features such as Auto Covariance (AC), conjoint triad (CT), Local descriptor (LD) and Moran autocorrelation (MAC) are extracted from each protein sequence to mine the interaction information in the sequence.

3.2 Local Similarity Based Algorithms

Local similarity based algorithms consider the local features of the protein interaction networks. The local features include common neighbors, degree of each node etc. There are several works based on these concepts.

3.2.1 Common neighbor index

One of the simplest local similarity based link prediction algorithm. The prediction is based on identifying the number of common neighbor between two candidate nodes. For a node *x*, let $\tau(x)$ denote the set of neighbors of node *x*. Then similarity between node x and node y is represented using the equation (1):

$$CN = |\tau(x) \cap \tau(y)| \qquad (1)$$

where $|X|$ is the cardinality of the set *X*. Authors of the paper[7] presents a novel work on link prediction on protein networks based on common neighbor index.

3.2.3 Jaccard Index

According to this work[8], the similarity index is calculated using the following equation (2):

$$S_{xy}^{Jaccard} = \frac{|\tau(x) \cap \tau(y)|}{|\tau(x) \cup \tau(y)|} \qquad (2)$$

3.2.4 Preferential attachment index

Preferential attachment is based on the fact that, the probability that a new link is connected to the node *x* depends on the degree k_x of the node. Motivated by these preferential attachment principles, the traditional preferential attachment index can be calculated using equation (3):

$$S_{xy} = k_x . k_y \qquad (3)$$

where k_x and k_y are the degrees of node *x* and *y* respectively. Authors of the paper[9] provides an enhanced version of preferential attachment scheme. They presents preferential attachment scheme corresponding to two levels. They are *(i) Links appear between the newly added nodes and the old ones ii) internal links appear between two old nodes*. The above work proved that, the proposed indices can provide more accurate predictions than the traditional preferential algorithm especially in the case of networks with large degree heterogeneity and disassortative degree correlation.

3.3 Global similarity based indices

Global similarity based algorithms considers the path between the nodes while constructing the similarity matrix. Different algorithms use different approaches while considering the path details. Some consider the shortest paths between two nodes, while some others are based on weighted paths.

3.3.1 Katz Index and Leich-Holme-Newman Index

Both these index are based on considering the effect of all paths between two node *x* and *y*. In Katz index a damping factor β is used to give the shorter paths more weights. The mathematical expression for Katz index represented using equation:

$$S_{xy}^{katz} = \sum_{i=1}^{\infty} \beta^i . |path_{xy}^i| \qquad (4)$$

$|path_{xy}^i|$ is the set of all paths connecting node *x* and node *y* with length *i*. β is the damping parameter selected such that, shorter paths will get more weight compared to longer paths. Leich-Holme-Newman Index is obviously a variant of katz index. It is based on the assumption that, two nodes are similar if their immediate neighbors' are similar.

3.3.2 Average commute time algorithms

Average commute time algorithms are based on the principles that, average commute time between two nodes *x* and *y* will be small, if they are similar. Let t*(x,y)* be the average time taken by a random walker to reach node *y* starting from node *x*. Then the average commute time is denoted by the equation (5):

$$T(x,y) = t(x,y) + t(y,x) \qquad (5)$$

The average commute time can be calculated by considering the pseudo inverse of Laplacian matrix as mentioned in the previous work [10]. Based on that the average commute time *T(x,y)* calculated using equation (6):

$$T(x,y) = L(l_{xx}^+ + l_{yy}^+ - 2l_{xy}^+) \qquad (6)$$

where l_{xx}^+ denote the corresponding entry in the pseudoinverse laplacian matrix. Here, $T(x,y)$ denote the average commute time between two nodes *x* and *y*. For two nodes are said to be similar, their commute time will be less. Based on that, the similarity matrix can be calculated using the equation (7):

$$S_{xy}^{CT} = \frac{1}{L(l_{xx}^+ + l_{yy}^+ - 2l_{xy}^+)} \qquad (7)$$

3.3.3 Random walk with restart

Random walk on a protein protein interaction network is defined as iterative walker's transition from its current node to any neighboring node with equal probability starting at a given source node. In a statistical point of view, random walk is a finite markov chain that is time-reversible. For example, if random walk begins with a node $n \in V$, the initial probability distribution P_0 defined as a vector in which the elements corresponding to *n* were 1 and 0 otherwise. The rule of the walk can be represented using the equation (8):

$$P_{t+1} = M . P_t \qquad (8)$$

where M is the column normalized adjacency matrix. The random walk performed till a steady state is reached. If the random walk is base on certain probability γ to return to the starting node at each step, then it is represented using equation (9):

$$P_{t+1} = (1-\gamma)M.P_t + \gamma P_0 \qquad (9)$$

The similarity score of two nodes is defined as the steady state probability between those two nodes. One disadvantage associates with random walk based link prediction is that its inherent complexity. Huge size and sparsity of protein protein data adds further complexity to random walk based methods. Authors of[11] proposed a novel method to reduce complexity by performing few random walk instead of using steady state probability.

3.3.4 Prediction using clustering

Clustering is an important research field in data mining where aim is to classify objects into groups based on similarity where, objects belongs to same groups are similar than objects belongs to other groups. Cluster analysis have been widely used in grouping related document, stock market to identify groups with similar price fluctuations etc. Clustering techniques have been widely using in biological data to group genes and proteins have similar functions. Some important clustering algorithms are K means clustering, DBScan, Spectral clustering.

K means clustering is based on partitional clustering approach, where data objects are classified into non overlapping subsets, such that each data object is exactly in one subset. Each cluster is associated with a centroid, and points are assigned based on the least distance from the cluster. Though k means clustering is easy to use and simple to implement, it is not widely used for link prediction purposes because of some problems associated with it. The first one is with the selection of initial centroid which have a great impact on the clustering efficiency. K mean tends to show problems when the clusters are of different size, densities and non-globular shapes.

In some work spectral clustering algorithms and their variants are being used for link prediction. Spectral clustering can be classified into two based on the number of eigen vectors they use. One based on the matrix of affinities between nodes and cluster these nodes based on the second smallest eigen vector of the laplacian matrix and this procedure is continued recursively. This method has been widely using in image segmentation and processing [12]. Another method of spectral clustering is based on top eigen vectors of laplacian matrix. In the paper[13][16], they provide a novel approach to link prediction using spectral clustering using the later method. They use spectral algorithm to cluster the data objects and then using k mean algorithm to find the centroid of each cluster. Then distance from each node to the centroid will be computed and represented in a matrix. Based on a triangular inequality similarity between two nodes will be computed.

Authors of the work [16], presents a novel method to extend the link prediction methods to signed networks for link prediction. In protein protein interaction network, proteins shows complex networking property, ranging from environmental conditions, hormones and performs various functions like cell growth, maintenance of cell survival, development and cell destruction [14]. In such cases it would be more appropriate to represent the protein protein interaction network as a signed graph. The laplacian matrix for the signed graph was computed as per the method described in the work [15], where they use incidence matrix to compute the laplacian matrix for spectral processing.

IV. CONCLUSION

In this article, we briefly summarized some techniques for link prediction in protein protein interaction network. Predicting links in protein interaction is of great significance in both disease prediction and drugs design. Here we summarized techniques for predicting interactions based on protein sequence similarity, and node similarity (proximity) based algorithms. In node similarity

based algorithms we briefed local, global and clustering based methods.

REFERENCES

Journal Papers:

[1] M. P. H. Stumpf, T. Thorne, E. de Silva, R. Stewart, H. J. An, M. Lappe, C. Wiuf, Estimating the size of the human interactome, *Proc. Natl. Acad. Sci. U.S.A. 105 (2008) 6959*

[2] C. Lei, J. Ruan, A novel link prediction algorithm for reconstructing protein protein interaction networks by topological similarity, *Bioinformatics29 (3) (2013) 355364*

[3] The relation between the divergence of sequence and structure in proteins. *Chothia C, Lesk AM EMBO J. 1986 Apr; 5(4):823-6*

[4] Xia JF, Zhao XM, Huang DS: Predicting protein protein interactions from protein sequences using meta predictor. *Amino Acids 2010, 39(5):1595-1599*

[5] Xia JF, Han K, Huang DS: Sequence-Based Prediction of Protein-Protein Interactions by Means of Rotation Forest and Autocorrelation Descriptor. *Protein Pept Lett 2010, 17(1):137-145.*

[6] You, Z.H., Lei, Y.K., Zhu, L., Xia, J.F., Wang, B.: Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. *BMC Bioinformatics 14(S10) (2013)*

[7] L. L, C.-H. Jin, T. Zhou, Similarity index based on local paths for link prediction of complex networks, *Phys. Rev. E 80 (2009) 046122.*

[8] P. Jaccard, tude comparative de la distribution florale dans une portion des Alpes et des Jura, Bull. Soc. Vaud. *Sci. Nat. 37 (1901) 547*

[9] K.; Xiang, J.; Yang, W.; Xu, X. , Tang, Y. (2012), 'Link Prediction in Complex Networks by Multi Degree Preferential-Attachment Indices', *CoRR abs/1211.1790*

[10] F. Fouss, A. Pirotte, J.-M. Renders, M. Saerens, Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation, *IEEE Trans. Knowl. Data. Eng. 19 (2007) 355*

[11] W. Liu, L. Lu, Link prediction based on local random walk, *EPL 89 (2010) 58007*

[12] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pp 888-905, Aug. 2000*

[13] Panagiotis Symeonidis, Nantia Iakovidou, Nikolaos Mantas, Yannis Manolopoulos, From biological to social networks: Link prediction based on multi-way spectral clustering, *Data Knowledge and Engineering, Volume 87, September 2013, Pages 226-242, ISSN 0169-023X*

[14] Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B et al. (2005) A human protein protein interaction network: a resource for annotating the proteome. *Cell 122: 957968*

[15] Y. Hou. Bounds for the least Laplacian eigen value of a signed graph. *Acta Mathematica Sinica, 21(4):955–960, 2005*

Proceedings Papers:

[16] Iakovidou, N.; Symeonidis, P.; Manolopoulos, Y., "Multiway spectral clustering link prediction in protein-protein interaction networks," *Information Technology and Applications in Biomedicine (ITAB), 2010 10th IEEE International Conference on , vol., no., pp.1,4, 3-5 Nov. 2010 doi: 10.1109/ITAB.2010.5687767*