Abundance of connected motifs in transcriptional networks, a case study using random forests regression

Khajamoinuddin Syed[*]
lnusk@.vcu.edu

Bhanu K. Kamapantula[†]
kamapantulbk@mymail.vcu.edu

Michael Mayo[‡]
Michael.L.Mayo@usace.army.mil

Edward Perkins[§]
Edward.J.Perkins@usace.army.mil

Preetam Ghosh[¶]
pghosh@vcu.edu

ABSTRACT

Biological network topologies are known to be *robust* despite internal and external perturbances. Motifs such as feed-forward loop and bifan have been marked to contribute to structural and functional significance. While network characteristics such as network density, average shortest path, and centrality measures etc., have been well studied, modular characteristics have not been explored in similar detail. Motif connectivity might play a major role in regulation under high perturbations. Connected motif abundance can skew network robustness as well. To test this hypothesis, we study the significance of the two connected feed-forward loop motifs using random forest regression modeling. We define thirty eight network features, fifteen of which are static and dynamic features and the other twenty three are two feed-forward loop connected motif features. We identify significant features among these using random forests regression and create models that can be used to train and predict the *robustness* of the biological networks. The performance of these models is measured using coefficient of determination metric and the significance of the features themselves is characterized using feature importances. Our experiments reveal that connected feed-forward loop motifs do not contribute to the *robustness* of network when models are created with all 38 features. For models with only connected motif features, the performance of a specific rhombus motif under high loss stands out.

[*]Corresponding author - lnusk@vcu.edu
Virginia Commonwealth University, 401 W Main St, Richmond, VA, USA

[†]Virginia Commonwealth University, 401 W Main St, Richmond, VA, USA

[‡]Environmental Laboratory, US Army Engineer Research and Development Center, Vicksburg, MS 39180

[§]Environmental Laboratory, US Army Engineer Research and Development Center, Vicksburg, MS 39180

[¶]Virginia Commonwealth University, 401 W Main St, Richmond, VA, USA

Categories and Subject Descriptors

D.2.8 [**Machine Learning**]: Metrics—*complexity measures, coefficient of determination, regression, feature ranking*

Keywords

motif connectivity, transcriptional networks, complex networks, vertex-shared motifs, connected motifs

1. INTRODUCTION

Motifs are often attributed to be one of the reasons for *robust* biological systems. A repetitive structure that occurs with a higher statistical significance in real networks than in random networks is termed to be a motif. In the past, researchers have identified feed-forward loop (FFL) motif to be an important motif in terms of abundance [13]. Further, functional significance such as response time speed-up and slow down has been attributed to FFL motif [12]. FFL structure is intriguing not only for its role in biological functionality but structurally as well Figure 1(b). It offers two ways of regulating the gene node (C) via two different transcription factor (A, B) nodes. In communication scenario, this becomes crucial when there is a network failure but information still needs to be transmitted. It is likely that the presence of higher FFL motifs will lead to better information transmission. In this work, we take a step further to study the connectivity between FFL motifs.

For the first time, this work aims to study the importance of the abundance of connected motifs. We use discrete event simulations and machine learning techniques to create a model, train and *learn* the feature data and predict *robust* behavior of biological network topologies. Discrete event simulations assist in modeling dynamic behavior of network interactions (information flow among the nodes in a network) under controlled conditions such as channel noise and congestion-based information loss. We assume that features in a biological network can be ranked. Does higher abundance of a connected motif pattern mean a *robust* network? Which of the considered network features contribute to *robustness*? Which machine learning model can accurately predict the *robust* behavior of biological network topologies? We explore these questions in the following sections. Answering these questions will reveal insights to the working of robust biological network topologies leading us to engineer specialized networks which are resilient under

heavy perturbations. Section 2 presents the methodology followed in this work. The definition of *robustness* varies from context to context. The metrics studied by researchers are predominantly static in nature [2, 3] as they do not consider the dynamic information flow within the network. [2] provides an in-depth review of existing metrics to measure *robustness*. None of the metrics consider features based on motifs or even connected motifs. *Robustness*, in our work, is measured in the aspect of successful information transmission as modeled by a discrete event network simulator. To this effect, we define network robustness as the ratio between the total number of packets received at the sink nodes to the total number of packets sent from the source nodes. We term this metric as packet receival rate. Packet receival rate is a dynamic metric as it models the network behavior at different perturbed conditions. This experiment setup has been detailed in our prior work and can be noticed in [11].

2. METHODOLOGY

The methodology followed in this work is illustrated in Figure 1. Subnetworks extracted (Section 2.2) from *E. coli* transcriptional regulatory network are passed to network simulator platform NS-2 (Section 2.3) to generate packet receipt rates and feature values are determined using Python programming language [18]. As a standard practice, features are scaled between 0 and 1. Section 2.4 describes Data processing followed in this experiment is described in Section 3.1. After processing the data in the correct format (as mentioned in the Step 1 in Figure 1) random forest regression machine learning technique is applied for feature ranking, and output prediction. Mean squared error metric is used to determine the optimal number of estimators (a key measurement used to estimate random forests) number (described in Section 3.2). Before feature ranking is actually performed, we perform feature selection which is a process to reduce feature set (from a thirty eight feature set). Features are ranked using feature importances (a technique used to determine feature significance in regression trees). Section 3.2 details the parameters used for creating random forests regression models followed by the performance of vertex-shared motif features.

2.1 Contributions

The major contributions of this work are as follows:

1. Define vertex-shared motifs which are potentially responsible for biological functionalities.

2. Using random forests regression to select important biological network characteristics.

2.2 Transcriptional subnetworks

Escherichia Coli and Saccharomyces cerevisiae are considered to be model organisms in the biological networks research community. For this work, we extract transcriptional subnetworks from *Escherichia Coli* to understand biological network characteristics and motif interactions from a structural perspective. To this effect, subnetworks of different sizes are considered: 100, 200, 300, 400, and 500 (size represents the number of nodes in a network). For each size, 1000 transcriptional subnetworks are extracted using
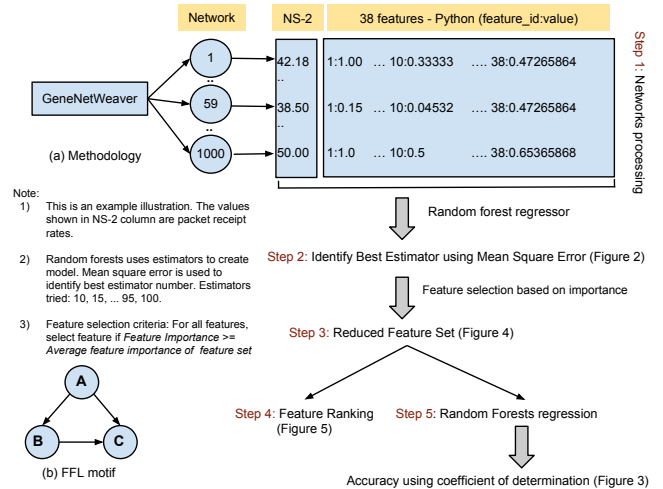

Figure 1: (a) Step-by-step methodology (b) FFL motif

GeneNetWeaver software [15]. During subnetwork extraction, GeneNetWeaver retains critical biological characteristics such as modularity. Specifically, these modules are responsible for distinct biological functionalities. Direction of the edges within these networks is retained as it captures regulation information of genes by transcription factors. Networks that are disconnected are not considered for further analysis. Self-edges (node with edges directed towards itself) in each network are discarded and the remaining network is reconstructed. This step pruned the dataset to 947, 943, 957, 932, and 941 networks for 100, 200, 300, 400, and 500 [1] network sizes respectively. This dataset is then used to explore network dynamics in two ways: a) model interactions using NS-2 (Section 2.3) and b) determine structural features from a static and dynamic perspective (Section 2.4 and Section 2.5). We study the significance of these features using machine learning techniques, specifically using regression modeling. This helps us identify the variation in feature importance from one network size to the other and under several lossy conditions.

2.3 Modeling network dynamics using NS-2

Network simulator platform, NS-2, allows researchers to explore the network characteristics. When compared to conventional graph analysis tools (UCINET, Gephi, Pajek, NodeXL etc.), NS-2 gives the advantage of exploring a complex system dynamically. NS-2 also models channel noise resembling perturbances within a biological system unlike other graph analysis tools. Previously, we mapped the problem of information flow in a biological network to that in a wireless sensor network [6, 8, 9]. This setup helps us understand the characteristics of biological networks uniquely using a framework used for wireless sensor networks. Following which, we established NS-2 as a robustness framework for biological networks. In a NS-2 network simulation, information is transmitted across the network via nodes and edges. Each node sends information in terms of packets

---

[1]871 networks were used for 500 network size at 10% loss. 941 networks were used for all other loss scenarios for 500 network size.

across its outgoing edges and these packets are collected at sink nodes. Transcription factors (also considered as source nodes here) and genes (also considered as sink nodes here) are both represented as nodes and the interactions among them are represented via edges. Packet transmission in each network is studied at various loss models: 10%, 20%, 35%, 50%, 60%, 75% and 90%. Packet receipt rate in the network is measured as the percentage of the number of packets received at sink nodes to the number of packets sent by all source nodes. Networks with higher packet receipt rate are considered to be more *robust*. Packet receipt rates of the networks range in between 0 (least *robust*) and 100 (most *robust*). Source nodes are considered to transmit or forward information (through packets) and sink nodes only receive the information. This situation is similar to a transcription network where gene is regulated (receiving information) by transcription factor(s).

2.4   Structural features

In order to understand the features contributing to *higher network robustness*, we studied several network characteristics. While some of these characteristics such as average shortest path, network density, and betweenness centrality have been explored by researchers under the context of robust networks, our definition of *what robust is* places emphasis on the study of network dynamics. In our earlier work, we identified fifteen different network features and ranked them using unsupervised learning techniques [10], [11]. These features include static characteristics such as average shortest path, network density, degree centrality and dynamic characteristics such as patterns derived from FFL-based direct and indirect paths [2]. These dynamic characteristics are derived after looking at the information flow using NS-2 simulation platform. This helps us identify the paths that were heavily used to transmit information and if these paths are related to FFL motifs. Some of the features use specific terminology from information communication theory (such as packet transmission).

The order of the features studied in this work is as follows: 1) network density, 2) average shortest path, 3) average degree centrality of the network, 4) transcription factors percentage, 5) genes percentage, 6) percentage of source to sink edges, 7) abundance of direct FFL motif edges, 8) abundance of indirect FFL motif edges, 9) percentage of FFL direct edges that contribute to successful packet transmission, 10) percentage of FFL indirect edges that contribute to successful packet transmission, 11, 12) number of direct and indirect FFL edges compared to the total successful (that contribute to successful packet transmission) direct and indirect edge paths in the network, 13) percentage of total edges in the network that participate in FFLs, 14) percentage of total edges that are actually FFL direct edges, 15) percentage of FFL direct edges that are source to sink edges. While our earlier work focused on identifying the impact of FFL, this work is focused on determining the impact of two FFLs that are connected. To this effect, we defined twenty three different connected FFL features that capture the abundance of connected FFL structures which are described in the following section. In total, we study thirty eight features to

---

[2]Consider an FFL ABC where C is regulated directly by A and indirectly by A via B. Here, the edge $A-C$ is considered to be a direct FFL edge and edge $A-B-C$ is considered indirect FFL edge

model the regression predictor. Hereafter, we refer to the connected feed-forward loop motifs as vertex-shared motifs.

2.5   Vertex-shared motif connectivity

It has been argued that interactions among modules are responsible for specific functionality in biological networks [4]. This is a deviation from another standpoint which states that the abundance of some structural patterns contributes to network robustness. While it is imaginable for both views to be correct, here we explore the structural role of specific modules in network robustness. Modules are essentially connected motifs at work. Here, we explore the vertex-shared feed-forward loop motifs for their structural role in attaining biological network robustness. In order to understand the significance of connected motifs, we first identified all possible ways two feed-forward loop motifs could be connected. Following the identification, we determined the abundance of each pattern in the above mentioned transcriptional networks.

The motif patterns can be divided into three categories first of which is *bow-tie* where one vertex is shared between two FFLs, second being *rhombus* where two vertexes are shared between two FFLs and third category being *bi-triangle* where all three vertexes are shared by two FFLs. All these patterns along with their respective abundance values are tabulated in Tables 1 and 2. Out of eighteen possible rhombus patterned motifs, there are six instances (RH-1/RH-8, RH-3/RH-14, RH-4/RH-11, RH-6/RH-17, RH-9/RH-13, RH-12/RH-16) where two patterns are found to be structurally isomorphic. All the isomorphic structures are shown in Table 3.

3.   RANDOM FOREST REGRESSION

Machine learning techniques prove quite useful in identifying significant features among a list of several features. Different strategies are employed for this task of significant feature identification. To perform machine learning tasks, we use the widely recognized *scikit* [14] module in Python. The aggregation of features defined in Section 2.4 and Section 2.5 combine to a total of thirty eight features. Abundance of connected motifs does not always contribute to *robust* network behavior. Data for connected motif abundance for different network sizes is suppressed here due to space considerations but provided in Section 6. The test for the correlation of feature abundance with robustness is performed in Section 3.4.

3.1   Data

Data is constructed similar to the procedure followed in our earlier work [11]. Each network is represented as a combination of feature values, feature ids and output labels. The output labels are determined using NS-2. In total, thirty eight features are studied in this experiment. These include the twenty three vertex-shared motif features introduced earlier apart from the fifteen features presented in [11]. As suggested in [7], we scale each feature between 0 and 1 for all the samples considered to create a model. Each network is represented as a combination of output labels and thirty eight network characteristics. This combination is known as a feature instance, in machine learning terminology. The results from NS-2 are used as output labels and the corresponding features are calculated using *networkX* [16] module in *Python* programming language. In our previous work

Table 1: Abundance of bow-tie and bi-triangle motifs in *E. coli* transcriptional network.

Pattern ID	Symbol	Abundance
BW-1		139827
BW-2		110505
BW-3		730
BW-4		24032
BW-5		1412
BW-6		1393
BT-1		17
BT-2		439
BT-3		4
BT-4		140
BT-5		3

Table 2: Abundance of rhombus motifs in *E. coli* transcriptional network.

Name	Symbol	Abundance
RH-1		623
RH-2		553
RH-3		788
RH-4		93
RH-5		7
RH-6		9
RH-7		69299
RH-8		516
RH-9		58364
RH-10		200
RH-11		656
RH-12		30

[10, 11], we considered the problem of ranking features to be an unsupervised one and used ANOVA [3] F-value to determine the significant features. But here, we consider the problem to be a supervised one and retained the output labels (range between 0 and 100) as floating points. In order to use classification techniques, one would have to group the output labels into bins which would mask the real data. Regression techniques are best suited for continuous data as output labels to predict new data. In order to avoid points that are equidistant from all the clusters (as noted in [11]), we increased the sample size for each network size from 100 to 1000 networks. By treating the problem as supervised instead of unsupervised one, we further take advantage of the output labels from NS-2. Further, we introduce feature selection here an improvement from our earlier work where the entire feature set was used to rank features. Before creating regression model, data is split into training and testing data in 75:25 ratio. Data split step is a common practice in machine learning tasks to 'train' the model on training data during which the model 'learns' the data and testing is performed on the test data. The accuracy of regression models presented in Figure 3 is based on testing of model created on the test data of all the 38 feature set.

3.2 Regression modeling

Firstly, network characteristics that are understood to capture the network robustness are defined. In our experiment, we have considered two scenarios, first one with a total of thirty eight features are considered in order to cap-

---

[3]analysis of variance

ture the network dynamics, and in second case twenty three features formed by the connected feed-forward loop motifs. However, before calculating an estimator that can be used to predict the performance of new network data, features need to be pruned. Some features might be correlated with each other and some might display higher variance than the rest.

We considered different feature selection methods to achieve the need of feature pruning. Randomized PCA was considered but ignored since it does not exploit the output label data to minimize feature space. LDA was also considered before being discarded. To this effect, feature selection step is performed using random forests with regression. Linear regression models such as Lasso and ElasticNet were considered before we discarded them for poor performance as measured by the coefficient of determination [4]. Recursive feature elimination techniques (with and without cross validation) were considered as well but were abandoned due to poor coefficient of determination values. These approaches involve removing one feature at a time and determining model performance on the remaining feature set at each step. The

---

[4]Coefficient of determination values were close to 0, far from being optimal.

Table 3: Isomorphic rhombus motifs in *E. coli* transcriptional network.

Name	Symbol
RH-1/RH-8	
RH-3/RH-14	
RH-4/RH-11	
RH-6/RH-17	
RH-9/RH-13	
RH-12/RH-16	

feature that impacts the model the best (i.e., model performance suffers upon that feature removal) is retained for future use. Random forests are used to solve classification and regression problems. The functioning of random forests is described in detail in [1].

Random forests is an ensemble machine learning technique which uses several trees (estimators) to predict the outcome of test data. A tree is constructed from sample data selected from the training data. At each terminal node of the tree, m features are selected out of the total features and a best feature is identified for the tree to be split at. The tree is then split into child nodes. This is repeated until the selected sample size from the training data is the least. By using several trees and averaging the predictions, the variance across the trees is reduced. Mean squared error (MSE) is used to determine the best number of estimators (number of decision trees) used in the random forests algorithm. Different number of estimators such as 10 to 100 in steps of 5 are used in creating different random forest models. MSE is determined for each estimator and the average of the number of estimators is used as the MSE value for that specific estimators' number. The variation in MSE after feature reduction is illustrated in Figure 2 (a) for one single case of network size 400 nodes at 90% loss and can be noticed that MSE is lowest when the number of estimators used in the random forest estimator is 70, The estimator for which MSE is the least is selected for calculating feature importances.

In Figure 2 (b) for feed-forward loop connected motifs model for network size 400 nodes at 90% loss, we can notice that before feature reduction MSE is lowest when the number of estimators used in the random forest estimator is 95. Detailed explanation for the feature importances is left out due to space considerations [1]. At every run, feature importances, coefficient of determination and corresponding mean squared error change due to the randomization in the algorithm. To negate this, we execute the entire process for hundred runs and take the average of the respective values. Our experiments reveal that the importance of features depends heavily on network size and loss it entails over time. Average of feature importances is used as a heuristic to select subset of thirty eight features. All the features with
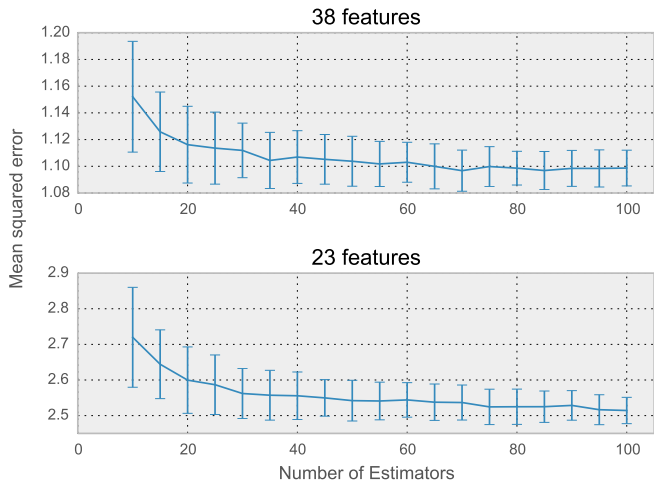

Figure 2: Mean squared errors (MSE) at different estimators for 400 network size at 90% loss. Measured for the model with 38 features and 23 features respectively. Errorbars capture the variation of MSE across hundred test runs. Note that the Y-axis does not start at 0. *Lower MSE is better.*

feature importance values greater than and equal to the average feature importance value are selected to model the final regressor for prediction.

3.3 Feature reduction

Coefficient of determination (COD) is used as a metric to measure a model's performance. For each of the thirty five random forest models, COD is determined before and after feature reduction. Each random forest regression model uses X number of estimators as shown in Figure 2. *Feature importance* of all the features is determined by averaging the total reduction in node impurity [5] across X estimators. We then create a random regressor to predict outcomes based on the model with reduced feature set which is tested using test data set.

COD measures the performance of predicted values by the model when compared to the real values. Good regressors will have a COD value close to 1 and the bad ones will have a COD close to 0. As evident from Figure 3, performing feature selection to reduce the feature set as explained in Section 3.2 does not improve the model accuracy. The majority of the models with all 38 features perform better than the models with a reduced feature set. The figure illustrating COD performance for models with 23 vertex-shared motif features is omitted as it follows similar trend.

Figure 4(a) presents the number of features selected by the feature selection process from all thirty eight features. It can be observed that the maximum number of features selected as important are 16 for the network size 200 at 50% loss and the least number of features that are selected as important are 3 for network 400 at loss 90%. At high loss (90%), few features (≤ 6) are responsible for network robustness.

Figure 4(b) shows important features selected from vertex-shared motifs for all network sizes at different loss scenarios. The number of significant features varies between 3 and 9.

---

[5] as used in scikit-learn toolkit

At high loss (90%), few features (≤ 5) are responsible for network robustness.

3.4 Feature value correlation with robustness

In order to test the hypothesis if high feature values directly correlate with high robustness, we perform the following tasks. These tasks are executed at a network level. That is, significant features are identified for all models at different loss types for a given network size.

1. First, we identify the top five features using random forest regression (feature importance as a metric).

2. We then calculate the number of times each of the features occurs in the top five ranks at different loss scenarios.

3. Further, we determine the mean of each feature for a given model and identify the top five features with highest mean.

4. We then compare these features with the features obtained in second step.

As a result, we found no correlation (direct or inverse) between feature value and its importance. Among the models with all 38 features, gene percentage, direct FFL edge abundance, FFL indirect edges that participate in successful packet transmission to sink nodes, and the occurrences of direct edges in feed-forward loop motif (IDs 6, 8, 11, 12 respectively in Figure 5) are strong indicators of robustness. Apart from these features, network density, average shortest path, average degree centrality, and percentage of transcription factors (IDs 0, 1, 2, 3) also correlate to robustness relatively well. It is important to note that certain features make their impact distinctively in specific network sizes or at specific loss scenarios. This can be attributed to the fact that these specific features might be expressed more during the network extraction step (Section 2.2). The distribution of feature importances (with feature IDs mentioned earlier) determined using random forest regression is shown in Figure 6. Each feature contains of hundred test runs to normalize the variations in feature importances due to randomization in regression algorithm. Outliers in the dataset are points that do not occur in the range of top and bottom whiskers and are identified by +.

4. VERTEX-SHARED MOTIFS

The importance of features as determined in Section 3.2 is charted in Figure 5. Heat maps are generated for all the networks at losses 10%, 20%, 35%, 50%, 60%, 75%, and 90%. Figure 5(a) represents one such case at 60% for model created with all 38 features. At one glance, it can be observed that features with IDs 1 to 13 and 28 stand out in all the networks. These features are *average shortest path, source to sink edge percentage, abundance of indirect FFL paths, percentage of direct FFL edges, percentage of indirect FFL edges, abundance of direct FFL edge occurrences,* and *abundance of indirect FFL path occurrences* respectively [6]. RH-7 (from Table 2) ranks as a significant feature in all network sizes and other connected motifs such as BW-4, RH-2 and BT-2 only stand out once.

---

[6] These features are described in our earlier work [11]
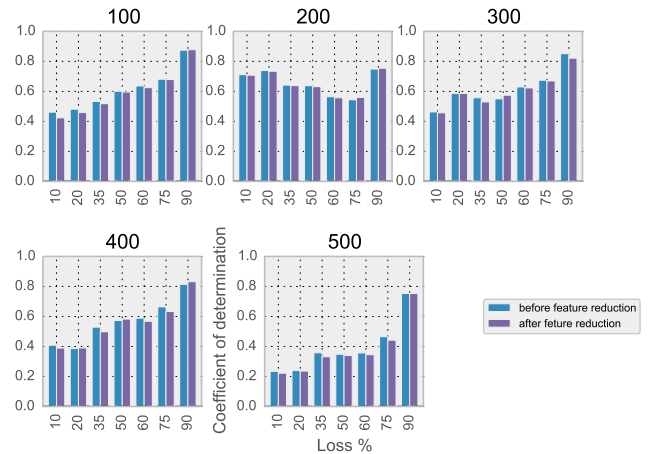

Figure 3: Coefficient of determination (COD) for regressors - different network sizes for 38 features model. *Higher COD is better.*

Extending the hypothesis test described in Section 3.4 to models with only vertex-shared motifs (23 features), we found no correlation between feature value and its importance. Here, BW-1, BW-2, BW-4 and RH-7 (Refer to Table 1 and Table 2) are the strongly expressed features with robustness in all network sizes at different loss models. The results indicate that controlling the presence of these features can significantly impact biological network robustness. These features can also assist in creating superior bio-inspired networks where signal transduction is influenced by selective features such as the ones derived from FFL motif and the network itself can be adaptive by activating different regions at different periods of time to conserve energy.

Figure 5(b) represents heat map of model created with twenty three features of feed-foreard loop connected motifs at loss 60%. This heatmap shows that Feature IDs BW-1, BW-2, BW-4, BW-6 (in two instances), RH-13 and BT-2 mark their presence in all the networks, but RH-7 ranks out as very important feature in all networks.

5. DISCUSSION

There is no one model that fits all data. We will extend the experiments to larger sized networks for *E. coli* transcriptional networks until maximum possible size is reached (i.e. number of nodes in *E. coli*) to explore if the trends in feature significance holds true. Further, we intend to extend the experiments to *Saccharomyces cerevisiae*. Our earlier experiments [11] revealed that feature significance varies from one model organism to the other and across network size and perturbation conditions. The higher ranking of FFL-derived features (IDs 7, 11, 12 in Figure 5) reveals the significance of motif derived features across different network sizes. Topological features such as network density, average shortest path remain important across all network sizes and under different loss conditions. The significance of vertex-shared motifs is relevant at high loss making them useful for constructing robust smart networks capable of surviving lossy conditions. New research has indicated the evolution of bow-tie motif under distinct conditions such as a limitation on number of edges in a network [5] and its potential

(a)
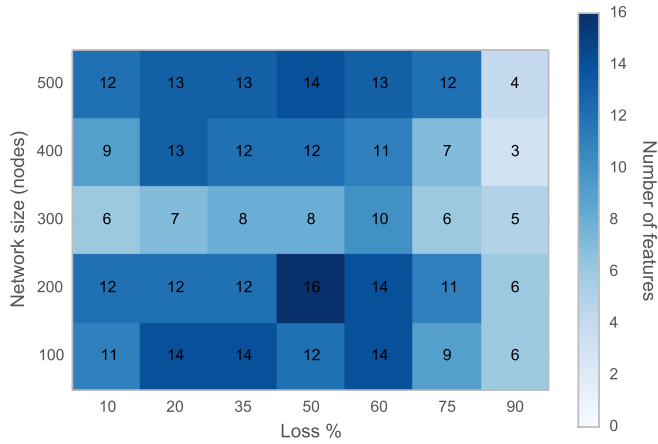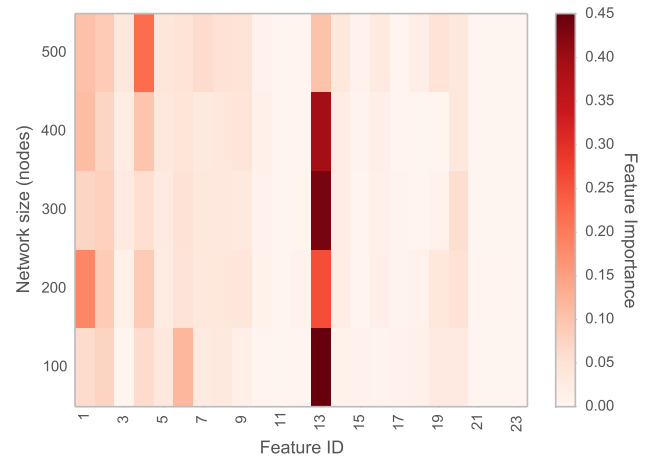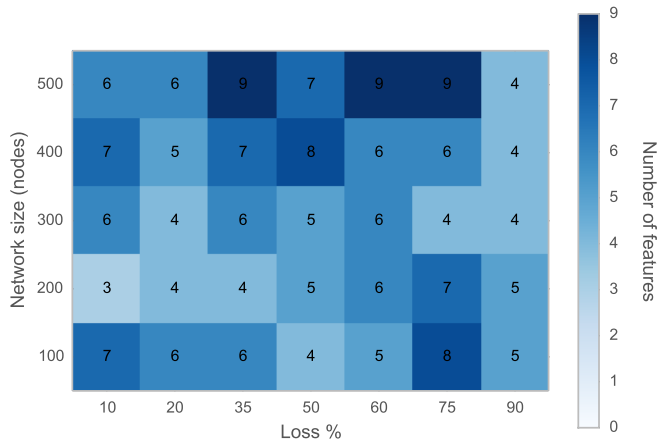
(b)

Figure 4: (a) Selected features (out of total 38) for every model at a given network size and loss model as described in Section 3.2. (b) Selected features (out of 23) feed-forward loop connected motifs. Criteria: select features that have higher than average feature importance using random forest regression.

(a)

(b)

Figure 5: (a) Feature significance in all the networks at 60% loss for model with all 38 features. (b) Feature significance of connected feed-forward loop motifs in all the networks at loss 60%. *The darker the color the higher the feature significance.* Additionally, numbers are included to indicate feature rank. *Higher the feature importance, better is the feature.*

role in maintaining biological network robustness [17].

This is an interesting proposition for designing engineered systems that exploit the principles seemingly intrinsic to the design of biological network topologies. The implications of specialized engineered systems cannot be ignored in the areas of disaster relief coordination.

6. ADDITIONAL MATERIAL

Entire dataset is made accessible for research purposes at http://bnet.egr.vcu.edu/data/bict2015. Results for all the loss models not presented in this paper due to space considerations are also made available at the same URL. Sensitivity analysis for variation in mean square error is also detailed.
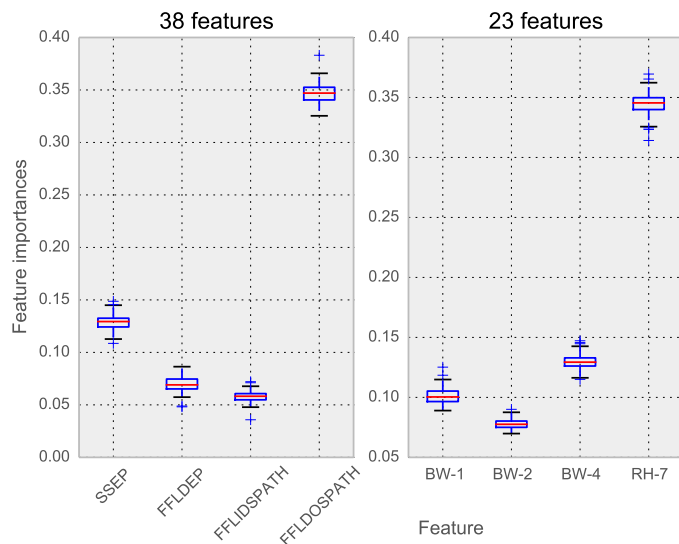
7. ACKNOWLEDGMENTS
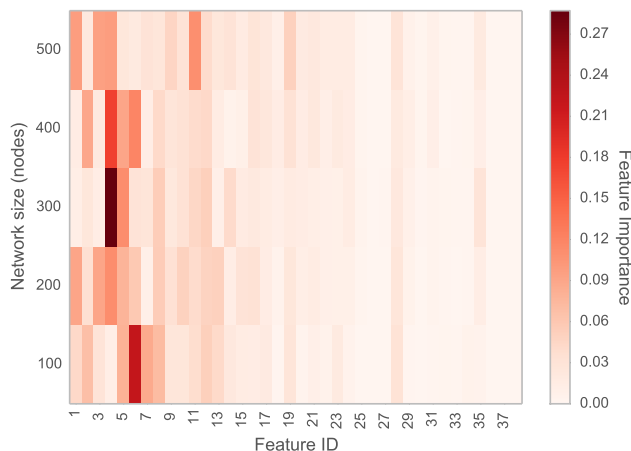
Figure 6: Feature importances as determined by random forest regression for models with 38 features and 23 features respectively for network size 100 at 75% loss. While the feature source to sink edge percentage is termed as SSEP, the percentage of FFL direct edges is termed FFLDEP. The features 10, 11 as explained in Section 2.4 are FFLIDSPATH and FFLDOSPATH. Refer to Table 1 and Table 2 for definitions of BW-1, BW-2, BW-4, and RH-7 *Higher feature importance is better.*

8. REFERENCES

[1] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[2] H. Chan, L. Akoglu, and H. Tong. Make it or break it: manipulating robustness in large networks. In *Proceedings of the 2014 SIAM Data Mining Conference*, pages 325–333. SIAM, 2014.

[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] M. R. Fellows, G. Fertin, D. Hermelin, and S. Vialette. Sharp tractability borderlines for finding connected motifs in vertex-colored graphs. In *Automata, Languages and Programming*, pages 340–351. Springer, 2007.

[5] T. Friedlander, A. E. Mayo, T. Tlusty, and U. Alon. Evolution of bow-tie architectures in biology. *PLoS computational biology*, 11(3):e1004055–e1004055, 2015.

[6] 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.

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

[8] 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.

[9] 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*, 5(3):323–339, 2014.

[10] B. K. Kamapantula, M. Mayo, E. Perkins, A. F. Abdelzaher, and P. Ghosh. Feature ranking in transcriptional networks: Packet receipt as a dynamical metric. In *Proceedings of the 8th International Conference on Bioinspired Information and Communications Technologies*, BICT '14, pages 1–8, 2014.

[11] B. K. Kamapantula, M. Mayo, E. Perkins, and P. Ghosh. Dynamical impacts from structural redundancy of transcriptional motifs in gene-regulatory networks. In *Proceedings of the 8th International Conference on Bioinspired Information and Communications Technologies*, BICT '14, pages 199–206, 2014.

[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] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[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] D. A. Schult and P. Swart. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conferences (SciPy 2008)*, volume 2008, pages 11–16, 2008.

[17] P. Tieri, A. Grignolio, A. Zaikin, M. Mishto, D. Remondini, G. C. Castellani, and C. Franceschi. Network, degeneracy and bow tie integrating paradigms and architectures to grasp the complexity of the immune system. *Theor Biol Med Model*, 7(32.10):1186, 2010.

[18] G. Van Rossum et al. Python programming language. In *USENIX Annual Technical Conference*, volume 41, 2007.