Biological Pathway Completion Using Network Motifs and Random Walks on Graphs

Maya El Dayeh and Michael Hahsler
Department of Computer Science and Engineering
Southern Methodist University
Dallas, TX, USA

***Abstract*—Enhancing our understanding of cellular processes will ultimately lead to the development of better therapeutic strategies. Completing incomplete biological pathways through utilizing probabilistic protein-protein interaction (PPI) networks is one approach towards establishing knowledge about cellular mechanisms. The existing complex/pathway membership methods are focused on uncovering candidate protein members from probabilistic PPI networks. In our previous work, we defined the pathway completion problem and developed a method that utilizes network motifs to complete incomplete biological pathways. Network motifs allow us to take into consideration the intrinsic local structures of pathways and identify the possible points of insertion of candidate proteins. However, our previous approach requires a complete and correct PPI network. In this paper, we extend our approach and use random walks on graphs to address the pathway completion problem with incomplete PPI networks. We evaluated our proposed method using three yeast probabilistic PPI networks and two yeast pathways from KEGG (Kyoto Encyclopedia of Genes and Genomes). Moreover, we compared the accuracy of our network motifs approach for pathway completion to the existing approach for pathway membership, which also utilizes random walks. Our experiments show that our new approach achieves similar or better accuracy. In addition, our method identifies the possible locations and connections of the candidate proteins in the incomplete pathway, which allows for more efficient experimental verification.**

Keywords- pathway completion; network motifs; local structures; pathway membership; protein networks; random walk

I. INTRODUCTION

Structured interactions among proteins are the backbone of the complex biological functions of organisms. Enhancing our knowledge of these interactions will help us better understand the causes of diseases and will ultimately lead to the development of advanced therapeutic strategies. Therefore, research effort has been directed towards discovering the complete set of interacting proteins [1], [2], [3]. Protein interactions are usually analyzed in the context of a protein-protein interaction (PPI) network rather than being studied as isolated pair-wise interactions. In a PPI network, nodes represent proteins and edges represent interactions between two proteins. With the help of high-throughput screens, like yeast-two-hybrid (Y2H) [4], [5] and tandem affinity purification-mass spectrometry (TAP-MS) [6], [7], [8], researchers were able to uncover large datasets of protein-protein interactions. However, further studies have shown that protein interaction data may contain a high number of false positives and false negatives [2]. Therefore, researchers have been incorporating multiple types of biological knowledge and evidence of interaction into the construction of PPI networks. The constructed networks are usually called probabilistic PPI networks. Such networks are weighted, whereby the weights on the edges represent the probability of interaction. Complex/pathway membership [2], [6], [9], protein function prediction and regulation [10], and pathway discovery [11] are examples of the areas that mine probabilistic PPI and PPI networks for knowledge.

Biological pathways are defined by [12] as "distinct, experimentally-validated sub-networks of proteins within the larger PPI network that interact with each other by well-defined mechanisms to regulate a specific biologic phenotype". Protein complexes can be defined as clusters of interacting proteins. Biologists are confident that a significant number of known protein pathways are incomplete [2]. The pathway membership problem, which is similar to complex membership, is defined by [2] as the problem of extracting a ranked list of candidate proteins from a given probabilistic PPI network based on the estimated probability of membership in a partly known pathway. Since PPI networks incorporate a large number of false negatives (missing interactions) and false positives (incorrect interactions), computational methods like network reliability [6], random walks on graphs [2], and Net-flow [9] are suggested to extract complex/pathway members from a probabilistic PPI network based on a measure of proximity. However, current research examines membership without looking at possible locations of insertion.

In our previous work [13], we have defined the pathway completion problem as the problem of uncovering an ordered set of candidate proteins from a given probabilistic PPI network and predicting their locations in an incomplete pathway. Location prediction utilizes the local structures (network motifs) of the pathways. In this paper, we extend our previous work, which was based on the simplifying assumption that PPI networks are complete and correct, and propose a random walk method, which employs network motifs, to tackle pathway completion. Network motifs are sub-network patterns that exist in networks at frequencies significantly greater than expected [14]. Our network motif-random walk method breaks

down an incomplete pathway into a set of potentially incomplete motifs and performs random walks on a given probabilistic PPI network to retrieve proteins that may complete the motifs. The proteins, which are retrieved from the network, are ranked according to their proximity to the proteins of the pathway motifs. An advantage of using a random walk based approach is computational efficiency, which is a requirement when dealing with large networks. Another advantage is the notion of proximity, which is essential when dealing with noisy protein interaction data. We evaluated our method using three yeast probabilistic PPI networks and two yeast KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways [15], [16]. Our experiments show that our network motif-random walk method for pathway completion has similar or better accuracy in comparison to the random walk method for complex/pathway membership while also providing positional information.

The rest of the paper is organized as follows. In Section II, we cover background and previous work. In Section III, we define the pathway completion problem, network motifs, and random walks on graphs. In Section IV, we discuss experiments and results. Finally, we conclude with Section V.

II. Background

Network reliability [6], random walks on graphs [2] and Net-flow [9] are computational methods that were developed to address the complex/pathway membership problem. These methods were applied to probabilistic PPI networks because the networks incorporate the probability of interaction amongst two proteins (weights on the edges). In the following paragraphs, we present a brief overview of the existing complex/pathway membership methods.

The concept of network reliability has been applied by [6] to a probabilistic PPI network and protein complexes to provide a solution to the complex membership problem. In network reliability, we have a network of connected components (proteins). The weights on the edges refer to the probability that the connection is functional, which is analogous to the probability of PPI. A measure of proximity between two components is determined by the probability that a path of functioning connections exists between the components.

In the random walk method in [2], the random walker begins the walk at a designated start node (member of the complex/pathway) and moves to a neighboring node based on the probabilities of the connecting edges in the probabilistic PPI network. The process is repeated until the walker decides to teleport to the start node given a certain restart probability. Proximity is inferred by the visit frequency since it is expected that the walker will visit strong candidate proteins more frequently than weak candidate proteins.

Maximum flow is another measure of proximity which has been utilized by [9] to identify possible candidates for a given protein complex from a probabilistic PPI network. To compute the proximity (maximum flow) of a protein in the network (sink) to the proteins of a given complex (sources), the capacities of the edges are set to one and the costs are set to the probability of protein interaction.

The complex membership methods were applied mainly to yeast MIPS (Munich Information Center for Protein Sequences) [17] benchmark protein complexes. Additionally, the authors of [2] tested the random walk method on a set of yeast KEGG pathways. Yeast is a model organism and its protein interaction data have been studied extensively. Since protein complexes form highly connected sub-graphs, membership methods are expected to provide plausible solutions to the pathway/complex membership problem. We believe that by exploiting the intrinsic local structures available in pathways, as we propose in this paper, we can provide a better solution to pathway completion. In the following section, we develop a method which tackles the pathway completion problem.

III. Methods

In this section, we provide a detailed description of the pathway completion problem and our proposed solution using network motifs and random walks on graphs. We define the pathway completion problem as: *the problem of retrieving an ordered set of candidate proteins for a given incomplete pathway from a probabilistic PPI network and predicting the locations and connections of the proteins in the incomplete pathway.*

Membership methods are more applicable to protein complexes because complexes form highly connected undirected sub-graphs; however, pathways are directed sub-graphs where typically not all components are directly connected to each other. When designing a method for the pathway completion problem, we must take advantage of the intrinsic local structures of the incomplete pathways. Therefore, we propose to address the pathway completion problem by utilizing network motifs to represent the local structures of the pathways. We approximate the proximity between proteins with random walks on a probabilistic PPI network. Next, we discuss network motifs and random walks on graphs. Then, we introduce our pathway completion algorithm.

A. Network Motifs

Network motifs are sub-network patterns that exist in networks at frequencies significantly greater than expected [14]. It has been shown that network motifs exist in biological networks [18] such as PPI [19], signal transduction [20], metabolic [21], and transcription-regulation [14], [19], [22]-[24]. The authors of [22] identified three significant motifs, which can be found in the transcription-regulation networks of Escherichia coli. The authors called the motifs feed forward loop (FFL), single input module (SIM), and dense overlapping regulons (DOR). The DOR motif is also known as multi-input motif (MIM) in [24]. In [13], we concentrated on a subset of the motifs identified in the literature. The subset consisted of three simple motifs, which are similar in structure to the motifs found in the literature. The motifs are called linear motif, single input motif, and multiple input motif. The simplicity of the motifs refers to the number of edges connecting the proteins of a given motif. For example, instead of dense overlapping edges in the case of a DOR with multiple input proteins connected to a single protein, the multiple input motif has only two input

proteins. We utilized the motifs in our Simple Search method in [13]. In Fig. 1, we illustrate the simple motifs together with the proposed possible extensions, which may complete an incomplete motif. We plan on addressing more complex motif structures and their extensions in future work.

In order to evaluate how likely it is that a motif or a completed motif occurs in a pathway, we devised a scoring mechanism in [13] for the motifs. We computed the score of a motif based on the weights we retrieved from a probabilistic PPI network for the edges in the motif. The proposed scores are the minimum, maximum, and average of the retrieved weights. The minimum score concentrates on the "weakest link'' in the motif while the maximum score takes only the strongest interaction into account. The average represents a tradeoff. Any scoring function which uses the retrieved weights and the motif structure can be used. We refer the reader to [13] for a more detailed treatment of scoring. However, since probabilistic PPI networks are known to be incomplete and also contain false positives and false negatives, computing a score directly from the weights found in these networks is problematic. Therefore, in this work, we use proximity information between proteins obtained from random walks on a given probabilistic PPI network instead of using the weights in the network directly.

B. *Random Walks on Graphs*

PPI and probabilistic PPI networks are noisy and incomplete; they suffer from a large number of false negatives and false positives (i.e. incorrect and missing edges). When mining PPI networks for knowledge, such as examining the neighborhoods of nodes, characterizing the relationships established through multiple paths among the nodes, and measuring the distance between them, it is important to design a technique that is robust to noise. It has been shown that proximity measures based on random walks with restart are robust and useful for incomplete and noisy PPI data [25]-[27].

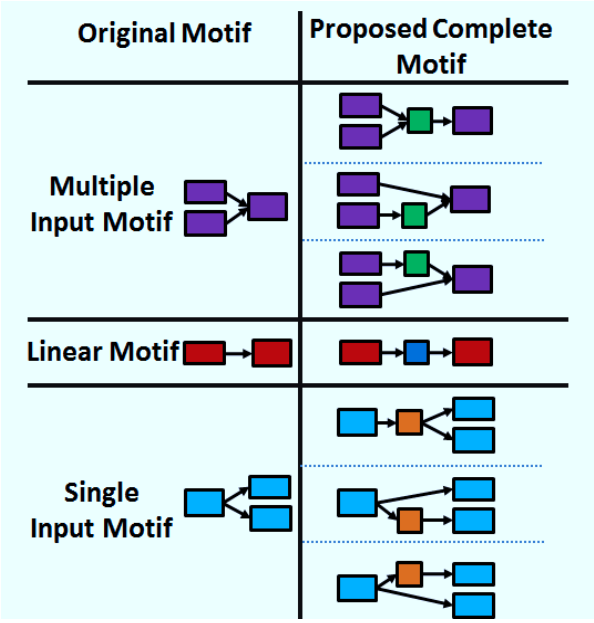


Figure 1. Original motifs and their proposed complete motifs. The proteins which may complete the motifs are represented by squares.

A probabilistic PPI network G, is represented as a weighted undirected graph $G = (V, E)$, where V is the set of nodes (interacting proteins) and E is the set of undirected weighted edges (pair-wise interactions). The weights on the edges represent the probability that two proteins interact. The random walk algorithm is applied to G to extract a list of candidate proteins for a given incomplete biological pathway. The random walk algorithm mimics a random walker and is described as follows: the random walker starts at a designated start node s (a protein in the incomplete pathway). The walker randomly selects an edge from the possible edges to transition to an adjacent node. Edge selection is biased such that edges with higher weights are preferred. The walker repeats the edge selection-transition process to move on the graph until it teleports back to the start node. The decision of whether or not to teleport back is made at each step by flipping a biased coin representing a restart probability c. The random walk is restarted n times from the start node and the proximity to other nodes is approximated by the number of times the random walker ends its walk at a node before teleporting back. The restart probability c controls how far the random walker will wander away from the start node s. If c is close to 0, then the algorithm will provide a more comprehensive view of the organization of the network around the start node s. However, if c is close to 1, we get a restricted view of the neighborhood of s [2]. The number of repetitions n required for the algorithm to converge to good proximity approximations increases as the restart probability gets closer to 0 [2]. The details of the random walk technique are given in [2] and [28].

C. *The Pathway Completion Method*

We tackle the pathway completion problem by retrieving possible complete motifs from a probabilistic PPI network with better scores (as defined previously) than the original motifs found in the pathway. Our Pathway Completion algorithm breaks down an incomplete pathway into a set of potentially incomplete motifs and performs a random walk on a given probabilistic PPI network to retrieve proteins that would complete the original motifs. Fig. 2 depicts the steps of our method. The Motif Extractor method breaks down an incomplete pathway into its constituent potentially incomplete motifs. The set of incomplete motifs and the probabilistic PPI network are used by the random walk based search method to extract possible protein candidates and their locations from the network. The proposed complete motifs are then ranked according to their scores. The ordered list is examined to suggest complete pathways. The formal steps of the Pathway Completion algorithm are also shown in Fig. 2. The algorithm accepts the weighted graph G (the probabilistic PPI network) and the directed sub-graph G' (the incomplete pathway). In step 1, the algorithm calls the Motif Extractor method which breaks down the incomplete pathway into L, the set of incomplete motifs that make up the pathway. Then for each incomplete motif l in L, and for each protein p_i in l, the algorithm retrieves the proteins with a proximity to p_i greater than zero using random walks on graphs (steps 2 to 4). In step 5, the algorithm uses the proximity measure to generate the list L' of the proposed complete motifs of l, the incomplete motif.

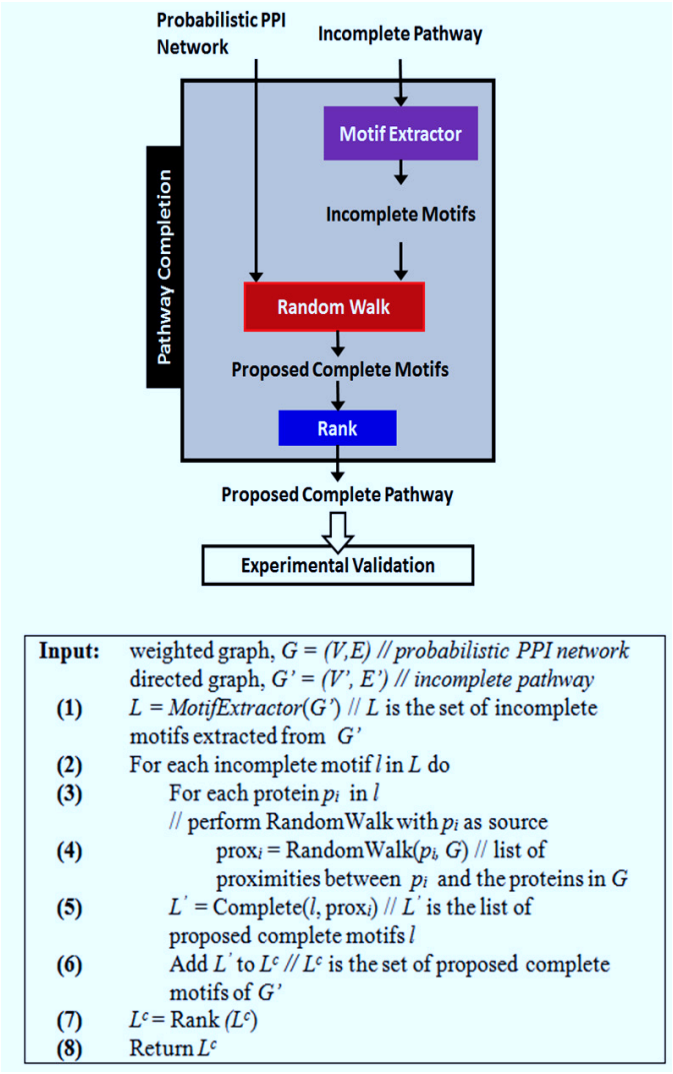
In step 6, L' is added to L^c, the set of proposed complete motifs of G'. In step 7, the algorithm ranks the completed motifs according to the score of the motif. In step 8, the algorithm returns the ranked list L^c.

IV. Experiments, Results, and Discussion

In this section, we look at the used datasets, the experimental setup, and the results. We evaluated our method using three yeast probabilistic PPI networks and two yeast KEGG pathways. It has been shown that the random walk based approach is at least as accurate as the network reliability approach [2]. Therefore, we compared the accuracy of our network motif-random walk pathway completion method to the random walk complex/pathway membership method.

A. Datasets

We used two yeast KEGG pathways, MAPK signaling and regulation of Autophagy, in our experiments. MAPK is made up of 61 proteins and 67 interactions, and regulation of Autophagy consists of 17 proteins and 14 interactions. In Table I, we show the numbers of linear motifs, multiple input motifs, and single input motifs that were retrieved by Motif Extractor from the two pathways.



Input: weighted graph, $G = (V,E)$ // *probabilistic PPI network*
directed graph, $G' = (V', E')$ // *incomplete pathway*
(1) L = *MotifExtractor*(G') // L is the set of incomplete motifs extracted from G'
(2) For each incomplete motif l in L do
(3) For each protein p_i in l
// perform RandomWalk with p_i as source
(4) $prox_i$ = RandomWalk(p_i, G) // list of proximities between p_i and the proteins in G
(5) L' = Complete(l, $prox_i$) // L' is the list of proposed complete motifs l
(6) Add L' to L^c // L^c is the set of proposed complete motifs of G'
(7) L^c = Rank (L^c)
(8) Return L^c

Figure 2. The Pathway Completion algorithm.

Yeast has been studied extensively due to the abundant information and experimental data available on its genome. Consequently, techniques to build PPI networks have been applied to yeast. Three important yeast probabilistic PPI networks are Naïve Bayes, ConfidentNet, and PIT-Network (Probabilistic Interactome Total Network). The Naïve Bayes probabilistic PPI network was constructed by [6] using the four high-throughput screens in [4], [5], [7], and [8] to identify protein interactions. Naïve Bayes consists of 3,112 proteins and 12,594 undirected probabilistic interactions. ConfidentNet was constructed by [29] through integrating five genomic datasets: mRNA co-expression, phylogenetic profiles, co-citation, gene fusions, and the high through-put screens in [4], [5], [7], and [8]. ConfidentNet contains 5,552 proteins and 235,222 undirected probabilistic interactions. PIT-Network was assembled by [30] through combining evidence from [4], [5], [7], and [8] and four genomic datasets: mRNA co-expression, MIPS function, GO (Gene Ontology) processes, and essentiality data. PIT-Network consists of 5,171 proteins and 49,640 interactions. The common sub-network of the three probabilistic PPI networks is made up of 993 nodes and 1,669 edges. In Fig. 3, we show examples of network motifs retrieved by our method from MAPK signaling and regulation of Autophagy pathways. In the figure, we demonstrate the nature of network connectivity, the edge weights, and protein availability in the three probabilistic PPI networks by giving some examples of the pathway motifs found in the networks. First, edge connectivity is not always similar across the networks. For example, the edges which connect the proteins of the MAPK multiple input motif in Fig. 3-A exist in ConfidentNet. However, none of the edges can be found in Naïve Bayes, and only one of the edges occurs in PIT-Network. Second, a weight on an edge in Naïve Bayes is the posterior probability of interaction. The weights on the edges in ConfidentNet and PIT-Network are log likelihood ratios and likelihood ratios, respectively. The edge weights cannot be compared directly because different types of biological data were integrated to compute the final edge weights for each network. Fig. 3 shows examples of the edge weights of the motifs we retrieved from the networks. For instance, the weight on the edge of a linear motif extracted from regulation of Autophagy (Fig. 3-B) is 6.80318 in ConfidentNet, 0.232092 in Naïve Bayes, and 4219.22 in PIT-Network. Last, there may be instances when one of the proteins of a given motif is not present in a given network. For example, the protein YLR240W is missing from Naïve Bayes. The protein is shown as an empty rectangle in the multiple input and the single input motifs taken from regulation of Autophagy in Fig. 3-B.

TABLE I. The Number of Motifs Extracted from MAPK and Regulation of Autophagy

Pathway	Linear	Multiple Input	Single Input
MAPK	13	15	13
Regulation of Autophagy	8	1	2

B. *Experiments and Results*

We evaluated our network motif-random walk approach to pathway completion by comparing its accuracy to the random walk based approach for complex/pathway membership devised by [2]. For evaluation, we used the two yeast KEGG pathways, MAPK and regulation of Autophagy, and the three probabilistic PPI networks. Note that the qualities of the probabilistic networks affect the results of the methods. We utilized leave-one-out cross validation (LOOCV) to measure the accuracy of our approach. In LOOCV, for each pathway one protein is left out from the pathway. Since removing a protein from the pathway causes an interruption, we chose to create a direct connection for the left-out protein. Then, the completion and membership algorithms are applied to the now incomplete pathway and a given probabilistic PPI network to retrieve candidate proteins. Afterwards, the rank of the left-out protein is observed, and the accuracy is assessed by examining the rank of the left-out protein. We would like to see the left-out protein among the highly ranked candidates retrieved from the probabilistic PPI network. To compare the results of the completion and membership approaches, we have to transform the completion information provided by our algorithm (completed motifs with scores) to a simple indication of membership. We accomplish this by identifying the proteins which form the highest ranked complete motifs when plugged into the motifs. Now, we can compare the accuracy of both algorithms. For our experiments, we set the restart probability to c = 0.2 and performed the random walk n = 100,000 times for each one of the proteins in the motifs. In Fig. 4, we compare the accuracy of the network motif-random walk method for pathway completion to the accuracy of the random walk method for complex/pathway membership using LOOCV. In Fig. 4-A, -B, and -C, we plot the results of LOOCV for MAPK with Confident, Naïve Bayes, and PIT-Network, respectively. In Fig. 4-D, -E, and -F, we plot the accuracy for regulation of Autophagy. Similar to [2], the x-axis shows the threshold rank and the y-axis shows the true positive rate (TPR), which is equivalent to the percentage of left-out proteins that were predicted correctly to be members of the pathway. In Fig. 4-A, for instance, we observe that around 40% of the left-out proteins from MAPK are ranked among the top-20 in the list of candidate proteins retrieved from ConfidentNet by our method. On the other hand, approximately 30% of the left-out proteins are ranked among the top-20 proteins for the membership method. Overall, our method has better or similar accuracy most of the time especially with respect to the top ranked proteins that were retrieved from the three probabilistic PPI networks for both pathways.

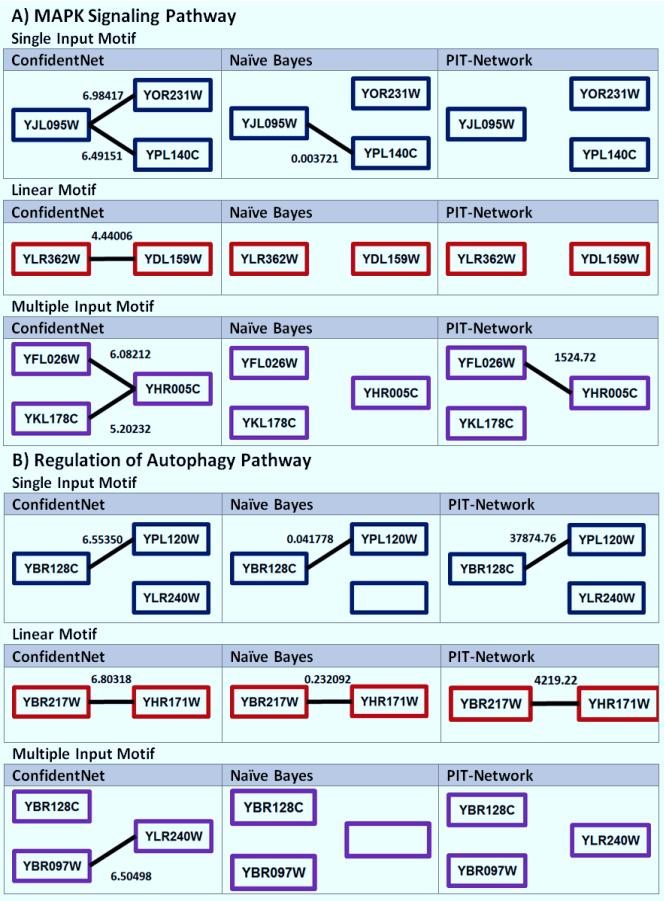


Figure 3. Examples of motifs and their representation in the three probabilistic PPI networks to demonstrate network connectivity, edge weights, and node availability. The motifs are taken from MAPK (A) and from regulation of Autophagy (B).

In Fig. 5, we show the percentage of leave-one-out queries returning the left-out protein with a rank in top-20 for different values of the restart probability c. The queries were performed on the three probabilistic PPI networks for each pathway using the two methods. The purpose of this analysis is to observe the effect of c on the rank of the left-out protein in the list of candidate proteins retrieved from a given network. In Fig. 5-A and -B, we plot the results obtained for MAPK after the application of the pathway completion and the complex/pathway membership methods to the three probabilistic PPI networks. In Fig. 5-A, we observe that around 30-40% of the leave-one-out queries returned the left-out protein in top-20 after we applied our pathway completion method to ConfidentNet (dotted curve). The percentage varies as c varies; however, the set of left-out proteins retrieved in top-20 remains the same with some proteins falling out of the set and back into the set as c changes. The percentages of the leave-one-out queries, which return the left-out protein in top-20 for PIT-Network (dashed curve) and Naïve Bayes (continuous curve), are roughly 11-23% and 14-16% respectively. In Fig. 5-B, we observe that the percentage of leave-one-out queries is around 20-30 % for MAPK and ConfidentNet with the application of the complex/pathway membership method. The percentages for Naïve Bayes and PIT-Network with the complex/pathway membership method are somewhat similar to the percentages achieved by pathway completion. On the right side of the plots, we utilize Venn diagrams to show the total numbers of left-out proteins that were returned in top-20 and their intersection across the three probabilistic PPI networks. For pathway completion (Fig. 5-A Venn diagram), the total numbers of the left-out proteins that

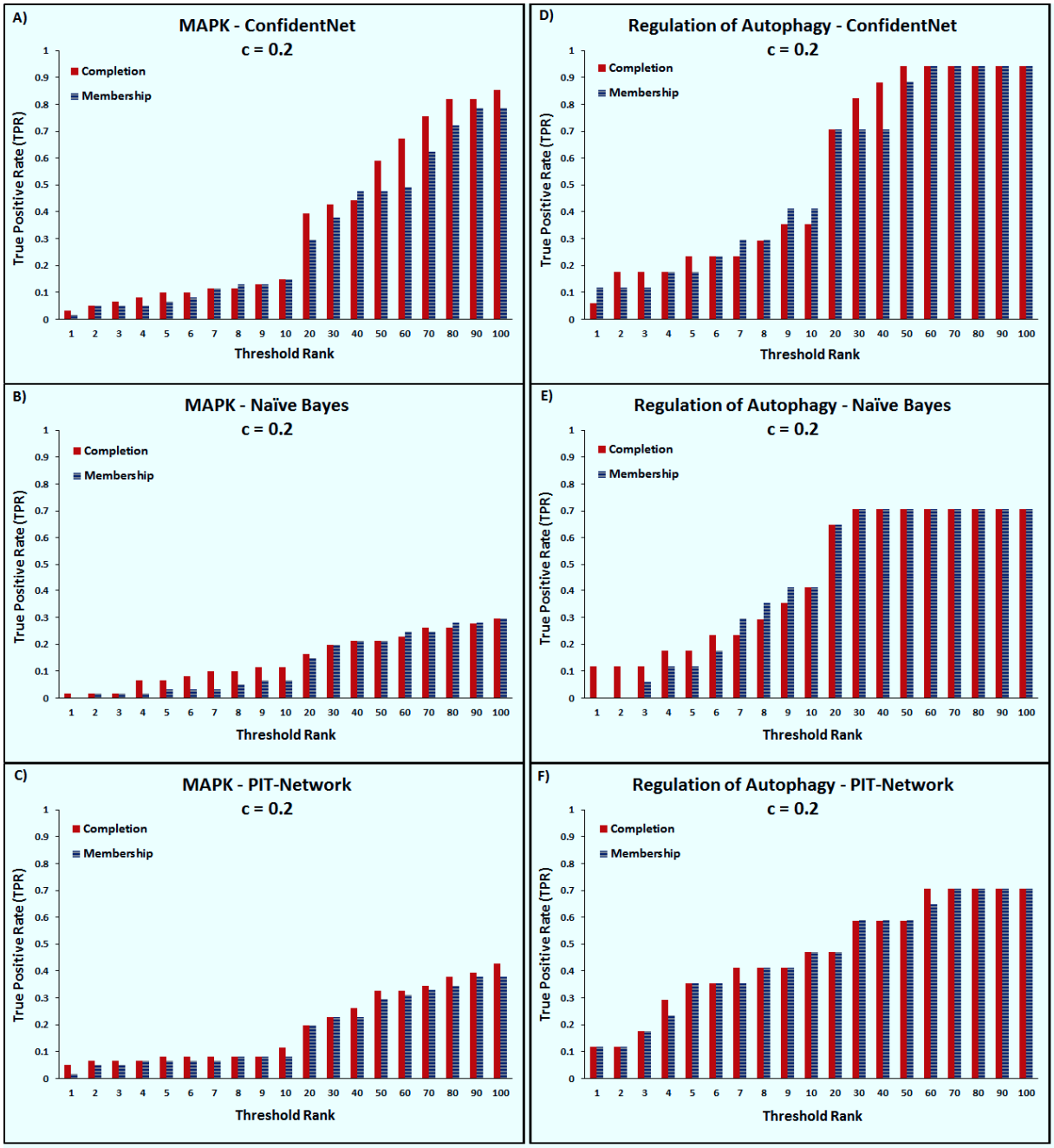


Figure 4. Accuracy assessment for pathway completion and complex/pathway membership when applied to MAPK and regulation of Autophagy and the three probabilistic PPI networks.

were retrieved in top-20 from ConfidentNet, Naïve Bayes, and PIT-Network are 23, 10, and 14, respectively, out of 61 proteins. For pathway membership (Fig. 5-B), the total numbers of the left-out proteins returned in top-20 from ConfidentNet, Naïve Bayes, and PIT-Network are 17, 9, and 14, respectively. In the Venn diagram for our pathway completion method (Fig. 5-A), we observe 5 left-out proteins retrieved in top-20 that are common across all the three networks. For pathway membership, the number of proteins is 4. Similar observations can be made with respect to the plots for regulation of Autophagy, which we show in Fig. 5-C and -D. For instance, approximately 59-76% of the leave-one-out queries returned the left-out protein in top-20 when we applied our network motif-random walk method for pathway completion to ConfidentNet. The percentage is roughly 65-70% with the application of the random walk method for complex/pathway membership. The percentages for Naïve Bayes are roughly 60-70% and 65-70% for pathway completion and complex/pathway membership, respectively, while for PIT-Network the percentages are 47-59% and 41-59%, respectively. It is important to note that the methods are not sensitive to the variations in c in terms of the set of left-out proteins that were returned in top-20. Larger values of c, which restrict the random walk to create a more local view of

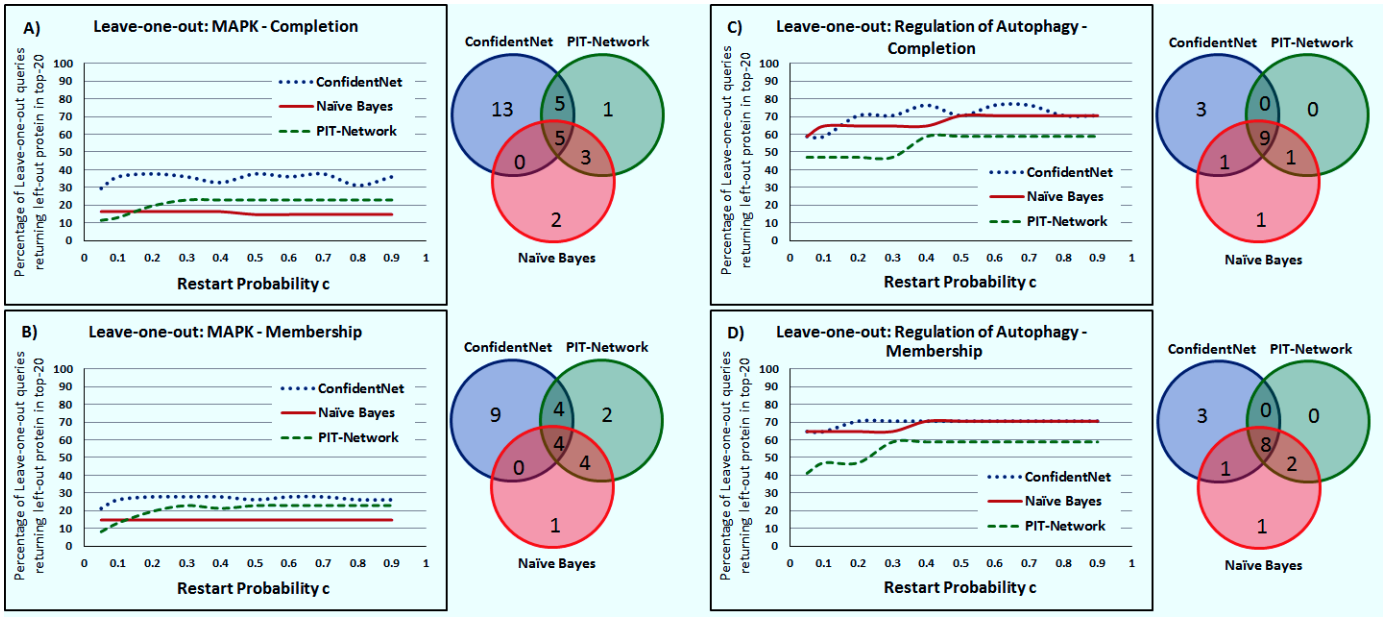


Figure 5. A comparison of the percentages of leave-one-out queries returning the left-out protein in top-20 with variations in the restart probability *c* for the three probabilistic PPI networks. A) and B) show MAPK with pathway completion and complex/pathway membership, respectively. C) and D) depict regulation of Autophagy. The Venn diagrams to the right of each plot show the intersection of the left-out proteins returned in top-20 across the three networks.

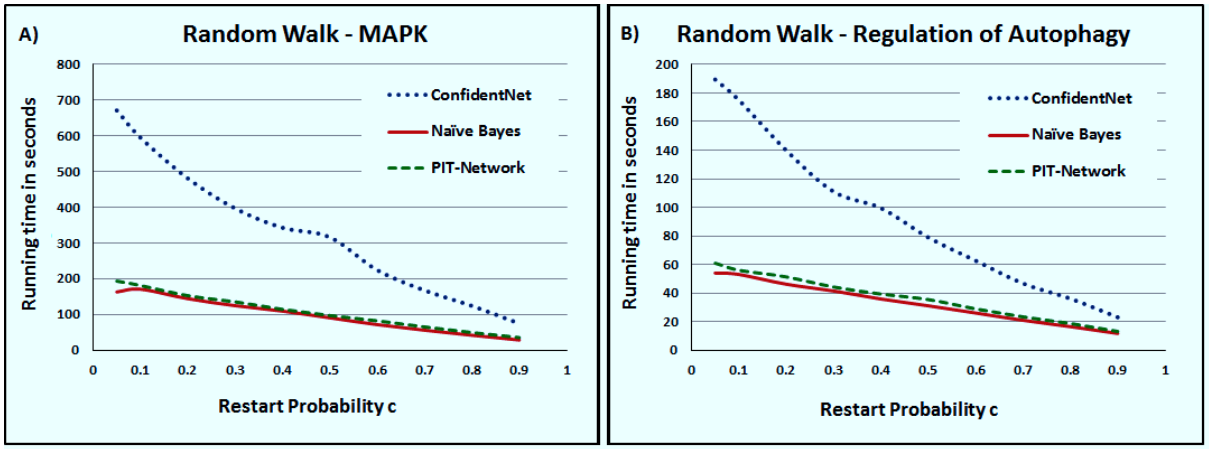


Figure 6. The running time of the random walk algorithm with varying restart probability *c* when applied to A) MAPK and B) regulation of Autophagy and the three probabilistic PPI networks.

the network to compute proximity, perform generally better which is important since larger values of *c* also mean faster computation (see Fig. 6).

We ran the algorithms, which we implemented using Jython, on a Dell R710 Dual Six Core Intel Xeon 3.4GHz 96GB machine. In Fig. 6 above, we show the running time for random walk after varying the restart probability *c*. Overall, the running time for MAPK is greater than that for regulation of Autophagy because MAPK is a larger pathway. In Fig. 6-A and -B, we observe the same time pattern for the random walk. The running time decreases as the restart probability increases. The running time is the highest for ConfidentNet (dotted curve) since it is a larger more connected network in comparison to Naïve Bayes (continuous curve) and PIT-Network (dashed curve).

V. Conclusion

In this paper, we proposed the network motif-random walk method to address the pathway completion problem of which the pathway membership problem is a sub-problem. Our method exploits the intrinsic local structures of the pathways, which are represented by network motifs. We studied the accuracy of our method using two KEGG pathways and three probabilistic PPI networks. Our experiments show that the approach is not sensitive to parameter choices, and that the accuracy of our method is very competitive with the existing random walk method for the complex/pathway membership problem. This indicates that the local structure information represented by the motifs has the potential to improve pathway membership prediction. However, the main advantage of our method is its ability to identify possible locations and

connections of candidate proteins in an incomplete pathway, which allows for more efficient experimental verification. Complex motifs, the evaluation of the accuracy of the predicted location, and experiments with more pathways are our next research goals.

ACKNOWLEDGMENT

This work was supported in part by research grant no. R21HG005912 from the National Human Genome Research Institute.

REFERENCES

[1] D. Bader and K. Madduri, "A graph-theoretic analysis of the human protein-interaction network using multicore parallel algorithms," *Parallel Computing*, vol. 34, no. 11, pp. 627-639, Nov. 2008.

[2] T. Can, O. Camoglu, and A. K. Singh, "Analysis of protein-protein interaction networks using random walks," in *Proc. 5th Int. Workshop Bioinformatics BIOKDD '05*, pp. 61-68, Aug. 2005.

[3] G. D. Bader and C. W Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC Bioinformatics*, vol. 4, no. 2, Jan. 2003.

[4] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 8, pp. 4569-4574, Apr. 2001.

[5] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg, "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae," *Nature*, vol. 403, no. 6770, pp. 623-627, Feb. 2000.

[6] S. Asthana, O. D. King, F. D. Gibbons, and F. P. Roth, "Predicting protein complex membership using probabilistic network reliability," *Genome Research*, vol. 14, no. 6 , pp. 1170-1175, Jun. 2004.

[7] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S.-L. L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sørensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. Hogue, D. Figeys, and M. Tyers, "Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry," *Nature*, vol. 415, no. 6868, pp. 180-183, Jan. 2002.

[8] A.-C. C. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A.-M. M. Michon, C.-M. M. Cruciat, M. Remor, C. Höfert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.-A. A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga, "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature*, vol. 415, no. 6868, pp. 141-147, Jan. 2002.

[9] O. Camoglu, T. Can, and A. K. Singh, "Accurate and scalable techniques for the complex/pathway membership problem in protein networks," *Advances in Bioinformatics*, 2009.

[10] S. Letovsky and S. Kasif, "Predicting protein function from protein/protein interaction data: a probabilistic approach," *Bioinformatics*, vol. 19, no. Suppl 1, pp. i197-i204, Feb. 2003.

[11] J. S. Bader, "Greedily building protein networks with confidence," *Bioinformatics*, vol. 19, no. 15, pp. 1869-1874, Oct. 2003.

[12] H. Ho, T. Milenković, V. Memišević, J. Aruri, N. Pržulj, and A. K. Ganesan, "Protein interaction network topology uncovers melanogenesis regulatory network components within functional genomics datasets," BMC Systems Biology, vol. 4, no. 84, Jun. 2010.

[13] M. El Dayeh and M. Hahsler, "Analyzing incomplete biological pathways using network motifs," *27th Symp. in Appl. Computing ACM SAC'12*, in press.

[14] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," Science, vol. 298, no. 5594, pp. 824-827, Oct. 2004.

[15] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27-30, Jan. 2000.

[16] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "KEGG for integration and interpretation of large-scale molecular datasets," *Nucleic Acids Research*, vol. 40, no. Database issue, pp. D109-D114 Jan. 2012.

[17] H. W. Mewes, A. Ruepp, F. Theis, T. Rattei, M. Walter, D. Frishman, K. Suhre, M. Spannagl, K. F. X. Mayer, V. Stümpflen, and A. Antonov, "MIPS: curated databases and comprehensive secondary data resources in 2010," *Nucleic Acids Research*, vol. 39, no. Suppl 1, pp. D220–D224, Jan. 2011.

[18] G. Ciriello and C. Guerra, "A review on models and algorithms for motif discovery in protein-protein interaction networks," *Briefing in Functional Genomic and Proteomics*, vol. 7, no. 2, pp. 147-156, Mar. 2008.

[19] E. Yeger-Lotem, S. Sattath, N. Kashtan, S. Itzkovitz, R. Milo, R. Y. Pinter, U. Alon, and H. Margalit, "Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 16, pp. 5934-5939, Apr. 2004.

[20] Z. Han, T. M. Vondriska, L. Yang, W. R. MacLellan, J. N. Weiss, and Z. Qu, "Signal transduction network motifs and biological memory," *J Theor Biol*, vol. 246, no. 4, pp. 755-761, Jun. 2007.

[21] M. Koyuturk, A. Grama, and W. Szpankowski, "An efficient algorithm for detecting frequent subgraphs in biological networks," *Bioinformatics*, vol. 20 suppl 1, no. 1, pp. 200-207, Aug. 2004.

[22] U. Alon, "Network motifs: theory and experimental approaches," *Nature*, vol. 8, no. 6, pp. 450-461, Jun. 2007.

[23] S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulaiton network of Escherichia coli," *Nature Genetics*, vol. 31, no. 1, pp. 64-68, Apr. 2002.

[24] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E.G. Jennings, H. L. Murray,D.B. Gordon, B. Ren, J. J. Wyrick, J.B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young, "Transcriptional regulatory networks in Saccharomyces cerevisiae," *Science*, vol. 298, no. 5594, pp. 799-804, Oct. 2002.

[25] H. Tong, Y. Koren, and C. Faloutsos, "Fast direction-aware proximity for graph mining," in *Proc. 13th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining KDD '07*, San Jose, CA, 2007, pp. 747-756.

[26] K. Voevodski, S. Teng, and Y. Xia, "Spectral affinity in protein networks", *BMC Systems Biology*, vol. 3, no. 112, Nov. 2009.

[27] J. Pandey, M. Koyuturk, and A. Grama, "Functional characterization and topological modularity of molecular interaction networks", *BMC Bioinformatics*, vol. 11 suppl 1, S35, Jan. 2010.

[28] L. Lovasz, "Random walks on graphs: a survey," *Combinatorics, Paul Erdos is Eighty*, vol. 2, pp. 353-398, 1993.

[29] I. Lee, S. V. Date, A. T. Adai, and E. M. Marcotte, "A probabilistic functional network of yeast genes," *Science*, vol. 306, no. 5701, pp. 1555-1558, Nov. 2004.

[30] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein, "A bayesian networks approach for predicting protein-protein interactions from genomic data," *Science*, vol. 302, no. 5644, pp. 449-453, Oct. 2003.