Feature ranking in E. coli and Yeast derived networks using the packet receival rate as robustness metric

Bhanu K. Kamapantula*
Virginia Commonwealth University
401 W Main st
Richmond, VA, USA
kamapantulbk@vcu.edu

Michael Mayo
Environmental Laboratory, US Army Engineer Research and Development Center
Vicksburg, MS 39180
Michael.L.Mayo@usace.army.mil

Edward Perkins
Environmental Laboratory, US Army Engineer Research and Development Center
Vicksburg, MS 39180
Edward.J.Perkins@usace.army.mil

Ahmed Abdelzaher†
Virginia Commonwealth University
401 W Main st
Richmond, VA, USA

Preetam Ghosh
Virginia Commonwealth University
401 W Main St
Richmond, VA, USA
pghosh@vcu.edu

ABSTRACT

Biological networks are now studied extensively by researchers to understand the working principles of nature. Machine learning techniques can be useful in realizing the features contributing to the robustness of biological systems. In this work, we compare subnetworks extracted from *E. coli* and yeast using *in silico* experiments. We use packet receival rate as a metric to quantify biological robustness. This metric is different from the usual structural metrics since it captures the dynamic behavior of the network. We define seventeen features based on structural significance such as motifs and conventional metrics such as average shortest path, network density among others. Then, in order to identify important features, feature ranking is performed based on grid search for best support vector machine classifier parameters using cross validation. Results show that feed-forward loop motif based features are important for *E. coli* networks. Network density, degree centrality based features and bifan motif based features are identified as significant for yeast derived networks. Also, results suggest that feature significance varies with network size (number of nodes). As a first, this study quantifies the impact of motif, feed-forward loop and bifan, abundance in natural networks.

General Terms

Machine learning, feature ranking

*Corresponding author - kamapantulbk@vcu.edu
†abdelzaheraf@vcu.edu

Keywords

biological robustness, e. coli, yeast

1. INTRODUCTION

We investigate the genetic regulatory network (GRN) of *E. coli* and yeast to understand the fundamental principles of robustness. The genetic networks of these organisms are known to be robust in *function* despite disruptions like gene failures and signal transmission. This behavior is attributed to their power-law degree distribution and the abundance of motifs. Frequently occurring substructures in a network are termed motifs. The frequency of a motif in a real network is determined by comparing its frequency in a random network. Following this, feed-forward loop (FFL) motifs are determined to be responsible for signal pulse generation and changing response times [12]. Additionally, these motifs are identified to contribute to robust *functioning* of genetic networks [9]. In this study on significant network features, we explore the performance of motif-derived features.

Researchers are also studying these organisms in a biological context. [17] traced the growth of two *E. coli* strands and addressed their stability over hundreds of generations of mother cell division. The study claims robust *E. coli* growth mechanism. Little is known about the motif formation or the regulatory interactions at a granular level. Our effort is to identify important features responsible for functional biological systems. These features can be structurally evident when studied or not evident clearly. Structurally significant features can be explored by mapping a biological system to a network graph problem. Here, we map the gene-gene and transcription factor-gene interactions in a GRN to a network of nodes and edges where genes and transcription factors are represented by nodes and edges represent the interactions between participating nodes. Once the system is mapped as a network, the principles of graph theory [1] allows us to study the GRN characteristics. Control theory is now used to change the properties of critical entities (nodes or communities) of a network to control the entire network

[11]. While it is intuitive to control high degree nodes, the study proves otherwise. In order to exploit this and create adaptive network topologies, we study the characteristics of subnetworks derived from the regulatory networks of *E. coli* and yeast. This work is built on our previous works on establishing robust bio-inspired wireless sensor network topologies [6, 7, 4] and quantifying robustness using NS-2 [8]. This paper is organized as described below.

Section 2 discusses the literature describing robustness in different network contexts. The methodology followed in this work is detailed in Section 3 including the extraction of networks, simulation setup and results. Section 4 describes the SVM model and feature ranking in the networks used. Section 5 describes several pointers to future work and challenges in comparison different biological systems.

2. BIOLOGICAL ROBUSTNESS

2.1 Definitions of robustness

Multiple definitions for *robustness* have been proposed by researchers in the last decade. Robustness is traditionally defined as the ability of a network to withstand disruptions and perform the tasks as intended. Kitano termed robustness to be "a property that allows a system to maintain its functions against internal and external perturbations" [Kitano 2004]. The idea of robustness varies from one complex system to another. For example, a robust biological system implies a functional system despite interruptions. In a social network context, traditional definition of robust functional network might not be applicable since information transmission is not based on physical signal pulses. In a financial network context where flow of debt is central to understanding functional economy, robustness depends on critical players such as banks, stronger economic countries and regulatory bodies. In the realm of complex networks and control theory, robustness of controllability is defined as the ability to control the network using a set of nodes [11]. Given this, several metrics have been proposed to quantify robustness in complex networks. Some of these metrics are centrality-based (degree, betweenness, closeness and eigenvector centralities) and path-based (average shortest path, communicability). The *strength* of a network has also been measured based on the strong and weak connected components. A recent study [2] which explored the idea of robustness in complex networks using a metric derived from *Estrada index* [3]. However, this is applicable to undirected networks. Further, this is a structural metric and hence cannot capture the dynamic behavior of a system. Since [2] reviews different metrics proposed earlier to explore robustness, we suggest interested readers to read that work. It can be noticed that all the metrics mentioned such as estrada index, variation in diameter, algebraic connectivity among others are static in nature. That is, the stated metrics measured are purely structural in nature and information flow in real-time has not been considered. Further, none of the measures consider and measure the impact of motif abundance which is suggested as one of the reasons for functionally robust biological systems. Our approach uses packet receival rate [1] as a measure for network robustness. Also, we use six features that are related to feed-forward loop and bifan motifs in the SVM classification model to measure their relative significance to network robustness.

[1]It is the ratio of the number of packets received at sinks to the number of packets sent by the source nodes

2.2 Packet receival rate as robustness metric

In order to quantify robustness for a given system, we use the network simulator NS-2 as a simulation framework. A biological system here is represented as a network of nodes and edges. To illustrate this scenario, we consider a transcription regulatory network as an example where transcription factors and genes are nodes and the interactions among them is represented by edges. Packets are transmitted from source nodes (transcription factors) and are received at sink nodes (genes). Gene nodes can only receive packets but TF nodes can send and forward packets. All nodes with zero out-degree are considered to be sink nodes. We quantify robustness as a ratio of number of packets received at sinks to the number of packets sent which essentially is the packet receival rate. It should be observed that this metric is dynamic in nature and is different from other topological metrics. The dynamic aspect of a biological system can be captured using packet receival rate by varying the parameters of the simulation such as network loss, packet transmission rates and queue limits.

2.3 Contribution

Our contribution is this work is identifying important features using cross validation and building a support vector machine (SVM) classification model for predicting performance of *E. coli* and yeast derived networks. The following are the contributions of this paper:

1. As a first, a comparison of networks derived from E. coli and yeast is presented. Packet receival rate is used as a robustness metric to measure the performance of the networks.

2. This work identifies features related to FFL and bifan motifs as significant for *E. coli* and yeast derived networks .

3. Another key finding of this work is that the significance of features for any specific type of GRN varies with network size.

3. METHODOLOGY

Figure 1 illustrates the procedure followed in this work. Section 3.1 describes the network extraction– as illustrated in Figure 1 (Step 1)– from E. coli and Yeast model networks using GeneNetWeaver software [15]. Section 3.2 details the simulation setup, determination of robustness –as illustrated in Figure 1 (Step 2)– using NS-2 software. Label determination using k-means clustering algorithm and feature calculation is discussed in Sections 4.1 and 4.2.

3.1 E. coli and Yeast derived networks

Hundred networks for each size of (number of nodes) 100, 200, 300, 400 and 500 are extracted from *E. coli* and yeast using GeneNetWeaver. The extracted networks are gene regulatory networks (GRN). The networks derived from *E. coli* and yeast are hereby termed as *E. coli* networks and yeast networks respectively for the rest of the paper. In this study, we compare networks for a given size, for example 200 nodes, even though they might vary in number

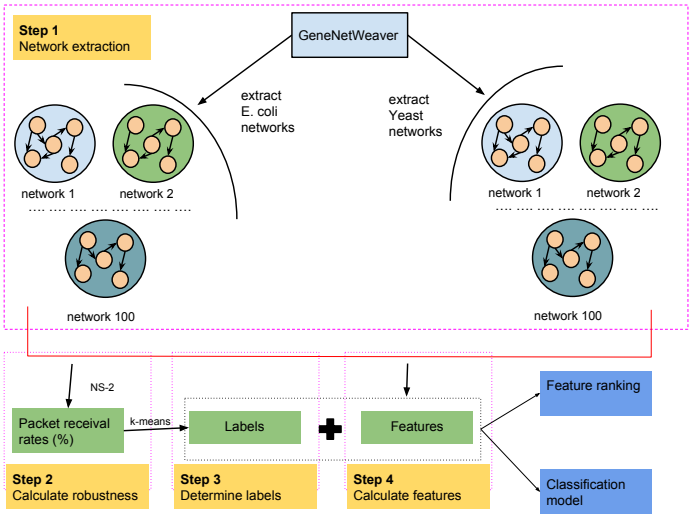


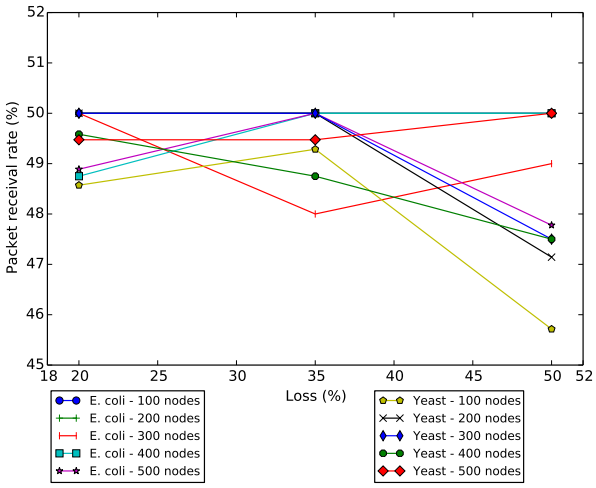
Figure 1: Illustration of the procedure followed in this work. **Step 1** - Extract subnetwrks of sizes 100, 200, 300, 400 and 500. For each network size, 100 network instances are extracted. This is done for both E. coli and Yeast. **Step 2** - Robustness is measured for all the networks. Robustness values can be in the range 0.0−100.0. **Step 3** - k-means clustering algorithm is used to determine integer labels. **Step 4** - Features listed in Section 4.2. Using the labels and features calculated in Steps 3 and 4, feature ranking is performed and a classification model is created.

of edges. This is due to the challenge in extracting networks of equal number of nodes and edges for both *E. coli* and yeast. It is suggested that most edges in real networks are regular [11] in the context of controllability of networks. For two subnetworks A and B extracted from *E. coli* and yeast networks, we hypothesize that even though subnetwork A differs from subnetwork B in terms of the number of edges, there potentially exists some topological feature(s) that makes subnetwork A robust than subnetwork B [2]. In order to compare networks derived from *E. coli* and yeast, we run *in-silico* experiments on each network and identify the best, average and worst performing (in terms of packet receival rates) networks. The respective comparison for the best, average and worst performing networks for sizes 100, 200, 300, 400 and 500 is presented in Figures 2a, 2b, and 2c.

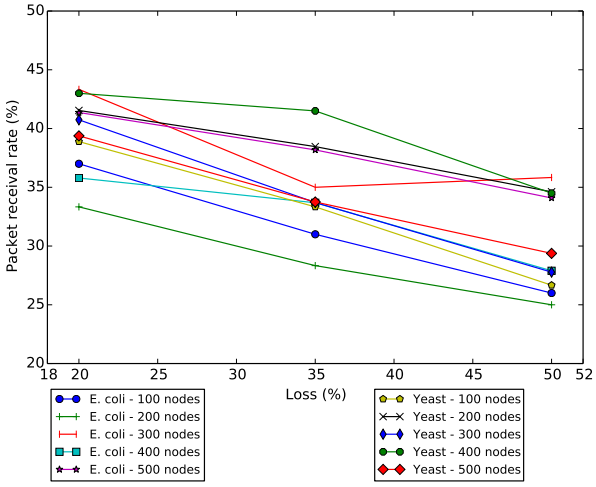
In order to identify an ideal GRN (within *E. coli* and yeast) to create robust systems, we compare the performance of respective network sizes. To this effect, the trapezoidal rule is used to measure the area under the curve [3]. Out of fifteen instances (five different network sizes and three different loss models) eleven *E. coli derived* networks performed better than their counterparts. In three cases (200, 400, 500 network sizes - best performing networks), Yeast derived networks performed better than their counterparts. In one case (100 network size - best performing networks), both performed the same. In order to identify significant proper-

[2]Networks can be manipulated by deleting (adding) existing (new) interactions. However, modifying these networks will damage the inherent structural properties.

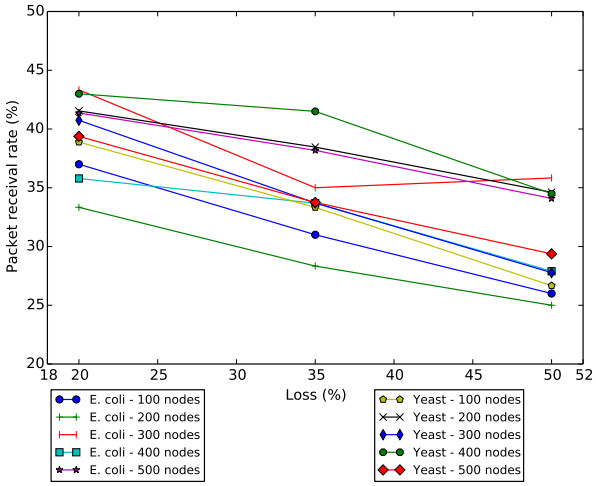
[3]It should be noted that area under the curve is not plotted in any of the figures. Here, our focus is to observe the trend of performance for different loss models.



(a) Comparison of *best* performing E. coli and Yeast derived networks of sizes 100, 200, 300, 400 and 500



(b) Comparison of *average* performing E. coli and Yeast derived networks of sizes 100, 200, 300, 400 and 500



(c) Comparison of *minimum* performing E. coli and Yeast derived networks of sizes 100, 200, 300, 400 and 500

Figure 2: Comparison of best, average and minimum performing E. coli and Yeast derived networks respectively for sizes 100, 200, 300, 400, and 500

ties responsible for this behavior, we use machine learning techniques and perform feature ranking which is described in Section 4.

3.2 NS2 simulation setup

This work uses our previous approach to quantify network robustness using packet receival rate in [6, 7, 8]. Network robustness is measured across three different loss models: 20%, 35% and 50%. Queue limit at a node is set at five (packets). All edges are considered to be directed. Nodes with zero out-degree are considered to be sinks and other nodes are considered to be source nodes. While sink nodes only receive packets, source nodes transmit and forward the packets. This scenario resembles a biological system where transcription factor(s) regulate gene(s). Packet receival rate is calculated as the ratio of number of packets received at all sinks to the number of packets generated at source nodes. We represent robustness of a network as a percentage *(packet receival rate)*$*$100. Higher robustness percentage suggests a *more* robust network compared to a network with lower robustness value.

3.3 Results

Our objective is to compare *E. coli* and yeast networks and identify reasons for good or bad performance. As stated in 3.1, *E. coli* networks performed better than the yeast counterparts in 11 out of 15 instances. We use support vector machine (SVM) modeling to identify features responsible for better performance. We measured a set of features identified as important by earlier literature and created additional features for a holistic picture. The next section describes the features used and the methodology used to identify significant ones among them.

4. SVM MODELING

Machine learning is now widely used by businesses to identify email spam, predict airline prices on a busy weekend, predict football game outcomes, predict national election outcomes, credit card fraud detection among a slew of other applications. Researchers recently built a feature detector to identify Human bodies and cat faces using unsupervised learning techniques [10]. Deep learning methods are in use to perform speech recognition tasks. As more data regarding cellular interactions within a GRN is available, such techniques can be employed to simulate regulation expression prediction models. On a broad perspective, data with labels require supervised learning techniques and data without labels require unsupervised methods. We take advantage of support vector machines using supervised learning method to identify significant features of *E. coli* and yeast networks. We tested our dataset for support vector machine (SVM) regression and classification models. SVM regression performed below par compared to SVM classification technique. For brevity, discussion on our regression implementation is avoided and will be addressed in a different context.

We follow the data preprocessing and model selection style defined by [5]. The features are scaled to the range $[-1, 1]$ in order to avoid unfair advantage to high valued features. This is applied to all data.

4.1 Label mapping using k-means algorithm

Robustness values of networks fall in the range of $0.0 - 100.0$. Regression technique is applied to continuous data. In order to build a SVM classification model, discrete labels are required. Hence, we map the robustness labels (in floating point) to integer values. This is performed using k-means clustering algorithm instead of arbitrary allocation. For a given set of n points, k-means algorithm partitions the points into k clusters. Initially, the points are clustered with a random center for each cluster. Then, the distance of each point to all the cluster centers is estimated and the point is reassigned to the cluster center nearest to it. This process is continued until the centers no longer change. For this work, we grouped the data into five clusters.

4.2 Features

To build a machine learning model, features are necessary. Features intuitively describe the properties of a network. Feature extraction is a critical aspect before choosing the learning model. We define certain features based on established research. It has been shown that average shortest path is crucial to the stability of the network. Network density captures the sparsity of nodes in the network. We also measure centrality metrics such as degree centrality, betweenness centrality and closeness centrality as they identify nodes that work as hub nodes for information flow in a network. Since packet receival rate depends on the paths from source nodes to sink nodes, we define features such as . Feed-forward loop (FFL) motif, which is a type of three-node motifs, has been identified to contribute to robustness–preserving system function despite internal and external perturbances–in genetic networks [9]. FFLs are also have been shown to be important for biological functions such as generating signal pulses, and speeding up or delaying response times [12]. Hence, three FFL-based metrics are defined as features. FFL motifs which despite being responsible for several biological functions are found to be less stable than bifan motif, which is a type of four-node motif [14]. Hence, three bifan-based metrics are defined as features.

A total of seventeen features are considered to build the SVM classification model. Consider a network of nodes and edges represented by $G(V, E)$ where G is graph and V is the set of vertices and E is the set of edges. We define each feature before we identifying the significant ones.

4.2.1 Network density

Network density is the amount of edges present in the network compared to the total number of edges possible in the network.

4.2.2 Average shortest path

Average shortest path of the network is the ratio of the sum of shortest paths for all pairs of nodes to the total number of possible edges.

4.2.3 Genes percentage

Genes percentage is the percentage of gene nodes with respect to the total nodes in the network.

4.2.4 Transcription factors percentage

TFs percentage measures the number of transcription factor nodes compared to the total number of nodes in the network.

4.2.5 Transcription factor network density (TFND)

TFND determines the percentage of total edges connected to transcription factor nodes.

$$TFND = \frac{|E_{TF}|}{|V|(|V|-1)} \quad (1)$$

where E_{TF} is the number of edges that are connected to transcription factor nodes.

4.2.6 Genes coverage

Genes coverage (GC) is the ratio of in-degree of all sink nodes to the sum of the number of source nodes with paths to the sink nodes [4].

4.2.7 Centrality measures

a) Degree centrality

The average degree centrality of gene nodes (ADCG) and transcription factor nodes (ADTF) are considered separately. Simultaneously, average degree centrality of the network (ADC) is considered.

b) Betweenness centrality

The average betweenness centrality of transcription factor nodes (ABTF) is considered. BC of a node is the number of shortest paths the node participates in compared to the total number of shortest paths.

c) Closeness centrality

Here, the average closeness centrality of transcription factor nodes (ACCTF) is considered. Closeness centrality measures the distance of each node to all other nodes. A normalized version of the closeness centrality is used as a feature.

It can be observed that betweenness centrality and closeness centrality for gene nodes is not considered as they do not participate as intermediate nodes in shortest paths since their out-degree is zero. Eigenvector centrality metric is not considered since the convergence using power method is not possible for all the networks.

4.2.8 FFL edge abundance

FFL edge abundance (FFLD) measures the number of edges participating in FFLs compared to the total number of edges in the network.

4.2.9 Bifan edge abundance

Bifan edge abundance (BFD) measures the number of edges participating in bifans compared to the total number of edges in the network.

4.2.10 FFLSSP

We determine the total edges such that each edge participates in an FFL and also goes from a transcription factor node (source) to a gene node (sink). Two features are derived from this metric: a) the count determined is compared to the total number of edges in the network (FFLSSP) and b) the count determined is compared to the total number of direct edges from transcription factor nodes to gene nodes (FFLSSPD). Figure 3 illustrates this scenario.

4.2.11 BifanSSP

We determine the total edges such that each edge participates in an bifan and also goes from a transcription factor (source) node to a gene node (sink). Two features are derived from this metric: a) the count determined is compared to the total number of edges in the network (BFSSP) and b) the count determined is compared to the total number of direct edges from transcription factor nodes to gene nodes (BFSSPD). Figure 3 can be extended to bifan motif.

[4]Definitions of other features have been removed for space considerations

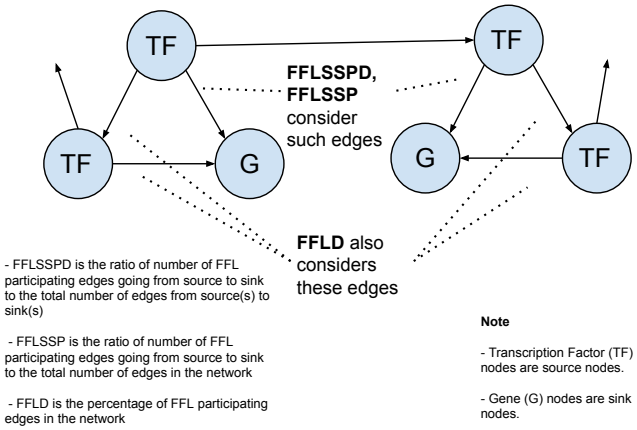


Figure 3: Demonstration of FFLSSP, FFLD and FFLSSPD features

All features are scaled from -1 to 1 based on the Equation 2.

$$F_{js} = (\frac{F_j - F_{min}}{F_{max} - F_{min}}) * 2 - 1 \quad (2)$$

where F is the set of features, F_{js} is the scaled jth feature, F_j is the jth feature, F_{max} and F_{min} are maximum and minimum values in the F.

4.3 Implementation and feature ranking

Python programming language [16] is used to implement the feature ranking and SVM classification model. We used *scikit-learn* [13] package developed using Python for SVM classification feature ranking and building a classification model. *scikit-learn* uses the popular *libsvm* and *liblinear* packages internally.

As recommended in [5], we parameters best classifier parameters are determined using grid search with ten-fold cross validation. Cross validation is performed to avoid overfitting the data. We considered linear, RBF and polynomial kernels as kernel options [5]. Initially, the classifier was modeled for ten-fold cross validation. In some instances, number of labels for a class was found to be less than the number of folds (ten). Hence, five-fold cross validation is performed. Best parameters are identified by taking the mean of accuracy across five-fold cross validation. For this study, training data is set at 85% of the entire data and remaining data is for testing purposes. Hence, 85% of the data is used for training purposes [6]. Here, the training set is divided into ten

[5]Due to limited space the parameters are described here. 1, 10, 100, 1000 are used as C values for Linear, RBF, Polynomial kernels. The set of values 0.0001, 0.001, 0.01, 0.1, 1 and 2 are used as γ for RBF kernel. A γ value of 1 is used for Polynomial kernel. 1, 2, 3, 4, 5 are used as *degree* values (applicable only to Polynomial kernel).

[6]Data was also modeled by using 75%:15% data ratio for

Table 1: Best grid search parameters using cross validation - *E. coli*, yeast-derived networks

Network size(s)	**Kernel**	**C**	**Gamma (γ)**	**degree**
yeast - 100, 300	Polynomial	1, 1	1, 1	3, 3
yeast - 200, 400, 500	RBF	10, 1, 10	1, 1, 2	-
E. coli - 100, 200, 400, 500	RBF	100, 10, 1, 100	0.1, 2, 1, 0.1	-
E. coli - 300	Polynomial	1	1	4

sub-datasets of equal size and each sub-dataset is tested using the classifier trained on the remaining nine sub-datasets. This is done for each C & γ pair. Once the best parameters (defined in Tables 1 for both yeast and *E. coli* networks) are identified, feature ranking and classification model building is performed. An SVM classification model is built for future purposes to predict the performance of extracted subnetworks. Accuracy score[7] is used to identify the accuracy of the classifier.

Features are ranked using analysis of variance (ANOVA) F-value metric. ANOVA F-value calculates the ratio of interclass variance to within-class variance. This metric is used from scikit-learn [13]. A higher F-value denotes higher significance of a feature. In this work, we filter top five features out of the defined seventeen features. While F-value calculates the feature significance individually, mutual feature dependence cannot be calculated by this metric. Significance of features can be estimated by considering the mutual impact of features on one another. We intend to address this in the future.

4.4 Feature significance

For each network size and specific network type, an SVM model is created and corresponding features are ranked. Considering yeast networks first, eight important features are plotted in Figure 4. Top five features are ranked for each network size. The superset of all features for five different network sizes is then selected for comparison. Further, the minimum ranked feature for each network size is identified and plotted as a threshold. This will help us observe the trend of each feature across all network sizes. For example, at network size of 200 features with higher F-value than minimum curve are more significant than the ones with lower F-value than minimum curve. Similarly, this is repeated for *E. coli* networks. For *E. coli* networks, from Figure 5 it can be observed that the features related to FFL motif (FFLSSPD, FFLD, FFLSSP) score consistently higher than the minimum value of top five features in all cases except one (network size - 100). Bifan motif-based features (BFSSPD, BFD, BFSSP) seem to score less than the minimum curve in majority of cases.

training and testing. No significant difference was noticed in estimating the features. A detailed study will be carried out elsewhere.

[7]It is the ratio of number of true predicted values to the true values.

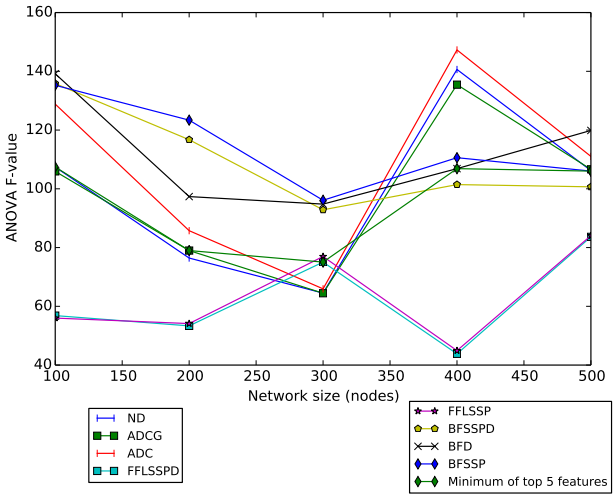


Figure 4: Comparison of 8 features across yeast-derived networks of sizes 100, 200, 300, 400 and 500.

4.4.1 *Weighted average features*

To understand the relative importance of features, weighted average of all the features is determined and illustrated in Figures 6a, 6b, 6c, 6d, 7 and 8. This is done as follows. **Identify features**: a list of features that ranked in the top five (from ANOVA Fvalue test described in Section 4.3) is created. Top features determined in both types (*E. coli* and yeast) of networks and all network sizes are considered. **Calculate weighted average**: for a given type of network and for particular size, feature weights are averaged across the all features for each feature.

4.4.2 *Feature comparison in* E. coli *and yeast networks*

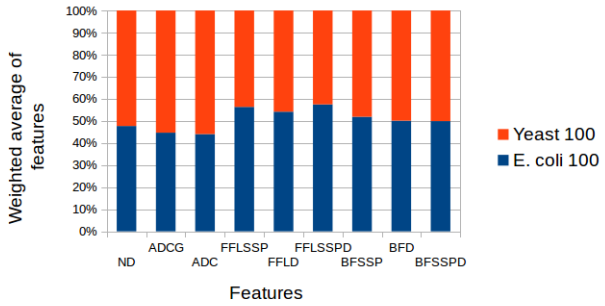
FFLSSPD, FFLSSP and FFLD rank higher for E. coli networks than yeast networks. In biological context, this information is crucial. Since FFLSSPD and FFLSSP consider both the number of edges in FFLs and direct edges from sources to sinks, the essence of network robustness is captured in these metrics.

4.4.3 *Feature trend*

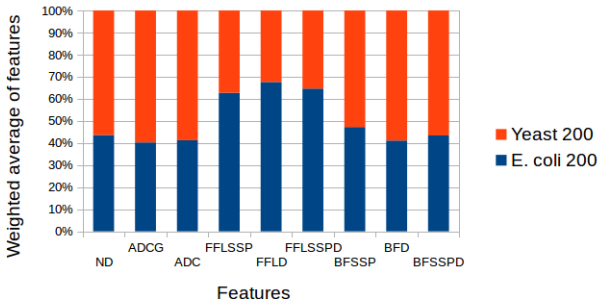
Figure 8 is used to observe the trend of individual features. ND, ADCG, ADC are, in all cases, ranked relatively higher in yeast networks than in *E. coli* networks. Similar trend can be realized for BFSSP, BFD, BFSSPD features as well. The trend reverses for FFLSSP, FFLD and FFLSSPD where they are ranked higher in *E. coli* networks compared to its counterpart. This study will help design flexible learning classifiers where features can be adaptively used as plug-ins depending on the network type. Since, bifan-based features work better for yeast networks, bifan interaction within these networks can offer insights for adaptive information transmission in a network.

5. DISCUSSION AND FUTURE WORK

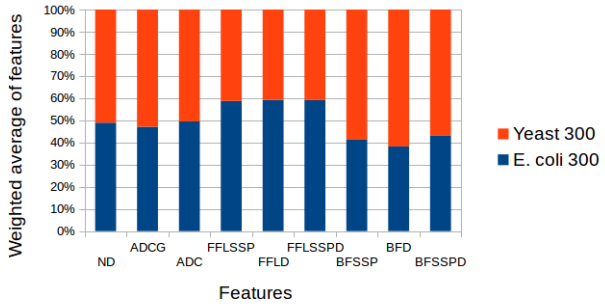
To compare the relative efficiency of *E. coli* and yeast networks, we use quantitative methods to simulate packet transmission in the subnetworks derived from their regula-



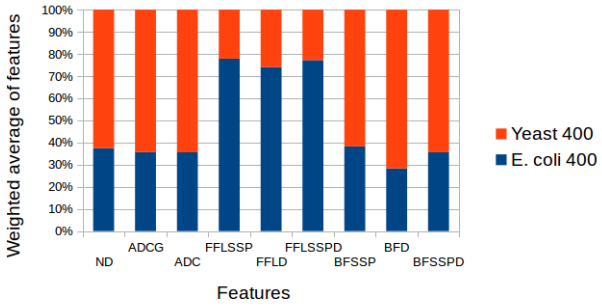
(a) Size 100



(b) Size 200



(c) Size 300



(d) Size 400

Figure 6: Feature comparison E. coli and Yeast derived networks for sizes 100, 200, 300 and 400.

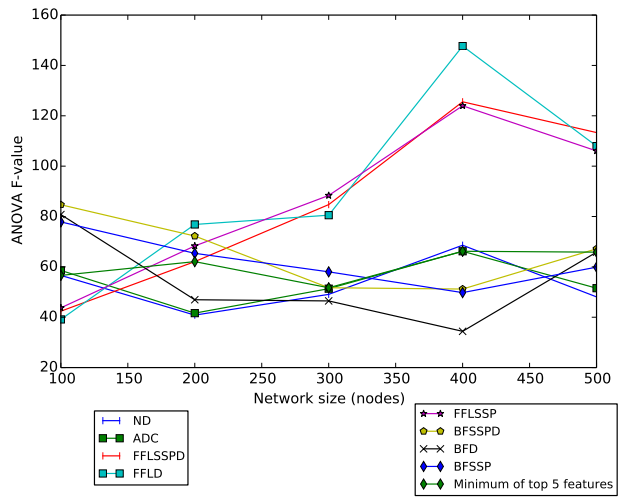
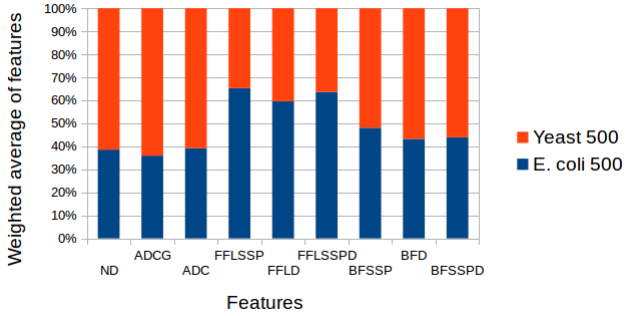


Figure 5: Comparison of 8 features across *E. coli*-derived networks of sizes 100, 200, 300, 400 and 500.



(a) Size 500

Figure 7: Feature comparison E. coli and Yeast derived networks for sizes 500.

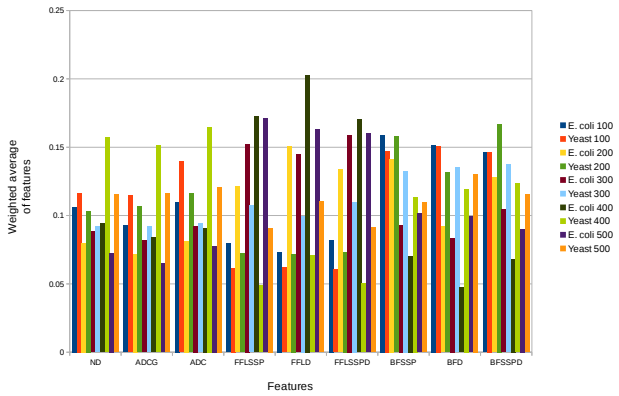


Figure 8: Weighted average of features calculated for a specific network type and size for *E. coli* and Yeast derived networks

tory networks. As a first, we identify several features that potentially contribute to the robustness of networks derived from both organisms. Feature ranking is performed to identify significant features using ANOVA F-value and weighted average ranking of top five identified features is performed. While [FFLSSP, FFLSSPD, FFLD] rank distinctly higher than other features for *E. coli* networks, [ND, ADCG, ADC, BFSSP, BFSSPD, BFD] rank outperform FFL-based features for yeast networks.

Machine learning can be a critical tool to understand biological principles. Our classifier will be improved in the future by pruning insignificant features. Extensive study using larger sample size will be carried out to understand the significance of label mapping using k-means clustering, choice of training and testing data split ratio. The full impact of cross validation fold size and number of labels in each class will be explored to design the best suitable classifier. This work paves a new way to compare biological systems and design bio-inspired topologies. Specialized networks can be designed which exploit the intuitive features such as ND, ADCG, ADC and biological features such as FFLSSP, FFLSSPD, FFLD, BFSSP, BFSSPD, BFD that are derived based on functionally important FFL and bifan motifs. Selective feature usage will help maximize information transmission and help realize network efficiency.

6. ACKNOWLEDGEMENTS

We would like to thank Ljiljana Zigic for the discussions on SVM classification and regression modeling.

7. REFERENCES

[1] J. A. Bondy and U. S. R. Murty. *Graph theory with applications*, volume 6. Macmillan London, 1976.

[2] H. Chan, L. Akoglu, and H. Tong. Make it or break it: Manipulating robustness in large networks.

[3] J. A. de la Peña, I. Gutman, and J. Rada. Estimating the estrada index. *Linear Algebra and its Applications*, 427(1):70–76, 2007.

[4] P. Ghosh, M. Mayo, V. Chaitankar, T. Habib, E. Perkins, and S. K. Das. Principles of genomic robustness inspire fault-tolerant wsn topologies: a network science based case study. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on*, pages 160–165. IEEE, 2011.

[5] C.-W. Hsu, C.-C. Chang, C.-J. Lin, et al. A practical guide to support vector classification, 2003.

[6] B. K. Kamapantula, A. Abdelzaher, P. Ghosh, M. Mayo, E. Perkins, and S. K. Das. Performance of wireless sensor topologies inspired by e. coli genetic networks. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on*, pages 302–307. IEEE, 2012.

[7] B. K. Kamapantula, A. Abdelzaher, P. Ghosh, M. Mayo, E. J. Perkins, and S. K. Das. Leveraging the robustness of genetic networks: a case study on bio-inspired wireless sensor network topologies. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–17, 2014.

[8] B. K. Kamapantula, A. F. Abdelzaher, M. Mayo, E. J. Perkins, S. K. Das, and P. Ghosh. *Quantifying robustness of biological networks using NS-2*. Springer (Under revision), 2014.

[9] H. Kitano. Towards a theory of biological robustness. *Molecular systems biology*, 3(1), 2007.

[10] Q. V. Le. Building high-level features using large scale unsupervised learning. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8595–8598. IEEE, 2013.

[11] Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabási. Controllability of complex networks. *Nature*, 473(7346):167–173, 2011.

[12] S. Mangan and U. Alon. Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences*, 100(21):11980–11985, 2003.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[14] R. J. Prill, P. A. Iglesias, and A. Levchenko. Dynamic properties of network motifs contribute to biological network organization. *PLoS biology*, 3(11):e343, 2005.

[15] T. Schaffter, D. Marbach, and D. Floreano. Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263–2270, 2011.

[16] G. Van Rossum et al. Python programming language. In *USENIX Annual Technical Conference*, 2007.

[17] P. Wang, L. Robert, J. Pelletier, W. L. Dang, F. Taddei, A. Wright, and S. Jun. Robust growth of< i> escherichia coli</i>. *Current biology*, 20(12):1099–1103, 2010.