Machine Learning in Biological Networks

Master Thesis Project

by

Monalisa
2009EE50484

under the guidance of

Dr. Sumeet Agarwal

Department of Electrical Engineering
Indian Institute of Technology, Delhi
May 2014

Abstract

Molecules and interactions among themselves can be modelled as a network(biological). In this project, I am trying to find out features which can help us in distinguishing between biological and non-biological networks. Non-linear dimensionality reduction technique and PCA have been used to reduce features. Various basic methods like SVM, interclass variance have been employed to find the classifying features. We also look at gene regulatory networks and if there is some similarities between them and biological networks (in this project, protein interaction networks). Fraction of k-core of a graph came out to be the most prominent feature. We tried to find out if there is some relation between existence k-core and number of edges in the graph.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

There have been many attempts to understand intricacies of human body, the biological interactions going on inside the body all the time. A simple activity is dependent on various interactions, factors and moderators which are in turn controlled by some other factors or moderators. For example, p53 (a 393-amino-acid protein sometimes called the guardian of genome) acts as tumour suppressor because of its position within a network of transcription factors. However, p53 is activated, inhibited and degraded by modifications such as phosphorylation, dephosphorylation and proteolytic degradation, while its targets are selected by the different modification patterns that exist. [3]. So, neither p53 nor the network work as a tumor suppressor by itself. There is a symbiotic kind of relation between them[3]. Now, the challenge is to use these kind of data (i.e. gene expression microarrays, protein-protein interaction) to understand underlying systems and mechanisms. Systems biology aims to tackle this by using the mathematial abstraction of graph to represent the system consisting of interacting component. The interactions and proteins or genes (or other participating molecules) are viewed as edges and nodes of a network(graph) to help understand them better. This project focuses on this only.

1.1 Systems Biology

Systems biology studies biological systems by systematically perturbing them (biologically, genetically, or chemically); monitoring the gene, protein, and informational pathway responses; integrating these data; and ultimately, formulating mathematical models that describe the structure of the system and its response to individual perturbations[1]

Basically, we are trying to understand system level of biology. In this approach, the understanding of structure and dynamics of a biological system is as important as the understanding of interacting genes and proteins. Just the knowledge of 'what' is not enough, we also need to know 'how' the interaction is happening if we want to understand how a gene or protein affects the system. We need a dynamic picture. We can have this understanding by insight in four key features namely [2]-

- System structures - These include genes and proteins, biological pathways and their interactions.

- System dynamics - It includes the behaviour of system over time and under different conditions. This mainly deals with the how part of systems biology

- Control method - The cell functions can be modulated by some mechanisms i.e. we can kind of control the environment and prevent malfunctioning of the cell

- Design method - We can use the information to design and modify biological systems instead of trial and error method.

Insight in any of the above features would help us to understand the biological networks, which has its own benefits. We can find out where exactly is the problem and what caused it. We can design new drugs and medicines which can be more precise and effective.

An accurate model of a biological network is highly desirable. We are looking for some patterns or similarities between biological networks which can guide or inference or reconstruction process. We are trying to find some structural signatures of the biological networks. For example, consider architecture. By looking at a building or its structure, an architect can easily tell what kind of building it is and can also list some salient features of that type of architecture. We want to have this kind of understanding of biological networks. We want the structural signature to be incorporated in the reconstruction process. To look this thing in more detail, we are trying to find features which distinguish biological networks from the non biological ones ans then use those structural differences as a prior (or given) in development of an algorithm.

1.2 Networks

1.2.1 Real World Networks

This project mainly concerns with real world networks[1] and their structural differences and similarities. They can be broadly divided into two types of networks -

- **Interaction** - These type of networks show the interaction between its nodes. Examples include social networks like facebook, protein interaction networks.

- **Correlation** - These networks denote similarities or correlation between different components of a system. Examples include political, language, financial networks.

Non Biological Networks

The non biological networks used in this project are generally social like facebook etc., political, language, financial etc. They are interaction as well as correlation type of networks. These are generally large networks as compared to biological ones. We have 167 of such networks in our database[4].

[1] List of these networks can be found ‘in appendix

Biological Networks

Proteins interact with each other to carry out various bodily and cell functions. Protein-protein interactions refer to intentional physical contacts established between two or more proteins as a result of biochemical events and/or electrostatic forces. These interactions combine to form a protein interaction network. We have 67 of biological networks in our database[4].

Gene Regulatory Networks

A gene is the molecular unit of heredity of a living organism. It is a segment of DNA and RNA. Gene forms functional gene product (which is mainly proteins and RNAs) by the process of gene expression. In this, the DNA containing the gene is transcribed to RNA and then it is translated from RNA to protein. The process of transcription and translation, in turn requires some external stimulation which comes in the form of enzymes or RNAs or some other molecules. These interactions, genes and molecules (or proteins) among themselves form a gene regulatory network. We have only 3 GRNs in our database. They are *bsubtilis, e. coli* and *dream4*. *Bsubtilis and e. coli* are real world networks while dream4 is one of the challenges of Dialogue for Reverse Engineering Assessments and Methods. These are the GRNs used in *Inferelator*[9] method.

1.2.2 Generative Model

In this project, we have looked at just one generative model, viz. *Erdös-Rényi model*, which is often denoted as G(n,p)[6], where n is the number of nodes and p is the probability of having an edge between any two nodes independent of other edeges. There's also very similar model denoted as G(n,m)[7], where n is same as above and m is the number of edges. The G(n,m) model randomly generates graph with n nodes and m edges, each of which have equal probability of existing.

1.3 Features

In this section, we define some of the features (or network diagnostics) which we use to understand the networks. These form columns of the design matrix (defined in later section). There are 253 features, some of which that recur in this report -

- **Density** - It is the ratio of number of edges in the network to the maximum number of possible edges.
- **Fraction-k core** (or k-core) - It is maximal connected subgraph of a given graph in which all vertices have degree at least k.
- **Clustering coefficient** - It is defined for each node of an unweighted graph as the ratio of number of edges between the neighbors of that node to the maximum possible number of edges between them.
- **Eigenvector centrality** - A centrality measure, whih gives weights to the edges according to the importance of their nodes.

- **Degree centrality** - It gives the ratio of a degree of node to the maximum possible degree of the node (i.e. n-1, if we ignore self edges.)

1.4 Literature Survey

Many efforts have been made by the researchers to understand biological networks and gene regulatory networks. There have been many previous works which focus on reconstruction of gene regulatory networks from the gene expression data. Some works have concentrated on how the networks interact and how are they coonected to evolution.

1.4.1 The Inferelator

As the title of paper says, it is an algorithm for learning parsimonious regulatory networks from systems biology[10]. The model solves a ordinary differential equation. Let us say that the gene expression level of a gene y is affected by other gene expression levels (or any other influencing factor) denoted by vector X = $(x_1, x_2, ..., x_N)$.

$$\tau \frac{dy}{dt} = -y + g(\beta.Z) \tag{1.1}$$

Here, Z = $(z_1(X), z_2(x), ..., z_m(X))$ is a set of functions on X. The coefficient $\beta = (\beta_1, \beta_2, ..., \beta_m)$ denote the influence of Z on y, i.e, if it is positive, we can say that it acts as an inducer while negative coefficient would mean it acts as an repressor. τ is the time constant of level y in absence of external influence. The function g acts ac an activation function and takes the form of sigmoidal or logistic function.

$$g(\beta.Z) = \frac{1}{1 + e^{-\beta.Z}} \tag{1.2}$$

Multivariate regression is used to find β. For model selection, LASSO is used.

1.4.2 Inductive Logic Programming (ILP)

This technique is concerned with the qualitative aspect of the network rather than the quantitative. It aims at predicting whether the influence is positive or negative. It doesn't tell how positive or negative it is. It has qualitative constraints and hence we get qualitative models. ILP is the best known technique to get qualitative models. As the name suggest, it is a logic based model and hence, we have yes or no as the response and we know a yes/no model can be easily visualised as a binary tree. The salient features of this technique is -

- Background knowledge - These are statements which define some qualitative constraints. It is the rule book for the model, so that we don't go on searching every path whether its feasible or not

- Examples - Examples are the observations (or data points). Given the definitions defined in the background knowledge, a model is said to be an explanation of an example if it yeilds true for the example

- Refinement Operator - This operator defines how the descendents of a node of tree would be. The descendents are mainly defined as whether they are generalisations or specialisations of the node.[11] Generalistion refers to removing one or more components or disconnecting them while specialisation means adding new components or connecting the existing ones.
- Cost function - It is a real-valued function for each node. It is basically a trade-off between the complexity and and explanation of model

ILP can be incremental, meaning, we can create a model of bigger and complex systems by breaking it into sub systems and creating a model for them and then sending it into another ILP which specialises for the complex model.

1.4.3 Comparative Network Analysis

This is an unpublished work which deals with the comparison of various networks and tries to find out if there is some correlation among them.

- Network classification - Some networks tend to display specific community structures and hence can be classified into communities.
- Hardness regression - It aims to identify the features of networks which makes TSP (travelling salesman) computation for the given graph easier and exploiting the result that some graphs can be solved more easily than others. Some features correlate highly with the solution length or runtime of the solver which can be calculated more easily and can be used to estimate the solution length or runtime.
- Phylogeny regression - It identifies if there are any signs of evolution i.e. phylogenetic signal in features as features evolve with evolution.

Approximate Bayesian Computation(ABC)

The purpose of finding prominent features is to help develop a model for generating networks. These features can help us limit our search space for required model and can also serve as a check for correctness of a model. Bayesian theory can be used here as the prominent features can help us define a likelihood of a model. The prior for a model can be defined using its parameters. Likelihood and prior together can give posterior which can be used for the comparison of models.

$$P(M,\theta|x) \propto P(x|M,\theta)P(M,\theta) \tag{1.3}$$

For complex models, the likelihood function might not be computable. In this scenario, we generate data points from the model given some parameters and prior. We define an error threshold ϵ. If the data point lies within the threshold, it is retained. So, the approximate likelihood $P_\epsilon(x_0|M,\theta)$can be defined as

$$P_\epsilon(x_0|M,\theta) = \frac{1}{\epsilon}\int_X I(|d(S(x),S(x_0))|) \leq \epsilon/2)P(x|M,)dx \tag{1.4}$$

where I(x) is an identify function which returns 1 if the condition is true and 0 if its false d(S(x), S(x_0)) is a distance function and can be Euclidean distance function.

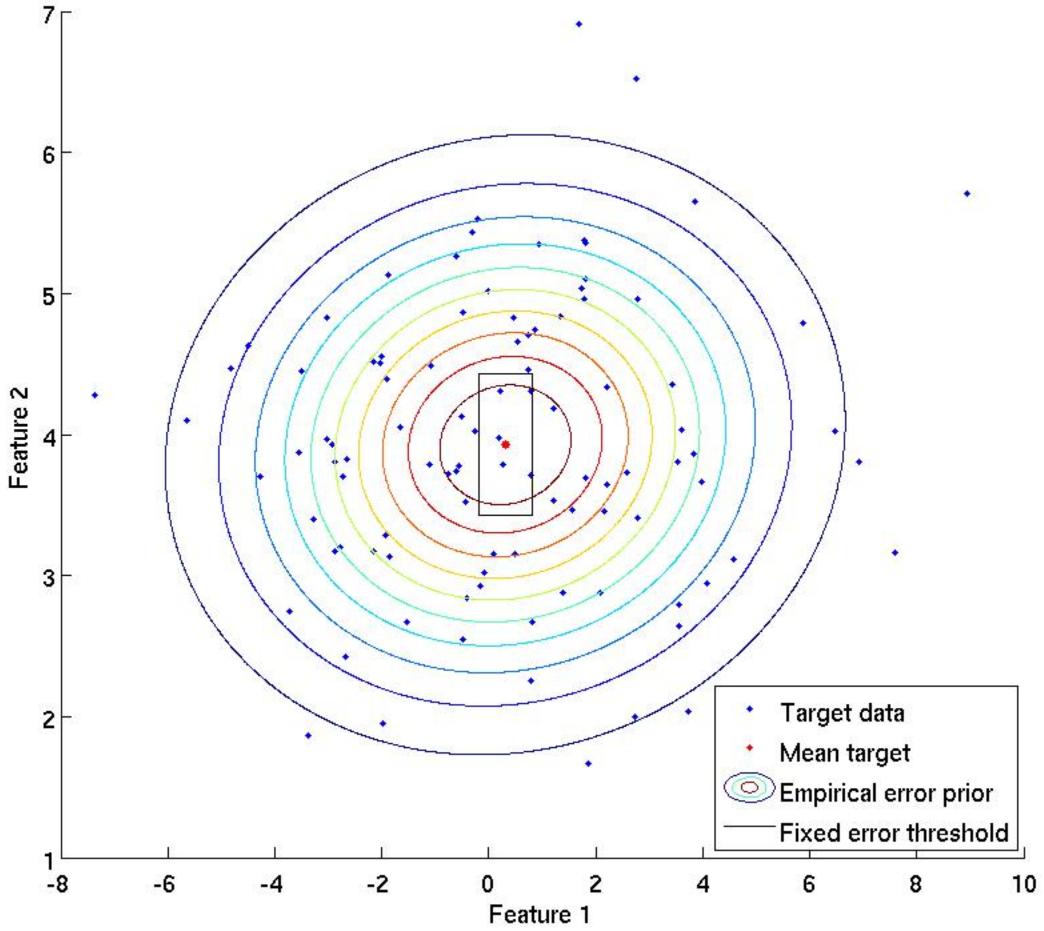


Figure 1.1: Obataining the empiral error prior from a given target data set[2]

[2] This image is courtesy unpublished paper by Dr. S. Agarwal and G. Villar, and N. S. Jones[5]

Chapter 2

Methods and Material

2.1 Network-Feature Matrix (Design Matrix)

For the purpose of analysing the networks and their properties, a network-feature matrix or design matrix, which has networks as its rows and features as its columns, is created. The matrix is formed by calculating value of each network diagnostic (feature) for every network[1]. It is possible that some diagnostics dont exist for some networks or aren't computable within the time constraint. The networks for which a particular feature value doesnt exist is set as NaN. To be able to compare these networks and features with each other, the matrix is normalised via the logistic function defined as follows[5] :

$$f(z) = 1 + e^{-z} \tag{2.1}$$

After normalisation, the features (columns) having less than 80% of their entries filled are discarded. This reduces our number of features to 211. For the columns with missing entries, the NaN values are replaced by the average of that column.
Since, the number of features is too large, this (normalised) matrix is reduced using Isomap and PCA. 10 principal features of each reduced matrix is taken into account for analysis.

Separate analysis is done for GRNs, as new matrix is created adding them into the list of networks already taken. After normalisation and discarding of features, we are left with 210 features. This matrix is also reduced using Isomap and PCA and 10 principal features are considered. Euclidean distance from the 3 GRNs is calculated for all the networks.

We tried to find out most significant feature(s) using various techniques. The features were ranked according to their interclass variances to see which feature is the most diverse. SVM was used to get a linear classifier for biological and non biological networks and then features were arranged according to their weights. Features, which were maximally correlated with the reduced matrix (as told earlier), were found out and ranked accordingly.

[1] The code for which is obatined from the unpublished paper by Dr. S. Agarwal and G. Villar, and N. S. Jones[5]

2.2 Relation between k-core of a graph and its number of edges

As discussed earlier, there exists a relation between the number of edges and the emergence of a k-core in an Erdos-Renyi random graph. It has been shown that the probability of emergence of a giant k-core (for k > 2) is high, when number of edges reaches a constant(c_k) multiplied by number of nodes. This constant varies with the value of k and can be found[17].

$$c_k = min_{\lambda > 0} \frac{\lambda}{\pi_k(\lambda)} \tag{2.2}$$

and

$$\pi_k(\lambda) = \mathbf{P}\{Poisson(\lambda) \geq k - 1\} \tag{2.3}$$

where, $c_k/2$ is the constant required and k is minimum degree of nodes in the core. c_k is the minimum value for which a k-core exists in G(n,m) (m > c_k). Below this value, it doesnt exist. As shown in the result by Erdos and Renyi, for large n, there is birth of a component when m reaches n/2 [7]. So, c$_k$ >1. The value of this constant for k = 3 is *3.35* and for k = 4 is *4.88*. We took ratio of number of edges(m) to c_kn/2 i.e.

$$\rho_k = \frac{m}{c_k n/2} \tag{2.4}$$

Value of ρ_k was calculated for each network for k = 3,4 and was added into the network-feature matrix as an extra feature.

Chapter 3

Results

3.1 Inter-class Variance

Inter-class variance for two classes X and Y, having p and q elements respectively, can be defined as

$$var = \frac{\sum_{i=1}^{p}\sum_{j=1}^{q}(X(i) - Y(j))^2}{p + q} \tag{3.1}$$

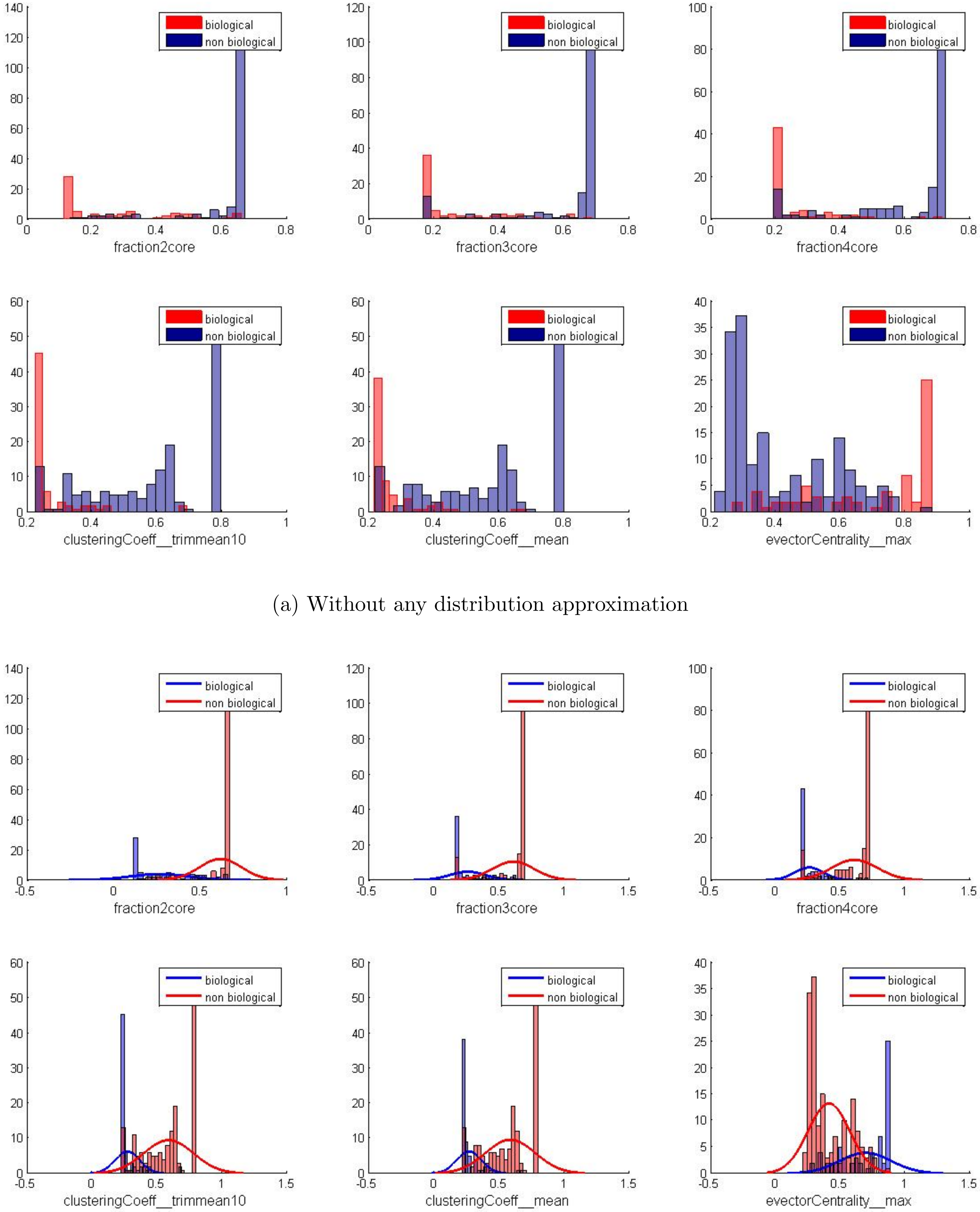
All the features were ranked according to their inter-class variances to find out which features vary the most in our two classes of networks viz. biological and non-biological. Fraction-k-cores turned out to be the best feature with highest inter-class variance . It was followed by clustering coefficient and evector centrality.

Features	Variance
fraction2core	8.006
fraction3core	7.944
fraction4core	7.870
clusteringCoeff__trimmean10	6.965
clusteringCoeff__mean	6.925
evectorCentrality__max	6.869

Table 3.1: Features with maximal interclass variance

As apparent from the figure 3.1, approximating distribution of networks for these feaure values using guassian distribution we can see that the two classes of networks are quite distinguishable. We, then, constrained the size and density of network to 1000 nodes and 0.25[1], respectively. We found out that the top contenders remained the same.

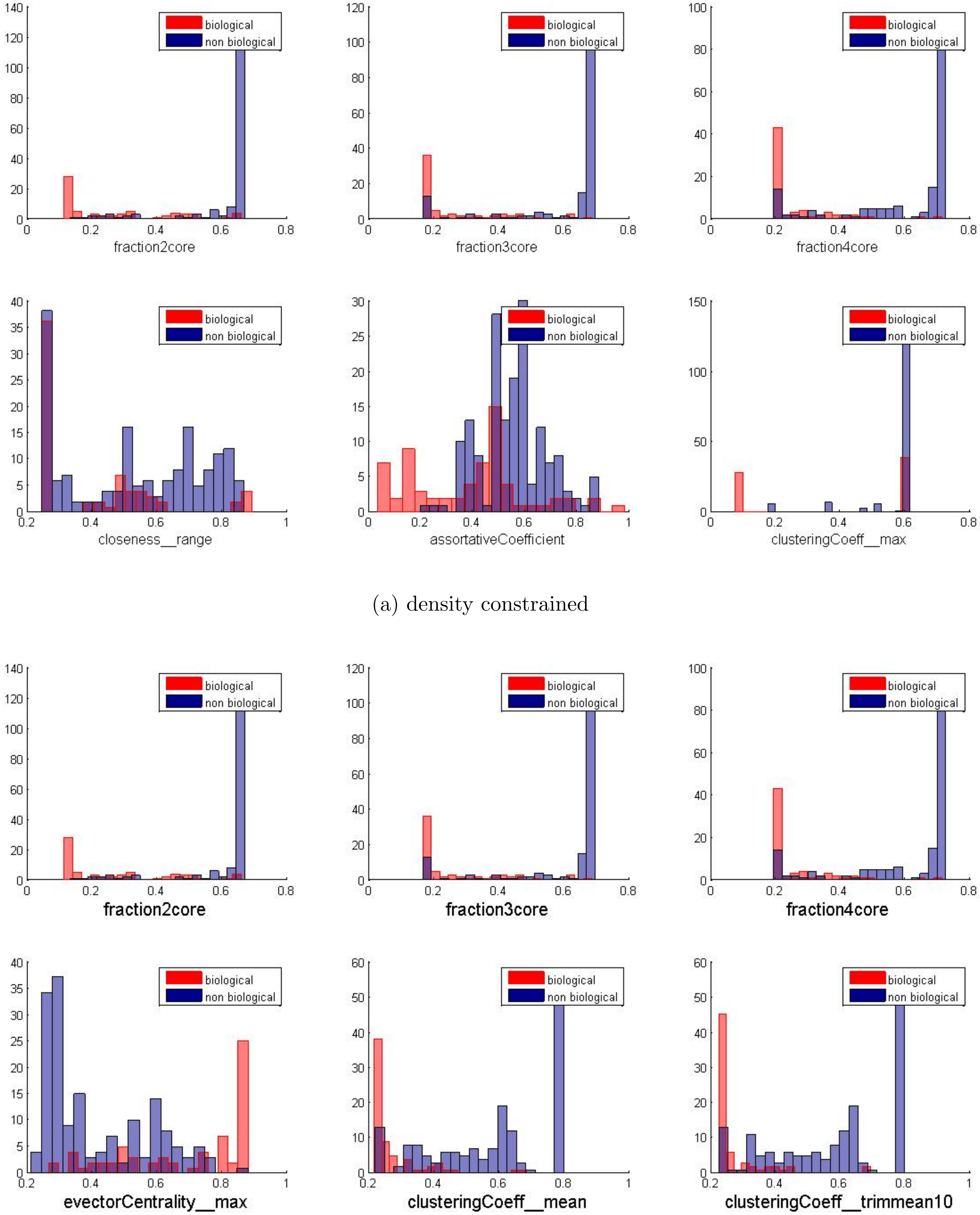
[1] density was on scale of 0 - 0.5



(a) Without any distribution approximation

(b) Approximated as gaussian distribution

Figure 3.1: Histogram of features in Table 3.1



(a) density constrained

(b) size constrained

Figure 3.2: Histogram features with max variance after some constraints

3.2 SVM

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.[14]
As told earlier, SVM was applied on the design matrix using linear kernel and cross-validation. It was able to classify normalised design matrix with 91.88%, isomap reduced design matrix with 74.36% and PCA reduced matrix with 85.47% accuracy.

Features	Weight
clusteringCoeff__medad	1.192
assortativeCoefficient__snowball100	0.996
fraction2core	0.879
betweenCentrality__max	0.878
betweenCentrality__range	0.878

Table 3.2: Features with maximal weights after applying SVM on normalised design matrix

3.3 Dimensionality Reduction

We used a linear and non-linear dimensionaltiy reduction technique viz. PCA and Isomap[16]. Principal component analysis (PCA) can be formally defined as a technique that is useful for the compression and classification of data. The purpose is to reduce the dimensionality of a data set (sample) by finding a new set of variables(principal components), smaller than the original set of variables, that nonetheless retains most of the sample's information (variation) [15].
We applied Isomap and PCA on the normalised design matrix on both with and without GRNs. The correlation of the reduced feautres with the normalised features was found out.

Dimension	Feature1	r1	Feature2	r2
1	degreeCentrality__harmmean	0.9226	degreeCentrality__geomean	0.9144
2	fraction2core__snowball100	-0.7354	numNodes__snowball100	-0.7246
3	evectorCentrality__fit_lognormal	0.7604	evectorCentrality__fit_wbl	0.7601
4	assortativeCoefficient__snowball100	0.6121	assortativeCoefficient	0.5624
5	clusteringCoeff__var	-0.5025	clusteringCoeff__iqr	-0.4973

Table 3.3: Maximum correlated features with Isomap reduced features (without GRN)

As we can see, the first feature is highly correlated with degree centrality (which is a measure of how the degree of nodes is spread) while the second feature is correlated with fraction-2-core and number of nodes. So, we can say that size and node variability of network capture most of the variance.

Dimension	Feature1	r1	Feature2	r2
1	clusteringCoeff__min	-0.9409	transitivity	-0.9304
2	evectorCentrality__posrms	-0.8413	numNodes__snowball100	0.7671
3	closeness__meanad	0.7969	closeness__medad	0.7901
4	degree__fit_gamma	-0.6231	degree__fit_wbl	-0.6230
5	betweenCentrality__posrms	-0.4997	degreeCentrality__range	-0.4886

Table 3.4: Maximum correlated features with PCA reduced features (without GRN)

The first two principal components of PCA capture 56% of variance among them. Since, it is a linear technique the variance captured is not high as the features themselves, can be quite correlated.

When we reduced the design matrix containing 3 GRNs using isomap and PCA, respectively, we find them among the biological networks. The correlation of the reduced design matrix with normalised design matrix was found out for both Isomap and PCA reduction. We find that the results are not that different from the design matrix which didn't have gene regulatorey networks. But, this can be very easily accounted for, by the fact that we didn't have many GRNs to work with.

Dimension	Feature1	r1	Feature2	r2
1	degreeCentrality__harmmean	-0.9224	degreeCentrality__geomean	-0.9141
2	fraction2core__snowball100	-0.7368	fraction3core__snowball100	-0.7251
3	evectorCentrality__fit_lognormal	-0.7531	evectorCentrality__fit_wbl	-0.7530
4	assortativeCoefficient__snowball100	0.6011	assortativeCoefficient	0.5496
5	clusteringCoeff__iqr	-0.4891	clusteringCoeff__var	-0.4615

Table 3.5: Maximal correlated features with Isomap reduced features

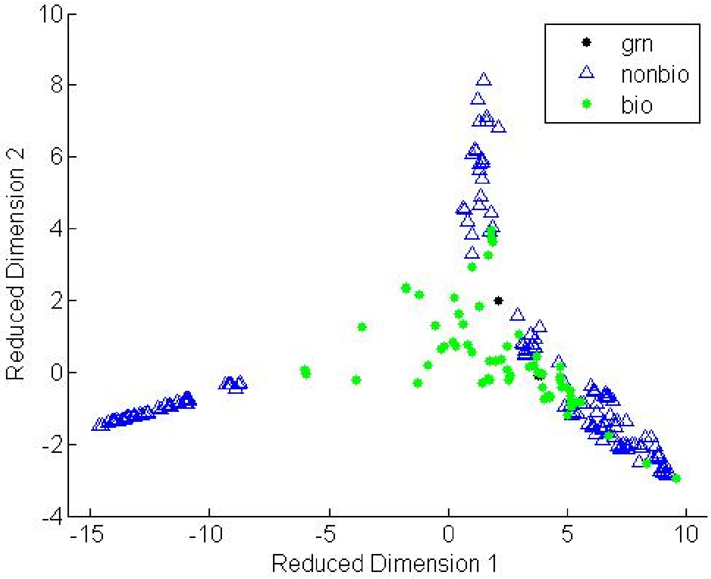


Figure 3.3: Network clustering via Isomap dimensionality reduction. Here, GRNs are shown in black color, biological networks in green and non-biological in black

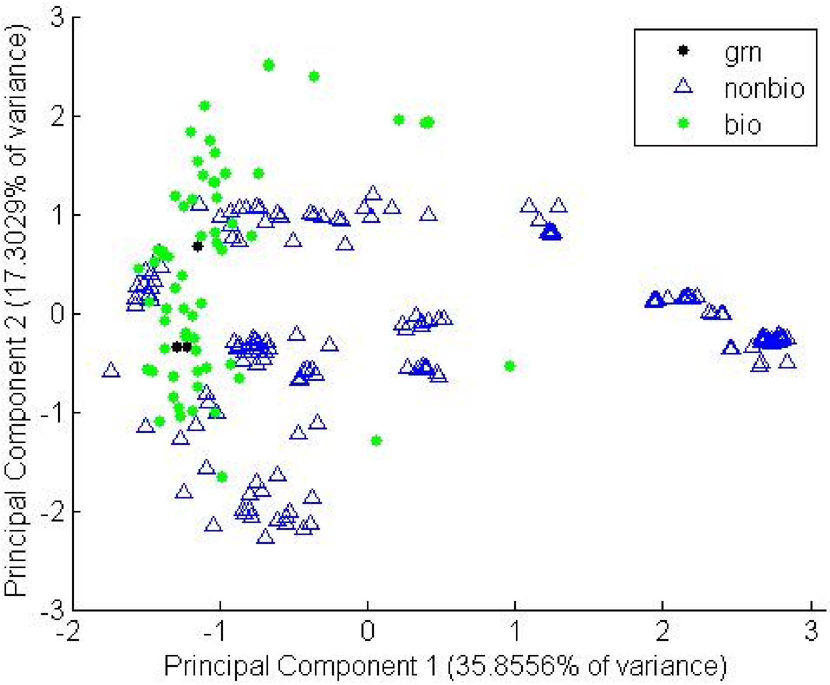


Figure 3.4: Network clustering via PCA dimensionality reduction. Here, GRNs are shown in black color, biological networks in green and non-biological in black

Dimension	Feature1	r1	Feature2	r2
1	clusteringCoeff__min	0.9409	transitivity	0.9307
2	evectorCentrality__posrms	0.8419	numNodes__snowball100	-0.7677
3	closeness__meanad	0.8109	closeness__medad	0.7997
4	degree__fit_gamma	-0.6329	degree__fit_wbl	-0.6327
5	betweenCentrality__posrms	0.4987	degreeCentrality__range	0.4858

Table 3.6: Maximum correlated features with PCA reduced features

We also tried to find out the distance between GRNs and the networks in our database. The measure you used to find the distance is Euclidean distance measure. As can be seen, the biological networks are closest to them.

Bsubtilis	Dream4	E.Coli
Rattus_norvegicus	Chlamydomonas_reinhardtii	Caenorhabditis_elegans
Caenorhabditis_elegans	Human_Herpesvirus_6	Rattus_norvegicus
DIP_Celeg_lcc	Leishmania_major	DIP_Celeg_lcc
biogrid_s_cerevisiae_lcc	Ustilago_maydis	biogrid_s_cerevisiae_lcc
fungal_4_11_lcc	fungal_17_2_lcc	fungal_4_11_lcc

Table 3.7: Networks with least Euclidean distance from GRNs

3.4 K-core and Number of edges

We saw that fraction-k-core is the most significant feature, so we dwelled deeper into this. We found that there are some networks which have low density but high fraction-k-core. In the Fig 3.5 and 3.6, we can see significant number of networks in the top right corner of the graph, which include a few biological networks.

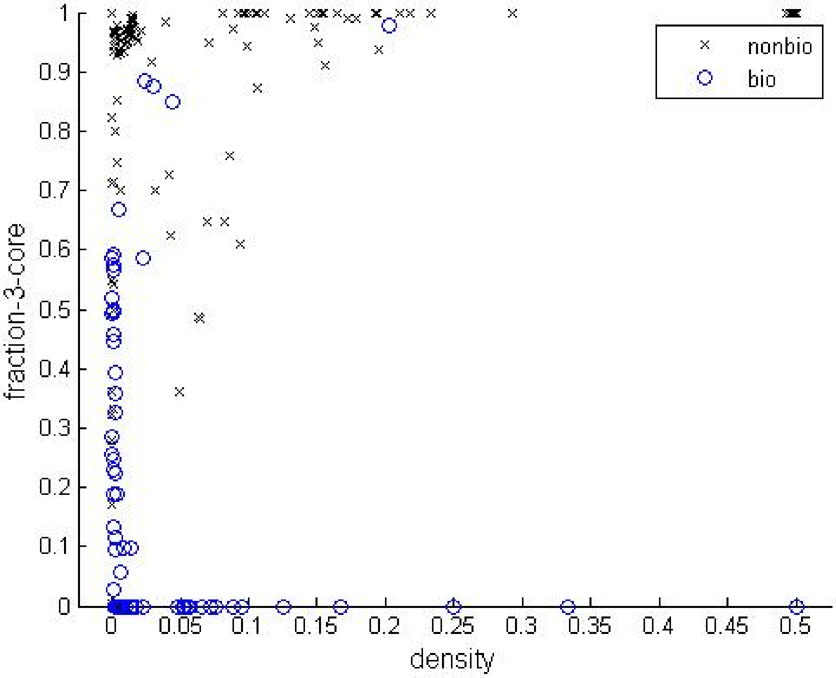


Figure 3.5: Network clustering in density and fraction-3-core space

So, we applied the method, which is defined for the genrative models, for our real world networks. We found out that this property was partially able to explain this behavior of real world networks. In Fig 7 and 8, we see that for $\rho_k = 10$, (which means number of edges as less as $\tilde{4}$ times the number of nodes) the fraction-k-core can be as high as 1.

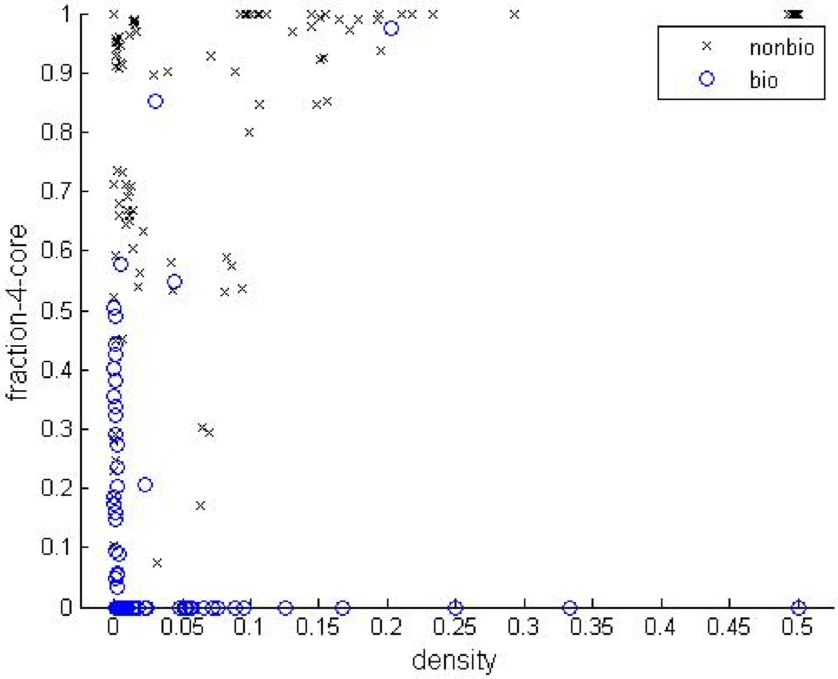


Figure 3.6: Network clustering in density and fraction-4-core space

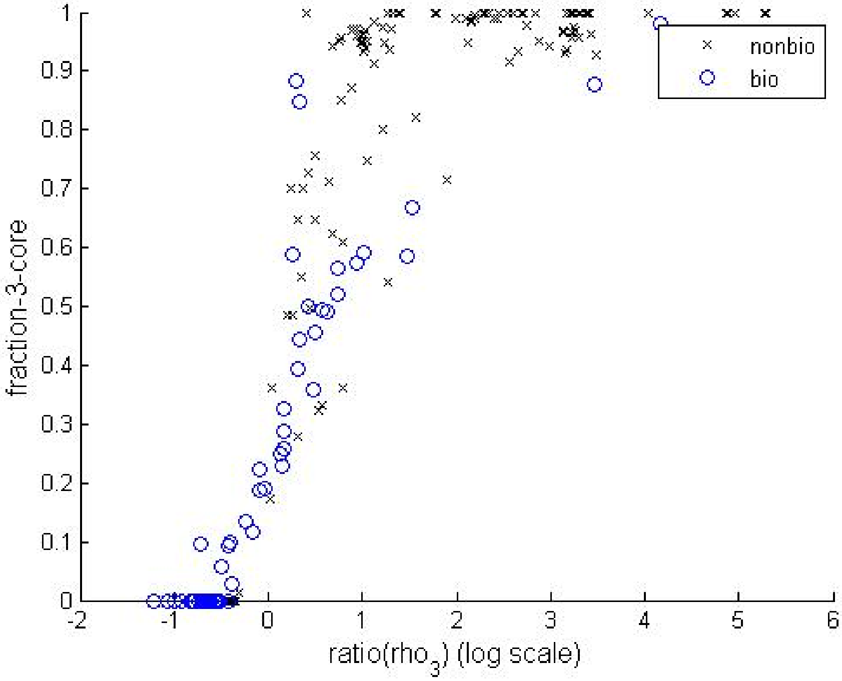


Figure 3.7: Network clustering in ρ_3 (log scale) and fraction-3-core space

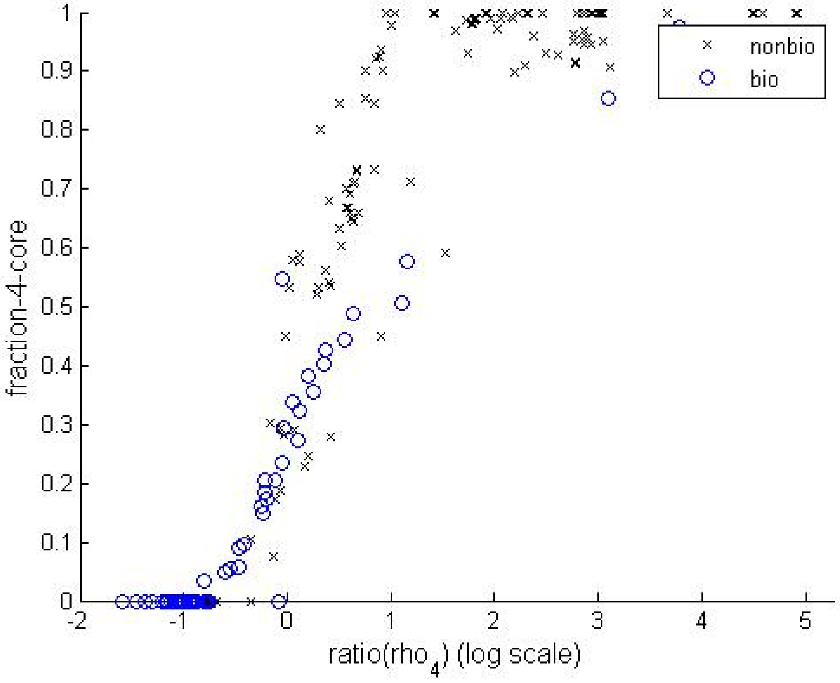


Figure 3.8: Network clustering in ρ_4 (log scale) and fraction-4-core space

Chapter 4

Conclusion

Till now, we saw that fraction-k-core features are the most signficant features that differ in biological and non biological networks. For biological networks, generally, the value of this feature is low. After that, there is clustering coefficient feature which has low value for biological networks. It is more scattered and sparse in nature. Then we have eigenvector centrality (max), which has large values for biological networks. But the mean of eigenvector centrality for biological and non biological networks seem to have similar distribution. From this, we can conclude that the biological networks may have some important nodes, i.e., it might have some proteins or molecules which is effecting the whole network the most. We apply SVM on the design matrix and it gives us clustering coefficient as the prominent feature.

All the methods we tried to find out the prominent features like svm, inter-class variance gave us inconsistent results. But we can say that the distinguishing factor between biological and non biological networks is that biological networks are sparse and small in size as compared to non biological networks.

Due to limited data available on GRNs, we can't conclude much. Seeing what we have, we can say that Bsubtilis and E.Coli have similar behaviours while Dream4 seems to deviate from them. Network clustering done using Isomap and PCA on design matrix with GRNs doesn't show much difference from what we observed for 234 networks. The maximal correlated features with first few principal components of Isomap reduced matrix seem to capture the degree distribution of nodes, fraction-k-core and size of the networks, while those of PCA seem to give more preference to clustering coefficient of a network. Since, the number of GRNs is just 3, we can't really say anything.

We observed an interesting feature in the networks. They had high fraction core even though the density was low. This can be partially explained by a research paper by *Pittel et al.*[17]. This information can also be helpful in thinking about what kind of networks they are and how can we develop an algorithm to generate models for the same.

Keeping the structural results that we got in mind, we can develop algorithms which constructs the network with the structural constraints. We can use ABC or some of the other methods like the Bayesian network[18] or ARACNE[19], where this can be used as a prior or can even act as a check for the generated networks.

Appendix A

List of Features and Statistics

Following is the list of all network diagnostics (features) used in this project. For each feature, shoet name is what that has been used in the text and code. For statistics, the short name is suffixed with the statistics name. (For example - clusteringCoeff__min means minimum clustering coefficient for the graph) The code for evaluating this project has been obtained from the thesis of Dr. Sumeet Agarwal [4].

Short Name	**Full Name**
Connectivity	
degree	Degree distribution
avgNearestNeighbourDegree	Average of degrees of adjacent nodes
assortativeCoefficient	Assortative coefficient
density fractionArticulation	Density Fraction of articulation nodes
erosionTime	Erosion Time
dilationTime	Dilation Time
fraction2core	Fraction of vertices comprising 2-core
fraction3core	Fraction of vertices comprising 3-core
fraction4core	Fraction of vertices comprising 4-core
richClub	Rich-club index
richClubNormalised	Normalised rich-club index
Centrality	
degreeCentrality	Degree centrality
degreeCentralityGroup	Group degree centrality
betweenCentrality	Betweenness centrality
betweenCentralityGroup	Group betweenness centrality
closeness	Closeness Group
closenessGroup	Closeness
evectorCentrality	Eigenvector centrality
subgraphCentrality	Subgraph centrality
subgraphCentralisation	Subgraph centralization
bipartivity	Estradas measure of bipartivity
infoCentrality	Information centrality
infoCentraliltyGroup	Group information centrality
vulnerability	Vulnerability
Community	

Short Name	**Full Name**
modularity	Spectrally optimized modularity
modularityFast	Louvain optimized modularity
greedyPartitionEntropy	Entropy of Louvain partition
spectral	Newmans spectral community detection
greedyComm	Louvain community detection
pottsModel	Potts model community detection
infomap	Infomap community detection
Clustering	
transitivity	Transitivity
clusteringCoeff	Clustering coefficient
clustSofferGlobalMean	Global mean Soffer clustering coefficient
clustSofferLocalMean	Local mean Soffer clustering coefficient
Distance	
diameter	Graph diameter
radius	Graph radius
szegedIndex	Szeged index
cyclicCoefficient	Cyclic coefficient
geodesicDistanceMean	Mean geodesic distance
geodesicDistanceVar	Variance of geodesic distance
harmonicMeanGeoDist	Harmonic mean geodesic distance
Complexity	
cyclomaticNumber	Cyclomatic number
edgeFraction	Edge fraction
connectivity	Connectivity
logNumSpanningTrees	log(number of spanning trees)
graphIndexComplexity	Graph index complexity
mediumArticulation	Medium articulation
efficiency	Efficiency
efficiencyComplexity	Efficiency complexity
offDiagonalComplexity	Off-diagonal complexity
chromaticNumber	Chromatic number
tspl	TSP length from cross-entropy algorithm
$tspl_{ga}$	TSP length from genetic algorithm
$tspl_{sa}$	TSP length from simulated annealing
Spectral	
largestEigenvalue	Largest eigenvalue
spectralScalingDeviations	Deviations from perfect spectral scaling
algebraicConnectivity	Algebraic connectivity
algebraicConnectivityVector	Algebraic connectivity vector
fiedlerValue	Fiedler value
Statistical physics energy	Energy
entropy	Entropy
Motif	
fraction3motifs	Fraction of 3-motifs
fraction4motifs	Fraction of 4-motifs

Short Name	**Full Name**
Size	
numNodes	Number of nodes
numEdges	Number of edges
totStrength	Sum of all link weights
Model	
ergm_edges	Exponential random graph model for edges
fitPowerLawAlpha	Fitted power law exponent for degrees
fitPowerLawP	p-value of power law fit to degrees

Table A.1: List of features

Central tendency	**Dispersion**	**Shape**	**Model fit log-likelihoods**
Mean	Minimum (min)	Kurtosis	Normal
Geometric mean (geomean)	Maximum (max)	Skewness	Log-normal
Harmonic mean (harmmean)	Variance (var)		Exponential
Mean excluding 10% outliers(trimmean10)	Range Inter-quartile range (iqr)		Extreme value
RMS of positive values (posrms)	Mean absolute deviation (meanad)		Gamma Weibull (wbl)
RMS of negative values (negrms)	Median absolute deviation (medad)		

Table A.2: List of statistics

Appendix B

List of 192 real world netowrks

This is the list of 192 real world networks used in this project obtained from the thesis of Dr. Sumeet Agarwal [4]

Name	Category
Human brain cortex: participant A1	Brain
Human brain cortex: participant A2	Brain
Human brain cortex: participant B	Brain
Human brain cortex: participant D	Brain
Human brain cortex: participant E	Brain
Human brain cortex: participant C	Brain
Cat brain: cortical	Brain
Cat brain: cortical/thalmic	Brain
Macaque brain: cortical	Brain
Macaque brain: visual/sensory cortex	Brain
Brain Macaque brain: visual cortex 1	Brain
Macaque brain: visual cortex 2	Brain
Co-authorship: astrophysics	Collaboration
Co-authorship: comp. geometry	Collaboration
Co-authorship: condensed matter	Collaboration
Co-authorship: Erdos	Collaboration
Co-authorship: high energy theory	Collaboration
Co-authorship: network science	Collaboration
Hollywood film music	Collaboration
Jazz collaboration	Collaboration
Facebook: Caltech	Facebook
Facebook: Cornell	Facebook
Facebook: Dartmouth	Facebook
Facebook: Georgetown	Facebook
Facebook: Harvard	Facebook
Facebook: Indiana	Facebook
Facebook: MIT	Facebook
Facebook: NYU	Facebook
Facebook: Oklahoma	Facebook
Facebook: Texas80	Facebook

Name	**Category**
Facebook: Trinity	Facebook
Facebook: UCSD	Facebook
Facebook: UNC	Facebook
Facebook: USF	Facebook
Facebook: Wesleyan	Facebook
NYSE: 1980-1999	Financial
NYSE: 1980-1983	Financial
NYSE: 1984-1987	Financial
NYSE: 1988-1991	Financial
NYSE: 1992-1995	Financial
NYSE: 1996-1999	Financial
Phanerochaete velutina control11-2	Fungal
Phanerochaete velutina control11-5	Fungal
Phanerochaete velutina control11-8	Fungal
Phanerochaete velutina control11-1	Fungal
Phanerochaete velutina control17-2	Fungal
Phanerochaete velutina control17-5	Fungal
Phanerochaete velutina control17-8	Fungal
Phanerochaete velutina control17-11	Fungal
Phanerochaete velutina control14-2	Fungal
Phanerochaete velutina control14-5	Fungal
Phanerochaete velutina control14-8	Fungal
Phanerochaete velutina control14-11	Fungal
Online Dictionary of Computing	Language
Online Dictionary Of Information Science	Language
Reuters 9/11 news	Language
Roget's thesaurus	Language
Word adjacency: English	Language
Word adjacency: French	Language
Word adjacency: Japanese	Language
Word adjacency: Spanish	Language
Metabolic: CE	Metabolic
Metabolic: CL	Metabolic
Metabolic: CQ	Metabolic
Metabolic: CT	Metabolic
Metabolic: DR	Metabolic
Metabolic: HI	Metabolic
Metabolic: NM	Metabolic
Metabolic: OS	Metabolic
Metabolic: PA	Metabolic
Metabolic: PG	Metabolic
Metabolic: PH	Metabolic
Metabolic: PN	Metabolic
Metabolic: SC	Metabolic
Metabolic: ST	Metabolic

Name	Category
Metabolic: TP	Metabolic
Bill cosponsorship: U.S. House 96	Political: cosponsorship
Bill cosponsorship: U.S. House 97	Political: cosponsorship
Bill cosponsorship: U.S. House 98	Political: cosponsorship
Bill cosponsorship: U.S. House 99	Political: cosponsorship
Bill cosponsorship: U.S. House 100	Political: cosponsorship
Bill cosponsorship: U.S. House 101	Political: cosponsorship
Bill cosponsorship: U.S. House 102	Political: cosponsorship
Bill cosponsorship: U.S. House 103	Political: cosponsorship
Bill cosponsorship: U.S. House 104	Political: cosponsorship
Bill cosponsorship: U.S. House 105	Political: cosponsorship
Bill cosponsorship: U.S. House 106	Political: cosponsorship
Bill cosponsorship: U.S. House 107	Political: cosponsorship
Bill cosponsorship: U.S. House 108	Political: cosponsorship
Bill cosponsorship: U.S. Senate 96	Political: cosponsorship
Bill cosponsorship: U.S. Senate 97	Political: cosponsorship
Bill cosponsorship: U.S. Senate 98	Political: cosponsorship
Bill cosponsorship: U.S. Senate 99	Political: cosponsorship
Bill cosponsorship: U.S. Senate 100	Political: cosponsorship
Bill cosponsorship: U.S. Senate 101	Political: cosponsorship
Bill cosponsorship: U.S. Senate 102	Political: cosponsorship
Bill cosponsorship: U.S. Senate 103	Political: cosponsorship
Bill cosponsorship: U.S. Senate 104	Political: cosponsorship
Bill cosponsorship: U.S. Senate 105	Political: cosponsorship
Bill cosponsorship: U.S. Senate 106	Political: cosponsorship
Bill cosponsorship: U.S. Senate 107	Political: cosponsorship
Bill cosponsorship: U.S. Senate 108	Political: cosponsorship
Committees: U.S. House 101, comms.	Political: committee
Committees: U.S. House 102, comms.	Political: committee
Committees: U.S. House 103, comms.	Political: committee
Committees: U.S. House 104, comms.	Political: committee
Committees: U.S. House 105, comms.	Political: committee
Committees: U.S. House 106, comms.	Political: committee
Committees: U.S. House 107, comms.	Political: committee
Committees: U.S. House 108, comms.	Political: committee
Committees: U.S. House 101, Reps.	Political: committee
Committees: U.S. House 102, Reps.	Political: committee
Committees: U.S. House 103, Reps.	Political: committee
Committees: U.S. House 104, Reps.	Political: committee
Committees: U.S. House 105, Reps.	Political: committee
Committees: U.S. House 106, Reps.	Political: committee
Committees: U.S. House 107, Reps.	Political: committee
Committees: U.S. House 108, Reps.	Political: committee
Roll call: U.S. House 101	Political: voting
Roll call: U.S. House 102	Political: voting

Name	**Category**
Roll call: U.S. House 103	Political: voting
Roll call: U.S. House 104	Political: voting
Roll call: U.S. House 105	Political: voting
Roll call: U.S. House 106	Political: voting
Roll call: U.S. House 107	Political: voting
Roll call: U.S. House 108	Political: voting
Roll call: U.S. Senate 101	Political: voting
Roll call: U.S. Senate 102	Political: voting
Roll call: U.S. Senate 103	Political: voting
Roll call: U.S. Senate 104	Political: voting
Roll call: U.S. Senate 105	Political: voting
Roll call: U.S. Senate 106	Political: voting
Roll call: U.S. Senate 107	Political: voting
Roll call: U.S. Senate 108	Political: voting
U.K. House of Commons voting: 1992-1997	Political: voting
U.K. House of Commons voting: 1997-2001	Political: voting
U.K. House of Commons voting: 2001-2005	Political: voting
U.N. resolutions 59	Political: voting
U.N. resolutions 60	Political: voting
U.N. resolutions 61	Political: voting
U.N. resolutions 62	Political: voting
Biogrid: A. thaliana	Protein interaction
Biogrid: C. elegans	Protein interaction
Biogrid: D. melanogaster	Protein interaction
Biogrid: H. sapiens	Protein interaction
Biogrid: M. musculus	Protein interaction
Biogrid: R. norvegicus	Protein interaction
Biogrid: S. cerevisiae	Protein interaction
Biogrid: S. pombe	Protein interaction
DIP: H. pylori	Protein interaction
DIP: H. sapiens	Protein interaction
DIP: M. musculus	Protein interaction
DIP: C. elegans	Protein interaction
Human: CCSB	Protein interaction
Human: OPHID	Protein interaction
Protein: serine protease inhibitor (1EAW)	Protein interaction
Protein: immunoglobulin (1A4J)	Protein interaction
Protein: oxidoreductase (1AOR)	Protein interaction
STRING: C. elegans	Protein interaction
STRING: S. cerevisiae	Protein interaction
Yeast: Oxford Statistics	Protein interaction
Yeast: DIP	Protein interaction
Yeast: DIPC	Protein interaction
Yeast: FHC	Protein interaction
Yeast: FYI	Protein interaction

Name	**Category**
Yeast: PCA	Protein interaction
Corporate directors in Scotland (1904-1905)	Social
Corporate ownership (EVA)	Social
Dolphins Family planning in Korea	Social
Unionization in a hi-tech firm	Social
Communication within a sawmill on strike	Social
Leadership course	Social
Les Miserables	Social
Marvel comics	Social
Mexican political elite	Social
Pretty-good-privacy algorithm users	Social
Prisoners	Social
Bernard and Killworth fraternity: observed	Social
Bernard and Killworth fraternity: recalled	Social
Bernard and Killworth HAM radio: observed	Social
Bernard and Killworth HAM radio: recalled	Social
Bernard and Killworth office: observed	Social
Bernard and Killworth office: recalled	Social
Bernard and Killworth technical: observed	Social
Bernard and Killworth technical: recalled	Social
Kapferer tailor shop: instrumental (t1)	Social
Kapferer tailor shop: instrumental (t2)	Social
Kapferer tailor shop: associational (t1)	Social
Kapferer tailor shop: associational (t2)	Social
University Rovira i Virgili (Tarragona) e-mail	Social
Zachary karate club	Social

Table B.1: List of 192-real world networks

Appendix C

List of 42 Biogrid Networks

These are 42 protien interaction metworks used in this project.

Name	**Category**
Anopheles_gambiae	Protein interaction
Arabidopsis_thaliana	Protein interaction
Aspergillus_nidulans	Protein interaction
Bacillus_subtilis	Protein interaction
Bos_taurus	Protein interaction
Caenorhabditis_elegans	Protein interaction
Candida_albicans_SC5314	Protein interaction
Canis_familiaris	Protein interaction
Cavia_porcellus	Protein interaction
Chlamydomonas_reinhardtii	Protein interaction
Cricetulus_griseus	Protein interaction
Danio_rerio	Protein interaction
Dictyostelium_discoideum_AX4	Protein interaction
Drosophila_melanogaster	Protein interaction
Equus_caballus	Protein interaction
Escherichia_coli	Protein interaction
Gallus_gallus	Protein interaction
Hepatitus_C_Virus	Protein interaction
Homo_sapiens	Protein interaction
Human_Herpesvirus_1	Protein interaction
Human_Herpesvirus_2	Protein interaction
Human_Herpesvirus_3	Protein interaction
Human_Herpesvirus_4	Protein interaction
Human_Herpesvirus_5	Protein interaction
Human_Herpesvirus_6	Protein interaction
Human_Herpesvirus_8	Protein interaction
Human_Immunodeficiency_Virus_1	Protein interaction
Human_Immunodeficiency_Virus_2	Protein interaction
Leishmania_major	Protein interaction
Macaca_mulatta	Protein interaction
Mus_musculus	Protein interaction

Name	Category
Neurospora_crassa	Protein interaction
Oryctolagus_cuniculus	Protein interaction
Oryza_sativa	Protein interaction
Pan_troglodytes	Protein interaction
Plasmodium_falciparum_3D7	Protein interaction
Rattus_norvegicus	Protein interaction
Ricinus_communis	Protein interaction
Saccharomyces_cerevisiae	Protein interaction
Schizosaccharomyces_pombe	Protein interaction
Simian-Human_Immunodeficiency_Virus	Protein interaction
Sus_scrofa	Protein interaction
Ustilago_maydis	Protein interaction
Xenopus_laevis	Protein interaction
Zea_mays	Protein interaction

Table C.1: List of 42 Biogrid networks

Bibliography

[1] T. Ideker, T. Galitski, and L. Hood. A New Approach To Decoding Life: Systems Biology. Annu. Rev. Genomics Hum. Genet. 2001. 2:34372

[2] H. Kitano. Systems Biology: A Brief Overview. Science 295, 1662 (2002); DOI: 10.1126/science.1069492

[3] H. Kitano. Computational Systems Biology. Nature 420, 206-210 (14 November 2002)

[4] S. Agarwal. Networks in Nature: Dynamics, Evolution and Modularity. Ph.D. thesis, University of Oxford (2012).

[5] S. Agarwal, G. Villar, and N. S. Jones. Comparative network analysis. In preparation.

[6] E. N. Gilbert. Random graphs. Annals of Mathematical Statistics, 30(4):11411144 (1959).

[7] P. Erdös and A. Rényi. On random graphs I. Publicationes Mathematicae, 6:290297 (1959).

[8] Albert-László Barabási and Zoltán N. Oltvai. Network biology: Understanding the cell's functional organization. Nature Reviews Genetics, 5:101-113, 2004.

[9] R. Bonneau. Learning biological networks: from modules to dynamics. Nature Chemical Biology 4:658-664 (2008).

[10] R. Bonneau, D. J. Reiss, P. Shannon, M. Facciotti, L. Hood, N. S Baliga and V. Thorsson. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets *de novo*. Genome Biology 2006, 7:R36

[11] A. Srinivasan, R. D. King, Incremental Identification of Qualitative Models of Biological Systems using Inductive Logic Programming.

[12] en.wikipedia.org/wiki/Gene

[13] en.wikipedia.org/wiki/Protein-protein_interaction

[14] http://docs.opencv.org/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html

[15] http://spider.ipac.caltech.edu/staff/fmasci/home/statistics_refs/PrincipalComponentAnalysis.pdf

[16] J. B. Tenenbaum, V. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. Science, 290(5500):2319-2323 (2000).

[17] B. Pittel, J. Spencer, N. Wormald. Sudden Emergence of a Giant k-core in a Random Graph. Journal of Combinatorial Theory, Series B 67, 111-151 (1996).

[18] M Vignes, J Vandel, D Allouche, N Ramadan-Alban, C Cierco-Ayrolles, et al. (2011) Gene Regulatory Network Reconstruction Using Bayesian Networks, the Dantzig Selector, the Lasso and Their Meta-Analysis. PLoS ONE 6(12): e29165. doi:10.1371/journal.pone.0029165

[19] A A Margolin, I Nemenman, K Basso, C Wiggins, G Stolovitzky, R D Favera and A Califano. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. BMC Bioinformatics 2006, 7(Suppl 1):S7