UNIVERSITY OF CALIFORNIA, SAN DIEGO

Community Detection in Biological Networks

A dissertation submitted in partial satisfaction of the requirements for the degree
Doctor of Philosophy

in

Electrical Engineering (Intelligent Systems, Robotics and Control)

by

Tejaswini Narayanan

Committee in charge:

Professor Shankar Subramaniam, Chair
Professor Sadik Esener, Co-Chair
Professor Todd Coleman
Professor Rajesh Gupta
Professor Daniel Tartakovsky

2013

The Dissertation of Tejaswini Narayanan is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

Co-Chair

_____

Chair

University of California, San Diego

2013

DEDICATION

I dedicate my dissertation to the following really important people without whom I wouldn't be writing this:

Dr. Shankar Subramaniam, the out-of-the world advisor who gave me the most wonderful opportunity to co-author papers with him, under his esteemed guidance

The world's best family: Narayanan and Megala, my adoring parents who have been extremely patient, and whose words of encouragement and prayers kept me going; My brother Dhyanesh, who constantly motivated me through every result– good or bad, every paper - accepted or rejected and every segment of code - that compiled and that didn't.

To the faculty from my Alma mater (National Institute of Technology, Surat) who constantly kept reminding me of how much they were looking forward to their student receiving a PhD; and to all the teachers who gave me the invaluable gift of education.

I also dedicate this dissertation to all those who helped me directly and indirectly to get me to this point, to those who taught me what to do and what not to, and to all those who made my life at UCSD, an experience of a life-time.

TABLE OF CONTENTS

LIST OF SUPPLEMENT FILES

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Shankar Subramaniam for his immense support as the chair of my committee. Through many thought provoking research discussions and multiple iterations of publication drafts, his guidance and feedback have been invaluable. I would also like to acknowledge Professor Sadik Esener for obliging to be the co-chair of my committee, and for his continuous support and confidence in me.

I would like to acknowledge Professor Rajesh Gupta, Professor Daniel Tartakovsky, and Professor Todd Coleman for agreeing to serve on my committee, and for their valuable feedback on my dissertation and defense.

I would like to thank the members of the Subramaniam Lab, for the many interesting and informative technical discussions. A special thanks to Julian, the lab administrator, who was always available to fix problems and install new programs in my workstation.

Chapter I, in full, is the material as it appears in Narayanan T, Gersten M, Subramaniam S, Grama A: Modularity detection in protein-protein interaction networks. BMC Research Notes 2011: 4-569. The dissertation author was the primary investigator and author of this paper.

Chapter II, in full, is the material as it appears in Narayanan T, Subramaniam S: Community Detection in Biological Networks Using a Variational Bayes Approach. In proceedings of the 3rd International Conference on Bioinformatics and Computational Biology: 23 – 25 March 2011; New Orleans, Louisiana USA. The dissertation author was the primary investigator and author of this paper.

Chapter III, in full, is the material as it appears in Narayanan T, Subramaniam S: Community Structure Analysis of Gene Interaction Networks in Duchenne Muscular Dystrophy. Public Library of Science (PLoS ONE) 2013. The dissertation author was the primary investigator and author of this paper.

Chapter IV, in full, has been submitted for publication of the material as it may appear in Narayanan T; Subramaniam S. A Newtonian Framework for Community Detection in Biological Networks. IEEE Transactions on Biomedical Circuits and Systems (TBioCaS) 2013. The dissertation author was the primary investigator and author of this paper.

VITA

2008	Bachelor of Technology (B. Tech), Electrical Engineering National Institute of Technology (NIT), Gujarat, India
2010	Master of Science (M.S.), Electrical Engineering (Intelligent Systems, Robotics and Control) University of California, San Diego (UCSD), California, USA
2011	Candidate of Philosophy (C.Phil), Electrical Engineering (Intelligent Systems, Robotics and Control) University of California, San Diego (UCSD), California, USA
2013	Doctor of Philosophy (Ph.D), Electrical Engineering (Intelligent Systems, Robotics and Control) University of California, San Diego (UCSD), California, USA

PUBLICATIONS

"A Newtonian Framework for Community Detection in Biological Networks" (submitted journal paper), IEEE Transactions on Biomedical Circuits and Systems (TBioCaS)

"Community Structure Analysis of Interaction Maps Derived from Gene Expression Signatures for Duchenne Muscular Dystrophy" (accepted journal paper), Public Library of Science (PLoS ONE)

"Modularity Detection in Protein-Protein Interaction Networks" (published journal paper), BioMed Central (BMC)

"The Pottery Informatics Query Database: A New Method for Mathematic and Quantitative Analyses of Large Regional Ceramic Datasets" (published journal paper), Journal of Archaeological Method Theory

"Modelling the Effect of Allergen Exposure on Sensitization in Relation to Atopy during Childhood: A Machine Learning Approach" (published poster), Women in Machine Learning (WiML) 2011, Granada, Spain

"Community Detection in Biological Networks Using a Variational Bayes Approach" (published paper), 3rd International Conference on Bioinformatics and Computational Biology, Mar 2011, New Orleans, LA

"SVM-SPLIT: A Divide and Conquer Approach to Protein Homology Detection using Support Vector Machines with Binary Tree Architecture" (published poster, also awarded Student Registration Scholarship), ISSNIP 07, Australia

"Towards Leveraging Inference Web to Support Intuitive Explanations in Recommender Systems for Automated Career Counseling" (published paper, IEEE pub), ACHI 08, Martinique

"Exploring the Support for Spoken Natural Language Explanation in Inference Web" (published paper) ECAI'08 Greece

ABSTRACT OF THE DISSERTATION

Community Detection in Biological Networks

by

Tejaswini Narayanan

Doctor of Philosophy in Electrical Engineering (Intelligent Systems, Robotics, and Control)

University of California, San Diego, 2013

Professor Shankar Subramaniam, Chair
Professor Sadik Esener, Co-Chair

Community Detection is an interesting computational technique for the analysis of networks. This technique can yield useful insights into the structural organization of a network, and can serve as a basis for understanding the correspondence between structure and function (specific to the domain of the network). In this dissertation, I have sought to leverage this technique for the study of biological networks of practical relevance and significance.

The study begins with an exploration of existing techniques for Community Detection, following which an optimization is proposed for one of the widely used graph-theoretic approaches. As the next step, an investigation is performed on the suitability of a machine-learning based algorithm for Community Detection in the context of biological networks. Subsequently, the use of Community Detection for understanding pathology with a specific focus on Duchenne Muscular Dystrophy (DMD), is explored. This illustrated key distinguishing features in the structural and functional organization of the constituent biological pathways as it relates to DMD. Finally, a novel algorithm for Community Detection is proposed, which is motivated by a physical systems analogy. An analysis of the algorithm's properties, together with its applications to biological networks, is also presented.

I believe that the techniques and algorithms developed as part of this dissertation in the context of biological networks, have the potential to open up new vistas for therapeutic applications such as targeted drug development.

Introduction

Community detection is a key problem of interest in network analysis, with widespread applications. For instance, community structure analysis is used in the context of networks that arise in domains such as social networks to understand the fundamental social structure in a community of interacting individuals. This provides insights about the influential individuals and the strongly-networked individuals in a community. Another domain where algorithms for community detection find useful application is the topological understanding of large scale connection networks such as Internet, and how one may use the insights from community structure analysis to design more resilient communication networks. In the context of biological networks, such insights can also be used to understand the biological significance of the underlying community structure and organization of the network. Furthermore, rich toolsets have also been developed for the purpose of understanding networks from a community structure perspective.

The Newman and Girvan (NG) algorithm has shown considerable promise for community detection. This is a divisive approach that used *edge-betweenness* as a metric to drive the algorithm. In Chapter I, a novel termination criterion based on a target edge-betweenness value is proposed. Protein-protein interaction networks (PPINs) are used to demonstrate that the proposed optimization results in communities comparable to those from the NG algorithm while significantly reducing runtime.

Chapter II presents an analysis of the applicability of the Variational Bayes (VB) approach to community detection in biological networks. VB results in communities comparable in quantity and quality to those from the optimal set of communities from the NG algorithm on a known PPIN, and yields a better distribution of community

membership.

Duchenne Muscular Dystrophy (DMD) is an important pathology associated with the human skeletal muscle. Gene expression measurements of DMD skeletal muscles provide the opportunity to understand the underlying mechanisms that lead to the pathology. The technique of community detection in combination with gene expression measurements from normal and DMD patient skeletal muscle tissue is leveraged to model DMD. The findings presented in Chapter III have the potential to serve as fertile ground for therapeutic applications involving targeted drug development for DMD.

In Chapter IV, a novel framework for community detection, motivated by a physical system analogy, is proposed. A network is modeled as a system of point masses and the process of community detection is driven by leveraging the *Newtonian* interactions between the point masses in the model. The applicability of the proposed approach is illustrated by applying the algorithm on PPINs, the results of which are comparable to that of the NG algorithm. A detailed analysis on the biological interpretation of the communities produced by the approach is also presented.

Chapter I

Modularity Detection in Protein-Protein Interaction Networks

Narayanan T, Gersten M, Subramaniam S, Grama A: Modularity detection in protein-protein interaction networks. BMC Research Notes 2011: 4-569.

Abstract

Many recent studies have investigated modularity in biological networks, and its role in functional and structural characterization of constituent biomolecules. A technique that has shown considerable promise in the domain of modularity detection is the Newman and Girvan (NG) algorithm, which relies on the number of shortest-paths across pairs of vertices in the network traversing a given edge, referred to as the *betweenness* of that edge. The edge with the highest betweenness is iteratively eliminated from the network, with the betweenness of the remaining edges recalculated in every iteration. This generates a complete dendrogram, from which modules are extracted by applying a quality metric called *modularity* denoted by *Q*. This exhaustive computation can be prohibitively expensive for large networks such as Protein-Protein Interaction Networks. In this paper, we present a novel optimization to the modularity detection algorithm, in terms of an efficient termination criterion based on a *target edge betweenness* value, using which the process of iterative edge removal may be terminated.

We validate the robustness of our approach by applying our algorithm on real-world protein-protein interaction networks of *Yeast*, *C.elegans* and *Drosophila*, and demonstrate that our algorithm consistently has significant computational gains in terms of reduced runtime, when compared to the NG algorithm. Furthermore, our algorithm produces modules comparable to those from the NG algorithm, qualitatively and quantitatively. We illustrate this using comparison metrics such as module distribution, module membership cardinality, modularity *Q*, and Jaccard Similarity Coefficient.

We have presented an optimized approach for efficient modularity detection in networks. The intuition driving our approach is the extraction of holistic measures of centrality from graphs, which are representative of inherent modular structure of the underlying network, and the application of those measures to efficiently guide the modularity detection process. We have empirically evaluated our approach in the specific context of real-world large scale biological networks, and have demonstrated significant savings in computational time while maintaining comparable quality of detected modules.

1.1 Background

The problem of modularity detection in networks has received considerable attention in recent literature [1-5]. Specifically, in the context of biological networks, identification of modules enables functional annotation of constituent biomolecules, discovery of targets for therapeutic intervention and screening etc. More generally, modular decomposition provides us with a higher-level understanding of the organization of networks and also serves as the basis for other network analysis tasks, such as hierarchical alignment, modular evolution, and orthology.

There are three primary approaches to modularity detection: (i) top down (or divisive) techniques, in which a series of network partitions hierarchically decompose a network into modules, (ii) bottom up (or agglomerative) techniques, in which modules are constructed by adding elements to an initial seed, and (iii) force directed methods, in which suitably designed parameters drive nodes belonging to the same module to

spatially proximate regions of space. There have also been investigations focused on relating various classes of methods [6].

1.1.1. Newman and Girvan algorithm

One such divisive technique of interest is the Newman and Girvan (NG) algorithm [1], which uses the notion of *edge-betweenness*, a metric that has received considerable recent research interest in the domain of modularity detection. Edge-betweenness is typically computed as the number of (pair-wise) shortest paths that traverse an edge in a network. This notion, which was first introduced by Anthonisse [7], can be used to compute modules by repeatedly identifying and eliminating the edge with highest betweenness. Note that since the elimination of a single edge (especially one with high betweenness) may cause significant perturbations to the shortest paths, the edge-betweenness of the remaining edges must be recomputed after each edge-elimination.

The output from the NG algorithm is a complete dendrogram, which decomposes a given graph down to individual nodes. Modules are extracted from this dendrogram by applying a quality metric called *modularity* (*Q*), which is defined as follows:

$$Q = \sum_i (e_{ii} - a_i^2) = Tr(e) - \| e^2 \|$$

where, e is a $k \times k$ symmetric matrix whose element e_{ij} is the fraction of all edges in the network that link vertices in module i to vertices in module j;k is the number of modules in the network;

$Tr(e) = \sum_i e_{ii}$, is the trace of e, which represents the fraction of edges in the network that connect vertices in the same module;

$a_i = \sum_j e_{ij}$, are the row (or column) sums, which represent the fraction of edges that connect to vertices in module *i*;

$\|E\|$ denotes the sum of the elements of matrix *E*.

We observe that, in a network in which edges fall between vertices without regard for the modules they belong to, $e_{ij} = a_i a_j$.

The *Q* value measures the fraction of the edges that connect vertices within the same module minus the expected value of the same quantity in the network. If the number of intra-modular edges is no better than random, we get $Q = 0$. Values approaching $Q = 1$, which is the maximum, indicate strong modular structure [1]. In practice, *Q* values for such networks with strong modular structure typically fall in the range from about 0.3 to 0.7. The modular decomposition of the network (from the dendrogram) with maximum *Q* value is considered to be the best split by the NG algorithm.

While the computation of modules using the NG algorithm has been shown to perform well in terms of quality of modules, its computational cost can be significant (particularly for large networks such as biological networks). This cost, in part, stems from repeated edge betweenness computations. Furthermore, a level of refinement in the output dendrogram to the individual nodes, is typically unnecessary from an application standpoint, often un-informative, and computationally expensive. Finally, the dendrogram requires additional post-processing to identify suitable modules based on quality measures associated with the modules. Computing the quality of each module corresponding to every node in the dendrogram is itself expensive. A stopping criterion

that identifies a near-optimal point at which the process of iterative edge-removal may be terminated would significantly reduce the time and space complexity of the NG algorithm.

The problem of terminating divisive clustering is an important one, especially when the clustering method is itself expensive. A number of other approaches have been proposed—including use of *p* values of clusters as termination criteria [8]. However, each of these methods assumes models for underlying data, or specific properties for quality measures applied to modules. For example, the divisive partitioning technique of Koyuturk et al. [8] stops the partitioning process when the *p* value of a module is lower than a user-specified threshold. This does not guarantee that the optimal *p* value modules are found. Similarly, for data-sets for which precise models are not available, estimation of number of clusters is difficult. Neither class of techniques is directly applicable for divisive partitioning based on the NG algorithm.

In this paper, we experimentally derive an optimized termination criterion for the NG algorithm (which we call the *target edge-betweenness*), based on initial values of edge-betweenness computed over the input network. In particular, we define the *target edge-betweenness* to be the *geometric mean* of edge-betweenness values of all edges in the input network (and hence refer to our algorithm as the *Gmean algorithm* in the discussion below). A detailed description of our algorithm is included in the Methods section.

1.2 Results and discussion

There are two computational problems with the NG algorithm:

1. The iterative removal of edges (preceded by recalculation of edge betweenness in every iteration) is performed until all the edges are removed, leading to a time complexity of O (ne^2) for a network of n vertices and e edges (using Brandes' algorithm, assuming connected networks as inputs). This computation becomes prohibitively expensive in the context of large biological networks.
2. The modularity Q is calculated for every partition of a network in the dendrogram. This is necessary for determining optimal splits.

The Gmean algorithm directly addresses these overheads in two fundamental ways: it terminates the process before all edges are removed, thus significantly reducing the first overhead. Since the termination criterion is computed just once (at the start of the algorithm), and does not rely on repeated Q value computations, we eliminate the second overhead altogether.

Furthermore, we demonstrate that our algorithm results in modules with Q values comparable to the maximum Q value from the NG algorithm—thus maintaining the quality of the identified modules, while significantly reducing runtime. We also use the *Jaccard Similarity Coefficient* (a measure of *similarity* between two sample sets) to show that the resulting modules from both the approaches are similar.

We validate our approach on the networks summarized in Table 1.1. For each of the networks, we eliminate multiple edges between pairs of nodes, self-loops, and mirrored

edges. Thus, the final number of edges/interactions considered is shown in #Edges (Network considered).

We perform our experimental evaluation using a parallelized approach [11] to implement the NG and Gmean algorithms. Our results (as shown in Figure 1.1) demonstrate excellent performance in terms of efficiency on moderate machine configurations (tens of processors).

1.2.1 Comparison of computational efficiency

For a specific network under consideration, let RT_{NG} and RT_{Gmean} denote the execution times for the NG and Gmean algorithms respectively. We define the percentage gain in computational time (τ) between the NG and Gmean algorithms, as follows:

$$\tau = \frac{\mathrm{RT_{NG}} - \mathrm{RT_{Gmean}}}{\mathrm{RT_{NG}}} \times 100$$

We observe significant and consistent savings in computational cost with our proposed optimization (for the networks in our biological test bed under consideration). Figure 1.1 presents a comparison of the execution times for the NG and Gmean algorithms.

1.2.2 Comparison of module size and distribution

In Figures 1.2 and 1.3, we present a broad quantitative comparison of the size and distribution of modules produced using the Gmean and NG algorithms. In particular, we

observe that, for all the three networks under consideration, the total number of modules produced by the two algorithms is comparable.

1.2.3 Comparison of modularity

In addition to quantitatively comparing and demonstrating that the modules resulting from our algorithm are comparable in number and distribution to the modules resulting from the NG algorithm, we also present a qualitative validation that the results are indeed statistically similar in terms of *quality* of the modules produced using the modularity value *Q*. Figure 1.4 shows the modularity value comparison for the set of modules produced by both the algorithms, for the different networks considered in this paper. We note that for all networks under consideration, our algorithm identifies modules with very similar modularity values as the NG algorithm.

1.2.4 Comparison of Jaccard similarity coefficient

Jaccard Similarity Coefficient or the *Jaccard Index* is a statistic used for comparing the similarity and diversity of sample sets. The Jaccard Index measures similarity between two sample sets (say *A* and *B*), and is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

The Jaccard Index is 1 if the two sample sets are exactly identical, and is equal to 0, if they have no overlap at all.

We use this metric to show the similarity of the modules produced as the output by the NG and the Gmean algorithms. Specifically, we consider the modules produced by the algorithms as sample sets constituted by vertices and calculate the Jaccard Indices *J* (*A*,*B*) for all pairs of modules *A* and *B* (one from the output of each algorithm).

We define the percentage similarity score (λ) as the following:

$$\lambda = \frac{\sum J(A,B)}{\sum J(A,B)^*} \times 100$$

where *J* (*A*,*B*) is the Jaccard Index for the modules *A* and *B*, one from the output of each algorithm;

J (*A*,*B*)* is the *ideal* Jaccard Index for the modules *A* and *B*, one from the output of each algorithm (note that *J* (*A*,*B*)* = 1, corresponding to perfect match, when the two modules *A* and *B* are exactly identical);

Σ is the summation over all pairs of modules, one from the output of each algorithm.

Table 1.2 shows the percentage similarity values for the modules produced by the two algorithms for all the networks considered. We observe that the modules produced by the two algorithms demonstrate a high degree of similarity.

1.3 Conclusions

In this paper, we have proposed a novel termination criterion for efficient modularity detection in networks. The intuition driving our approach is the extraction of holistic measures of centrality from graphs, which are representative of inherent modular structure, and the application of those measures to efficiently guide the modularity

detection process. We have empirically evaluated our approach against existing techniques for modularity detection in the context of biological networks, and have demonstrated significant savings in computational time while maintaining comparable quality of detected modules.

1.4 Methods

1.4.1 Existing NG method

In the NG algorithm, the edge-betweenness is computed for each edge in the network under consideration. The edge with the maximum edge-betweenness is identified and eliminated, followed by a recalculation of the edge-betweenness values of all the remaining edges in the resultant network. This process is iteratively repeated till no edges are remaining, thus generating a complete dendrogram which is then traversed to identify the partition with best modularity value Q.

1.4.2 Proposed Gmean method

Figure 1.5 presents a flow diagram that illustrates the general framework of the proposed Gmean algorithm. Our motivation is to compute a *target edge betweenness T* that is used to determine termination of the algorithm. In particular, we propose that the recalculation of edge-betweenness and removal of the edges be stopped when the edge to be removed has a betweenness value less than T. More intuitively, we propose that for an edge to be considered to be an inter-modular edge, it must have betweenness value of at least T.

Based on extensive experimentation, we propose the following definition of *T*:

$$T = G(e)$$

where *G* (*e*) is the geometric mean (*gmean*) of edge-betweenness values of all edges in the input network. Validation on real networks shows that this choice serves as a robust and high-quality termination criterion. Specifically, as stated in the results section, this choice produces a set of modules comparable in quality and quantity to those produced by the NG algorithm. We show this for a number of biological networks of interest. All biological network data used for the experimental study are from publicly available data sources [9,10].

1.5 List of abbreviations

C.elegans: Caenorhabditis elegans; gmean: Geometric Mean

1.6 Competing interests

The authors declare that they have no competing interests.

1.7 Authors' contributions

TN investigated the problem of modularity detection and associated literature, proposed the optimization to the existing Newman and Girvan algorithm, and empirically evaluated the approach. MG helped with refining the proposed optimization and perform quantitative comparison. SS and AG provided guidance relative to the theoretical and practical aspects of designing/evaluating the algorithm. All authors read and approved the final manuscript.

Acknowledgements

We acknowledge NSF grant awards Science and Technology Center Grant 0939370, DBI 0835541 and DBI 0641037 which supported this work.

Chapter I, in full, is a reprint of the material as it appears in Narayanan T, Gersten M, Subramaniam S, Grama A: Modularity detection in protein-protein interaction networks. BMC Research Notes 2011: 4-569. The dissertation author was the primary investigator and author of this paper.

Table 1.1: Summary of Networks that were used to validate our approach

Network	**Source**	**#Vertices [Original Network]**	**#Edges**	
			Original Network	**Network Considered**
C.elegans	[9]	453	4596	2025
Yeast*	[10]	3654	15316	9946
Drosophila	[10]	7666	25649	25433

** The entire Yeast network contains 160,566 interactions. We restrict the dataset to interactions determined by Co-purification or Yeast Two-hybrid experiments. This yields a network of 15,316 interactions*

Table 1.2: Summary of % similarity for biological networks considered

	C.elegans	**Yeast**	**Drosophila**
$\Sigma J(A,B)$	4.5472	47.973	40.5089
$\Sigma J(A,B)$	5	48	46
λ	90.94%	99.94%	88.06%

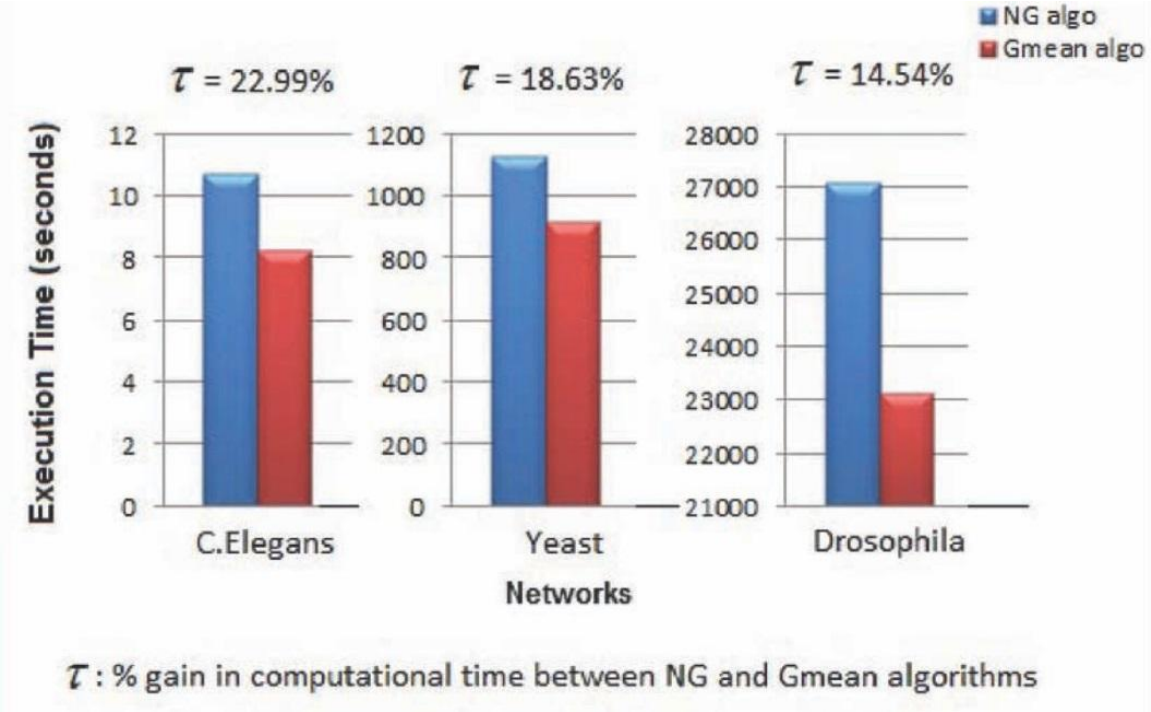


Figure 1.1: Comparison of Runtimes for NG and Gmean algorithms for C.elegans, Yeast and Drosophila networks

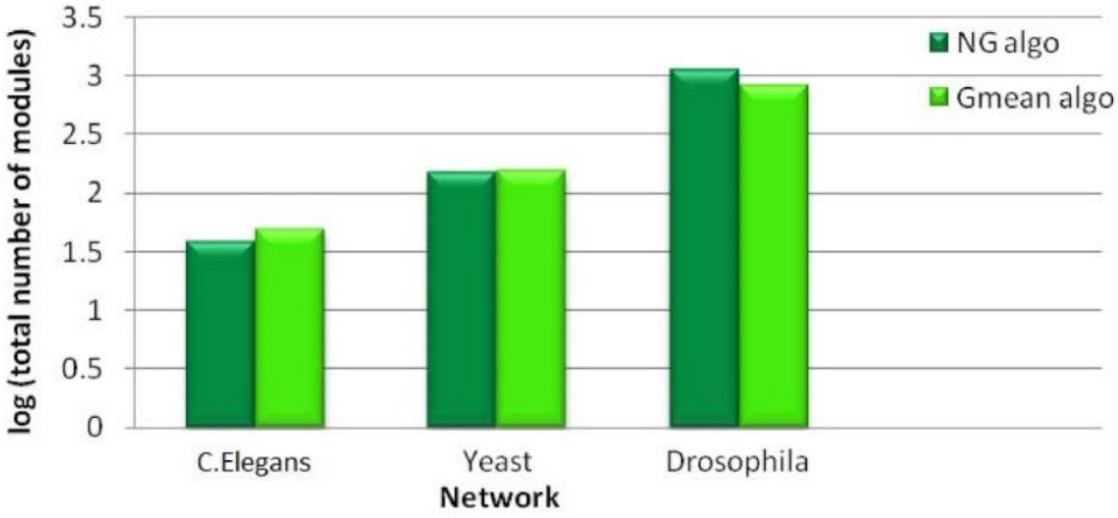


Figure 1.2: Log scale comparison of total number of modules identified by NG and Gmean algorithms for C.elegans, Yeast and Drosophila networks

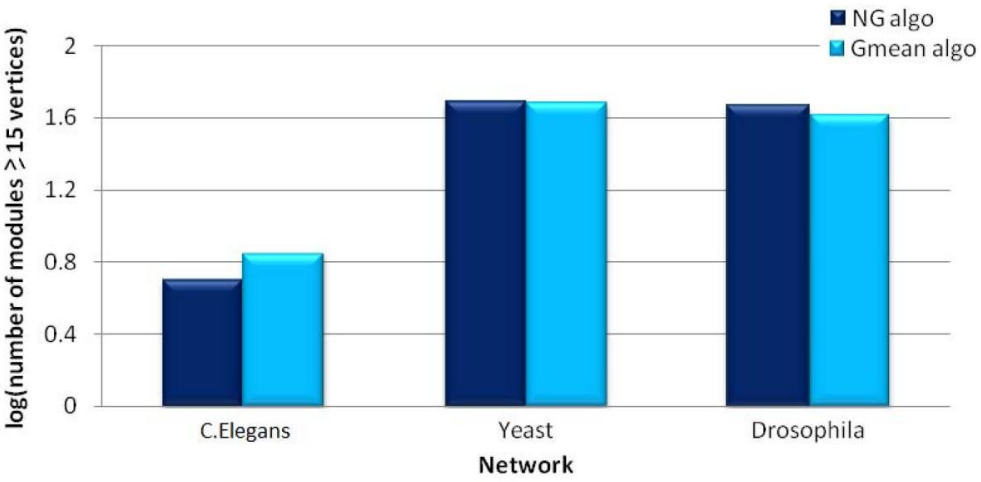


Figure 1.3: Log scale comparison of number of modules with at least 15 vertices identified by NG and Gmean algorithms for C.elegans, Yeast and Drosophila networks

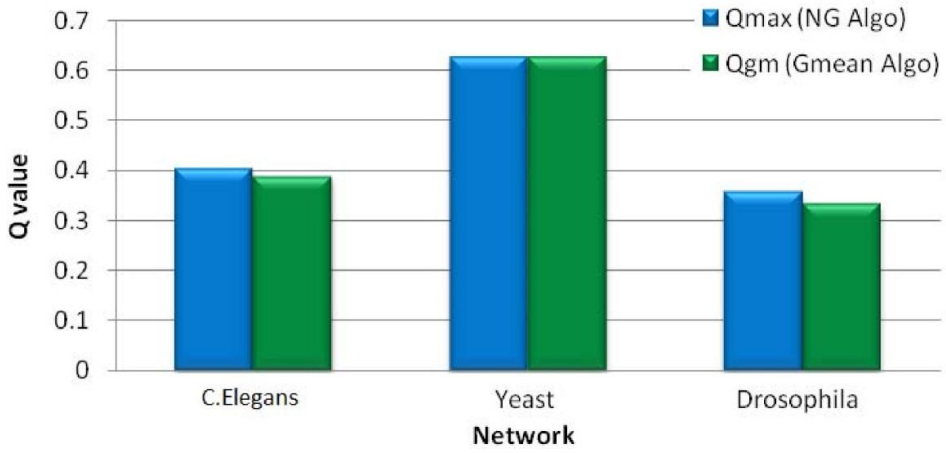


Figure 1.4: Comparison of Modularity (Q) values from NG and Gmean algorithms for C.elegans, Yeast and Drosophila networks

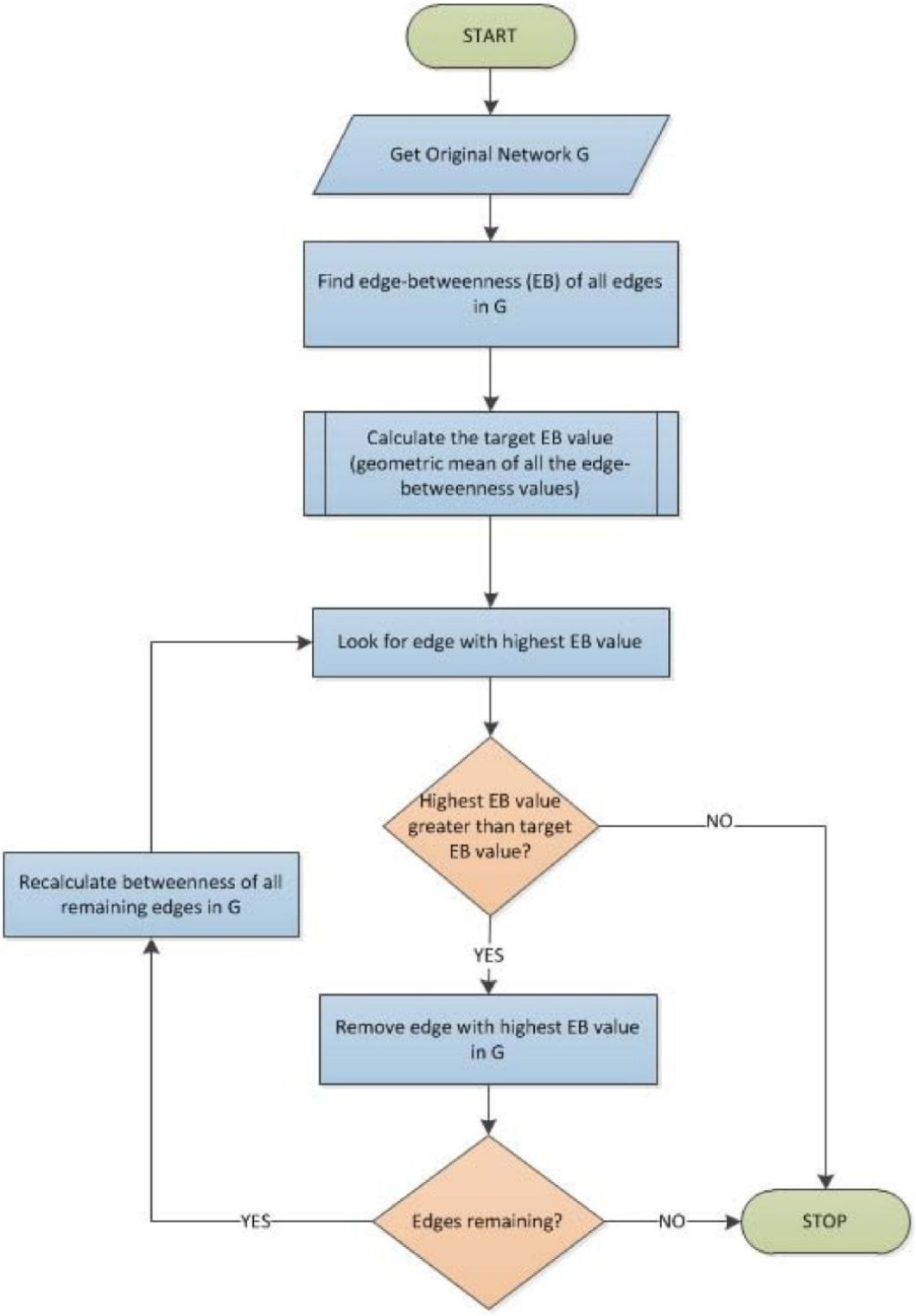


Figure 1.5: Flow diagram illustrating the general framework of the proposed Gmean algorithm

References

1. Newman MEJ, Girvan M: Finding and evaluating community structure in networks. *Phys Rev E* 2004, 69:026113.

2. Bader G, Hogue C: An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinforma* 2003, 4:2.

3. Dunn R, Dudbridge F, Sanderson C: The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinforma* 2005, 6:39.

4. Rives A, Galitski T: Modular organization of cellular networks. *PNAS* 2003, 100:1128–1133.

5. Sharan R, Ideker T, Kelley B, Shamir R, Karp R: Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. In *Proceedings of CM RECOMB*; 2004:282–289.

6. Sharan R., Ideker, T., Kelley, B., Shamir, R., Karp, RM. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *J Comput Biol.* 2005, 12(6):835-46.

7. Quigley A, Eades P: *FADE: Graph Drawing, Clustering, and Visual Abstraction.* Springer-Verlag; 2001.

8. Anthonisse JM: *The Rush in a Directed Graph.* Technical Report BN 9/71, Stichting Mathematicsh Centrum. Amsterdam; 1971.

9. Koyuturk M, Grama A, Szpankowski W: Pairwise local alignment of protein interaction networks guided by models of evolution. In *Proceedings of ACM RECOMB*; 2005:48–65.

10. Duch J, Arenas A: Community identification using extremal optimization. *Phys Rev E* 2005, 72:027104.

11. The Biogrid [http://thebiogrid.org/]

12. Yang Q, Lonardi S: A parallel Edge-betweenness clustering tool for protein-protein interaction networks. *Int J Data Min Bioinform* 2007, 1(3):241–247.

Chapter II

Community Detection in Biological Networks

Using a Variational Bayes Approach

Narayanan T, Subramaniam S: Community Detection in Biological Networks Using a Variational Bayes Approach. In proceedings of the 3rd International Conference on Bioinformatics and Computational Biology: 23 – 25 March 2011; New Orleans, Louisiana USA.

Abstract

A number of recent studies have investigated the problem of Community Detection in networks. In this paper, we consider one such approach called Variational Bayes, that utilizes a Bayesian formulation of the community detection problem, and evaluate its applicability in the context of biological networks. We perform a quantitative comparison of our results using a widely employed divisive algorithm for community detection (viz. the Newman and Girvan (NG) edge-betweenness algorithm). This algorithm generates as its output, a complete dendrogram from which suitable modules are extracted by applying a quality test, called the Q-value. We demonstrate that the Bayesian approach results in modules comparable in quality to those from the optimal-split of the dendrogram as determined by the NG algorithm (based on Q-value comparison) and yields a better distribution of module membership.

2.1 Introduction

The problem of computing modularity in networks has received considerable attention in recent literature [1, 2, 8, 9]. In the specific context of biological networks, the identification of modules enables functional annotation of constituent biomolecules (nodes in the same module are likely to be associated with identical function).

Most biological networks of interest tend to be large and complex, both in terms of the number of nodes and the interactions between them. For example, the complete Protein- Protein Interaction (PPI) network of *H.Sapiens* consists of over 10000 proteins and 81000 interactions [3]. Secondly, biological networks also tend to be sparse from a

connectivity perspective. These characteristics of biological networks pose interesting challenges from a community detection standpoint.

Some of the desirable features of a community detection algorithm (especially in the context of biological networks) are:

- The algorithm should be able to produce modules or some form of a representation of modules, as its output.
- It should be broadly applicable or easily extensible to different types of networks. For example, weighted and un-weighted networks, directed and un-directed graphs, static and time-varying networks, etc.
- It should be efficient and converge to a reasonable number of modules (with acceptable vertex memberships) within finite amount of computation time.
- The resulting modules should be uniformly distributed (i.e. reasonable vertex cardinality) since *singletons* (that are 'modules' with just one vertex) and modules with just 2-3 vertices are not expected to yield much biological insight.
- The modules produced represent functional elements of pathways, giving rise to *phenotype* prediction.

The aforementioned characteristics of an algorithm often are considered good metrics for *evaluating* a community detection algorithm. In this paper, we explore the applicability of the Variational Bayes approach to community detection in biological networks.

2.2 Variational Bayes Algorithm

The Variational Bayes (VB) model [1] is a computationally efficient framework for inferring the number of modules, model parameters, and module assignments for such a model. It attempts to pose module detection as inference of a latent variable within a probabilistic model.

In this framework the problem of module detection can be stated as follows: given an adjacency matrix A, determine the most probable number of modules $K^* = \text{argmax}_K$ $p(K|\mathbf{A})$ and infer posterior distributions over the model parameters and the latent module assignments. Here, $p(K|\mathbf{A})$ is referred to as the evidence.

In this section, we summarize the Variational Bayes algorithm for modularity detection. Using the notation in [1], if $F\{q; \mathbf{A}\}$ denotes the variational free energy, the algorithm provably converges to a local minimum of $F\{q; \mathbf{A}\}$ and provides controlled approximations to the *evidence* $p(\mathbf{A}|K)$ as well as the posteriors $p(\vec{\pi}, \vec{\upsilon} \mid \mathbf{A})$ and $p(\vec{\sigma} \mid \mathbf{A})$:

Initialization- Initialize the N-by-K matrix $\mathbf{Q} = \mathbf{Q}_0$ and set pseudocounts.

Main Loop: Until convergence in $F\{q; \mathbf{A}\}$:

(i) Update the expected value of the coupling constants and chemical potentials.

(ii) Update the variational distribution over each spin σ_i.

(iii) Update the variational distribution over parameters from the expected counts and pseudocounts.

(iv) Calculate the updated optimized free energy.

As this provably converges to a local optimum, VB is best implemented with multiple randomly-chosen initializations of $\mathbf{Q}_0$ to find the global minimum of $F\{q; \mathbf{A}\}$:

Convergence of the above algorithm provides the approximate posterior distributions and simultaneously returns K*, the number of non-empty modules that maximizes the evidence. As such, one needs only to specify a maximum number of allowed modules and run VB; the probability of occupation for extraneous modules converges to zero as the algorithm runs and the most probable number of occupied modules remains.

2.3 Alternate Approaches to Community Detection

Modularity computations have focused on modeling (modularity computation as an optimization problem), method development (suitable algorithms and data structures), and validation (characterizing the purity of modules). Modularity computations have also been posed as Graph Clustering problems, which explicitly compute dense sub-components in graphs.

There are three primary approaches to modularity computations – (i) top down (or divisive) techniques, in which a series of network partitions hierarchically decompose a network into modules, (ii) bottom up (or agglomerative) technique, in which modules are constructed by adding elements to an initial seed, and (iii) force directed methods, in which suitably designed parameters drive nodes belonging to the same module to spatially proximate regions of space. There have also been investigations focused on relating various classes of methods [7].

One of the promising state-of-art divisive algorithms is the *edge-betweenness algorithm* proposed by Newman and Girvan (NG). The *betweenness* of an edge is computed as follows: the shortest paths between all pairs of vertices are computed. Each edge in the network is tagged with the number of shortest paths passing through the edge. In the classical version of the Newman-Girvan method, the edge-betweenness of each edge is computed. The edge with the maximum edge-betweenness is eliminated. The algorithm is executed to completion and the resulting dendrogram is then traversed to identify the partition with best (highest) *Q-value.*

The Q-value or the *modularity* measures the fraction of the edges that connect vertices within the same cluster (module) minus the expected value of the same quantity in the network. Theoretically, Q varies from Q=0 (when the number of within-community edges is no better than random) and Q = 1 (indicating strong community structure [50]). In practice, Q-values for such networks with strong community structure typically fall in the range from about 0.3 to 0.7. The modular decomposition of the network with maximum Q-value is considered to be the best split.

2.4 Results

In an effort to leverage the VB approach to community detection in biological networks, we consider the *Yeast PPI network* as a case study. The yeast network data was collected from *Biogrid*. The full network contained 160566 interactions, but we restricted the dataset to interactions determined by co-purification or yeast two-hybrid experiments,

thus giving only 15316 interactions. Moreover, the edge betweenness algorithm that we work with neglects the following:

a. Multiple edges between a pair of nodes
b. Self-loops and
c. Mirrored edge representations in the network- i.e. an A-B interaction is considered the same as a B-A interaction. Thus, the final yeast network that we worked with, contained 9946 edges or interactions and 3654 vertices.

In this section, we present the results of our evaluation of the VB technique for community detection using the biological datasets described above. Specifically, we use the metric of *modularity* introduced by [2] to quantify the quality of communities produced by the algorithm to compare the communities produced by both approaches. We have used it as a metric of comparison of quantitative equivalence of the modules produced by the NG and the VB approaches because of the following reasons:

1. The Q-value is considered to be a good metric of modularity as it measures the difference between the fraction of the intra-community / intra-modular edges and the expected value of the same quantity in the network.
2. It is used as a measure to define the *best / ideal* partition of the dendrogram produced as output by the NG algorithm. Hence by using the *same* metric to quantify the quality of modules produced by the VB approach, we seek to perform a comparison of the VB approach against the NG approach, using a performance baseline defined by the latter.

The simulations were performed in Matlab on a 64-bit Intel dual-core 2.13GHz workstation, with 2 GB RAM.

The Q value corresponding to the set of modules produced by the VB algorithm was 0.5431 which we observed was very close to the Qmax or the Q value that corresponded to the *best* partition of the NG algorithm's output dendrogram (0.6254).

We provide in Figure 2.1, the distribution of Yeast modules obtained using the Variational Bayes approach and the Newman and Girvan's *edge-betweenness* algorithm. In particular, the figure shows a distribution of the total number of 150 modules that we obtained from the NG algorithm (as against 11 from the Variational Bayes algorithm) across bins defined by vertex cardinality range. This distribution of modules across the bins is represented in terms of the percentage of total number of modules produced by the respective algorithms.

It can be noted that the VB algorithm results in what we believe to be a more *robust* partitioning of the Yeast PPI network into modules. Specifically, there is a denser distribution of vertices into a smaller number of structural modules, which suggests better *functional cohesion* between vertices within a module (as discussed in Section 6). In contrast, while the NG algorithm does result in some partitions with denser memberships, there are also a significant number of modules that are sparsely populated.

In particular, we see that there are only 8 modules (constituting 5.33% of the total number of 150 modules produced by the NG algorithm) with greater than 100 vertices, as against the VB approach in which more than 50% of the modules produced have memberships resulting in more than 100 vertices per module. Similarly, it can be observed that, while the VB approach produces only 2 modules that have a vertex cardinality of less than 20 vertices, there are 114 modules which constitute 76% of the total number of modules produced by the NG algorithm that exhibit much smaller vertex

membership (including a significant number with only a pair of vertices constituting a module). From a functional standpoint, most of these modules may be viewed as *noise* with little information content from a biological perspective, in the broader context of the large number of vertices present in the original network distributed across different functional categories. This is discussed in greater detail in the next section.

2.5 Biological Interpretation

In this section, we evaluate the correspondence between the topological modules we identified using the Variational Bayes approach, with functional units in the yeast network. In particular, we leverage the GO Slim [6] functional annotations for the proteins constituting the network, and examine the distribution of the proteins in each structural module relative to their GO Slim annotations. The approach we employ in evaluating the correspondence is motivated by the analysis outlined in [5].

As noted in [5], we expect a nonrandom distribution of proteins of a given functional category across modules (if the structural modules correspond to functional units). Furthermore, the primary composition of genes in a given module is expected to correspond to a single or a few functional mappings. Stated differently, the expectation is that a single or a few functional categories are highly expressed in each module.

We validated our results against the expected outcome by simulating the probability of expression of a given functional category in a module of a given size. For example, the total vertex cardinality of the first module from the yeast network is 109, of which 32 belong to the GO slim functional category *transcription regulator activity*

(Figure 2.3). In order to estimate the expected value of this membership distribution, we selected 109 genes from the original yeast network uniformly at random, and identified the ones with the functional annotation *transcription regulator activity* (repeating the process a large number of times, using a different random seed each time, for statistical convergence of expectation). In particular, we are interested in the probability that the *transcription regulator activity* occurrences in the random selection of 109 proteins is greater than or equal to 32. Figure 2.2 illustrates these color-coded probabilities.

We observe that at least one functional category is enriched in 10 out of 11 modules (90.91%). In other words, we have only one module that has no distinct functional category uniquely expressed. Furthermore, 9 functional categories are each *uniquely* or *highly* expressed in *exactly* one module. It can also be noted that 5 out of 11 clusters each have one functional category, *uniquely* expressed. These observations are supportive of nonrandom distributions of protein functions across structural modules.

For each of the 17 occurrences of very high confidence measures of enhanced GO slim functional category representation, (the green cells representing probability p lesser than 10^{-5}) we analyzed the percentage of vertices that constituted the specific GO slim functional activity in each module. This was calculated for every module M using the ratio of number of vertices in M that corresponded to a functional activity A, to the total number of vertices in that module M. Figure 2.3 gives a color coded representation of such percentages and we observe that all *enhanced representations* (except for one) have at least 10 % of the total module size constituted by vertices corresponding to the *enhanced* functional category.

In Figure 2.4, we perform a similar analysis for studying the relative distribution of vertices in a given GO Slim functional category across modules. In particular, for each of the 10 *highly expressed* functional categories, we calculate a measure representing the fraction of vertices that constitute the functional category, across different modules. This is performed for each functional category F, by computing the ratio of the number of vertices from each module M (in which F is highly expressed), to the total number of vertices that correspond to F, across all modules. We observe that, for all *enhanced* functional representations (except for one) there exists at least one module which contains at least 10 % of vertices corresponding to that functional category across all modules.

2.6 Conclusions and Future Work

In this paper, we have studied the applicability of a Bayesian inference approach to the problem of community detection in the context of biological data sets. Our initial results demonstrate that, for the datasets under consideration, a VB approach produces results that are comparable to the NG algorithm. Furthermore, in the case study of the Yeast network considered, we have established that the VB approach results in modules that correspond to functional units based on the GO Slim mapping of constituent proteins in the *yeast* network. Not only is the distribution of function non-random across structure, but also supports specificity in functional enrichment across modules.

Some of the areas that we plan to deal with in our future work, which can be considered as an extension to the current subject of interest are: comparison of the VB

and the NG approaches based on time complexities and run times, analysis of modules produced by the VB approach for other PPI networks (such as the *H. Sapiens* network) and a comprehensive study of performance characteristics across other existing modularity detection algorithms.

Acknowledgements

We acknowledge NSF grant awards Science and Technology Center Grant 0939370, DBI 0835541 and DBI 0641037 which supported this work.

Chapter II, in full, is a reprint of the material as it appears in Narayanan T, Subramaniam S: Community Detection in Biological Networks Using a Variational Bayes Approach. In proceedings of the 3rd International Conference on Bioinformatics and Computational Biology: 23 – 25 March 2011; New Orleans, Louisiana USA. The dissertation author was the primary investigator and author of this paper.

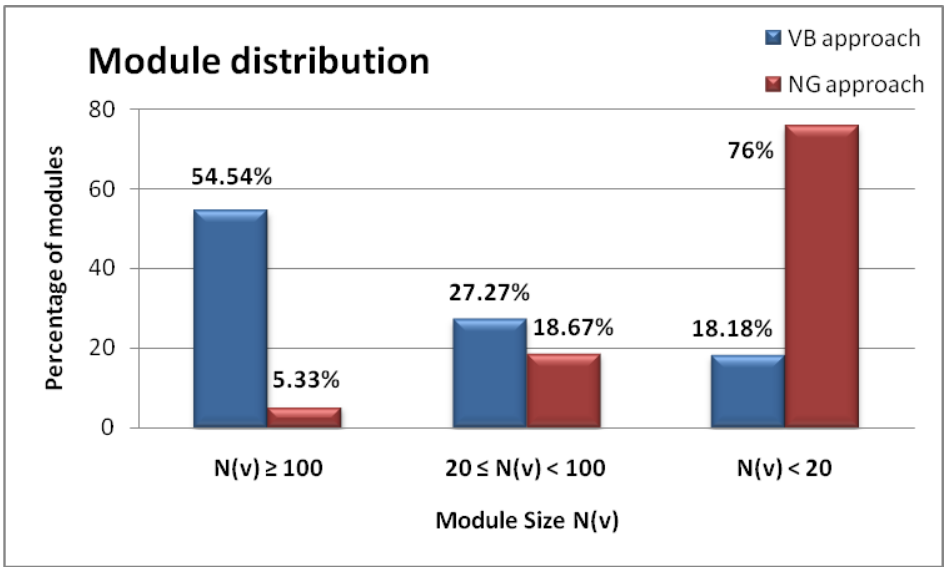


Figure 2.1: Distribution of Yeast modules using the VB and the NG approaches

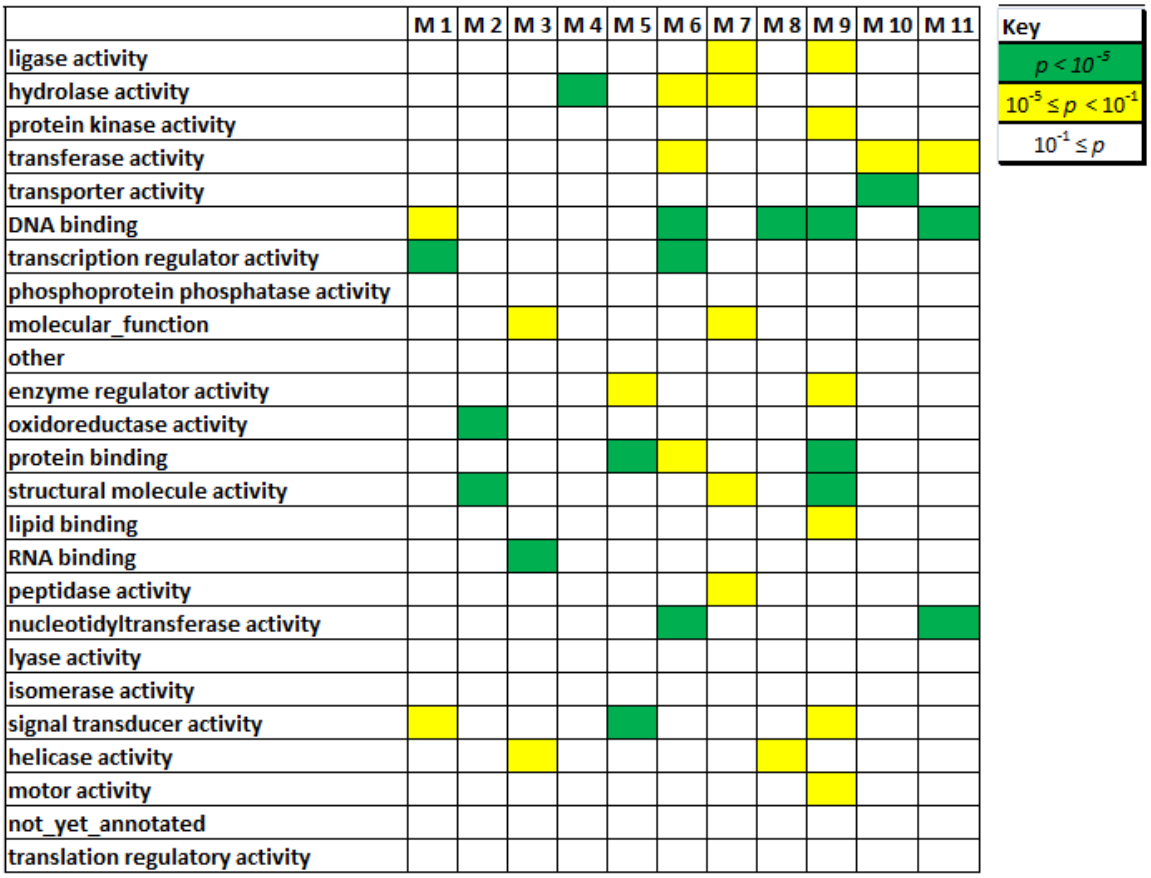


Figure 2.2: Biological Interpretation for Yeast modules using GO slim functional annotation

Module	Module size	Enhanced GO slim functional category	Vertex Cardinality	Percentage
M1	109	transcription regulator activity	32	
M2	21	oxidoreductase activity	10	
	21	structural molecule activity	7	
M3	363	RNA binding	62	
M4	10	hydrolase activity	10	
M5	29	signal transducer activity	10	
	29	protein binding	4	
M6	308	DNA binding	53	
	308	transcription regulator activity	80	
	308	nucleotidyltransferase activity	30	
M8	8	DNA binding	7	
M9	739	DNA binding	88	
	739	protein binding	161	
	739	structural molecule activity	82	
M10	499	transporter activity	154	
M11	34	DNA binding	11	
	34	nucleotidyltransferase activity	8	

Key
r > 30
30 ≥ *r* > 10
10 ≥ *r*

Figure 2.3: Percentage of module size constituting the enhanced GO slim functional annotation

GO slim Functional Activity	Vertex Cardinality (across all modules)	Module	Vertex Cardinality (per module)	Percentage
hydrolase activity	330	M4	10	
transporter activity	199	M10	154	
DNA binding	193	M6	53	
	193	M8	7	
	193	M9	88	
	193	M11	11	
transcription regulator activity	169	M1	32	
	169	M6	80	
oxidoreductase activity	70	M2	10	
protein binding	336	M9	161	
	336	M5	4	
structural molecule activity	145	M9	82	
	145	M2	7	
RNA binding	135	M3	62	
nucleotidyltransferase activity	47	M6	30	
signal transducer activity	26	M5	10	

Key
r > 30
30 ≥ *r* > 10
10 ≥ *r*

Figure 2.4: Percentage of module size constituting the enhanced GO slim functional annotation

References

1. Jake M. Hofman and Chris H. Wiggins, *A Bayesian Approach to Network Modularity*, Phys. Rev. Lett. 100, 258701 (2008)

2. Newman, M. and Girvan, M. (2004) *Finding and Evaluating Community Structure in Networks*, Physical Review E, Vol. 69, pp.026113 (15 pages).

3. Database of Interacting Proteins, http://dip.doe-mbi.ucla.edu/dip/Main.cgi

4. Community Identification using Extremal Optimization J. Duch and A. Arenas, Physical Review E , vol. 72, 027104, (2005).

5. Wang Z, Zhang J. In search of the biological significance of modular structures in protein networks. PLoS Comput. Biol. 2007;

6. The Gene Ontology, www.geneontology.org/GO.slims.shtml

7. Quigley A. and Eades P. FADE: Graph Drawing, Clustering, and Visual Abstraction, Springer-Verlag (2001).

8. Bader, G. and Hogue, C. (2003) 'An Automated Method for Finding Molecular Complexes in Large Protein Interaction Networks', BMC Bio-informatics, Vol. 4, No. 2.

9. Dunn, R., Dudbridge, F. and Sanderson, C. (2005) 'The Use of Edge-betweenness Clustering to Investigate Biological Function in Protein In-teraction networks', BMC Bioinformatics, Vol. 6, No. 39.

Chapter III

Community Structure Analysis of Gene Interaction Networks in Duchenne Muscular Dystrophy

Chapter III, in full, is the material as it appears in Narayanan T, Subramaniam S: Community Structure Analysis of Gene Interaction Networks in Duchenne Muscular Dystrophy. Public Library of Science (PLoS ONE) 2013. The dissertation author was the primary investigator and author of this paper.

Abstract

Duchenne Muscular Dystrophy (DMD) is an important pathology associated with the human skeletal muscle and has been studied extensively. Gene expression measurements on skeletal muscle of patients afflicted with DMD provides the opportunity to understand the underlying mechanisms that lead to the pathology. Community structure analysis is a useful computational technique for understanding and modeling genetic interaction networks. In this paper, we leverage this technique in combination with gene expression measurements from normal and DMD patient skeletal muscle tissue to study the structure of genetic interactions in the context of DMD.

We define a novel framework for transforming a raw dataset of gene expression measurements into an interaction network, and subsequently apply algorithms for community structure analysis for the extraction of topological communities. The emergent communities are analyzed from a biological standpoint in terms of their constituent biological pathways, and an interpretation that draws correlations between functional and structural organization of the genetic interactions is presented. We also compare these communities and associated functions in pathology against those in normal human skeletal muscle. In particular, differential enhancements are observed in the following pathways between pathological and normal cases: Metabolic, Focal adhesion, Regulation of actin cytoskeleton and Cell adhesion, and implication of these mechanisms are supported by prior work. Furthermore, our study also includes a gene-level analysis to identify genes that are involved in the coupling between the pathways of interest.

We believe that our results serve to highlight important distinguishing features in the structural / functional organization of constituent biological pathways, as it relates to normal and DMD cases, and provide the mechanistic basis for further biological investigations into specific pathways differently regulated between normal and DMD patients. These findings have the potential to serve as fertile ground for therapeutic applications involving targeted drug development for DMD.

3.1 Keywords

Duchenne Muscular Dystrophy, Human skeletal muscle, Community structure analysis, Biological pathways, Gene expression

3.2 Background

Community structure analysis is an interesting computational technique for studying interaction networks. Analysis of community structure in networks can yield useful insights into the structural organization of the network. For instance, community structure analysis is used in the context of networks that arise in domains such as social networks to understand the fundamental social structure in a community of interacting individuals [1-7]. This provides insights about the influential individuals and the strongly-networked individuals in a community. Another domain where algorithms for community structure analysis find useful application is the topological understanding of large scale connection networks such as Internet, and how one may use the insights from community structure analysis to design more resilient communication networks [6, 8-10].

In the context of biological networks, such insights can also be used to understand the biological significance of the underlying community structure and organization of the network. There is existing work that discusses the use of community structure analysis in networks that are observed in biological contexts [4-5, 11-13]. For example, [4] presents the application of an algorithm for community structure analysis to a food web of marine organisms living in the Chesapeake Bay, a large estuary on the east coast of the United States. Furthermore, rich toolsets have also been developed for the purpose of understanding biological networks from a community structure perspective [13-17].

In this paper, we explore the application of community structure analysis as an effective technique to understand the topological structure and biological behavior of human skeletal muscle. Skeletal muscles are a form of striated muscle tissue existing under the control of the somatic nervous system, which are attached to bones by tendons. This muscle category has been clinically associated with diseases such as Myopathy, Muscular Dystrophy, Paralysis, and a host of other diseases. DMD is a group of inherited disorders that involve muscle weakness and loss of muscle tissue, which get worse over time [18] and results in death before the individual reaches adulthood. Given the genetic nature of this disorder, techniques that leverage the underlying genetic interactions are expected to yield useful insights, and this is the primary focus of our study.

3.2.1 Community structure analysis: Newman and Girvan Algorithm

The Newman and Girvan (NG) algorithm is a popular algorithm for community structure analysis in networks [7]. It is a divisive approach that selects and removes edges based on its *betweenness* value. The betweenness of an edge is defined as the number of shortest paths between all vertex pairs in the network, which run along that edge. The steps involved in the NG algorithm are as follows: The betweenness values of all edges are computed. The edge with the largest betweenness is removed (in case of ties with other edges, one of them is picked at random). This is followed by the recalculation of betweenness values of the remaining edges in the network. The entire process is repeated iteratively till all edges are removed.

The output from this algorithm is a dendrogram capturing the possible division of the network into communities. In order to select the *optimal* split from these possible candidates, Newman and Girvan introduce the concept of *modularity*, which is a measure of the quality of a particular division of a network into communities [7]. Given a specific division of a network into k communities, let us define a $k \times k$ symmetric matrix e whose element e_{ij} is the fraction of all edges in the original network that link vertices in community i to vertices in community j. The row (or column) sums $a_i = \sum_j e_{ij}$ represent the fraction of edges that connect to vertices in community i. *Modularity* is then defined as follows:

$$Q = \sum_i \left(e_{ii} - a_i^2 \right) = Tr\ e - \left\| e^2 \right\|$$

where $Tr\ e = \sum_i e_{ii}$, denotes the trace of the matrix e and $\|e\|$ denotes the sum of the elements of the matrix e. Typically, Q is calculated for each split of a network into communities as the algorithm moves down the dendrogram, with the optimal split corresponding to the peak value of Q. For a network with n vertices and m edges, the worst-case time complexity for this algorithm is $O(m^2 n)$ (or $O(n^3)$ for a sparse network).

3.3 Results and Discussion

Consequence of DMD pathology manifests in the state of muscle cells. The physiological state and cellular state of muscles are altered, involving concomitant changes in the expression of genes associated with the physiological function. In particular, gene expressions in DMD patients have the potential to provide information on distinguishing characteristics of pathology, relative to normal muscle (since altered gene expressions could aid in identification of functional communities). In this work, we have devised a novel approach to analyze human DMD patient gene expression data using a combination of techniques from linear algebra and network theory. Specifically, we posit that the correlation of gene expression data from DMD patients captures salient characteristics of pathology. Accordingly, we build the correlation network from the gene expression data for the normal and DMD muscles. Under the assumption that correlation implies mechanistic causality, we take the approach of community structure analysis, to identify functional communities from the correlation network, to display known functional and pathway mechanisms.

3.3.1 Derived Interaction Networks

In this section, we present an analysis of the global properties of the *derived interaction networks* (defined in the Methods section) for the normal and DMD muscle data from a descriptive statistics standpoint. We use well known global network properties such as density, average degree etc. to inform our analysis. This analysis aims to highlight the key similarities and differences between the derived interaction networks for normal and DMD muscle data, in order to enable a structural understanding of the underlying genetic interactions at a macro level.

Figure 3.1 illustrates the key structural differences in the normal and the DMD interaction networks. As can be noted from Table 3.1, the number of vertices and edges in the DMD interaction network is much smaller than those of the normal interaction network. Thus, as one would expect, the density and the average degree of the DMD interaction network are also lesser than the normal network (as shown in Figure 3.1). However, it is interesting to note that *both* interaction networks have turned out to be *sparse* from a network-theoretic standpoint.

From the planar-layout visualization of the normal and the DMD interaction networks generated using Cytoscape [19], we observe that the pre-processed networks containing 7685 vertices are by themselves disconnected into many independently connected components. Table 3.2 summarizes the key network parameters for the normal and DMD cases for the whole interaction map.

Since we are interested in finding communities from the networks, we consider the largest connected component in both networks. Table 3.3 shows the number of

vertices and edges considered for community structure analysis in both the networks (i.e. the parameters defining the largest components in the respective interaction networks).

3.3.2 Community structure analysis

In this section, we present our results from running the NG algorithm on the largest components of the derived interaction networks for the normal and DMD muscle datasets. Table 3.4 presents the number of communities identified in the dataset, along with the corresponding modularity values (Q_{max}). We provide in Figure 3.2, a comparison of the distribution of communities in both networks (obtained using the NG algorithm), across bins defined by vertex cardinality range.

3.3.3 Pathway Analysis

We perform an analysis of the communities obtained from the NG algorithm from the perspective of its constituent pathways, by generating *pathway projection networks* (PPNs). The motivation, technique and color-coding convention of PPNs are detailed in the Methods section. Figure 3.3 illustrates the PPNs that are considered for analysis.

3.3.4 Biological Interpretation and Discussion

While we have included a representative set of PPNs in the Supporting Information (Figure S1- Figure S11), we consider the 4 PPNs shown in Figure 3.3 to elucidate the significance of the pathways of interest (shown in Table 3.5) in each community and the correlation to their presence in the PPNs. Specifically, the pathways

we consider are Metabolic, Focal adhesion, Regulation of actin cytoskeleton and Cell adhesion. We are interested in finding evidence from past work that can potentially help with triangulating our algorithmic findings about the specific pathway enhancements that we have identified. For example, if a specific pathway is determined to be enhanced by our algorithmic technique, we would expect the evidence corresponding to that pathway to correlate well with such an enhancement (for the network under consideration).

Conversely, for a pathway that is determined to not have a pronounced enhancement using our algorithmic approach, we are interested in finding whether the experimental evidence surrounding that pathway is aligned with our finding. We believe that this analysis will help us validate our algorithmic findings with evidence from existing research. We also perform a gene-level analysis on the PPNs to identify genes that are involved in the coupling between the corresponding pathways of interest, and summarize sample gene pairs with their corresponding correlation scores. We leverage UniProtKB [20] for identifying the functional information associated with the sample genes we consider in the discussion below.

As background for rest of the discussion, we note that *dystrophin* is a key protein of interest in the study of dystrophy. Specifically, the absence of *dystrophin* is associated with DMD and was identified as the source of pathology in humans using positional cloning [21]. Mice lacking dystrophin have high serum levels of muscle enzymes and possess histological lesions similar to human muscular dystrophy [22-24].

3.3.4.1 Metabolic pathways and DMD

Our results emphasize an interesting connection between metabolic pathways and DMD, and we leverage Figures 3.3A-3.3C (PPN 1-PPN 3) to explore these connections in greater detail. We summarize the key observations from our analysis here. The first observation from Figures 3.3A-3.3C is that, PPN 1 – PPN 3 exhibit enhanced representation of metabolic pathways. Furthermore, Figure 3.3A (PPN 1) illustrates a strong coupling between metabolic pathways and regulation of actin cytoskeleton. Similarly, we observe a direct coupling of metabolic pathways to calcium signaling from Figure 3.3C. Finally, we reiterate the importance of metabolism as a key differentiator in pathology, in terms of glycolytic and oxidative variations of metabolic pathways. The rest of this section provides evidence from prior work in this domain to support our observations.

Our first observation around metabolic pathways and their connection to DMD, is in alignment with prior work. In particular, [22] identifies that a dystrophin-dependent cytoskeletal organization in skeletal muscles is directly related to the efficiency of cytoplasmic and mitochondrial metabolic pathways in situ. More generally, the lack of dystrophin or a functionally mildly defective dystrophin is connected with subnormal rates of muscle energy conversion and the subnormal energy status of sarcoplasm. In other words, enhancement of metabolic pathways is a canonical characteristic in normal muscle, and our findings (Figure 3.3A- 3.3C) are consistent with this result. Also, from a computational standpoint, the observed specificity in enhancement validates the algorithm for community structure analysis used in our approach, since the algorithm grouped the genes corresponding to metabolic pathways in cohesive communities.

Furthermore, a similar exercise of pathway projection performed on the DMD network had no significant representation of the metabolic pathways.

Secondly, the observation of strong coupling between metabolic pathways and regulation of actin cytoskeleton is corroborated by prior experimental work, which has identified that, muscles from the dystrophic *mdx* mouse show reduced maintenance metabolic rates [22]. The authors of [22] also propose that the in vivo efficiency of metabolic pathways may depend on stabilization of enzyme complexes by dystrophin-associated elements of the cytoskeleton. By performing a gene-level analysis on PPN 1 (Figure 3.3A), we found that many genes were involved in the coupling between the two pathways of interest. Table 3.6 presents five sample gene pairs and the corresponding correlation scores between them.

Specifically, Leukotriene A4 hydrolase is an epoxide hydrolase that catalyzes the final step in the biosynthesis of the proinflammatory mediator leukotriene B4 [20]. This gene is highly correlated with cell division cycle 42 which is involved in epithelial cell polarization processes. It also plays a role in the extension and maintenance of the formation of thin, actin-rich surface projections called filopodia**.** Phosphoglycerate mutase 1 is highly correlated with Cofilin 1 which regulates actin cytoskeleton dynamics and plays a role in the regulation of cell morphology [20]. It is interesting to note that a similar correlation was observed between these genes in astrocytomas involved in pathogenesis of radioresistance [25]. There is existing evidence of association between Iduronate 2-sulfatase and integrin, alpha V from a Gene Set Enrichment Analysis point of view (which is in accordance with the results shown in Table 3.6, in terms of their correlation) [26]. Iduronate 2-sulfatase plays a role in the lysosomal degradation of

heparan sulfate and dermatan sulfate. integrin, alpha V is a receptor for fibronectin and fibrinogen [20]. Finally, referring to the high correlation between PIK3CA and PDGFRB, there is existing evidence that reports an interaction between these genes [27].

Similarly, we note that there is evidence from past research that aligns with our observation around the coupling of metabolic pathways to calcium signaling. In particular, [28] suggests that high intracellular Ca 2+ (linked to calcium signaling) in dystrophic fibers, may be the cause of the inefficiency of mitochondrial metabolic pathways. Table 3.7 provides five sample gene pairs with their corresponding correlation scores, from among the many genes that we found to be highly correlated in function between the metabolic and calcium signaling pathways.

While CYP2C6 plays a role in drug metabolism [29], CYP2C9 localizes to the endoplasmic reticulum and its expression is induced by rifampin. From Table 3.7, we observe that both CYP2C6 and CYP2C9 are highly correlated to Phosphodiesterase 1C, calmodulin-dependent 70kDa. Members of the Cyclic nucleotide phosphodiesterases (PDE1) family, are calmodulin-dependent PDEs [CaM-PDEs] that are stimulated by a calcium-calmodulin complex [30]. This gene is also highly correlated to Cysteine conjugate-beta lyase, cytoplasmic (from Table 3.7). ErbB-4 protein binds to and is activated by neuregulins and induces a variety of cellular responses including mitogenesis and differentiation [20]. It is interesting to note that this gene is highly correlated to Fructose-1,6-bisphosphatase 1, deficiency of which is associated with hypoglycemia and metabolic acidosis [31].

Analysis of functional communities that are differentially regulated, demonstrates metabolism as the most important mechanistic change in DMD muscle. In particular, glycolysis and oxidative metabolism play significant roles in muscle energetics including remodeling of the muscle into fast and slow fiber forms responding to the nature of the energy demands. Experiments that have been performed on normal muscle showed accumulation of glycolytic and oxidative metabolism capacity with increased age, but this accumulation failed in DMD [32]. The data used in [32] shows stage-specific remodeling of human dystrophin-deficient muscle, with inflammatory pathways predominating in the presymptomatic stages and failure of metabolic pathways later in the disease [32-33].

In the slow twitch (type I) fibers, the slow muscles are more efficient at using oxygen to generate more fuel (known as ATP) for continuous, extended muscle contractions over a long time. In other words, these are the fibers that correspond to oxidative phosphorylation. Whereas, because fast twitch (Type II) fibers use anaerobic metabolism to create fuel, they are much better at generating short bursts of strength or speed than slow muscles. These typically correspond to glycolysis / gluconeogenesis, which is involved in converting glucose into pyruvate. We performed an analysis on the number of genes that contributed to the fast and slow twitch fibers, in the three communities in which metabolic pathways were enhanced (PPN 1-PPN 3). The results are summarized in Table 3.8.

3.3.4.2 Regulation of actin cytoskeleton and DMD

The discussion on Regulation of actin cytoskeleton and its relationship to DMD is centered around Figure 3.3D (PPN 4). Specifically, PPN 4 illustrates that in normal skeletal muscle, the actin cytoskeleton pathways are enhanced, whereas they are less utilized in DMD muscle. This is consistent with prior work as follows. Dystrophin links the actin cytoskeleton to the dystroglycan complex (which is a part of an adhesion receptor complex [34]) in the plasma membrane as part of the linkage between the cytoskeleton and the extracellular matrix [36-37]. This link helps maintain sarcolemmal integrity in a muscle [38]. Damage to or absence of or mutations in dystrophin causes DMD [21, 37-38].

The skeletal muscle L-type Ca2+ channel (CaV1.1), which is responsible for initiating muscle contraction, is regulated by phosphorylation by cAMP-dependent protein kinase (PKA) in a voltage-dependent manner [39]. Furthermore, the role of the actin cytoskeleton in channel regulation was investigated in skeletal myocytes cultured from mdx mice that lack the cytoskeletal linkage protein dystrophin, and a skeletal muscle cell line, 129 CB3. Results of the experiments detailed in [39] show that regulation of Ca2+ channel activity by hormones and neurotransmitters that use the PKA signal transduction pathway may interact in a critical way with the cytoskeleton and may be impaired by deletion of dystrophin, contributing to abnormal regulation of intracellular calcium concentrations in dystrophic muscle.

We see that most pathways in PPN4 are well-coupled to each other. From the sample correlation scores provided in Table 3.9, we infer that there is strong correlation

[40] that exists between the genes, which signifies the coupling between the regulation of actin cytoskeleton and focal adhesion pathways.

3.3.4.3 Focal adhesion and DMD

We use Figure 3.3D (PPN 4) to motivate the discussion around the focal adhesion pathway, and its relationship to pathology. In particular, PPN 4 shows the expected level of association of focal adhesion pathways in normal muscle and this is consistent with the evidence presented below. The representation of focal adhesion kinase (FAK) in dystrophy networks has been studied previously [23, 41]. For example, the authors of [41] find that at 12 weeks of age, both hind limb muscles of dystrophic mice possessed a lower FAK protein than normal mice. It is proposed that FAK is a part of the pathway that would be of potential importance in transducing mechanical signals from cell membranes to skeletal muscle fiber nuclei [42-43]. Focal adhesion pathway is coupled tightly not only to regulation of actin cytoskeleton (as shown in the Table 3.9), but also to cell adhesion molecules, with high correlation scores, some of which are shown in Table 3.10.

Referring to genes in Table 3.9, Laminin alpha-4 is a protein thought to mediate the attachment, migration and organization of cells into tissues by interacting with other extracellular matrix components, by binding to cells via a high affinity receptor [20]. Integrin alpha-6 is a receptor for laminin in epithelial cells and it plays a critical structural role in the hemidesmosome. Laminin alpha4 and integrin alpha6 are upregulated in regenerating dy/dy skeletal muscle [20]. Furthermore, laminin alpha4 and integrin alpha6 expression patterns are notably different in dy/dy when compared to normal muscle. This

is especially pronounced in the interstitium of regenerating areas and on newly formed myotubes [44]. Our observation about the high correlation between Laminin alpha4 and integrin alpha6 (Table 3.9) is in alignment with these findings.

We also present a brief description (collated from [20]) of other genes in Table 3.9 amongst which we observe a high correlation. Moesin is conjectured to be involved in connections of major cytoskeletal structures to the plasma membrane. Kinase insert domain receptor (a type III receptor tyrosine kinase) is a vascular endothelial growth factor (VEGF) receptor. Beta-actin is one of six different actin isoforms which have been identified in humans. This is one of the two nonmuscle cytoskeletal actins. Actins are highly conserved proteins that are involved in cell motility, structure and integrity. Type IV collagen is the major structural component of glomerular basement membranes, forming a 'chicken-wire' meshwork together with laminins, proteoglycans and entactin/nidogen.

From Table 3.10, we observe that Platelet/endothelial cell adhesion molecule 1 (PECAM-1) and Cadherin 5, type 2 (vascular endothelium) genes from cell adhesion molecules pathway are highly correlated to the genes from the focal adhesion pathway. PECAM-1 is a transmembrane protein in the inter-endothelial cell contacts [20]. PECAM-1 is a homophilic adhesive molecule that is diffusely distributed on subconfluently growing endothelial cells, but concentrates at cell-cell borders upon cell-cell contact [45]. Our observation of high correlation between PECAM-1 and genes in the focal adhesion pathway (shown in Table 3.10) is corroborated by [46] which illustrates the co-localisation of some of the ECM components viz. laminin α1, collagen type IV with the endothelial cell marker PECAM-1. Cadherin 5, type 2 (vascular endothelium)

are calcium-dependent cell adhesion proteins. They play an important role in endothelial cell biology through control of the cohesion and organization of the intercellular junctions [20]. From Table 3.10, we see that it is highly correlated with Integrin, alpha 6 and Laminin, alpha 4.

3.3.4.4 Cell adhesion and DMD

Figure 3.3D (PPN 4) illustrates that the cell adhesion pathway is not enhanced significantly in the normal network (given that it is a relatively small sized node, representing smaller pathway cardinality). When we performed a detailed analysis of the genes that constitute this pathway in the network, we find that most genes are a form of the Class I and Class II type major histocompatibility complex (MHC). There exists enough evidence that MHC proteins in normal skeletal muscle fibers show lower expression levels, when compared to DMD [47]. Prior work also shows that for every MHC protein, the fold change for DMD muscle is greater than one [48], which represents a higher expression in DMD than in normal. Thus, we see that the algorithm, not only highlights the more enhanced pathways in the communities, but also identifies the lowly expressed pathways in the normal muscle. This evidence provides more confidence to the robustness of the communities detected. Table 3.10 shows a few genes from cell adhesion that are correlated to focal adhesion pathway.

3.4 Methods

3.4.1 Muscular Dystrophy: Dataset Description

We used the skeletal muscle gene expression data, *Series GSE6011* from the Gene Expression Omnibus [35]. The gene expression dataset consisted of measurements on probes for genes with a many-to-many mapping between probes and genes. In order to obtain one-to-one equivalence between the probes and genes, we perform a series of pre-processing steps, which are included in the Supporting Information (see Appendix S1). Table 3.11 summarizes the parameters of the pre-processed dataset.

3.4.2 Derived interaction networks

We introduce the notion of an *interaction network* that is *derived* from an underlying gene expression dataset. This is one of the novel contributions in our paper. We consider a gene expression dataset $A_{m \times n}$ (consisting of measurements on m probes for genes across n experiments) that has been pre-processed to represent one-on-one mappings between probes and genes. Let ρ denote the *correlation matrix* for the dataset, containing the pairwise linear correlation coefficient between each pair of columns in the matrix $A^T_{n \times m}$, where A^T denotes the transpose of the matrix A

$$\rho = [\rho_{ij}]_{m \times m}$$

We define the *interaction network* for the dataset as an undirected network $\Delta = \Delta(V, E)$, such that the set of vertices V corresponds to the set of genes in the

underlying dataset (i.e. $|V| = m$) and the interactions between them are captured by the set of edges E via an adjacency matrix as follows:

$$adjMat(\Delta) = [\rho_{ij}]_{m \times m}$$

$$a_{ij} = \begin{cases} 1, & |\rho_{ij}| > \tau \\ 0, & |\rho_{ij}| \leq \tau \end{cases}$$

where $0 < \tau < 1$ is a pre-defined threshold

Our intuition behind the definition of the interaction network was to capture the *inherent associations* between genes in a dataset, by using the correlation of expression measurements as a representative surrogate for the interactions between the underlying genes. In other words, the hypothesis is that a stronger correlation is likely to signify a stronger interaction between the genes exhibiting the correlation (modeled by the presence of an edge between the genes in the interaction network), while a weaker correlation is likely to correspond to a weaker interaction between the genes (modeled by the absence of an edge).

3.4.3 Derived interaction networks for the GSE6011 Dataset

We generated the derived interaction networks for the pre-processed GSE6011 dataset for both the normal and DMD data. We used a threshold of $\tau = 0.8$ as the correlation cut-off, applying the guidelines from [40]. Hence, an edge was present between two genes in the generated interaction network if and only if the absolute value of correlation between those genes was greater than 0.8. We note that due to the post-

processing steps described in the Supporting Information (see Appendix S1), the actual number of vertices considered for subsequent analysis in this paper is less than the initial number of vertices in the raw interaction networks generated for both normal and DMD data (summarized in Table 3.1).

3.4.4 Pathway Analysis

Consequence of DMD pathology manifests in the state of muscle cells. The physiological state and cellular state of muscles are altered, involving concomitant changes in the expression of genes associated with the physiological function. In particular, gene expressions in DMD patients have the potential to provide information on distinguishing characteristics of pathology, relative to normal muscle (since altered gene expressions could aid in identification of functional communities). In this work, we have devised a novel approach to analyze human DMD patient gene expression data using a combination of techniques from linear algebra and network theory. Specifically, we posit that the correlation of gene expression data from DMD patients captures salient characteristics of pathology. Accordingly, we build the correlation network from the gene expression data for the normal and DMD muscles. Under the assumption that correlation implies mechanistic causality, we take the approach of community structure analysis, to identify functional communities from the correlation network, to display known functional and pathway mechanisms.

In this section, we present an analysis of the communities from the perspective of the pathways that the constituent genes represent. The goal is to understand the communities from derived interaction networks through functional analysis, since

functions help elucidate alterations in pathological conditions [49-50]. Furthermore, we expect that the analysis of normal and DMD interaction networks from a pathway perspective is likely to yield more holistic insights into the correlation between functional and structural organization of the underlying genetic interactions.

We describe below, the transformation technique we employed to generate an equivalent network in terms of the constituent pathways for each community [also schematically presented in the flowchart in Figure 3.4]. We call this a *Pathway Projection Network* (PPN). For each community from the normal muscle interaction network, we extract a sub-network consisting of only those genes present in the normal muscle network and not in the DMD muscle network. From these sub-networks, we identify those that have a minimum vertex cardinality of 100 (we found four such candidates), and performed pathway analysis for these candidates using the KEGG mapper [51-52].

It is important to note that there is a one-to-many mapping between genes and pathways. Hence there are multiple pathway assignments that are possible for a given gene and this would lead to a combinatorial explosion in the number of pathway projection networks. To avoid this, we prune the space of gene-pathway mappings by employing a heuristic that we call the *maximum spanning pathway reduction heuristic.* This heuristic works as follows: From all candidate pathways that a gene from a sub-network belongs to, we choose that pathway **p** which maximizes the number of *other* genes spanning the sub-network which can also be assigned the pathway **p**.

We use Cytoscape to visualize the PPNs and these are shown in Figures 3.3A-3.3D (denoted as PPN1 – PPN4). The PPNs 1-4 use the following convention. The

pathway-nodes are color coded from Green to Red, with increasing degree of the node. This is a measure of connectivity between the pathways. A second attribute (pathway cardinality) defines the size of the node- a larger node signifying a larger pathway cardinality, which is the number of genes from the community that correspond to that pathway). Thus, strong connections between two large, red nodes imply a strong coupling between the set of genes in one pathway to the set that correspond to another.

From among the pathways represented in the PPNs, we are specifically interested in further analyzing pathways that are enhanced in each community and / or are known to be relevant to DMD from prior work [21-24, 32-39, 41-43]. These are summarized in Table 3.5. The pathway interactions analysis for the resultant PPNs is presented in the Results and Discussion section.

3.5 Conclusion

In this paper, we have proposed a principled approach for transforming gene expression datasets into interaction networks, which serve as a useful representation for downstream analysis of pathology. Furthermore, we have illustrated the utility of community structure analysis applied to the interaction networks, as a sound computational technique for gaining insights about the underlying topology and function. We have leveraged this approach to study the characteristics of normal and DMD human skeletal muscle tissues, in terms of functional communities. In addition to providing a topological perspective on the differential regulation of transcripts between normal and DMD skeletal muscle, the derived communities provide extensive information on

functional pathways and their association with pathology. Not only does our analysis provide clear evidence of the role of altered metabolic, calcium signaling and cytoskeletal remodeling pathways in DMD, but also identifies novel cross-talk between them. We believe that our work provides the steps for biomarker identification, as well as systems level information for therapy of the DMD skeletal muscle.

3.6 List of abbreviations

DMD: Duchenne Muscular Dystrophy

PPN: Pathway Projection Network

ATP: Adenosine Triphosphate

NADH: Reduced Nicotinamide Adenine dinucleotide).

FADH2: Flavin Adenine Dinucleotide (hydroquinone form)

FAK: Focal Adhesion Kinase

PKA: Protein Kinase

3.7 Authors' contributions

TN investigated the problem of community structure analysis and associated evidence from past work, and proposed its applicability as a useful computational technique for understanding and modeling genetic interaction networks in the specific context of muscular dystrophy. TN also made substantial contributions to the acquisition of data, and empirically evaluated the approach. SS provided guidance relative to the theoretical and practical aspects of designing/evaluating the analysis and applications. SS

also validated TN's interpretation of the results and provided deep biological insights into the same. TN drafted the first version of the manuscript and both authors revised it iteratively. All authors read and approved the final manuscript.

3.8 Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We thank Dr. Wang from the University of California, San Diego, for the valuable discussions.

Chapter III, in full, is the material as it appears in Narayanan T, Subramaniam S: Community Structure Analysis of Gene Interaction Networks in Duchenne Muscular Dystrophy. Public Library of Science (PLoS ONE) 2013. The dissertation author was the primary investigator and author of this paper.

Table 3.1: Summary of interaction networks for normal and DMD muscle

Dataset	Number of vertices considered*	Number of edges
Normal muscle	7453	130225
DMD muscle	3332	16445

*The original number of vertices after pre-processing the GSE6011 dataset was 7685

Table 3.2: Summary of network parameters

Full Network	Normal	DMD
Vertices	7453	3332
Edges	130225	16445
Clusters	670	283
Clustering coefficient	0.278	0.216
Connected components	27	219
Network diameter	16	24
Network radius	1	1
Network centralization	0.056	0.052
Shortest paths	98%	71%
Characteristic path length	5.302	7.852
Avg. number of neighbors	34.946	9.871
Network density	0.005	0.003
Network heterogeneity	1.819	2.042

Table 3.3: Parameters of networks' largest component used for community structure analysis

Dataset	Number of vertices	Number of edges
Normal muscle	7389	130185
DMD muscle	2823	16142

Table 3.4: Communities from the GSE6011 dataset

Dataset	Number of communities	Q_{max}
Normal muscle	670	0.498339
DMD muscle	283	0.535499

Table 3.5: Pathways of interest in each community

Pathway projection network	Pathway of interest
PPN1 – PPN3	Metabolic pathways
PPN4	Focal adhesion, Regulation of actin cytoskeleton and Cell adhesion molecules

Table 3.6: Sample correlation scores of highly correlated genes (Metabolic and Regulation of actin cytoskeleton pathways)

Highly correlated genes		Correlation Score
Metabolic pathway	**Regulation of actin cytoskeleton**	
Leukotriene A4 hydrolase	Cell division cycle 42	0.876235649
Phosphoinositide-3-kinase, class 2, alpha polypeptide	Platelet-derived growth factor receptor, alpha polypeptide	0.870928383
Phosphoglycerate mutase 1	Cofilin 1	0.860115666
Iduronate 2-sulfatase	integrin, alpha V	0.849397619
dCMP deaminase	Actinin, alpha 4	0.82944117

Table 3.7: Sample correlation scores of highly correlated genes (Metabolic and Calcium signaling pathways)

Highly correlated genes		Correlation Score
Metabolic pathway	**Calcium signaling pathway**	
Cytochrome P450, family 2, subfamily B, polypeptide 6	Phosphodiesterase 1C, calmodulin-dependent 70kDa	0.970659478
Cytochrome P450, family 2, subfamily C, polypeptide 9	Phosphodiesterase 1C, calmodulin-dependent 70kDa	0.945775382
Gamma-glutamyltransferase 1	Calcium/calmodulin-dependent protein kinase IV	0.912367742
Cysteine conjugate-beta lyase, cytoplasmic	Phosphodiesterase 1C, calmodulin-dependent 70kDa	0.906362395
Fructose-1,6-bisphosphatase 1	v-erb-a erythroblastic leukemia viral oncogene homolog 4	0.900885014

Table 3.8: Summary of muscle fibers' cardinality

ID	Pathway	PPN1 (Fig 3.3A)	PPN2 (Fig 3.3B)	PPN3 (Fig 3.3C)
hsa00010	Glycolysis / Gluconeogenesis (fast twitch)	4	1	5
hsa00190	Oxidative phosphorylation (slow twitch)	18	13	6

Table 3.9: Sample correlation scores of highly correlated genes (Focal adhesion and Regulation of actin cytoskeleton pathways)

Highly correlated genes		Correlation Score
Focal adhesion pathway	**Regulation of actin cytoskeleton**	
Kinase insert domain receptor (a type III receptor tyrosine kinase)	Moesin	0.930732193
Laminin, alpha 4	Integrin, alpha 6	0.916349568
Collagen, type IV, alpha 2	Actin, beta	0.914346045
Collagen, type IV, alpha 1	Actin, beta	0.910817128
Laminin, alpha 4	Actin, beta	0.9039736

Table 3.10: Sample correlation scores of highly correlated genes (Focal adhesion and Cell adhesion molecules pathways)

Highly correlated Genes		Correlation Score
Focal adhesion pathway	**Cell adhesion molecules (CAMs)**	
Laminin, alpha 4	Platelet/endothelial cell adhesion molecule 1	0.964950625
Laminin, alpha 4	Cadherin 5, type 2 (vascular endothelium)	0.939171386
Collagen, type IV, alpha 1	Platelet/endothelial cell adhesion molecule 1	0.937020588
Integrin, alpha 6	Cadherin 5, type 2 (vascular endothelium)	0.929446836
Actin, beta	Platelet/endothelial cell adhesion molecule 1	0.929072591

Table 3.11: Summary of pre-processed GSE6011 dataset parameters

Dataset	Number of Probes / Genes	Number of Experiments
Normal muscle	7685	13
DMD muscle	7685	23

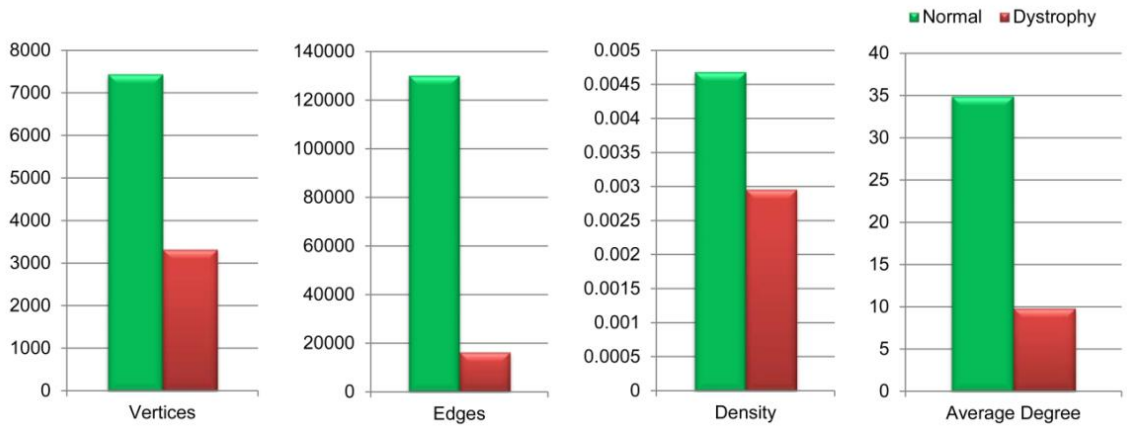


Figure 3.1: Structural Properties- Normal vs. Dystrophy Interaction Networks. Plots of the number of Vertices, number of Edges, Density and Average Degree of the Normal and DMD interaction networks that were constructed from the GSE6011 dataset [discussed in Methods Section]. The scales (y-axis) for these structural properties are different and the data for the networks are color coded as green and red for the Normal and DMD muscles respectively.

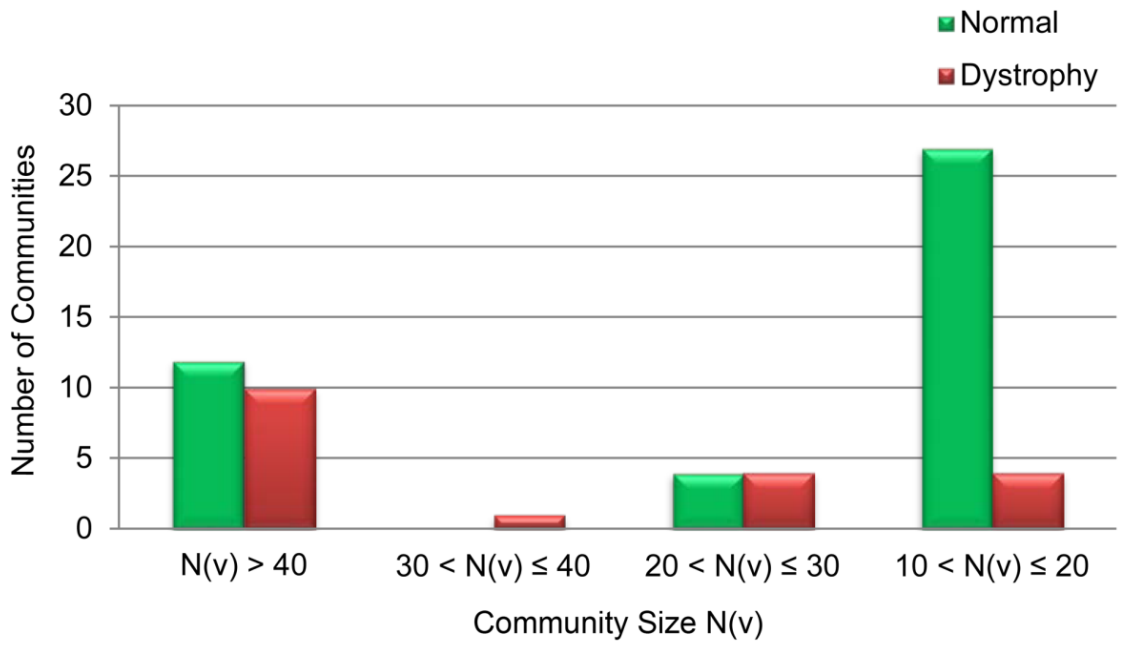
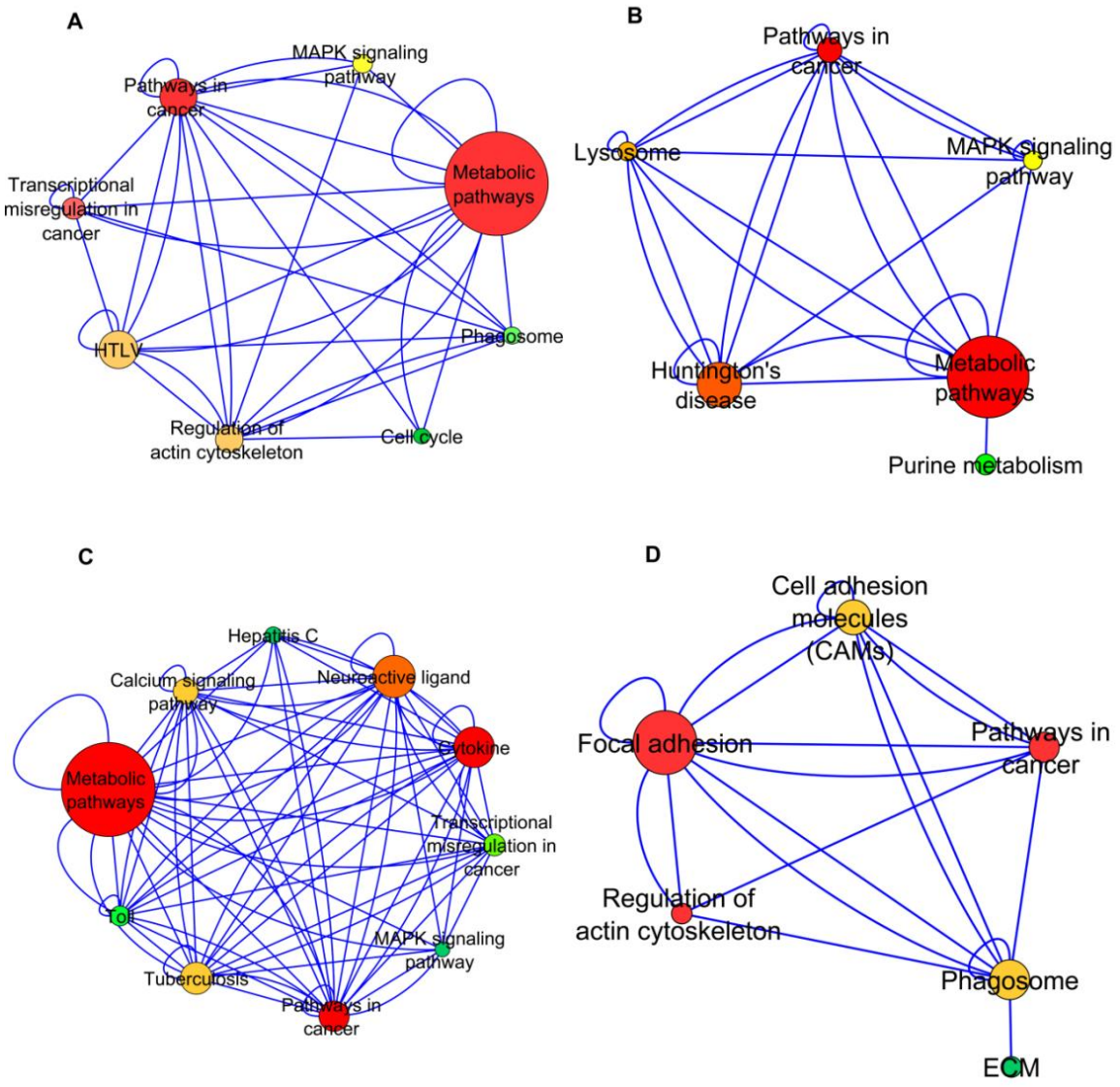


Figure 3.2: Distribution of Communities. A comparison of the distribution of communities in both the Normal and DMD networks, obtained using the Newman and Girvan's edge-betweenness algorithm. The green bars show the distribution of the total number of 644 communities obtained from the Normal network, across the four bins of community size, and the red bars represent the distribution of the 283 communities from the DMD network.



Figures 3.3: Pathway Projection Networks. Representation of communities (of interest) from the perspective of the pathways. Nodes in the PPNs are derived from (and are representative of) the pathway(s) that the constituent genes correspond to. The edges between the pathway-nodes represent the connections between the underlying genes in the original network. The nodes are color-coded according to the degree (measure of connectivity between the pathways) and size-coded according to the pathway cardinality of the node (number of genes from the community that correspond to that pathway). The transformation technique that was employed to generate an equivalent network in terms of the constituent pathways for each community is described in the Methods section [also schematically presented in the flowchart in Figure 3.4].

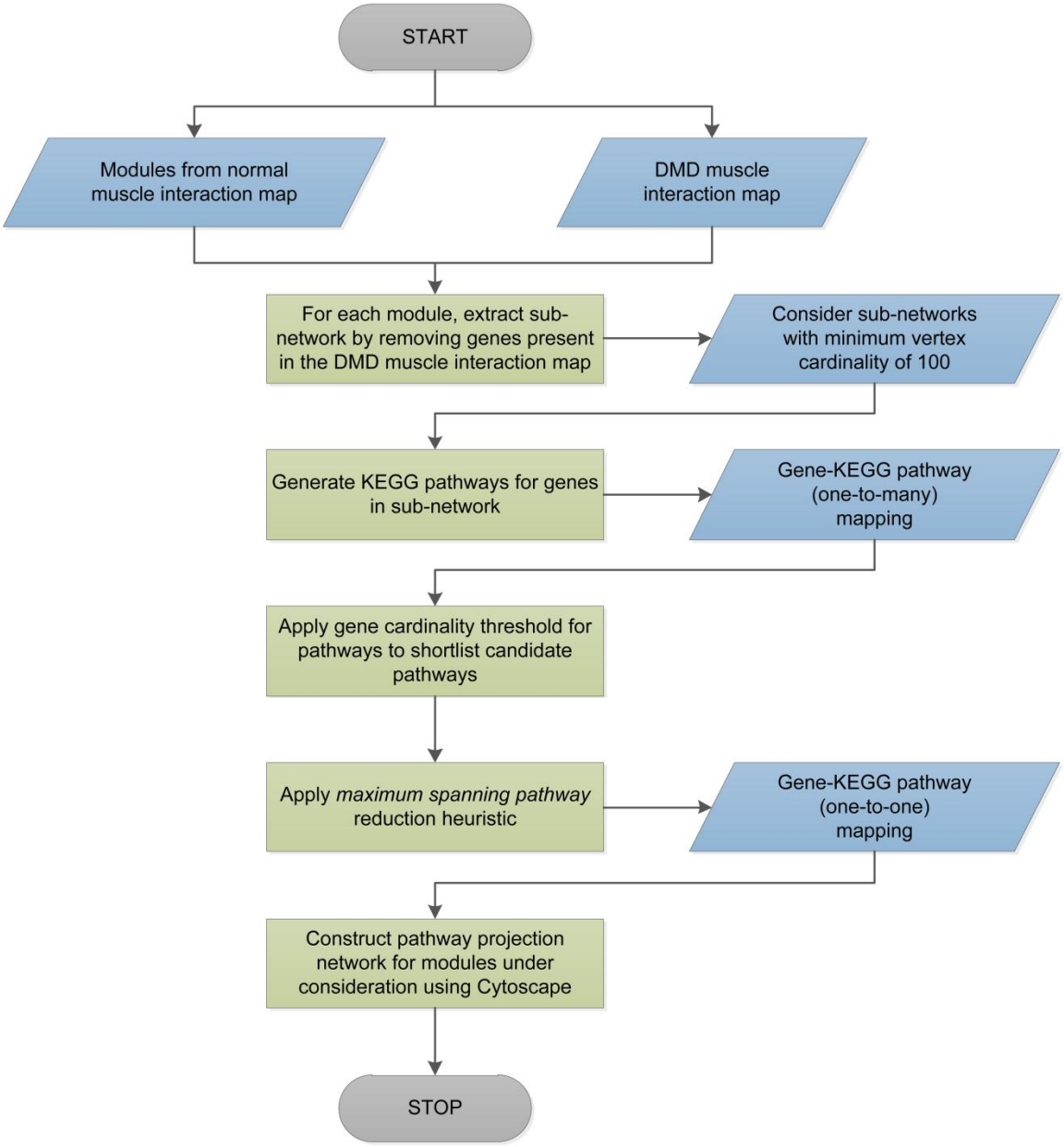


Figure 3.4: Schematic representation of transformation technique employed to generate PPNs**.** A schematic representation of the transformation technique that was employed to represent the communities from the perspective of the pathways that the constituent genes correspond to. This technique is described in detail in the Methods section.

References

1. Steinhaeuser K., Chawla N.V: Community detection in a large real-world social network. Social Computing, Behavioral Modeling, and Prediction. Springer 2008:168–175.

2. Leskovec J, Lang K, Dasgupta A, Mahoney M: Statistical properties of community structure in large social and information networks. In Proceedings of the 17th International Conference on World Wide Web, pages: 21-25 April 2008; Beijing.

3. Wasserman S., Faust K., 1994: Social Network Analysis: Methods and Applications. Cambridge University, in press.

4. Girvan, M., Newman, M: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA, 2002, 99, 7821–7826.

5. Ruan, J., Zhang, W. An efficient spectral algorithm for network community discovery and its applications to biological and social networks. In Proceedings of International Conference on Data Mining: 28-31 October 2007; Nebraska, USA

6. J. Chen, O. R. Zaiane, R. Goebel, Detecting communities in social networks using max-min modularity. In proceedings of SIAM Data Mining Conference: 30 April- 2 May, 2009; Nevada, USA.

7. Newman, M. and Girvan, M: Finding and Evaluating Community Structure in Networks, Physical Review E, 2004

8. K. Eriksen, I. Simonsen, S. Maslov, K. Sneppen Modularity and extreme edges of the Internet Phys. Rev. (2003).

9. T. N. Dinh, Y. Xuan, M. T. Thai. Towards social-aware routing in dynamic communication networks. In Proceedings of International Performance Computing and Communications Conference: 14-16 December, 2009; Phoenix, USA.

10. Nguyen NP, Dinh TN, Xuan Y, Thai MT. Adaptive algorithms for detecting community structure in dynamic social networks. In Proceedings of IEEE International Conference on Computer Communications: 10-15 April 2011; Shangai.

11. Narayanan T, Gersten M, Subramaniam S, Grama A: Modularity detection in protein-protein interaction networks. BMC Research Notes 2011: 4-569.

12. Narayanan T, Subramaniam S: Community Detection in Biological Networks Using a Variational Bayes Approach. In proceedings of the 3rd International Conference on Bioinformatics and Computational Biology: 23 – 25 March 2011; New Orleans, Louisiana USA.

13. Picard F, Miele V, Daudin JJ, Cottret L, Robin S: Deciphering the connectivity structure of biological networks using MixNet. BMC Bioinformatics, 2009.

14. Yang Q, Lonardi S: A parallel edge-betweenness clustering tool for Protein-Protein Interaction networks. International Journal of Data Mining and Bioinformatics 2007, 1(3):241-247.

15. Jake M. Hofman and Chris H. Wiggins: A Bayesian Approach to Network Modularity, Physical Review Letters, 2008.

16. The Database for Annotation, Visualization and Integrated Discovery (DAVID) [http://david.abcc.ncifcrf.gov/]

17. The Gene Ontology [http://www.geneontology.org/]

18. Muscular dystrophy [http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0002172/]

19. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2011 February: 431–432.

20. The UniProt Consortium: Reorganizing the protein space at the Universal Protein Resource (UniProt) Nucleic Acids Res. 40: D71-D75 (2012).

21. Nowak K, McCullagh K, Poon E, Davies KE: Muscular dystrophies related to the cytoskeleton/nuclear envelope. Novartis Found Symp 2005, 264:98-111.

22. Chinet AE, Even PC, Decrouy A: Dystrophin-dependent efficiency of metabolic pathways in mouse skeletal muscles. Experientia 1994.

23. Fadic R: Cell surface and gene expression regulation molecules in dystrophinopathy: mdx vs. Duchenne. Biol Res 2005, 38: 375-380.

24. Medical Dictionary Online [http://www.online-medical-dictionary.org/]

25. Yan H, Yang K, Xiao H, Zou YJ, Zhang WB, Liu HY: Over-expression of cofilin-1 and phosphoglycerate kinase 1 in astrocytomas involved in pathogenesis of radioresistance. CNS Neuroscience & Therapeutics 2012, 18(9):729-36

26. Subramanian A, Tamayo P, et al.: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS 2005, 102: 15545-15550

27. Domin J, Dhand R, Waterfield MD: Binding to the platelet-derived growth factor receptor transiently activates the p85alpha-p110alpha phosphoinositide 3-kinase complex in vivo. The Journal of Biological Chemistry, 1996, 271(35):21614-21

28. Wrogemann K, Pena SD: Mitochondrial calcium overload: a general mechanism for cell-necrosis in muscle diseases. Lancet 1976, 1:672-4.

29. The Rat Genome Database [http://rgd.mcw.edu/]

30. Repaske DR, Swinnen JV, Jin SL, Van Wyk JJ, Conti M: A polymerase chain reaction strategy to identify and clone cyclic nucleotide phosphodiesterase cDNAs. Molecular cloning of the cDNA encoding the 63-kDa calmodulin-dependent phosphodiesterase. The Journal of Biological Chemistry 1992, 267(26): 18683-8

31. RefSeq, Jul 2008 Pruitt KD, Tatusova T, Klimke W, and Maglott DR: NCBI Reference Sequences: current status, policy and new initiatives. Nucleic Acids Research 2008, doi: 10.1093/nar/gkn721

32. Chen YW, Nagaraju K, Bakay M, et al.: Early onset of inflammation and later involvement of TGFbeta in Duchenne muscular dystrophy, Neurology 2005: 826-834.

33. Bianchi ML, Morandi L: Evaluating Bone and Mineral Metabolism in Patients with Duchenne Muscular Dystrophy. European Musculoskeletal Review, 2008.

34. Spence HJ, Chen YJ, Batchelor CL, Higginson JR, Suila H, Carpen O, Winder SJ: Ezrin-dependent regulation of the actin cytoskeleton by beta-dystroglycan. Human Molecular Genetics 2004, 13(15):1657-68.

35. Pescatori M, Broccolini A, Minetti C, Bertini E et al. Gene expression profiling in the early phases of DMD: a constant molecular signature characterizes DMD muscle from early postnatal life throughout disease progression. FASEB J 2007 (4):1210-26

36. Tinsley JM, Blake DJ, Pearce M, Knight AE, Kendrick-Jones J, Davies KE: Dystrophin and related proteins. Current Opinion in Genetics & Development 1993,484-90.

37. Keep NH: Structural comparison of actin binding in utrophin and dystrophin. Neurological Sciences 2000, 929-37.

38. Warner LE, DelloRusso C, Crawford RW, Rybakova IN, Patel JR, Ervasti JM, Chamberlain JS: Expression of Dp260 in muscle tethers the actin cytoskeleton to the dystrophin-glycoprotein complex and partially prevents dystrophy. Human Molecular Genetics. 2002, 11(9):1095-105.

39. Johnson BD, Scheuer T, Catterall WA: Convergent regulation of skeletal muscle Ca2+ channels by dystrophin, the actin cytoskeleton, and cAMP-dependent protein kinase. Proceedings of the National Academy of Sciences 2005, 102(11):4191-6.

40. Cohen, J: Statistical power analysis for the behavioral sciences (2nd ed.), Erlbaum; 1988.

41. Sakuma K, Nakao R, Inashima S, Hirata M, Kubo T, Yashuram M: Marked reduction of focal adhesion kinase, serum response factor and myocyte enhancer factor 2C, but increase in RhoA and myostatin in the hindlimb dy mouse muscles. Acta Neuropathol (Berl) 2004, 108: 241-249.

42. Wei L, Zhou W, Wang L, Schwartz RJ: ß1-Integrin and PI 3- kinase regulate RhoA-dependent activation of skeletal a-actin promoter in myoblasts. Am J Physiol 2000, 278:H1736–H1743.

43. Carson JA, Wei L: Integrin signaling's potential for mediating gene expression in hypertrophying skeletal muscle. J Appl Physiol 2000, 88:337–343

44. Sorokin LM, Maley MA, Moch H, von der Mark H, von der Mark K, Cadalbert L, Karosi S, Davies MJ, McGeachie JK, Grounds MD: Laminin alpha4 and integrin alpha6 are upregulated in regenerating dy/dy skeletal muscle: comparative expression of laminin and integrin isoforms in muscles regenerating after crush injury. Experimental Cell Research 2000 256(2):500-14

45. Albelda SM, Muller WA, Buck CA, Newman PJ: Molecular and cellular properties of PECAM-1 (endoCAM/CD31): a novel vascular cell-cell adhesion molecule. The Journal of cell biology 1991, 114(5):1059-68

46. Irving-Rodgers HF, Hummitzsch K, Murdiyarso LS, Bonner WM, Sado Y, Ninomiya Y, Couchman JR, Sorokin LM, and Rodgers RJ: Dynamics of extracellular matrix in ovarian follicles and corpora lutea of mice. Cell and Tissue Research 2010 339(3): 613–624

47. Torrente Y, Camirand G, Pisati F, Belicchi M, Rossi B, Colombo F, Fahime ME, Caron NJ, Issekutz AC, Constantin G, Tremblay JP: Identification of a putative pathway for the muscle homing of stem cells in a muscular dystrophy model. Nereo Bresolin J Cell Biol 2003, 162(3): 511–520.

48. Engvall E: Cell adhesion in muscle. Braz J Med Biol Res 1994, 27(9):2213-27.

49. Wang Y, Winters J, Subramaniam S: Functional classification of skeletal muscle networks I: Normal physiology. Journal of Applied Physiology 2012

50. Wang Y, Winters J, Subramaniam S: Functional classification of skeletal muscle networks. II. Applications to pathophysiology. Journal of Applied Physiology 2012, 113(12): 1902-1920

51. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: KEGG for integration and interpretation of large-scale molecular datasets. Nucleic Acids Res 2012, 40: D109-D114.

52. Kanehisa M, Goto S: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 2000, 28: 27-30.

Chapter IV

A Newtonian Framework for Community Detection in Biological Networks

Chapter IV, in full, has been submitted for publication of the material as it may appear in Narayanan T; Subramaniam S. A Newtonian Framework for Community Detection in Biological Networks. IEEE Transactions on Biomedical Circuits and Systems (TBioCaS) 2013. The dissertation author was the primary investigator and author of this paper.

Abstract

Community detection is a key problem of interest in network analysis, with applications in a variety of domains such as biological networks, social network modeling, and communication pattern analysis. In this paper, we present a novel framework for community detection that is motivated by a physical system analogy. We model a network as a system of point masses, and drive the process of community detection, by leveraging the Newtonian interactions between the point masses. Our framework is designed to be generic and extensible relative to the model parameters that are most suited for the problem domain. We illustrate the applicability of our approach by applying the Newtonian Community Detection algorithm on protein-protein interaction networks of E. coli, C. elegans, and S. cerevisiae. We obtain results that are comparable in quality to those obtained from the Newman and Girvan algorithm, a widely employed divisive algorithm for community detection. We also present a detailed analysis of the structural properties of the communities produced by our proposed algorithm, together with a biological interpretation using E. coli protein network as a case study. A functional enrichment heat map is constructed with the GO functional mapping, in addition to a pathway analysis for each community. The analysis illustrates that the proposed algorithm elicits communities that are not only meaningful from a topological standpoint, but also possess biological relevance. We believe that our algorithm has the potential to serve as a key computational tool for driving therapeutic applications involving targeted drug development for personalized care delivery.

Index Terms— Biological pathways, Biological systems modeling, Community detection, Protein networks, Proteomics

4.1 Introduction

Community detection is a key problem of interest in network analysis [1] [2] [3] [4] [5]. The techniques for identifying communities in networks have widespread applications in the analysis of biological networks [6] [7] [8] [9] [10], social networks [6] [11] [12] [13] [14] [15] [16], and internet traffic patterns [17] [18] [19]. The problem of community detection is especially challenging when dealing with large-scale networks with thousands or tens of thousands of nodes and edges.

Algorithms used for community detection have the following characteristics as their desiderata. The fundamental expectation is that the algorithm should be able to produce communities (or a representation thereof), which are contextually meaningful, as its output. Additionally, the algorithm would have wider applicability if it is schema-agnostic i.e. it lends itself to being adapted / extended to different types of networks (for example, weighted and unweighted networks, directed and undirected networks, static and time-varying networks etc.). Furthermore, not only must the computation time of the algorithm be bounded, but the algorithm must also be efficient with convergence guarantees on the number of communities (with acceptable node memberships).

Typically, we also expect that the communities produced by a community detection algorithm reflect underlying semantic structures in the original network. For example, in the context of biological networks, we expect that the communities produced

by a community detection algorithm represent some functionally related elements of the underlying biological pathways in the original network (which could potentially be used for applications such as phenotype prediction). Moreover, singleton-communities (with just one node) or communities with just 2-3 nodes are not expected to yield much biological insight. Thus, an effective community detection algorithm should yield communities with reasonable vertex cardinality.

Applications of community detection in biological networks include functional annotation of constituent biomolecules, since nodes in the same community are likely to be associated with similar functions. This can serve as a key tool in the analysis of pathology [20] and enable therapeutic applications, with potential for personalized care delivery. Most biological networks of interest tend to be large and complex, both in terms of the number of nodes and the edges between them. For instance, the complete Protein-Protein Interaction Network (PPIN) of H.Sapiens consists of over 10000 proteins and 81000 interactions [21]. Another characteristic of biological networks that augments the complexity of community detection is that, such networks also tend to be sparse from a connectivity perspective [22].

4.2 Community Detection: A Newtonian Framework

In this paper, we present a general framework for com-munity detection that is motivated by a physical system analogy. The intuition driving our approach is to model a network with nodes and edges as a physical system of point-masses and consider the interactions between them from a Newtonian standpoint. This is illustrated in Fig. 1.

Conceptually, we consider these interactions to be acting "along" the edges connecting nodes (for example nodes v_i and v_j shown in Fig. 1). We quantify the "strength" of these interactions using structural properties of the underlying network, together with a set of configurable model parameters and the Newtonian laws applicable to interaction between point-masses in the model. Finally, we leverage the quantified interactions to guide the community detection process on the underlying network. The design of our framework is also informed by the observation of a key characteristic of communities in real world networks (such as biological networks) viz. the intra-community edges are typically denser than the (sparse) inter-community edges. We formalize our approach in subsequent sections.

4.3 Definitions

In this section, we present the definitions of terminology that we use in the remainder of the paper.

4.3.1 Community [K]

Given a graph $G = (V, E)$, we define a Community K as a sub-graph of G such that $K = (v, e)$ where $|v| \leq |V|$ and $|e| \leq |E|$.

4.3.2 Distance Matrix [D]

Given a graph $G = (V, E)$, we define the Distance Matrix D as a $|V| \times |V|$ matrix where every element d_{ij} is the length of the shortest path between nodes v_i and v_j.

4.3.3 Mass Mapping [μ]

Given a graph $G = (V, E)$, we define a Mass Mapping μ as a function that maps every node v_i of G to a real number m_i

$$\mu(\boldsymbol{v}): V \rightarrow$$

$$\mu(v_i) = m_i \qquad v_i \in V \; and \; m_i \in R$$

m_i is said to be the mass-point of node v_i. We note that the Mass Mapping $\hat{\mu}$ is a configurable model parameter of our algorithmic framework.

We further define the Unit Mass Mapping $\hat{\mu}$ as follows:

$$\hat{\mu}(\boldsymbol{v}): \boldsymbol{V} \rightarrow$$

$$\hat{\mu}(v_i) = 1 \qquad v_i \in V$$

4.3.4 Newtonian Field [Ƒ]

Given a graph $G = (V, E)$ and a Mass Mapping μ, we define the Newtonian Field of G under μ as a $|V| \times |V|$ matrix:

$$= [f_{ij}] \; [f_{ij}]_{|V| \times |V|}$$

$$f_{ij} = \begin{cases} \frac{\dot{G}\mu(v_i)\mu(v_j)}{d_{ij}^2}, & i \neq j \\ 0, & i = j \end{cases} \qquad v_i, v_j \in V$$

Here, $\dot{G}$ is a constant configurable model parameter of our algorithmic framework, and d_{ij} denotes the corresponding element from the Distance Matrix D of G. Since the length of the shortest path between a node and itself (d_{ii}) is zero, the corresponding element in the Newtonian Field is taken to be zero.

4.3.5 Vertex Field Projection [Φ]

Given a graph $G = (V, E)$ and a Mass Mapping μ, we define the Vertex Field Projection of G under μ as a $|V|$-dimensional vector:

$$\Phi = [\varphi_i]_{1\times|V|}$$

$$\varphi_i = \sum_j f_{ij}$$

4.3.6 Edge Field Projection [Ω]

Given a graph $G = (V, E)$ with an adjacency matrix A, a Mass Mapping μ, and a real valued function $\lambda(x, y)$, we define the Edge Field Projection of G under the parameters μ and λ as a $|V|\times|V|$ matrix:

$$\boldsymbol{\Omega} = [\omega_{ij}]_{|V|\times|V|}$$

$$\omega_{ij} = \begin{cases} \lambda(\varphi_i, \varphi_j), & A_{ij} = 1 \\ 0, & A_{ij} = 0 \end{cases}$$

We call $\lambda(x, y)$ a projection transformation function since it serves the purpose of transforming projections defined over nodes (φ_i) to projections defined over the edges (ω_{ij}). We note that the projection transformation function $\lambda(x, y)$ is a configurable model parameter of our algorithmic framework.

4.3.7 Modularity [Q]

Modularity is a measure of the quality of a particular division of a network into communities. We use the definition of this term from [1]. In particular, given a specific division of a network into k communities, we define a $k \times k$ symmetric matrix $\boldsymbol{e}$ whose

element e_{ij} is the fraction of all edges in the original network that link nodes in community i to nodes in community j. We further define the row (or column) sums $a_i = \sum_j e_{ij}$, which represent the fraction of edges that connect to nodes in community i. Finally, we define modularity as follows:

$$Q = \sum_i (e_{ii} - a_i^2) = Tr\ \mathbf{e} - \|\mathbf{e}^2\|$$

where $Tr\ \boldsymbol{e} = \sum_i e_{ii}$, denotes the trace of the matrix $\boldsymbol{e}$ and ‖e‖ denotes the sum of the elements of the matrix $\boldsymbol{e}$.

4.4 Algorithm: Newtonian Community Detection (NCD)

In this section, we present our algorithm for community detection.

INPUT

Graph $G = (V, E)$, Mass Mapping μ, $\dot{G}$, Projection Transformation Function $\lambda(x, y)$

OUTPUT

Set of communities { K_1, K_2, …, K_n }

STEPS

1. Create an empty ordered list $L \leftarrow \emptyset$

2. Repeat while the edge set E of $G \neq \emptyset$

{

2.1 Compute the distance matrix D of G

2.2 Compute the Newtonian Field F of G under μ

2.3 Compute the Vertex Field Projection Φ of G under μ

2.4 Compute the Edge Field Projection Ω of G under parameters μ and λ

2.5 Let e_{ij} denote the edge corresponding to the maximum ω_{ij} from Ω

2.6 Append e_{ij} to L

2.7 $E \leftarrow E - \{ e_{ij} \}$

}

3. L defines a dendrogram on G. Compute the modularity Q for every split of G

4. Output communities $\{ K_1, K_2, \ldots, K_n \} \leftarrow$ the set of communities when Q is maximum.

4.5 Algorithm Analysis

We describe the semantic connection between the algorithm (as outlined above) and how it strives to address the problem of community detection (by adopting a physical systems analogy). As noted in section II, we view the input graph (G) as a connected physical system of point-masses interacting under Newtonian influences. Specifically, we model these interactions to be acting along the edges of the input graph, and determine how the nodes in the graph would congregate into communities (under the influence of these interactions).

Step 2 constitutes the main loop of the algorithm where the iterative process of edge-removal is performed until no further edges remain. In each iteration of the algorithm, the edge which is subjected to the maximum magnitude of Newtonian interaction, defined by the Edge Field Projection, is removed (Step 2.5). The magnitude of interaction along an edge is modeled as the projection of the aggregated interactions on

the nodes connected by the edge, denoted by Vertex Field Projection (Step 2.4). The aggregated interaction on a given node is obtained by considering every other node in the graph and computing the Newtonian interaction between the two nodes (Steps 2.3 and 2.2). To aid in the computation of the interactions between any two nodes, the distance matrix for the whole graph is calculated (Step 2.1).

Our algorithm for community detection is divisive in nature and works by iteratively removing edges that satisfy a certain criterion within the Newtonian framework defined in section III. The order of removal of edges defines a dendrogram on *G*. At every possible candidate split of the network into communities (as defined by the dendrogram), we calculate the modularity *Q*. Finally, we select that candidate split which corresponds to the maximum value of Q as the one that produces the output set of communities. The NCD algorithm will terminate deterministically. The proof follows from the fact that in every iteration of the main loop of the algorithm, we consider successively smaller sub-graphs of *G* to search for communities that meet a certain criterion within the Newtonian framework defined in this paper.

If m and n denote the number of edges and nodes in the input graph *G* and we assume that there is a constant computational cost in evaluating the functions μ and λ in every iteration of the algorithm, we can bound the worst-case time complexity of the NCD algorithm by $O(m^2n + n^2m)$ or $O(n^3)$ for sparse graphs. We observe that the bound matches the worst-case complexity for the Newman and Girvan (NG) algorithm [1] and is generally accepted to be computationally tractable for most real world networks. Note that we assume that the distance matrix computation is performed using the Johnson's algorithm [23] which has a time complexity of $O(n * \log(n) + ne)$.

4.6 Algorithm Evaluation

In this section, we evaluate our proposed algorithm using biological PPIN of common interest. In particular, we compare the results from our algorithm to those obtained from running the NG algorithm, on the same networks. The networks we considered are summarized in Table I.

4.6.1 Dataset Description

The full network of S. cerevisiae (yeast), obtained from Biogrid [24], contained 160566 interactions, but we restricted the dataset to interactions determined by co-purification or yeast two-hybrid experiments, thus giving only 15316 interactions. We also pre-processed the networks in Table I to collapse multiple edges between a given pair of nodes into a single edge, eliminated self-loops on nodes, and also ignored mirrored-edge representations in the network (i.e. an interaction between nodes A and B is considered the same as an interaction between nodes B and A). The resulting S. cerevisiae network had 9946 interactions (edges) between 3654 proteins (nodes) with multiple components, of which we consider the largest component (with 9890 edges and 3551 nodes as shown in Table I). Similarly, the complete E. coli network (obtained from [25]) had 3989 edges between 1941 nodes, but only the largest component (with 1274 nodes and 3124 edges) was considered for our analysis. The C. elegans network with 199 nodes and 251 edges was obtained from [26].

The parameters we used when applying the NCD algorithm on the PPINs are summarized below:

- $\dot{G} \equiv 1$

- Mass Mapping (μ) ≡ Unit Mass Mapping ($\hat{\mu}$)
- Projection Transformation Function $\lambda(x, y) \equiv x + y$

4.6.2 Modularity (Q) comparison

Fig. 2 presents the results of the execution of our algorithm in terms of the modularity (Q) corresponding to the resultant division of the network into communities. We also present the corresponding results from the NG algorithm. As noted in [1], the modularity of networks with a strong community structure typically fall in the range from about 0.3 to 0.7 in practice. Accordingly, we conclude that the networks under consideration exhibit modular structures. Furthermore, we observe that our approach to community detection yields communities which are comparable in quality (as defined by the modularity measure) to the NG algorithm.

4.7 Biological Interpretation

In this section, we present a detailed analysis of the topological properties and biological interpretation of the communities produced by the NCD algorithm using the PPIN of E. coli* as a case study. E. coli is an extensively used model organism in biological analysis. The discussion in this section is intended to illustrate the applicability of our approach to detecting communities in large scale biological networks of practical relevance and significance. We believe that this can serve as a foundation for understanding the correspondence between structure and function of biological networks,

which could serve as a key step in the process of "-omics" based diagnostics and disease modeling.

* While the biological interpretation of the communities produced by the NCD algorithm from the PPIN of *S. cerevisiae* was also performed, we present only the results from the E. coli PPIN in this Chapter. The E. coli network is not only a widely used model organism, but is also more annotated than the *S. cerevisiae* network, both in terms of the functional and pathway annotations of the constituent genes. Hence, it serves as a more illustrative example for the purposes of the discussion in this section.

4.7.1 Analysis of network properties

When evaluating a community detection algorithm, it is important to analyze the communities produced by the algorithm from a topological standpoint, to ascertain that they reflect the underlying structural organization of the network. In the following sections, some common structural properties of interest that can be used to characterize real-world community structures are defined and discussed in the context of the communities produced by the NCD algorithm from the E. coli network. We restrict our analysis to communities with a minimum vertex cardinality of 30 (as shown in Table II).

4.7.1.1 Network Density

The density ρ of a community is defined as the ratio of edges it actually contains (m_c), to the number of edges it could contain if all its nodes were connected. In the case of an undirected network, the latter is $n_c(n_c - 1)/2$, where n_c is the number of nodes in

the community. Thus,

$$\rho = \frac{2*m_c}{(n_c(n_c-1))} \quad .$$

When compared to the overall network density, the density allows assessing the cohesion of the community: by definition, a community is supposed to be denser than the network it belongs to [27]. In other words, an effective community detection algorithm must produce as output, communities that are more cohesive relative to the input network.

Fig. 3 presents the log of network density of the com-munities produced by the NCD algorithm from the E. coli network (represented by the circles). The figure also includes the corresponding data point for the input E. coli network (solid square). It is evident that the network densities of the communities produced are significantly higher (representing cohesion) than the network density of the entire E. coli network, which serves to reinforce the effectiveness of our algorithm.

4.7.1.2 Node Degree Distribution

Prior work has shown that, while in a random network most nodes have comparable degrees, real networks tend to have a significant number of highly connected nodes and large differences in node degrees [28], typically attributed to the network's scale-free property. In other words, the degree distribution DD of many real-world networks (such as biological networks) approximates a power law: $DD(k) \sim k^{-\propto}$ [29].

Fig. 4 shows the node degree distribution for the E. coli network and three other sample communities from the ones shown in Table II. We observe that these follow a

power law distribution (it should be noted that the plots in Fig. 4 have a logarithmic axis and a straight line on such log–log plots indicates a power-law scaling). This observation illustrates that the NCD algorithm conserves the underling topological semantics of the input network (such as the scale-free property), when generating output communities.

4.7.1.3 Clustering Coefficient Distribution

The clustering coefficient C_i captures the density of edges in node i's immediate neighborhood: C = 0 means that there are no edges between i's neighbors; C = 1 implies that each of the i's neighbors are connected to each other . For a random network of size N, the average clustering coefficient depends on the network size as N^{-1}, whereas, it is largely independent of the network size in real networks [28].

Fig. 5(a) shows the distribution of the average clustering coefficient of the communities from the E. coli network (black dots) shown in Table II. The solid line corresponds to the predicted trend for random networks, with the average clustering coefficient decreasing as N^{-1} . It is interesting to note that the average clustering coefficients of the communities produced by the NCD algorithm is independent of N, illustrating that the communities identified by our algorithm exhibit characteristics of real networks.

Furthermore, authors of [28] note that, unlike random networks, the clustering coefficient distribution C(k) for real networks decreases with the node degree k. C(k) is measured by averaging the clustering coefficient of all nodes with the same degree k. Fig. 5(b)-(d) plots the C(k) function for the E. coli network and two other sample communities from the ones shown in Table II. We observe a decreasing trend of C(k) with the node degree k, across all communities.

These observations reinforce our belief in the effectiveness of the NCD algorithm. Specifically, these results illustrate that the communities produced by the proposed algorithm are not just random, but that the underlying real-world network topological properties are captured and preserved such as the scale-free property), when generating output communities.

4.7.1.4 Average distance

The distance between two nodes in a network is defined as the length of the shortest path connecting them. When averaged over all pairs of nodes in a community, it allows assessing the cohesion of the community. In real-world networks, the average distance of small communities ($n_c \leq 10$) increases logarithmically with the community size n_c. For larger communities, the increase in average distance is even less pronounced [27]. Fig. 6 shows the distribution of the average distance of the communities listed in Table II. We observe that the trend line in this figure is consistent with the aforementioned behavior. This serves as testimony to the ability of the NCD algorithm to elicit communities that possess properties which are aligned with real world networks.

4.7.2 Functional Enrichment Analysis

In this section, we evaluate the correspondence between the topological communities we identified using the NCD algorithm, with functional units in the E. coli network. In particular, we leverage the functional annotations from Gene Ontology (GO) [30] for the genes constituting the network, and examine the distribution of the genes in

each structural community relative to their functional annotations. We use DAVID [31] [32] for performing this analysis.

DAVID uses a modified Fisher Exact p-value, for gene-enrichment analysis, which ranges from 0 to 1; a lower p-value is considered to be indicative of stronger enrichment in the annotation categories [33]. Fig. 7 presents a color-coded "heat-map" representation of the relative enrichment of each functional annotation in the communities from the E. coli PPIN. As noted in [9], we expect a nonrandom distribution of proteins of a given functional category across communities (as reflected by a lower p-value), if the structural communities correspond to functional units. Furthermore, the primary composition of genes in a given community is expected to correspond to a small set of functional annotations.

4.7.3 Heat-map: Observations / Results

For the functional enrichment analysis, we consider only those communities from the E. coli network, that have a vertex cardinality (V_c) of at least 20 genes. The NCD algorithm yielded 14 such communities (C_1-C_{14}). These are represented in the columns of Fig. 7. The rows of the heat-map represent the union of the highest and second highest enrichments (in terms of the functional annotations) across C_1-C_{14}. The highest and the second highest enrichments are color-coded black and gray respectively.

It is worthwhile noting that the median p-value of the highly enriched functional annotations represented in the heat-map is ~ 10^{-8}. This is strongly indicative of statistical significance of functional enrichment in the communities identified by the NCD

algorithm (typical p-values used for determining statistical significance of functional categories is ~ 10^{-2}) [31] [32].

The first observation we make from the heat-map is that, within a given community, the highest (black) and second highest (gray) functional enrichments tend to be related from a biological standpoint (for most communities). For example, in C_2, *structural constituent of ribosome* and *structural molecule activity* are color-coded black and gray respectively and from [34] we know that they are functionally related (*is-a* relationship). Similarly, it is known from [34] that *symporter activity* and *cation: sugar symporter activity* are functionally related and in Fig. 7, we observe that they are the highest and second highest enrichments respectively in C_{10}. These observations are in alignment with our expectation that similar functional annotations are likely to cluster in a community, thus enhancing the specificity of functional enrichment.

Secondly, we observe that across communities, the set of highly expressed functional annotations are predominantly disjoint. In other words, for a given row (functional annotation) in the heat-map, there is utmost one cell that is color-coded with black or gray, across all columns (communities), with the only exception of *structural molecule activity*, which occurs in C_2 and C_5, as the second highest enrichment. For example, *protein-N(PI)-phosphohistidine-sugar phosphotransferase activity* is exclusively enriched in community C_9 only. Similarly, community C_{14} exhibits high enrichment in *Cell surface antigen activity, host-interacting* and this annotation is not highly expressed in any other community.

Furthermore, not only are the enriched functional annotations across communities syntactically disjoint, but also correspond to semantically different functions from a

biological standpoint (in most cases). For example, *motor activity* and *symporter activity* (that are highly enriched in C_5 and C_{10} respectively) correspond to *catalytic activity* and *transporter activity* respectively which are disjoint paths in the GO functional hierarchy [34]. This serves to reinforce our belief that, not only do communities exhibit specificity of functional enrichment within themselves, but also that such enrichment occurs with inherent diversity across communities.

4.7.4 Pathway Analysis

In this section, we present an analysis of the communities from the E. coli network from the perspective of the pathways that the constituent genes represent. We expect that, this analysis is likely to yield more holistic insights into the correlation between functional and structural organization of the underlying genetic interactions.

We restrict our analysis to those communities from the E. coli protein interaction network, that have a vertex cardinality of at least 20. The pathway analysis was performed for these candidate communities (C_1-C_{14}) using DAVID Pathway Viewer, a feature in the DAVID Functional Annotation Tool that provides KEGG pathway enrichments [35] [36] for each community. Not only does the DAVID pathway viewer provide the list of pathways that the constituent genes in each community corresponds to, but also provides the associated p-values for every pathway. The technique for p-value computation is based on the hypergeometric distribution and is detailed in [31] [32].

Table III summarizes the highly expressed pathways (by gene cardinality) for communities C_1-C_{14} and also presents their corresponding p-values. For each community, the *percentage of vertex cardinality* is a measure of the percentage of genes from that

community which corresponded to the associated pathway. Furthermore, the highly expressed pathways summarized in Table III (except one) also turned out to be the ones that corresponded to the least p-value, indicating statistically significant enrichment.

A desirable property of a community detection algorithm is to detect communities that correspond to some underlying semantic organization of the network [9]. The results from the pathway analysis illustrate that the NCD algorithm exhibits this property. In particular, Table III illustrates specificity of unique pathway enrichment across communities, which showcases the underlying biological / functional organization of the network.

We also note from Table III that many of the enriched pathways are functionally related to the enriched GO term for that corresponding community represented in the heat map (as shown in Fig. 7). For example, the *Ribosome* pathway is the most expressed in C_2 (as shown in Table III). We can also see from the heat map that C_2 is enriched (color coded by a black cell) in a related biological function, namely, *Structural constituent of ribosome.* Similarly, the *Two-component response regulator activity* is enriched in C6 (from the heat map) and we can observe that the pathway that the community C_6 is enriched in, is the *Two-component system.* This consistency in related biological functions and pathway enrichments from the corresponding communities reinforces our confidence in the results produced by the NCD algorithm.

4.8 Conclusions and Future Work

In this paper, we have presented a novel approach to community detection which

is motivated by a physical system-modeling of a network. We believe that our approach has the following merits. Firstly, our algorithmic framework is designed to be non-prescriptive relative to model parameters, and thus allows for a flexible and extensible approach to community detection. Secondly, the construct of a Mass Mapping (μ) enables us to reason about both weighted and unweighted graphs using the same model for the purposes of community detection.

Our results from the E. coli network case study illustrate that the NCD algorithm yields communities that possess structural properties that align with real world scale-free networks. Furthermore, we find that the resultant communities possess biological significance in terms of functional enrichments and pathway specificity. These results reinforce our belief in the applicability of the NCD algorithm for effective use in analyzing biological networks. We expect that this can have a significant impact on the creation of tools and techniques for therapeutic intervention and drug targeting.

Acknowledgement

Chapter IV, in full, is a reprint of the material has been submitted for publication of the material as it may appear in Narayanan T; Subramaniam S. A Newtonian Framework for Community Detection in Biological Networks. IEEE Transactions on Biomedical Circuits and Systems (TBioCaS) 2013. The dissertation author was the primary investigator and author of this paper.

Table 4.1: Summary of Networks. Summary of the Protein-Protein Interaction Networks (PPIN) used in our study to evaluate the NCD algorithm.

Network	**Number of Nodes**	**Number of Edges**	**Source of Data**
E. coli	1274	3124	[25]
C. elegans	199	251	[26]
S. cerevisiae	3551	9890	[24]

Table 4.2: E. coli Network Communities. Summary of communities (with a minimum vertex cardinality of 30) produced by the NCD algorithm from the E. coli network.

Community	**Number of Nodes**	**Number of Edges**
C_1	77	391
C_2	97	160
C_3	115	299
C_4	77	180
C_5	308	594
C_6	115	160
C_7	36	54
C_8	33	46
C_9	46	57

Table 4.3: Pathway Enrichments of E. coli Network Communities. Summary of the highly expressed pathways (by gene cardinality) for communities C1-C14 with their corresponding p-values. Percentage of vertex cardinality is a measure of the percentage of genes from that community which corresponded to the associated pathway.

Community	Pathway	% Vertex Cardinality	P-Value
C_1	Purine metabolism	5.24	2.34E-08
C_2	Ribosome	29.41	3.77E-47
C_3	Sulfur metabolism	5.75	6.69E-07
C_4	Pyrimidine metabolism	20.25	2.37E-20
C_5	Flagellar assembly	41.56	2.59E-62
C_6	Two-component system	52.00	1.07E-57
C_7	Lysine biosynthesis	15.79	1.85E-09
C_8	Valine, leucine, isoleucine biosynthesis	11.11	7.41E-04
C_9	Phosphotransferase system	25.81	2.31E-12
C_{10}	Starch and sucrose metabolism	25.93	4.78E-11
C_{11}	Nitrogen metabolism	33.33	4.88E-14
C_{12}	Citrate cycle (TCA cycle)	54.17	1.62E-26
C_{13}	ABC transporters	78.95	1.79E-23
C_{14}	Lipopolysaccharide biosynthesis	82.35	1.50E-32

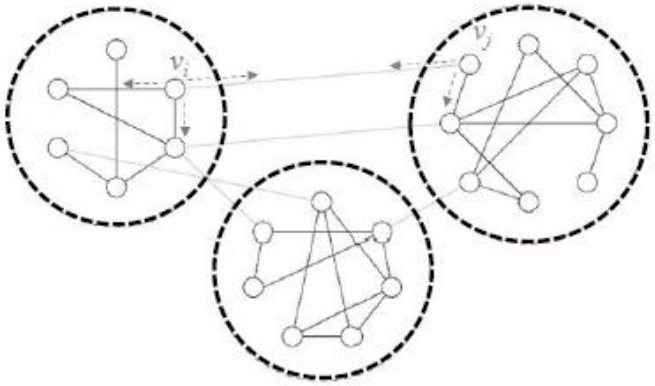


Figure 4.1: Newtonian Framework for Community Detection. This figure illustrates the intuition driving our approach, which is to model a graph with nodes and edges as a

physical system of point-masses and consider the interactions between them from a Newtonian standpoint.

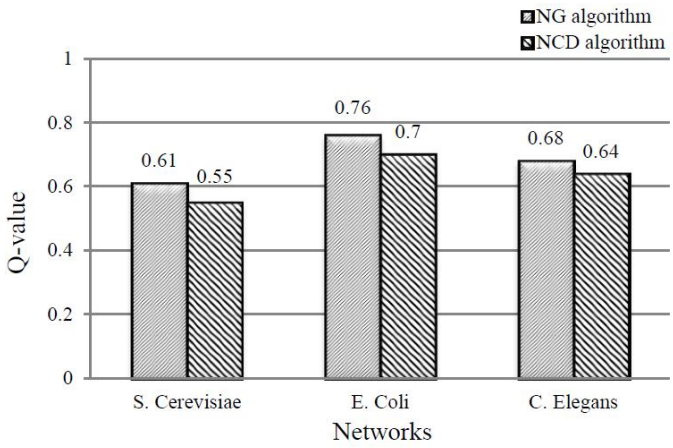


Figure 4.2: Comparison of Q-value. The figure presents the modularity (Q) of the resultant divisions from our proposed algorithm and of the NG algorithm. We observe that our approach to community detection yields communities which are comparable in quality (as defined by the modularity measure) to the NG algorithm.

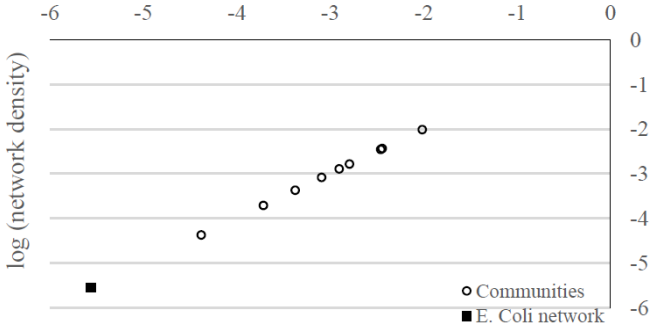


Figure 4.3: Log network density of E. coli network communities. Black circles represent the significantly higher network density (log) of communities produced by the NCD algorithm from the E. coli network. The network density (log) of the input E. coli network is represented as the solid square.

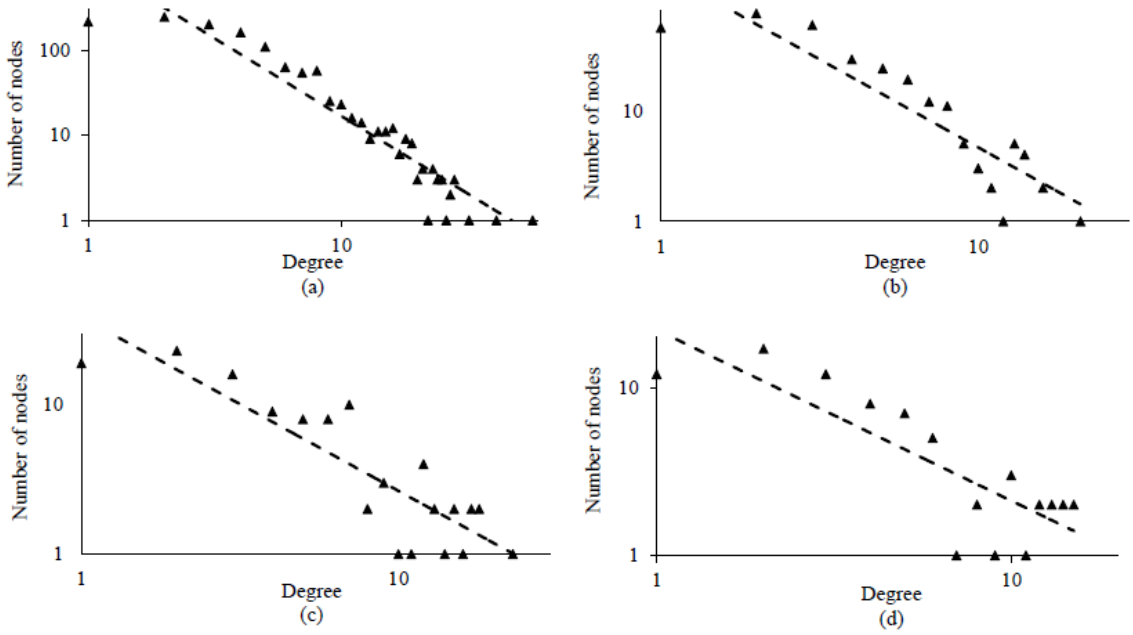


Figure 4.4: Node degree distribution. Log-log plots of node degree distribution of (a) the E. coli network and (b)-(d) three other sample communities from the E. coli network, follow a power law distribution. It should be noted that the plots have a logarithmic axis and a straight line thus indicates a power-law scaling.

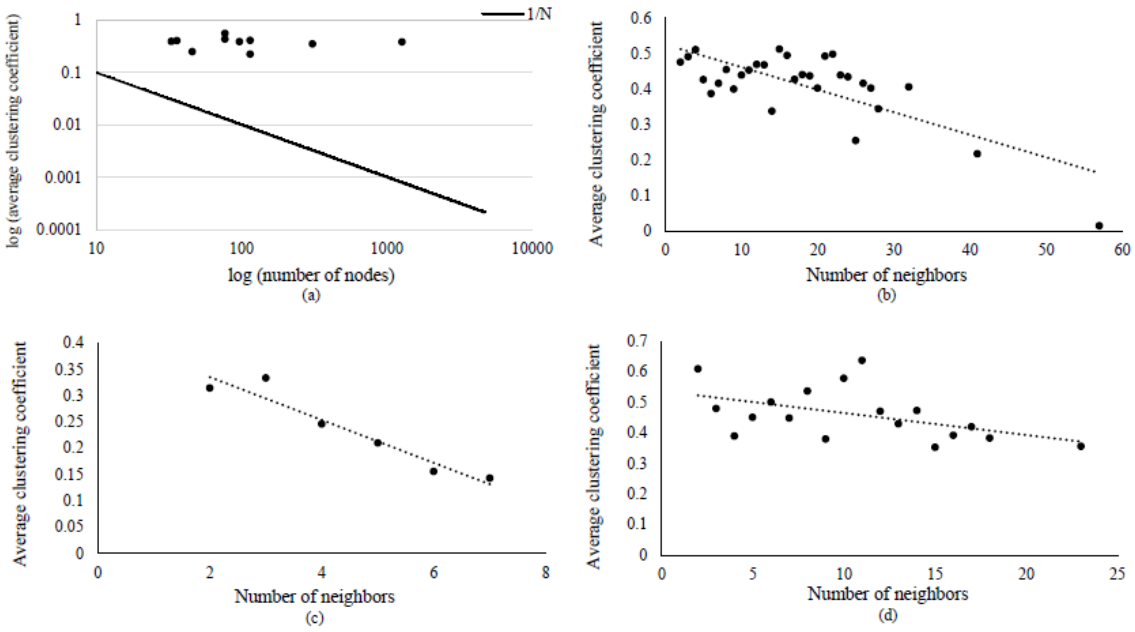


Figure 4.5: Average clustering coefficient distribution. (a) shows the distribution of the average clustering coefficient of the communities from the E. coli network (black dots). The solid line corresponds to the predicted trend for random networks, with the average clustering coefficient decreasing as N-1. The average clustering coefficients shown are independent of N. (b)-(d) plots the clustering coefficient distribution for the E. coli network and two other sample communities. We observe a decreasing trend of clustering coefficient distribution with the node degree k, across communities.

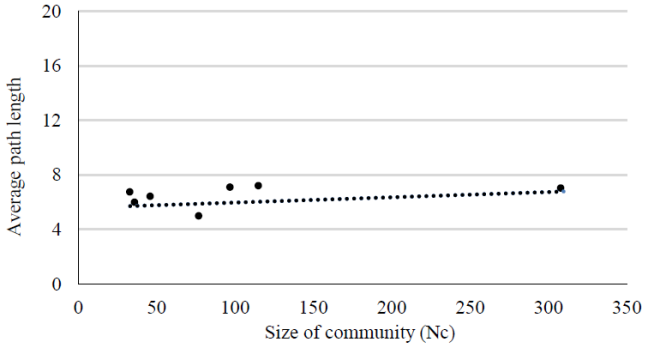


Figure 4.6: Distribution of average distances. The slow logarithmic increase of average distance of communities with the size of the community, is even less pronounced for larger communities (nc>10). The black circles represent the average distances of E. coli network communities of interest, and the trend line is consistent with the aforementioned behavior.

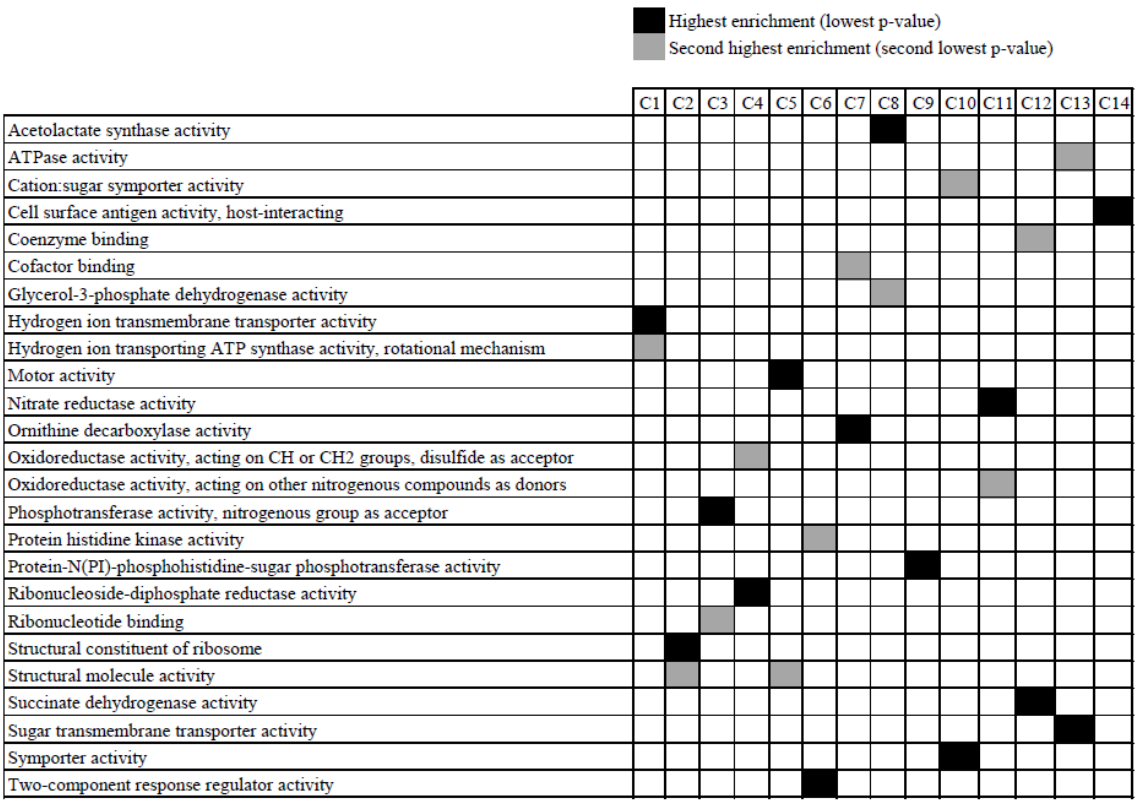


Figure 4.7: "Heat-map" representation of functional enrichment. The "heat-map" is a representation of the relative enrichment of each functional annotation in the communities from the E. coli PPIN. The columns of the heat-map represent the 14 communities (C1-C14) that have a vertex cardinality (Vc) of at least 20 genes. The rows represent the union of the highest and second highest enrichments (in terms of the functional annotations) across C1-C14. The highest and the second highest enrichments are color-coded black and gray respectively.

References

1. M.E.J Newman, M. Girvan. (2004). Finding and evaluating community structure in networks. *Phys Rev E*. 69:026113.

2. G. Bader, C Hogue. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinforma*, 4(2).

3. R. Dunn, F. Dudbridge, C. Sanderson. (2005). The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinforma*. 6:39.

4. A. Rives, T. Galitski. (2003). Modular organization of cellular networks. *PNAS. 100,* pp. 1128–1133.

5. R. Sharan, T. Ideker, B. Kelley, R. Shamir, R.M. Karp. (2005). Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *J Comput Biol. 12(6),* pp. 835-46.

6. M. Girvan, M. Newman. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. 99,* pp. 7821–7826.

7. J. Ruan, W. Zhang. "An efficient spectral algorithm for network community discovery and its applications to biological and social networks". in *Proc. of ICDM,* Nebraska, 2007.

8. T. Narayanan, M. Gersten, S. Subramaniam, A. Grama. (2011). Modularity detection in protein-protein interaction networks. *BMC Research Notes.* 4-569.

9. T. Narayanan, and S. Subramaniam, "Community Detection in Biological Networks Using a Variational Bayes Approach," presented at the 3rd Int. Conf. International Conference on Bioinformatics and Computational Biology, New Orleans, Louisiana, USA, 23 – 25 March 2011.

10. F. Picard, V. Miele, JJ. Daudin, L. Cottret, S. Robin. (2009). Deciphering the connectivity structure of biological networks using MixNet. *BMC Bioinformatics.*

11. K. Steinhaeuser, N.V. Chawla. (2008). Community detection in a large real-world social network. *Social Computing, Behavioral Modeling, and Prediction. Springe.* pp. 168–175.

12. J. Leskovec, K. Lang, A. Dasgupta, M. Mahoney, "Statistical properties of community structure in large social and information networks", in *proc. WWW,* Beijing, 2008.

13. S. Wasserman, K. Faust. (1994). Social Network Analysis: Methods and Applications. *Cambridge University, in press.*

14. J. Leskovec, K. Lang, M.W. Mahoney, "Empirical Comparison of Algorithms for Network Community Detection" in *proc. WWW,* 2010, Raleigh, USA

15. J. Chen, O. R. Zaiane, R. Goebel, "Detecting communities in social networks using max-min modularity", in *proc. SIAM DMC,* 2009, Nevada, USA.

16. M. Newman, M. Girvan (2004). Finding and Evaluating Community Structure in Networks. *Physical Review E*.

17. K. Eriksen, I. Simonsen, S. Maslov, K. Sneppen. (2003). Modularity and extreme edges of the Internet. *Phys. Rev*.

18. T. N. Dinh, Y. Xuan, M. T. Thai, "Towards social-aware routing in dynamic communication networks", in *proc. IPCCC*, 2009, Phoenix, USA.

19. N.P. Nguyen, T. N. Dinh, Y. Xuan, M.T Thai, "Adaptive algorithms for detecting community structure in dynamic social networks" in *proc. IEEE ICCC*, 2011, Shangai.

20. T. Narayanan and S. Subramaniam, "Community Structure Analysis of Gene Interaction Networks in Duchenne Muscular Dystrophy", to be published.

21. L. Salwinski, C.S. Miller, A.J. Smith, F.K. Pettit, J.U. Bowie, D. Eisenberg. (2004). The Database of Interacting Proteins. *NAR 32(D)*, pp. 449-51

22. U. Alon, "The feed-forward loop network motif," in Introduction to Systems Biology: And the Design Principles of Biological Circuits, CRC Press, 2007

23. Johnson, B. Donald. (1977). Efficient algorithms for shortest paths in sparse networks. *Journal of the ACM 24 (1)*, pp. 1–13.

24. C. Stark, B. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, et al. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Research, 34: D535*.

25. J.M. Peregrín-Alvarez, X. Xiong, C. Su, J. Parkinson. (2009). The Modular Organization of Protein Interactions in Escherichia coli. *PLoS Comput Biol, 5(10).*

26. J. Duch, A. Arenas (2005), Community identification using external optimization, *Phys Rev E Stat Nonlin Soft Matter Phys, 72*

27. G.K. Orman, V. Labatut, C. Hocine. (2012). Comparative Evaluation of Community Detection Algorithms: A Topological Approach. *Journal of Statistical Mechanics: Theory and Experiment, P08001*.

28. A.L. Barabási. (2012, November). *Network Science.* [Online]. Available: http://barabasilab.neu.edu/networksciencebook/downlPDF.html

29. E. Ravasz et al. (2002). Hierarchical organization of modularity in metabolic networks. *Science 297*, pp. 1551-1555

30. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium (2000) *Nature Genet. 25:* pp. 25-29

31. D.W. Huang, B.T. Sherman, R.A. Lempicki. (2009). Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc. 4(1), pp.* 44-57.

32. D.W. Huang, B.T. Sherman, R.A. Lempicki. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res. 37(1),* pp. 1-13.

33. Z. Wang, J. Zhang. (2007). In search of the biological significance of modular structures in protein networks. *PLoS Comput. Biol.*

34. D. Binns, E. Dimmer, R. Huntley, D. Barrell, C. O'Donovan, R. Apweiler. (2009). QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics, 25(22).* pp. 3045-6.

35. M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, M. Tanabe. (2012). KEGG for integration and interpretation of large-scale molecular datasets. *Nucleic Acids Res, 40.* pp. D109-D114.

36. M. Kanehisa, S. Goto. (2002). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res, 28*. pp. 27-30.

Conclusions

In this dissertation, community detection has been explored as an effective computational tool to analyze and understand biological networks, and gain valuable insights about the networks from a structural and functional standpoint. A variety of biological datasets have been used as case studies to illustrate the applicability of community detection in the domain of biological network analysis.

In Chapter I, a novel optimization for a widely used community detection algorithm is proposed. This is empirically evaluated in the context of biological networks, illustrating significant savings in computational time, while maintaining comparable quality of detected communities. The applicability of a Bayesian inference approach to the problem of community detection in the context of biological datasets has been studied in Chapter II. The results from this study demonstrate that a machine-learning based approach lends itself well for analyzing biological networks and produces communities that correlate strongly with the underlying structure and function of such networks.

Chapter III focuses on leveraging community detection for the study of pathology, with a specific emphasis on DMD. A novel approach of transforming gene expression datasets into interaction networks is proposed. A detailed pathway analysis of communities from these networks illustrates the underlying differences between structural organization of the networks from normal muscle and pathology. An innovative algorithm for community detection is presented in Chapter IV, which is motivated by a physical system-modeling of a network. PPINs of model organisms such as E. coli are used as a case study to illustrate that the proposed algorithm is an effective tool for studying biological networks. Not only do the communities produced by the algorithm

have scale-free properties, but also exhibit biological relevance in terms of pathway specificity and functional enrichments.

The algorithms considered in this dissertation take different approaches to the problem of community detection. The NG algorithm is a widely used graph theoretic (divisive) approach to community detection, and is directly applicable only to undirected, unweighted and static networks. It is also a computationally intensive algorithm and so may not be easily applicable to very large datasets. However, the distribution of the vertex cardinality of communities that the algorithm yields is varied and so, the NG algorithm can be used as a tool in applications where a more granular functional representation of constituent genes of communities is desired (i.e. in the context of biological networks). The Variational Bayes approach to community detection is a statistical approach which is more efficient while dealing with large datasets. However, the communities that it yields (for the Yeast dataset considered in this dissertation) are more cohesive – and so, VB can be used as a technique for detecting communities where a more coarse grained division of network community functions are desired. The Newtonian approach proposed in the dissertation is also a divisive approach to community detection with the advantage of not only being directly extensible to weighted networks, but also providing a variety of model parameters which can be tuned by the user, depending on the domain of application / dataset under consideration. Thus, the NCD algorithm has the potential to be leveraged as an effective technique for community detection in a variety of domains of application.

The techniques and algorithms developed as part of this dissertation in the context of biological networks, serve as a solid foundation for future exploration in the following

areas. Specifically, the results from the study of DMD from a community detection standpoint provide the mechanistic basis for further biological investigations into specific pathways differently regulated between normal and DMD patients. Furthermore, this study reinforces the suitability of community detection as an effective tool for disease modeling with potential diversification to analyze different forms of pathology. The NCD algorithm proposed in this dissertation provides a framework for effective analysis of networks from a community structure perspective and generalizes to encompass weighted networks. Additionally, in the context of biological networks, the algorithm could serve as a key computational tool for driving therapeutic applications involving targeted drug development for personalized care delivery.

In addition, the methodologies and associated design principles presented in this dissertation have the potential to be extended for handling time-varying networks, which often arise in biological contexts and are the subject of active research (such as cancer pathology studies). Specifically, community detection could provide key insights into the temporal evolution of the underlying structure of such biological networks. Another prospective domain of application of these computational techniques is the study of neuronal networks. Connections between cortical areas of the primate brain have demonstrated small-world network properties, suggesting that community detection could aid in the identification of cortical hubs that correspond to key neural functions.

The applications of community detection are not limited to the biological domain. Some non-biological applications of community detection include social network modeling, search engine recommendations, and electrical / electronic circuit design and management. The applications of community detection in social network modeling are

multifold. Community detection can be used as a technique for not only defining ‘friend circles’, but also as a tool for making friend recommendations, based on the number of messages exchanged among a group of friends (which can be modeled as edges) and a new friend (node). Secondly, in the domain of search engine recommendations, personalized search results are desired (results based on past searches, pages visited and inferred interests and geographic location of the user). Weighted networks based on these parameters can be constructed. Communities detected from such networks can be studied and used as inputs to rank the pages that the search engine extracts from the key words in the search. In the domain of electrical / electronic circuit design and management, community detection offers a variety of crucial applications. With the exponential growth of technology and industries, the proportional increase in demand for power resources is inevitable. On the other hand, with the development of optical communication systems and nano-chip technology, the need for optimal small scale circuit design with minimum power loss is also sought after. These systems can be modeled as networks, with the electrical / electronic components as nodes and connection between them as edges. Specifically in the electrical systems, the arterial connections of the networks can be determined using network analysis techniques. In the event of power surges from production plants, these important connections could be severed first to protect hubs of homes and industries that are connected to the plant. Similarly, in electronic circuits where power losses have to be minimum, the analysis of electronic circuits modeled as networks could provide insights into the holistic functionality of different groups of electronic components. Network and shortest path analyses could provide insights into optimal placement of circuit components, resulting in minimum power loss.