Graph Theoretic Concepts as the Building Blocks for Disease Initiation and Progression at Protein Network Level: *Identification and Challenges*

Ananda Mohan Mondal*
School of Computing and Information Sciences
Florida International University
Miami, USA
amondal@fiu.edu

Cornelia Ada Schultz
Department of Mathematics and Computer Science
Claflin University
Orangeburg, USA
Cornelia.ada.schultz@gmail.com

Markea Sheppard
Department of Mathematics and Computer Science
Claflin University
Orangeburg, USA
markeasheppard@gmail.com

Jasmine Carson
Department of Mathematics and Computer Science
Claflin University
Orangeburg, USA
jacarson50@gmail.com

Raihanul Bari Tanvir
School of Computing and Information Sciences
Florida International University
Miami, USA
rtanv003@fiu.edu

Tasmia Aqila
School of Computing and Information Sciences
Florida International University
Miami, USA
taqil001@fiu.edu

***Abstract* --- Protein networks that mirror the transitions between disease stages hold the key to early diagnosis and make it easy to understand the essential mechanisms of disease progression at protein network level. But, identifying critical transitions between disease stages and corresponding protein networks during the initiation and progression of a complex disease like cancer is a challenging task. This preliminary work identifies the possible building blocks for disease initiation and progression at the protein network level based on biological rationale that a group of proteins are localized at a specific subcellular location to accomplish a function, which could be beneficial to human body or adversarial to cause a disease. We discovered that three graph-theoretic concepts – i) Clique-like structures, ii) Bipartite-like structures, and iii) Diffusion Kernels could be possible building blocks for disease progression at the protein network level. Using these building blocks, disease progression can be modeled as an event-schedule-like structure, meaning that each of the disease stages corresponds to an event, where each event is completed by a set of proteins by forming a clique-like structure. Once an event or disease stage is completed by a group of proteins, disease signals go to the next group of proteins to cause the next event or disease stage and so on. The transfer of signals can be represented by bipartite-like structure and diffusion kernels can be used to find the strength of disease signals. Further study is required to fully explore the application of these building blocks to analyze the disease progression.**

Keywords --- bipartite graph, clique, diffusion kernel, disease progression, protein network.

I. Introduction

Studies on disease progression [1-5] for different diseases using time-series gene expression profiles on human and mouse genomes show that there exists a dynamical network biomarker (DNB), a group of proteins whose behavior, unlike other groups of proteins, changes at the pre-disease state of a three-state (normal, pre-disease, and disease) model for disease progression. The major limitation of studies based on DNB is that the researchers hypothesized that disease progression is composed of three states only - normal state, pre-disease state, and disease state. In reality, disease progression may have more than three states. According to the sixth edition of the cancer staging system by American Joint Committee on Cancer (AJCC), the disease state of colon cancer has 7 different stages (I, IIa, IIb, IIIa, IIIb, IIIc, and IV) [6]. Similarly, lung cancer also has 7 stages (Ia, Ib, IIa, IIb, IIIa, IIIb, and IV) in the disease state devised by International System for Staging Lung Cancer [7]. The second limitation of the studies based on the three-state model is that, in the disease state, the member proteins of a DNB behave normally like the rest of the proteins in the network. Thus, it is clear that DNBs fail to differentiate among different stages of the disease state.

Our work is motivated by the prospective applications of protein-protein interaction (PPI) networks or, simply, protein networks to diseases [8]. Ideker and Sharan [8] enumerated four different applications of protein networks to diseases: i) identifying new disease genes, ii) studying the network properties of disease genes, iii) classifying diseases based on protein network, and iv) identifying disease-related subnetworks. Genome-wide protein-protein interaction (PPI) networks come with rich information about the dynamic processes such as the behavior of genetic networks in response to DNA

Funding Agency: NSF-iAAMCS, NSF CAREER.

damage [9] and exposure to arsenic [10], the prediction of protein function [11], genetic interaction [12], protein subcellular localization [13-18], the process of aging [19], and protein network biomarkers [20, 21]. Based on these literatures, it is quite reasonable to claim that the signature of disease progression, which is dynamic, is also left behind in static PPI network. In this paper, we used PPI network as the backbone for discovering the building blocks for disease initiation and progression. We leverage the motivation of a group of proteins to be localized at a specific subcellular location to accomplish a common goal or function which is similar for a group of proteins to be involved in initiating a disease and subsequent progression from one stage to the next.

Limitations of Computational Studies

According to the supplementary document of [2], a DNB network is neither a set of disease genes nor a driving factor. It only provides early-warning signals of the pre-disease state based on its dynamical features from the observable data such as time-series gene expression. The second limitation of the studies based on the three-state model is that, in the disease state, according to [1-5], the member proteins of a DNB behave normally like the rest of the proteins in the network. The third limitation is that the disease state itself has more stages; for example, both colon cancer and liver cancer have seven stages. This means that DNBs fail not only to identify the genes/proteins that initiate the disease but also genes/proteins responsible for each stage of disease progression.

II. Hypothesis and Challenges

To overcome the limitations of the state-of-the-art computational studies on disease progression, one needs to have a network biomarker that is capable of representing the whole disease progression from its initiation. This is possible if the network biomarker has an event-schedule-like structure, meaning that each of the disease stages corresponds to an event, where each event is completed by a set of proteins as shown in Fig. 1.

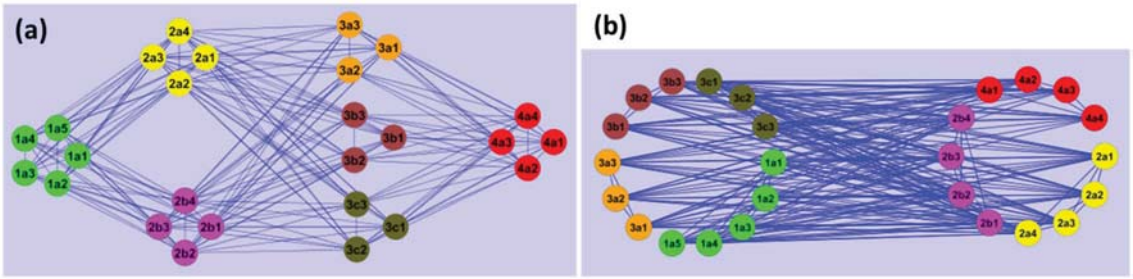


Fig. 1. Hypothetical protein network model for colon cancer. (a) Event-schedule-like protein network model for colon cancer. Seven stages (I, IIa, IIb, IIIa, IIIb, IIIc, IV) of colon cancer are represented by seven different colors. (b) Event-schedule-like structure collapsed into a single clique-bipartite-like graph.

Fig. 1a represents the hypothetical event-schedule-like protein network depicting seven stages (I, IIa, IIb, IIIa, IIIb, IIIc, and IV) of colon cancer. Once the green event (Stage I) is completed by the green group of proteins, signals go to the yellow group of proteins to cause the yellow event (Stage IIa) and to the purple group of proteins to cause the purple event (Stage IIb). Once the blue and purple events are complete, signals go to the orange group of proteins to cause the orange event (Stage IIIa), to the brown group of proteins to cause the brown event (Stage IIIb) and to the olive group of proteins to cause the olive event (Stage IIIc). Finally, signals go to the red group of proteins to cause the red event (Stage IV). For an event, a group of proteins works together forming a clique or clique-like structure and transfer of signals from a stage to the next form a bipartite graph. It is noticeable that the whole disease progression represented by multiple clique-bipartite graphs (Fig. 1a) can be collapsed into a single clique-bipartite-like graph as shown in Fig. 1b. This phenomenon leads to the algorithmic challenge of identifying the most likely event-schedule-like structure for disease progression given a protein network for a disease.

III. Methodology

Datasets Preparation: Two sets of data, namely i) list of biomarkers or single protein biomarkers (SPBs) and ii) protein-protein interaction data are required for identifying protein subnetwork biomarkers for a disease. The list of biomarkers, 84 key genes commonly involved in the dysregulation of signal transduction and other normal biological processes during disease, is obtained from SABiosciences of Qiagen [22]. Genome-wide PPI networks for human are obtained from STRING database [23]. Protein

subcellular locations, needed to annotate the proteins of protein network biomarker, are obtained from the cellular components of GO (Gene Ontology) database [24]. The details of cleaning these data can be found in [21].

Original PPI dataset, downloaded from STRING database version 9.0, contains 3,281,414 PPIs. For the present study, direction of interaction is not important. After removing direction and some erroneous data (860 in total: some are missing scores, some do not conform to STRING names etc.), final dataset contains 1,640,129 PPIs with 18,595 proteins.

STRING PPIs do not come with official protein names but disease proteins procured from Qiagen [22] are in official protein names. A mapping between STRING and official protein names is required. Another file from STRING database with GO annotation contains both STRING and official protein names, which is used as the mapping file. Original mapping file contains 17919 unique records. After cleaning some erroneous data (some protein names are in numbers i.e., not in official protein names), left with 17839 unique records. Finally, STRING PPIs are converted to PPIs in official protein names and working network is composed of 1,568,065 PPIs and 16,614 proteins. So, on an average, there are 94 interactions per protein. PERL program was used to clean the data.

Constructing Protein Network Biomarker: The disease genes obtained from Qiagen were overlaid on top of PPI network obtained from STRING database to construct the protein network biomarker for a disease.

Filtering Proteins Using Cytoscape: Cytoscape [25], an open-source software, is a tool for analyzing biomolecular networks. Protein network biomarker (list of PPIs) obtained above and a list of protein annotated with subcellular locations are loaded in Cytoscape. Then a filter was created by grouping the proteins based on their locations. The rationale of using this approach is that the group of proteins localized at the same subcellular location is more likely to interact with each other to cause a function, which could be beneficial to our body or adversarial to initiate a disease.

IV. Results and Discussions

Clique and Bipartite Graph as Building Blocks: Fig. 2 shows the filtered proteins as grouped by locations from a protein network biomarker for liver cancer. It is clear that the groups of proteins at different locations form two distinct network structures, namely, clique-like structure and bipartite graph. The largest group (group-1) of 21 proteins is located at Cytoplasm. The 2nd, 3rd, and 4th groups of proteins are located at Nucleus, Plasma Membrane, and Extracellular, respectively.

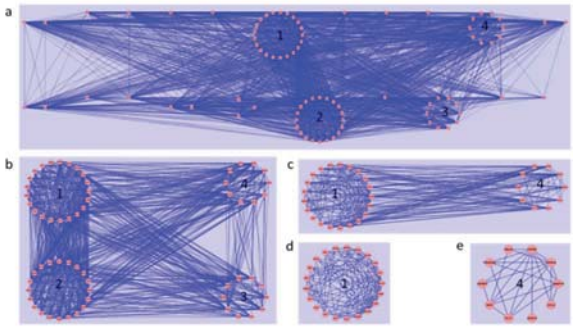


Fig. 2. Clique and bipartite components of protein network biomarker for liver cancer based on protein locations: (a) Protein network biomarker, extracted from whole-genome PPI network using the 84 key genes adopted from QIAGEN [22]. The Largest group (group-1) of 21 proteins is located at Cytoplasm. The 2nd, 3rd and 4th groups of proteins are located at Nucleus, Plasma Membrane, and Extracellular, respectively. **(b)** A clique-bipartite-like structure with four clique-like and six bipartite-like structures, isolated from (a). **(c)** A clique-bipartite-like structure with two clique-like and one bipartite-like structure. **(d)** A clique-like structure among the proteins in the 1st group. **(e)** A clique-like structure among the proteins in the 4th group.

First, intra-group proteins form a clique-like structure, Fig. 2(d, e), meaning they form a cluster in the protein network and interact with each other, usually, to accomplish a function. Second, each group of proteins is connected with other groups by forming a bipartite-like structure, Fig. 2c. Four groups of proteins together form a clique-bipartite-like structure composed of four clique-like and six bipartite-like structures among themselves, as isolated in Fig. 2b.

Usually, proteins at a location, work together as a group, are responsible for a specific function or event to occur, maybe a specific disease stage. The bipartite structure between two groups of proteins at two different locations can be thought of as cross-talks or the flow of signals between two groups or events or disease stages. These together with the literatures mentioned in introduction motivate the authors to come up with the hypothesis that different stages of a disease or whole disease progression process can be represented in terms of clique-like and bipartite-like structures at the protein network level. ***Assumption-1:*** Each of these clusters of proteins or clique-like structures corresponds to one disease stage, which can also be thought of as an event. Since the disease is a complex phenomenon, the formation of a clique-like structure representing a disease stage should not be based on location only. Other factors, both genetic and epigenetic, such as gene expression, mutation, DNA methylation, histone modification, and miRNA dysregulation should be accounted for. ***Assumption-2:*** Once an event is complete or a disease stage is complete by a cluster of proteins, they send the signal, by forming a bipartite graph, to the next group of proteins to start the next event or next disease stage and so on. ***Assumption-3:*** The signal or potential to

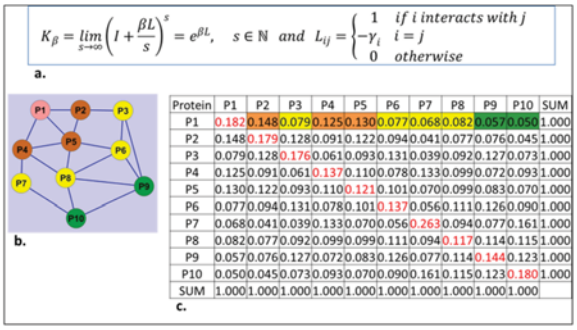
cause a disease associated with individual gene/protein will be evaluated considering both genetic and epigenetic factors mentioned in assumption-1.

Diffusion Kernel as the Building Block: We discovered two possible building blocks, clique-like structure and bipartite graph, for disease progression based on protein localization. In case of actual disease initiation and progression, the formation of clique-like structures and bipartite-like structures will be based on disease-causing factors/signals both genetic (gene expression and mutation) and epigenetic (DNA methylation, histone modification, and microRNA dysregulation). In a genome-wide PPI network, there will be a lot of clique-like and bipartite-like structures. The **overarching question** is – how to find clique-like and bipartite-like structures that are related to a specific disease given the disease-causing factors/signals associated with each protein of the genome-wide PPI network? To address this question, the factors and signals can be thought of as potentials that can travel/diffuse in any random direction in a graph or protein network.

A diffusion kernel, explained later, on a protein network is equivalent to a random walk on a graph [26]. The kernel values are used by different investigators as a measure of information flow between two proteins in a network [11-18]. So, a diffusion kernel can be considered as the representation of flow of disease signal between proteins and the kernel value between two proteins can be considered as the strength of this signal. An appropriate threshold on kernel values can be used to identify the edges that will form possible clique-like structures responsible for different stages of a disease including initiation as well as bipartite-like structures among the identified clique-like structures. The rationale for using some threshold on kernel values is also evidenced from [27], where the authors used a threshold on kernel values in identifying the missing connections in a protein network biomarker.

Diffusion Kernel on a Protein Network: Diffusion kernel is a computational framework that is based on the physical phenomenon of gas diffusion in a medium, which is also equivalent to the Computer Science concept of random walk on a graph [26]. PPI network or protein network is a graph where each node represents a protein and a connection or an edge between two nodes represents the existence of an interaction between two proteins. A genome-wide PPI network comes with rich information about the signature of the disease process [8], protein functions [11], genetic interaction [12], protein subcellular localization [13-18], etc. The randomness of the flow of this information from one protein to another is hidden inside the complex structure of the PPI network, which makes it difficult to decipher this information. A diffusion kernel, since it is based on random walk on a graph, provides a suitable computational framework to extract meaningful biological information from the PPI network. Application of a diffusion kernel provides improved results compared to the state-of-the-art methods for predicting protein functions [11], genetic interaction [12], and protein subcellular localization [13-18]. These factors motivate using a diffusion kernel to find the strength of disease signal between two proteins in a protein network.

The formal definition of diffusion kernel on a PPI network [26], Fig. 3a, corresponds to a random walk with an infinite number of infinitesimally small steps. In the formula of Fig. 3a, I is the identity matrix, β is the diffusion constant, L is a Laplacian matrix, and γ_i is the number of interaction partners of protein i. Fig. 3b shows an example of a protein network and Fig. 3c is the corresponding kernel matrix.



Protein	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	SUM
P1	0.182	0.148	0.079	0.125	0.130	0.077	0.068	0.082	0.057	0.050	1.000
P2	0.148	0.179	0.128	0.091	0.122	0.094	0.041	0.077	0.076	0.045	1.000
P3	0.079	0.128	0.176	0.061	0.093	0.131	0.039	0.092	0.127	0.073	1.000
P4	0.125	0.091	0.061	0.137	0.110	0.078	0.133	0.099	0.072	0.093	1.000
P5	0.130	0.122	0.093	0.110	0.121	0.101	0.070	0.099	0.083	0.070	1.000
P6	0.077	0.094	0.131	0.078	0.101	0.137	0.056	0.111	0.126	0.090	1.000
P7	0.068	0.041	0.039	0.133	0.070	0.056	0.263	0.094	0.077	0.161	1.000
P8	0.082	0.077	0.092	0.099	0.099	0.111	0.094	0.117	0.114	0.115	1.000
P9	0.057	0.076	0.127	0.072	0.083	0.126	0.077	0.114	0.144	0.123	1.000
P10	0.050	0.045	0.073	0.093	0.070	0.090	0.161	0.115	0.123	0.180	1.000
SUM	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	

c.

Fig. 3. Diffusion kernel on a PPI network. a) Definition of diffusion kernel. b) PPI network with 10 proteins/nodes and 18 PPIs/edges on which diffusion kernel is applied. c) Output of diffusion kernel, a global similarity matrix with weights on all possible edges.

A diffusion kernel generates edge weights (interpretable as similarity) between two proteins of all possible protein pairs as seen in Fig. 3c, which is based on a global perspective of the network. For example, based on the direct use of a graph, proteins with the same shortest path distance will have the same similarity, while a diffusion kernel will produce a different similarity. This property makes the diffusion kernel perform better than the direct use of a graph. For example, from the protein P1, the green proteins (P9 and P10) at the shortest path distance of 3 (Fig. 3b) will have the same similarity value of 1/3 (inverse of distance) with the protein P1, but the diffusion kernel produces different values of similarity (P9: 0.057and P10: 0.050), as seen in Fig. 3c. Similarly, the diffusion kernel produces different values of similarity for the brown proteins at the shortest path distance of 1 as well as for the yellow proteins at the shortest path distance of 2.

Big Data and Precision Medicine Perspective: Though the protein network (for example – 84 nodes and 1900 edges) for a disease is small in size, finding the most likely event-schedule-like structure representing disease progression is combinatorial or complex in nature, which makes this problem a big data problem. At the same time, any of the combinations (a specific event-schedule-like structure) can be related to a specific patient, which could be utilized for designing the right medicine and right dose for a specific patient. An event-schedule-like structure that matches the disease-causing parameters for a specific patient both genetic (gene expression and mutation) and epigenetic (DNA methylation, histone modification, and miRNA dysregulation) will be used for representing actual model for the disease progression for that person.

Fig. 4c shows a clique-bipartite-like structure developed from a protein network for breast cancer composed of 84 proteins and 1900 PPIs using the two-color technique [28]. Using breadth-first-search (BFS), Fig. 4a, and depth-first-search (DFS), Fig. 4b, a network can be colored using only two colors such that any two adjacent nodes will have different colors. In Fig. 4, A and B represent the two colors; thus, proteins in alternate levels are designated as A and B.

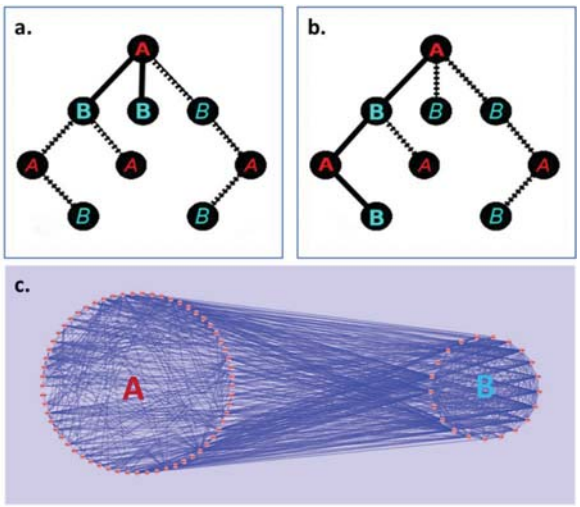


Fig. 4. Clique-bipartite-like structure using BFS and DFS algorithms. a) Diagram of BFS. b) Diagram of DFS. c) Clique-bipartite-like structure for protein network for breast cancer.

Two sets of proteins, set A and set B in Fig. 4c, representing clique-bipartite-like structure, can be obtained using both BFS and DFS algorithms. Since any protein can be the root node, a maximum of 168 (84x2) different clique-bipartite-like structures can be generated for the given protein network.

Challenge-1: The desired single clique-bipartite-like structure representing the collapsed network for disease progression could be any of 168 different single clique-bipartite-like structures. ***Challenge-2:*** Once the single clique-bipartite-like structure representing the collapsed network for disease progression (Fig. 1b) is identified, it needs to be unfolded to represent the event-schedule-like structure (Fig. 1a) for the progression of disease. These two challenges, which are combinatorial in nature, make this problem a big data problem.

V. Conclusion and Future Remarks

This paper discovered three graph theoretic concepts – clique-like structures, bipartite-like structures, and diffusion kernels – that can be used as the building blocks for disease progression from stage to stage including initiation. Biological rationale that a group of proteins are localized at a subcellular compartment to accomplish a specific function is used in discovering cliques or clique-like structures to represent a disease stage. Cross-talks among these clique-like structures are used in discovering bipartite-like structures as the second building block. Bipartite-like structures represent the transfer of disease signals from one stage to the next. Finally, the physical phenomenon of gas diffusion is used to discover the third building block, diffusion kernels, which represent the strength of disease signals to be transferred from one stage to the next. Further experiment is required for validating these building blocks.

Acknowledgment

This work was partially supported by NSF-iAAMCS as a subcontract from Winston-Salem State University to Claflin University and NSF CAREER Award no - 1651917.

References

[1] R. Liu, M. Li, Z.-P. Liu, J. Wu, L. Chen, and K. Aihara, "Identifying critical transitions and their leading biomolecular networks in complex diseases," *Scientific reports,* vol. 2, 2012.

[2] L. Chen, R. Liu, Z.-P. Liu, M. Li, and K. Aihara, "Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers," *Scientific reports,* vol. 2, 2012.

[3] M. Li, T. Zeng, R. Liu, and L. Chen, "Detecting tissue-specific early warning signals for complex diseases based on dynamical network biomarkers: study of type 2 diabetes by cross-tissue analysis," *Briefings in bioinformatics,* vol. 15, pp. 229-243, 2014.

[4] R. Liu, K. Aihara, and L. Chen, "Dynamical network biomarkers for identifying critical transitions and their driving networks of biologic processes," *Quantitative Biology,* vol. 1, pp. 105-114, 2013.

[5] X. Liu, R. Liu, X.-M. Zhao, and L. Chen, "Detecting early-warning signals of type 1 diabetes and its leading biomolecular networks by dynamical network biomarkers," *BMC medical genomics,* vol. 6, p. S8, 2013.

[6] J. B. O'Connell, M. A. Maggard, and C. Y. Ko, "Colon cancer survival rates with the new American Joint Committee on Cancer sixth edition staging," *Journal of the National Cancer Institute,* vol. 96, pp. 1420-1425, 2004.

[7] C. F. Mountain, "Revisions in the international system for staging lung cancer," *Chest Journal,* vol. 111, pp. 1710-1717, 1997.

[8] T. Ideker and R. Sharan, "Protein networks in disease," *Genome Res,* vol. 18, pp. 644-52, Apr 2008.

[9] S. Bandyopadhyay, M. Mehta, D. Kuo, M.-K. Sung, R. Chuang, E. J. Jaehnig*, et al.*, "Rewiring of genetic networks in response to DNA damage," *Science,* vol. 330, pp. 1385-1389, 2010.

[10] A. C. Haugen, R. Kelley, J. B. Collins, C. J. Tucker, C. Deng, C. A. Afshari*, et al.*, "Integrating phenotypic and expression profiles to map arsenic-response networks," *Genome biology,* vol. 5, p. R95, 2004.

[11] H. Lee, Z. Tu, M. Deng, F. Sun, and T. Chen, "Diffusion kernel-based logistic regression models for protein function prediction," *OMICS,* vol. 10, pp. 40-55, Spring 2006.

[12] Y. Qi, Y. Suhail, Y.-y. Lin, J. D. Boeke, and J. S. Bader, "Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions," *Genome research,* vol. 18, pp. 1991-2004, 2008.

[13] A. M. Mondal and J. Hu, "NetLoc: Network based protein localization prediction using protein-protein interaction and co-expression networks," in *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on*, 2010, pp. 142-148.

[14] A. M. Mondal, J.-r. Lin, and J. Hu, "Network based subcellular localization prediction for multi-label proteins," in *Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on*, 2011, pp. 473-480.

[15] A. M. Mondal and J. Hu, "Protein Localization by Integrating Multiple Protein Correlation Networks," in *Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP)*, 2012, pp. 82-88.

[16] J.-R. Lin, A. M. Mondal, R. Liu, and J. Hu, "Minimalist ensemble algorithms for genome-wide protein localization prediction," *BMC bioinformatics,* vol. 13, p. 157, 2012.

[17] A. M. Mondal and J. Hu, "Scored Protein-Protein Interaction to Predict Subcellular Localizations for Yeast Using Diffusion Kernel," in *International Conference on Pattern Recognition and Machine Intelligence*, 2013, pp. 647-655.

[18] A. M. Mondal and J. Hu, "Network based prediction of protein localisation using diffusion Kernel," *International journal of data mining and bioinformatics,* vol. 9, pp. 386-400, 2014.

[19] F. E. Faisal and T. Milenković, "Dynamic networks reveal key players in aging," *Bioinformatics,* p. btu089, 2014.

[20] K. Charles, A. Afful, and A. M. Mondal, "Protein Subnetwork Biomarkers for Yeast Using Brute Force Method," in *Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP)*, 2013, pp. 218-223.

[21] P. Timalsina, K. Charles, and A. M. Mondal, "STRING PPI Score to Characterize Protein Subnetwork Biomarkers for Human Diseases and Pathways," in *Bioinformatics and Bioengineering (BIBE), 2014 IEEE International Conference on*, 2014, pp. 251-256.

[22] (May 16, 2013). *Biomarkers*. Available: http://www.sabiosciences.com/Biomarker.php

[23] C. von Mering, L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini*, et al.*, "STRING: known and predicted protein-protein associations, integrated and transferred across organisms," *Nucleic Acids Res,* vol. 33, pp. D433-7, Jan 1 2005.

[24] M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger*, et al.*, "The Gene Ontology (GO) database and informatics resource," *Nucleic Acids Res,* vol. 32, pp. D258-61, Jan 1 2004.

[25] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage*, et al.*,

"Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome research,* vol. 13, pp. 2498-2504, 2003.

[26] R. I. Kondor and J. Lafferty, "Diffusion kernels on graphs and other discrete structures," in *Proceedings of the 19th international conference on machine learning*, 2002, pp. 315-322.

[27] D. K. Bett and A. M. Mondal, "Diffusion kernel to identify missing PPIs in protein network biomarker," in *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, 2015, pp. 1614-1619.

[28] A. Levitin, *Introduction to the design & analysis of algorithms*, 3rd ed. Boston: Pearson, 2012.