Multiway Spectral Clustering Link Prediction in Protein-Protein Interaction Networks

Nantia Iakovidou, Panagiotis Symeonidis and Yannis Manolopoulos

Abstract— **An increasing number of observations support the hypothesis that the vast majority of biological functions involve interactions between proteins and that the complexity of living systems arises as a result of such interactions. In this paper we apply a multiway spectral clustering technique to protein-protein interaction networks to perform accurate link prediction between proteins. We provide experimental evidence about the accuracy and the performance of the proposed method in real datasets, compared to other known algorithms such as k-means and RWR.**

I. INTRODUCTION

Inspired from the recent surge of research on large, complex networks and their properties, a considerable amount of attention has been devoted to the computational analysis of protein networks - structures whose nodes represent proteins and whose edges represent interaction, or influence between them. Interactions between proteins are important for numerous - if not all - biological functions. Given a natural example from the area of biology, signals from the exterior of a cell are mediated to the inside of that cell by protein-protein interactions of the signaling molecules. This process, called signal transduction, plays a fundamental role in many biological processes and in many diseases (e.g. cancers).

Understanding the mechanisms by which protein-protein interactions work and evolve, is very important for decoding every process in a living cell. This fundamental question, that is still not clearly understood, forms the motivation for our work in this paper. We study a basic computational problem underlying protein networks, the link prediction problem. Given a part of a protein network we seek to accurately predict the rest of the network's edges, by performing multiway spectral clustering analysis, a technique that uses information obtained from the top few eigenvectors and eigenvalues of the normalized laplacian matrix. We compare our method to approaches which are based on local features of a network, such as FOAF [1], to global approaches, such as RWR [2] and we also compare our method with traditional clustering algorithms, such as k-means [3]. As will be shown in the experimental section our proposed method outperforms, having many fine properties that make it more suitable for protein-protein interaction networks.

The contributions of our approach are summarized as follows:

- We provide more accurate predictions, as far as interactions between proteins are concerned, than other previously well-known algorithms such as k-means.
- Our approach, by performing dimensionality reduction of the normalized laplacian matrix, results to a smaller and more compact graph matrix than the original one, as will be shown experimentally. Thus, our method succeeds higher efficiency than the global approaches.
- We define two node similarity measures that exploit local and global characteristics of a network. In particular, we calculate the similarity between nodes that belong in the same cluster and similarity between nodes that belong in different clusters.
- Using a real human protein data set we perform extensive experimental comparison of the proposed method against existing link prediction algorithms and k-means clustering algorithm.

The ultimate goal in our work is to use multiway spectral clustering analysis to discover new information from a protein-protein interaction network and to suggest new roles for already known proteins.

The rest of the paper is organized as follows. Section 2 summarizes previous related work in this area. The next section provides some preliminaries in graphs. The proposed method is presented in section 4, while experiments and results follow on to the next section. Finally, this paper concludes in section 6.

II. RELATED WORK

Protein-protein interactions (PPIs) are the most intensely analyzed networks in biology and there are a multitude of biochemical and biophysical methods to detect them [4], [5]. But since molecular biology techniques are quite expensive and costly and very often time-consuming, it is by far preferable to apply graph theory techniques to study such kind of problems.

This section presents previous related work about spectral clustering techniques and about other algorithms that have been used in the past to extract information from graphs that represent protein networks.

There are two main categories of spectral clustering algorithms based on the number of eigenvectors they use. The first category [6] uses a matrix of affinities between nodes in order to cluster these nodes based on the second smallest eigenvector of the Laplacian matrix. Then, recursively uses the second smallest eigenvector to further partition these clusters. The second category, which is more similar to our approach, directly computes a myltiway partition of the

This work was not supported by any organization

N. Iakovidou, Panagiotis Symeonidis and Yiannis Manolopoulos are with Department of Computer Science, Aristotle University, 54124 Thessaloniki, Greece {niakovid, symeon, manolopo}@csd.auth.gr

data [7], [8]. In particular, it selects the largest k eigenvectors and their corresponding eigenvalues. Then, it extracts the clusters by finding the approximate equal elements in the selected eigenvectors using any clustering algorithm e.g. k-means.

Authors in [9] use sequence data to apply spectral clustering techniques. They prove that their algorithm offers competitive performance on the clustering of biological sequence data. Authors in [10] also present a simple and unified derivation of the spectral algorithms and they apply it to microarray datasets. They illustrate the performance of spectral algorithms by providing numerous experimental results.

Stelzl et al. [11] also studied a human protein-protein interaction network and they developed a tool for the identification of PPIs, which can be used to detect interactions across the entire proteome of an organism. Another tool, named Local Protein Community Finder has also been developed from the authors in [12]. This tool finds a community close to a queried protein in any network specified by the user.

Generally, a variety of computational methods have been investigated so far for the protein network inference problem [13] , [14]. Authors in [15] present a local path index to estimate the likelihood of the existence of a link between two nodes. Authors in [16] introduce a method based on a variant of kernel canonical correlation analysis to predict the protein network of a yeast. Other methods try to predict protein interactions from evolutionary similarities [17], while others combine different sources of data to infer the network [18].

In contrast to the above methods we develop a multiway spectral clustering technique that focuses only on predictions based on the link structure of a protein network and we compare it to other well-known approaches which are based on local features of a network, such as FOAF [1], to global approaches, such as RWR [2] and also with traditional clustering algorithms, such as k-means [3]. These algorithms are described in the related experimental section.

III. Preliminaries in Graphs

A graph $G = (V,E)$ is a set of vertices V and a set of edges E. Vertices represent proteins and edges represent interactions between proteins. In this paper, G will always be an undirected and unweighed graph as shown in Figure 1.

The adjacency matrix A of an undirected graph G is a square matrix with rows and columns labelled by graph vertices. For undirected graphs, the adjacency matrix is always symmetric. An element in the adjacency matrix has value equal to 1 in position (v_i, v_j) if proteins v_i and v_j are interacting with each other and 0 otherwise. For instance, the element (v_1, v_2) of the adjacency matrix A derived from the graph depicted in Figure 1 will have value 1, while the element (v_1, v_6) will have value 0. Note that all elements along the principal diagonal of A are zeros, indicating that a node is not connected to its self.

The spectral algorithms are based on eigenvectors of Laplacians, which are a combination of the adjacency and the degree matrix. The normalized laplacian matrix of graph G is computed by Equation $L = I$ - $D^{-\frac{1}{2}} \times (D$ - $A) \times D^{-\frac{1}{2}}$, where D and I is the degree and the identity matrix of graph G, respectively. The normalized laplacian matrix L is positive semi-definite and has n non-negative real-valued eigenvalues $0 = \lambda_1 \leq \ldots \leq \lambda_n$. Moreover, the number of 0 eigenvalues equals the number of the connected components in a graph.

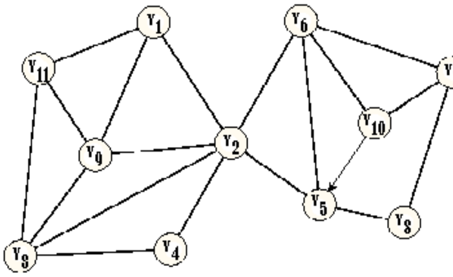

Fig. 1. An example of undirected and unweighed graph.

IV. The Proposed Method

Our proposed clustering approach, computes similarities between nodes in an undirected graph, that represents an already known protein network. Our clustering algorithm uses as input the connections of a graph G and outputs a similarity matrix between any two nodes in G. Therefore, two proteins can be assumed that they interact with each other, according to their values (weights) in the similarity matrix. For reasons of convenience, our method will be denoted as BioSpectral for the rest of this paper.

As illustrated in Figure 1, proteins are connected in a graph. If we have to predict an interaction between a new protein and a protein from the already known part of the graph, then there is no direct indication for this task in the adjacency matrix A. However, after applying the BioSpectral algorithm, we can get a similarity matrix between any two nodes of graph G and recommend connections between proteins according to their weights.

Firstly, BioSpectral computes the first k eigenvectors $u_1, \ldots, u_k$ with the corresponding $\lambda_1, \ldots, \lambda_k$ eigenvalues of the normalized laplacian matrix L based on Equation $L \times U = \lambda \times U$. The U matrix has columns the eigenvectors $u_1, \ldots, u_k$ and nodes $v_i \in R$, with $i = 1, \ldots, n$, that correspond to the i-row of U. We compute the first k eigenvectors and the first λ eigenvalues of the normalized laplacian matrix L.

Secondly, we cluster nodes v_i with the k-means algorithm into clusters $C_1, \ldots, C_k$, based on the aforementioned eigenvectors u_k. In our graph example (Figure 1), k is equal to 2 and thus, we divide the nodes in two clusters C_1 and C_2, where $C_1 = \{v_1, v_2, v_3, v_4, v_9, v_{11}\}$ and $C_2 = \{v_5, v_6, v_7, v_8, v_{10}\}$. The node assignment information is kept in a matrix that will be named *IDX* for the rest of the paper. Having defined the clusters we can now compute the centroids of each cluster and then compute the distances of each node from each cluster centroid. Let matrix D be the one that keeps the distances of each node from the centroid of each cluster. We will use this information in the following step.

In the third step, BioSpectral uses Equation 1 to quantify the similarity between nodes that belong in the same cluster. Moreover, BioSpectral uses Equation 2 to quantify the similarity between nodes that belong to different clusters.

$$SimSC(i,j) = 1 - |min(D(i)) - min(D(j))| \quad (1)$$

$$SimDC(i,j) = \frac{1}{D(i,IDX(j)) + D(j,IDX(i))} \quad (2)$$

Finally, for a test protein (node) v_i we rank the calculated similarities with other proteins and recommend to it the top ranked nodes as its possible interacting proteins.

V. Experiments and Results

In this section, we experimentally compare our approach with other existing link prediction and clustering algorithms. We use the k-means [3] algorithm, already described above, the Random Walk with Restart [2] algorithm, and the Friend of a Friend [1] algorithm, denoted as k-means, RWR, and FOAF, respectively.

The FOAF algorithm is based on the notion that two nodes are more likely to form a link in the future, if they have many common neighbors. FOAF considers only pathways of maximum length 2 between a protein and its possible interactors, which obviously harms accuracy prediction. The RWR algorithm starting from a node v_i chooses randomly among the available edges to transmit to another node v_j. At each step, RWR has some probability c to return to the node v_i. Thus, the relevance score of node v_i with respect to node v_j is defined as the steady-state probability r_{v_i,v_j} that the random walker will finally stay at node v_j, as shown by Equation 3:

$$\vec{r}_{v_i} = c \cdot A \cdot \vec{r}_{v_i} + (1-c) \cdot \vec{e}_{v_i}, \quad (3)$$

where $\vec{e}_{v_i}$ is the $n \cdot 1$ starting vector with the v_i^{th} element equal to 1 and 0 for the other elements of the vector, and A is the adjacency matrix of graph G.

The data set[1] used in our paper contains a total of 3269 unique interactions between 1925 different human proteins. We formed the adjacency matrix of the data set based on the interactions and we applied the four algorithms.

We use the classic precision/recall metric as performance measure for protein link predictions. For a test protein receiving a list of n candidate interactors (top-n list), precision and recall are defined as follows: Precision is the ratio of the number of relevant proteins in the top-n list to n and recall is the ratio of the number of relevant proteins in the top-n list to the total number of relevant proteins.

To address the topological properties of our interaction network we calculated the distribution of the shortest path between pairs of proteins, depicted in Figure 2, as well as the degree distribution of the protein data set, depicted in Figure 3. The mean shortest path length between any two proteins of the network was found to be equal with 5.34. This means that most proteins are very closely linked, a phenomenon that has been described as small world property of networks [19]. Also, as shown in Figure 3 the degree distribution of the network decreases slowly, closely following a power-law. On average, we found that proteins in the network have 3.4 interaction partners. However, we detected 982 proteins with only one partner, as well as 22 hubs-proteins with more than 30 partners.

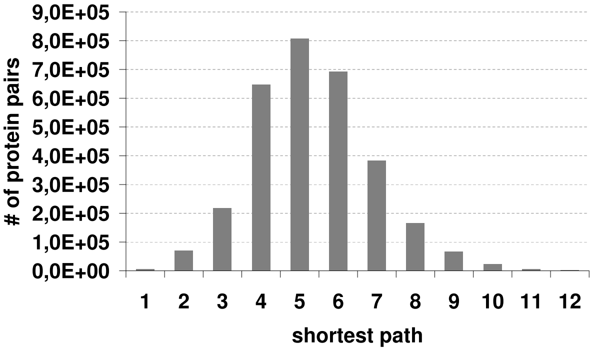

Fig. 2. Distribution of the shortest path between pairs of proteins in the network. On average, any two proteins in the network are connected via 5.34 links.

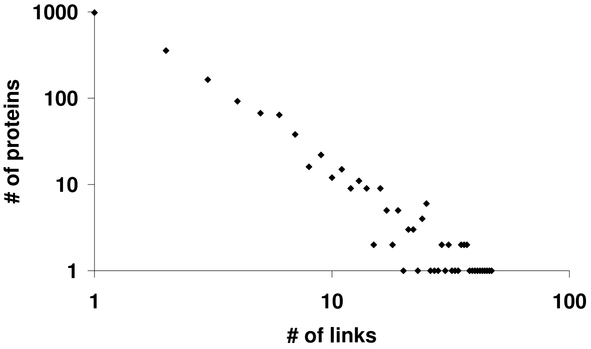

Fig. 3. Degree distribution of the network proteins. Number of proteins with a given link k in the network approximates a power law (P(k) $\sim k^{\gamma}$, γ=3.4)

We also studied the sensitivity analysis of BioSpectral's accuracy performance. As mentioned before, a required input of BioSpectral algorithm is the number k of clusters. To improve our predictions in terms of effectiveness, it is important to fine-tune the k variable. As shown in Figure 4 the best precision performance is obtained when k equals 1400 clusters. In the following we keep k=1400 as the default initial value for the BioSpectral algorithm. Figure 5 shows the precision diagram for the human protein data set and presents the increase in precision when a larger amount of protein-neighbours is known. As expected, with increasing the percentage of observed links, the precision increases too. Thus, BioSpectral can predict more effectively new links between proteins for larger node degree values, since in such cases the network density is increased.

We proceed with the comparison of BioSpectral with k-means, RWR and FOAF algorithms in terms of precision and recall. We examine the ranked list, which is recommended to an examinee protein, starting from the top one. In this situation, the recall and precision vary as we increase the number of recommended proteins. For the human protein data set a we plot a precision versus recall curve for all four

[1] Downloaded from `http://www.cell.com/cgi/content/full/122/6/957/DC1/`

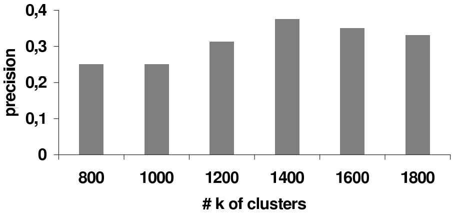

Fig. 4. Precision performance when tuning variable k (number of clusters).

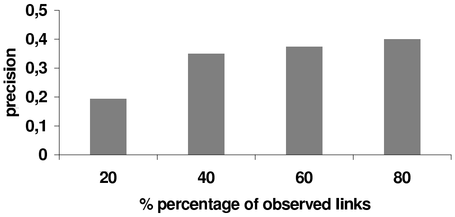

Fig. 5. Precision diagram for the data set.

algorithms, as shown in Figure 6. BioSpectral's precision value is equal to 0.375 when recommending the top protein, while the respective values for k-means, RWR and FOAF are 0.2, 0.114 and 0.06.

This experiment shows that BioSpectral and k-means are more robust in predicting relevant proteins and the reason is that BioSpectral and k-means, identify clusters which present high within-cluster nodes similarity and low between-cluster similarity. Thus, the high within-cluster node similarity captures effectively the notion of the local characteristics of a graph, whereas the low between-cluster dissimilarity captures effectively the notion of the global characteristics of a graph. In contrast, RWR traverses globally the protein network, missing to capture adequately the local characteristics of the graph. Moreover, FOAF fails to provide accurate predictions because it exploits only length-2 paths, missing to capture the notion of the global characteristic of a graph. BioSpectral outperforms k-means, because it is based on the Normalized Laplacian matrix, whereas k-means is based on the Adjacency matrix. This means, that BioSpectral takes into consideration also the degree of connectivity of a graph. Moreover, BioSpectral is more flexible than k-means, in capturing a wider range of cluster geometries and shapes.

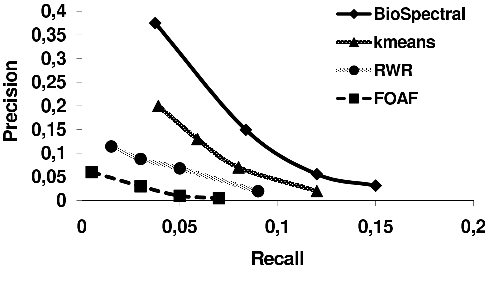

Fig. 6. Comparison of Algorithms.

VI. Conclusions

In this paper, we introduced a framework that uses multiway spectral clustering to provide protein link predictions in a protein-protein interaction network. We performed extensive experimental comparison of the proposed method against existing well-known link prediction algorithms and k-means, using a real human protein data set. We have shown that our proposed method provides more accurate and competitive results.

In the future, this work can be extended by examining and comparing the current results with another more dense protein network with different topological properties.

References

[1] J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy. "Make new friends, but keep the old: recommending people on social networking sites", *In Proceedings of the 27th international conference on Human factors in computing systems*, 2009, pp 201-210.

[2] J. Pan, H. Yang, C. Faloutsos, and P. Duygulu, "Automatic multimedia cross-modal correlation discovery", *In Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp 653-658.

[3] J. MacQueen, "Some methods for classification and analysis of multivariate observations", *In Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 1967, pp 281-297.

[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, "Prediction of protein-protein interactions based on PseAA composition and hybrid feature selection", *Biochemical and Biophysical Research Communications*, vol. 380, 2009, pp 318-322.

[6] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation", *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, 1997, pp. 731-737.

[7] A. Ng, M. Jordan and Y. Weiss, "On spectral clustering: Analysis and an algorithm", *Advances in Neural Information Processing Systems*, vol. 14, 2001, pp 849-856.

[8] M. Maila, J. Shi, "A Random Walks View of Spectral Segmentation", *International Conference on AI and Statistics (AISTAT)*, 2001, pp 1-4.

[9] 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.

[10] 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.

[11] 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.

[12] K. Voevodski, S. Teng and Y. Xia, "Finding local communities in protein networks", *BMC Bioinformatics*, vol. 10, 2009, pp 297-310.

[13] S. Brohee and J. van Helden, "Evaluation of clustering algorithms for protein-protein interaction networks" *BMC Bioinformatics*, vol. 7, 2006, pp 488-506.

[14] C. Brun, F. Chevenet, D. Martin, J. Wojcik, A. Guenoche and B. Jacq, "Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network" *Genome Biology*, vol. 5, 2003, pp R6.1-R6.13.

[15] 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.

[16] 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.

[17] F. Pazos and A. Valencia, "Similarity of phylogenetic trees as indicator of protein-protein interaction" *Protein Eng.*, vol. 14, 2001, pp 609-614.

[18] E.M. Marcotte, M. Pellegrini, M.J. Thomson, T.O. Yeates and D. Eisenberg, "A combined algorithm for genome-wide prediction of protein function" *Nature*, vol. 14, 1999, pp 849-856.

[19] S.H. Strogatz, "Exploring complex networks", Nature, vol. 410, 2001, pp 268-276.