University of Tennessee, Knoxville

Trace: Tennessee Research and Creative Exchange

Masters Theses

Graduate School

8-2008

Comparative Analysis of Thresholding Algorithms for Microarray-derived Gene Correlation Matrices

Bhavesh Ram Borate

University of Tennessee - Knoxville

To the Graduate Council:

I am submitting herewith a thesis written by Bhavesh Ram Borate entitled "Comparative Analysis of Thresholding Algorithms for Microarray-derived Gene Correlation Matrices." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Life Sciences.

Michael Langston, Major Professor

We have read this thesis and recommend its acceptance:

Arnold Saxton, Elissa Chesler, Brynn Voy

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

I am submitting herewith a thesis written by Bhavesh Ram Borate entitled "Comparative Analysis of Thresholding Algorithms for Microarray-derived Gene Correlation Matrices." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Life Sciences.

Prof. Michael Langston,
Major Professor

We have read this thesis
and recommend its acceptance:

Prof. Arnold Saxton

Prof. Elissa Chesler

Prof. Brynn Voy

Accepted for the Council:

Carolyn R. Hodges,
Vice Provost and Dean of the Graduate School

(Original signatures on file with student records.)

Comparative Analysis of Thresholding Algorithms for Microarray-derived Gene Correlation Matrices

A Thesis Presented for
the Master of Science
Degree
The University of Tennessee, Knoxville

Bhavesh Ram Borate

August 2008

ACKNOWLEDGEMENTS

Dr. Mike Langston and Dr. Arnold Saxton for their encouragement, ideas and constant support. Dr. Elissa Chesler and Dr. Brynn Voy for their insight and ideas when things blurred out for me.

John Eblen, Andy Perkins, Gary Rogers, Yun Zhang and all the wonderful students under Dr. Langston. For assisting me around and making me feel at home. Especially John, a wonderful friend and colleague. For helping me out with Perl and UNIX programming. And giving me sufficient insight in graph theory so as to be able to write about it.

Dr. Bing Zhang and Dr. Roumyana Yordanova for their help on certain topics of the study.

The GST program and its current and former Directors, Dr. Peterson and Dr. Becker, for giving me an opportunity to study at University of Tennessee, Knoxville.

ABSTRACT

The thresholding problem is important in today's data-rich research scenario. A threshold is a well-defined point in the data distribution beyond which the data is highly likely to have scientific meaning. The selection of threshold is crucial since it heavily influences any downstream analysis and inferences made there from. A legitimate threshold is one that is not arbitrary but scientifically well grounded, data-dependent and best segregates the information-rich and noisy sections of data. Although the thresholding problem is not restricted to any particular field of study, little research has been done. This study investigates the problem in context of network-based analysis of transcriptomic data. Six conceptually diverse algorithms – based on number of maximal cliques, correlations of control spots with genes, top 1% of correlations, spectral graph clustering, Bonferroni correction of p-values and statistical power – are used to threshold the gene correlation matrices of three time-series microarray datasets and tested for stability and validity. Stability or reliability of the first four algorithms towards thresholding is tested upon block bootstrapping of arrays in the datasets and comparing the estimated thresholds against the bootstrap threshold distributions. Validity of thresholding algorithms is tested by comparison of the estimated thresholds against threshold based on biological information. Thresholds based on the modular basis of gene networks are concluded to perform better both in terms of stability as well as validity. Future challenges to research the problem have been identified. Although the study utilizes transcriptomic data for analysis, we assert its applicability to thresholding across various fields.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I
INTRODUCTION

Advancements in technology have helped churn out a great amount of data in all sectors of industry and education. Computers and statistical procedures are indispensable for the handling and analysis of this vast amount of data and to parse the signal from noise.

One significant problem frequently encountered by data analysts is the identification of a threshold above which most of the data is highly likely to have scientific meaning. This is especially so when values at a particular end of data distribution are more informative. The application of a threshold also limits analysis to only meaningful regions of data and thus helps to make huge datasets manageable.

Many times a threshold assists in making good use of limited resources to solve difficult problems. For example, regardless of advancements in computer technology, there are some problems, which are *NP*-complete, that most likely will remain difficult to solve in real time. A threshold, besides eliminating meaningless data, also can make such problems tractable.

The applications of thresholding are vast and span a wide spectrum of fields (biology, business, economics). However, the issue needs to be addressed under light of the characteristics of data being analyzed.

This study focuses thresholding as an application for network-based analysis of *transcriptomic* data. *Transcriptomics* is the systematic and simultaneous analysis of expression profiles of thousands of genes. DNA microarray technology was developed to carry out such analysis [Schena et al. 1995]. With continual improvements in this technology, the field of transcriptomics has been successful in making a significant contribution to medical health research [Simon et al. 2002, Tefferi et al. 2002, Lorentz et al. 2002, Elkin 2003]. This, in turn, has attracted a wide range of statistical concerns [Smyth et al. 2003, Mayo et al. 2006].

An important effort behind a microarray experiment is towards discerning sets of co-expressed genes. In order to do this, a network-based approach is routinely used to represent the complexity of gene interactions [Dehmer and Emmert-Streib 2008]. Such a representation with genes as nodes and co-expression measures as edges between them is both intuitive and straightforward. However, the application of a threshold to such a network is not easy and involves the complex task of balancing the number of false positives and false negatives in the data. The problem gets worse as the level of noise in the data increases.

Two philosophies for thresholding have been investigated in reference to biological networks: hard and soft [Zhang and Horvath 2005]. The principle difference between them is that hard thresholding utilizes correlations between gene pairs as edge-weights and thus takes into account individual pair-wise relationships between genes, while soft thresholding involves the assignment of connection edge-weights based on graph characteristics such as ‘scale-free topology criterion’ and considers modular relationships between genes. Zhang and Horvath (2005) have shown that threshold based on aggregate, modular relationships between genes yields more robust results than individual pair-wise relationships.

In this study, we compare and analyze six conceptually different algorithms – based on number of Maximal Cliques, correlations of Control Spots with genes, Top 1% of correlations, Spectral graph clustering, Bonferroni correction of pvalues and statistical Power – used towards thresholding the gene correlation matrix derived from microarray data that was pre-verified statistically to be of high quality. Importantly, two of the methods (Maximal Clique algorithm and Spectral graph clustering) consider aggregate gene relationships to arrive at a threshold while the rest of them consider only pair-wise relationships. The ***objectives*** of this study were 1) to evaluate thresholding method/s for stability and reliability: identifying ones that exhibit a high level of robustness and 2) to evaluate thresholding method/s for validity: identifying those that accommodate maximum biological information with a relatively low noise component.

The results of our analysis help to assess the relative performance of each of the methods for thresholding the gene correlation matrix. We hope to apply the conclusions from this study in our quest towards generating ‘combinatorial’ algorithms for threshold determination. The general applicability of the thresholding methods used in this study should serve as a guide to data analysts into choosing a suitable threshold not only for transcriptomic but also for data in other fields.

CHAPTER II
BACKGROUND

Gene expression is a dynamic process that is tightly linked to activities within the cell. Genes are either turned on (increase in expression) or turned off (decrease in expression) such that the resulting products of gene expression (proteins) can drive complex cellular pathways to satisfy the continuous needs of the cell. Since cellular pathways rely on a spectrum of proteins to be activated or deactivated, genes expressing such proteins tend to display similar expression patterns [van Noort et al. 2003]. Thus, gene expression is orchestrated in aggregates and the exploration of such aggregate relationships provides important insight for the dissection of cellular pathways.

Biological relationships are complex. The thousands of genes within a cell can take part in more than one cellular pathway. Further, cellular pathways are intricately linked to each other and assuming them to be autonomous is a drastic over-simplification. Thus, extracting meaningful biological relationships from gene expression data is difficult. A significant level of noise routinely present within the data further complicates the picture.

Early microarray studies derived conclusions from simplifying this picture. Some studies considered only pair-wise relationships between genes [Stuart et al. 2003, Moriyama et al. 2003, Sanoudou et al. 2003]. Other studies considered small-scale networks by limiting analysis to only genes of interest [Bredel et al. 2005]. However, with advent of genome-scale transcriptomic studies [Szodoray et al. 2006, Anisimov et al. 2007], microarray analysis has matured to model large sets of aggregate relationships between genes.

Many reports have highlighted the network architecture as an abstract schematization of biological systems [Alon 2003, Barabasi and Oltvai 2004, Oltvai and Barabasi 2002]. The depiction of genes as nodes and edges as relationships between them [Bader and Enright 2005] revealed the scale-free nature of biological networks [Jeong et al. 2000, Bray 2003, Albert 2005, Aloy and Russell 2004]. A typical large-scale gene network displays important hubs and sub-graphs with high connectivity. Sub-graphs within a network represent genes with similar patterns of expression and afford valuable clues to intra-cellular pathways [Wolfe et al. 2005, Eisen et al. 1998, Wu et al. 2002, Stuart et al. 2003]. Graph and network analysis techniques have been utilized to extract such sub-graphs and derive biologically meaningful relationships that look beyond just pair-wise associations [Voy et al. 2006, Yan et al. 2007, Freeman et al. 2007].

The analysis of such genome-scale gene networks involves a classical thresholding problem. First, given a particular weight to the edges in the graph, only edge-weights at the higher end of the distribution tend to contain significant biological meaning. Second, genome-scale data is huge and storage and analysis of it as a whole encounters tremendous difficulties. Thus, thresholding becomes an important issue in the analysis of gene networks. Butte et al. (2000) first highlighted the issue by introduction of the concept of relevance networks.

Relevance Networks

An expression data matrix – the outcome of a typical microarray experiment – is an n-by-p matrix, where each of the n rows corresponds to a gene and each of the p columns corresponds to an array [Mayo et al. 2006]. Similarity metric measures like Spearman's rank or Pearson's correlation coefficient and Euclidean distance are used by various algorithms to quantify co-expression between pairs of genes, producing an n-by-n gene correlation matrix [Slonim 2002, Allison et al. 2006, Voy et al. 2006]. Relevance networks are created after thresholding the matrix of similarity metric such that the resulting graph – with vertices as genes and similarity metric as edge-weights – has only edge-weights that exceed the threshold value [Butte et al. 2000].

Relevance networks incorporate a higher number of edges that would imply significant biological meaning [Butte and Kohane 2000]. Subsequent extraction of sub-graphs from such a network has the advantage of producing more biologically meaningful results besides being faster from elimination of insignificant data. Extraction of sub-graphs from within relevance networks has been very well documented in recent literature to yield sets of co-expressed genes [Voy et al. 2006, Yan et al. 2007, Freeman et al. 2007].

Amongst these, the utility of cliques as sub-graphs in biology deserves special mention. Many studies have demonstrated the use of cliques to depict important relationships in biological systems [Wu and Li 2007, Setubal and Meidanis 1997] and specifically to extract “putative sets of co-expressed genes” from microarray data [Voy et al. 2006, Manfield et al. 2006].

Clique

Clique is a sub-graph in which all the nodes are connected to each other. Within a relevance network, cliques represent “putative sets of co-expressed genes” [Voy et al. 2006]. Solving such a network for cliques however is a “classic graph-theoretic problem” [Bomze et al. 1999] and is *NP* (Non-deterministic Polynomial time)-complete [Zhang et al. 2005, Garey and Johnson 1979]. Researchers have successfully used vertex cover to solve the clique problem on massive scales [Zhang et al. 2005, Langston 2004, Fellows and Langston 1994].

Cliques or any sub-graphs are found on an unweighted graph, which is derived from a weighted graph. This transformation represents a binary decision problem and requires the selection of a threshold. Also, many other graph-theoretic problems require the application of a threshold to a weighted graph and analyze the subsequent unweighted graph. A few of them with applications towards microarray analysis are enumerated below:

1. Enumeration of maximal cliques: required for gene expression network analysis, cis-regulatory motif finding, investigation of QTL’s for high-throughput molecular phenotypes [Zhang et al. 2005, Abu-Khzam et al. 2005].
2. Finding a maximum clique: used to find paraclique, a “noise-adaptive graph algorithm” [Chesler and Langston 2005] that also addresses the

issue of false-negatives encountered from using cliques as a clustering technique.

3. Vertex cover: used to identify transcripts that relate individual phenotypes to QTL regulatory models [Chesler and Langston 2005].
4. CAST (Cluster Affinity Search Technique): a clustering technique by Ben-Dor et al. (1999) that has been reported to cluster gene expression data well.

Though algorithms to the above-mentioned graph-theoretical problems have been developed and applied to microarray data analysis with certain amount of success, the results of such sub-graph analysis (clique or otherwise) heavily depend on the selection of an appropriate threshold. In case of a correlation matrix of genes, the dense distribution of correlations gives rise to a completely different picture of the graph (with addition or removal of many edges/nodes) even with a slight change in threshold, which in turn, impacts any sub-graphs extracted from the graph. As the final results of any microarray data analysis heavily influence further scientific investigation in biological laboratories, selection of an appropriate threshold becomes an important step in a network-based approach that needs to be addressed adequately.

Figure 1 (all figures and tables are located in the appendix) illustrates the flow of microarray data processing in a typical graph-based analysis. For details regarding the approach, the reader can refer to Voy et al. (2006).

Non-thresholding alternatives to analyze gene correlation matrix

Many clustering algorithms have been used to segregate transcriptomic data into distinct, closely related units [Bellaachia et al. 2002, Ben-Dor et al. 2000, Ben-Dor et al. 1999, Hansen and Jaumard 1997, Hartuv et al. 1999]. A good review is presented in Quackenbush (2001). A variety of metrics could be used to cluster transcriptomic data [Quackenbush 2001], some similar to the metrics used for relevance networks like Spearman’s rank or Pearson’s correlation coefficient and Euclidean distance [Slonim 2002, Allison et al. 2006, Allocco et al. 2004].

However, one principle difference between clustering algorithms and relevance networks is the application of a threshold in case of the latter.

Although some clustering algorithms can circumvent the thresholding problem, there are many arguments that can be raised against them or in favor of relevance networks based analysis (eg. cliques and other subgraphs).

The clusters generated by most clustering algorithms are disjoint [Voy et al. 2006]. So a particular gene can lie only in one or the other cluster. This is contrary to what is observed in biological networks, where any particular gene (gene-product) could participate in more than one network [Rajasekaran et al. 2005, Kim and Chung 2002, Lopez and Martinez 2002]. Though researchers have utilized Singular Value Decomposition (SVD) as a solution to the disjoint issue [Alter et al. 2000], the factors involved in SVD are not easily assigned a biological interpretation [Girolami and Breitling 2004]. Cliques, on the other hand, are not disjoint and so are able to mimic biological networks better.

Cliques also represent negative correlations [Chesler and Langston 2005, Voy et al. 2006], which signify biologically inverse gene relationships. Many clustering algorithms do not do so.

Cliques actually are a way to represent overlapping sets of highly connected nodes in a graph. Thus, clique can be thought of as a type of clustering technique. However, the occurrence of a clique entails an extremely stringent criterion: for the removal of even a single edge from the clique and the clique is lost. In the context of microarrays, this stringency brings forth a tremendous advantage as it serves to brace the research analysis from very high level of variability and noise factors routinely observed in microarray data, and consequently reduces the number of biological false positives [Baldwin et al. 2005].

In graph theory, a cluster's edge density is assessed upon solving the k-subgraph problem [Feige et al. 2001, Rougemont and Hingamp 2003, Watts and Strogatz 1998, Baldwin et al. 2005]. Since a clique is a cluster in which all nodes are connected, it serves to maximize this edge density. However, a drawback of this is the occurrence of high number of false negatives. This drawback is addressed upon introduction of the “paraclique” concept [Chesler and Langston 2005].

Recent studies have highlighted the stochastic (instead of deterministic) nature of biological networks [Quackenbush 2007, Elowitz et al. 2002, Ozbudak et al. 2002]. Once an interesting clique is found, other genes interacting with any of the nodes in the clique, but having an edge-weight lower than the threshold, could be easily recovered from the expression profile thus accommodating for stochasticity. This also serves to recover genes that may show *transitive co-expression* with other genes in a clique and so do not have sufficient correlation to be part of the clique [Zhou et al. 2002].

Studies have shown that metric-based clustering algorithms tend to cluster genes that have very low similarities in expression [Allocco et al. 2004]. Thus, it is not surprising to find unregulated genes clustered together. Relevance networks, through the simple use of a threshold, discard insignificant data from consideration thus reducing the probability of such occurrences.

Besides, individual clustering methods contribute their own set of inherent drawbacks. K-means clustering requires prior knowledge of number of clusters into which the data needs to be segregated [Quackenbush 2001]. Dougherty et al. (2002) have shown that the algorithm will generate clusters even in random data.

Hierarchical clustering ends up clustering every data point thus generating errors. Also, the final results tend to be biased by the properties of the genes that have defined the clusters initially [Quackenbush 2001]. Visual interpretation of the resulting dendrograms also has its subjective flaws [Voy et al. 2006].

Self-Organizing Maps require a geometric configuration for partitioning nodes into clusters, a problem similar to that linked with K-means clustering [Quackenbush 2001].

Thus, the inherent limitations of traditional clustering techniques prevent them from depicting biologically meaningful relations between genes. Cliques and other sub-graphs, on the other hand, have been found to obtain interpretable results from mining of relevance networks.

Current approaches to Thresholding

The thresholding problem we address concerns the application of a cut-off point to the similarity measure matrix so as to only consider co-expression between pairs of genes that are greater or equal to the threshold value. Many studies have chosen an arbitrary threshold of 0.8 [Bredel et al. 2005, Sanoudou et al. 2003]. The problem with selecting such an arbitrary threshold is that it does not take into account the inherent properties of the data.

Allocco et al. (2004) conducted a microarray study with *Saccharomyces cerevisiae* using 611 arrays over a wide range of conditions to show that gene coexpression is linked to the sharing of common transcription factor binding sites. Specifically, they concluded that at a correlation of 0.84, there is a 50% chance of sharing a common transcription factor binding site between two genes. However, due to high variability arising from multiple sources in microarray data, results from a single lab cannot be taken as a standard and applied across all datasets.

Moriyama et al. (2003) obtained random correlation distributions for gene pairs by permuting their expression values. They defended their choice of threshold based on the statistical significance levels (p-value < 0.001, 0.01, 0.05). Although such a method is statistically strong, it may not necessarily yield biologically significant relationships [Quackenbush 2003]. Voy et al. (2006) also discussed a similar picture: Biologically meaningful relationships from small experiments (low number of arrays) could fail to have a statistically strong base due to insufficient power; conversely, biologically insignificant relationships from large experimental designs may display statistically significant relationships as a result of a higher statistical power. Thus, purely statistical methodologies may not work best for extracting biologically meaningful relationships from relevance networks.

Lee et al. (2004) considered only the top 1% of correlations (absolute value) for each dataset and built a co-expression network for multiple human microarray datasets.

Zhang and Horvath (2005) selected parameters for ‘soft’ thresholding based on the scale-free topology criterion that serves to optimize the biological signal. Such a criterion is based on the fact that gene co-expression networks often appear to satisfy approximate scale-free topology [Jeong et al. 2000, Bergman et al. 2004].

Langston et al. (2006) recommend the use of ontological distance, statistical significance and various graph structural attributes to arrive at a correlation threshold.

Voy et al. (2006) used distribution of correlations of genes with buffer spots on the arrays to select a threshold of 0.875, at which the correlation values dropped down to a mere few. They supported their selection by evaluating a statistically significant confidence level at this threshold by using Fisher’s z-transform and Bonferroni correction for multiple testing.

However, many of the above reports mention the need of a thorough analysis of the issue.

Major issues with Thresholding

Two important philosophies have been investigated in relation to thresholding. Hard thresholding considers gene affiliations as independent pairs. Soft thresholding, on the other hand, takes account of aggregate modular gene relationships, which closely mimics the real-world biological network model. The weak biological basis of hard thresholding makes it very susceptible to loss of information, besides being extremely sensitive to the chosen threshold [Carter et al. 2004]. Zhang and Horvath (2005) have shown that threshold based on aggregate, modular relationships between genes yields more robust results than individual pair-wise relationships.

Thresholding has also been studied in the statistical significance framework. This applies more in relation to hard thresholding: the selection of threshold value has been based by many investigators [Davidson et al. 2001, Butte and Kohane 2000, Carter et al. 2004] on the significance level of correlation coefficient rather than directly on the correlation coefficient. In the context of genome-wide studies, such an issue predominantly involves the problem of multiple hypothesis testing [Dudoit et al. 2003]. The family-wise-error-rate (FWER) becomes too conservative in defining the critical value of rejection region, especially when the number of tests is very large [Storey 2002]. On the other hand, the false discovery rate (FDR: the expected proportion of false-positives amongst all the rejected hypotheses) is relatively liberal and more powerful measure of error [Benjamini and Hochberg 1995, Storey and Tibshirani 2003]. Thus, the FDR provides a valuable alternative to the FWER in discovery-based settings like microarrays, where scientists are willing to accommodate a few false positives provided their numbers are very small as compared to the total number of rejected hypotheses. Similar to the widely used p-value, the q-value [Storey 2002, Storey 2003] is also a measure of statistical significance. However, it is based on the FDR unlike the p-value, which is based on the false positive rate [Storey and Tibshirani 2003]. There is an important difference here: the false positive rate is the rate that truly null features (a feature being any attribute of the genome-wide study that needs to be statistically evaluated, eg. correlation coefficient measure between genes) are identified to be significant while the FDR is the rate that significant features are truly null. A q-value assesses statistical significance on the basis of significant features while a p-value does the same on the basis of features that are truly null. Thus, a q-value provides a measure of false positive to true positive results and in context of genome-wide studies, offers a statistical significance that has a better practical interpretation [Storey and Tibshirani 2003].

Developments in computer science and data visualization are also influencing the way researchers pursue the issue of thresholding. New et al.

(2008) have developed dynamic visualization tools to track changes occurring in large-scale, real-world gene co-expression networks as threshold is raised or lowered. This gives a definite advantage to researchers in visualizing dynamic, real-time developments taking place in interesting gene modules within the network and work around with different thresholds before ultimately deciding on any particular one. Such a dynamic visualization can be applied in both hard as well as soft thresholding scenarios.

CHAPTER III
METHODS

Datasets chosen for study

Microarray data for the yeast *Saccharomyces cerevisiae* was chosen for this study. The very complete annotation reported for the *S.cerevisiae* genome (around 80% as reported on the *Saccharomyces cerevisiae Genome Database website: http://www.yeastgenome.org/*) influenced this selection. Moreover, information on transcription factors for many *S. cerevisiae* genes is already available, which could be used in future studies to assess biological information more accurately.

Anoxia/Reoxygenation data

This dataset was obtained from the *Saccharomyces cerevisiae Genome Database website: http://www.yeastgenome.org/.* Lai et al. (2006) have carried out microarray analysis in the yeast to identify gene networks that show metabolic-state dependent differences when yeast cells are exposed to anaerobic conditions with subsequent aerobic revival. One set of cells was grown in glucose-containing medium while the other in galactose-containing medium to bring out the metabolic-state differences. 31 arrays were used for the anoxic state (16 under galactose and 15 under glucose) while 21 arrays were used for the reoxygenation state (11 under galactose and 10 under glucose).

Yeast Cell Cycle analysis data

This is a standard dataset from the Eisen laboratory and is part of the Yeast Cell Cycle analysis project [Spellman et al. 1998].

For this dataset, we have:

1. 2 arrays (40 min and 30 min) from induction with G1 cyclin Cln3p
2. 2 arrays (both 40 min) from induction with B-type cyclin Clb2p

3. 18 arrays (every 7 min) from yeast cultures synchronized using Alpha-factor arrest
4. 14 arrays (every 30 min) from yeast cultures synchronized using elutriation
5. 24 arrays (every 10 min) from arrest of a cdc15 temperature-sensitive mutant
6. Also includes data from Cho et al., which has 17 arrays (every 10 min) from arrest of cdc28 temperature-sensitive mutant [Cho et al. 1998].

The 18 arrays from yeast cultures synchronized using Alpha-factor arrest have been used in our study. The 24 arrays from yeast cultures arrested with a cdc15 temperature-sensitive mutant were not considered for this study (even though the number of arrays in this dataset was higher) since the Gene Ontological (GO) Score evaluated for this dataset failed to show a rise at high positive correlations (Figure 18). A rise in GO Score at high correlations, which is expected to be seen in microarray datasets and was seen for the three datasets used in the study (Figure 17), was important to identify the inflection point for biological threshold determination using Gene Ontology.

Preliminary data processing

For all three datasets, the Locus Tags for genes were converted to GeneIDs using the gene_info.gz file available at NCBI ftp location (*ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/)*. A GeneID is a unique NCBI gene identifier assigned to annotated genes. This reduced the Anoxia and Reoxygenation datasets from 6212 to 5525 genes, while the Alpha dataset shrunk from 6178 to 5466 genes. The elimination of un-annotated genes prevented skewing of the threshold estimated on the basis of Gene Ontology, which we use as a measure to assess performance of the other thresholding methods. For the Control-Spot method, the datasets were modified by adding in control spot information to the end of data.

Exploratory data analysis carried out using Principal Components Analysis, Box-and-Whiskers, Normal Quantile plots and evaluating pair-wise correlations between arrays failed to identify any outlier arrays. This also served to validate the good quality of the data used in the study. Pearson's correlation coefficient metric was evaluated between genes as it has been shown to contain greater amount of information as compared to Spearman's correlation metric [Voy et al. 2006].

Thresholding Algorithms

Six different algorithms for thresholding the gene correlation matrix have been analyzed and compared in this study. Software written by Langston and colleagues (University of Tennessee) including Datagen version 1.4a [Jon Scharff, private communication], Maximal clique enumeration code version 2.0.1 [Zhang et al. 2005], spectral analysis code [Perkins 2008] and GO Pairwise Similarity analysis code version 1.0, was used. Matrix calculations for spectral graph analysis were carried out in MATLAB 7.0. P-values were calculated in SAS version 9.1. Statistical power was calculated using PASS statistical software [*http://www.ncss.com/pass.html*].

Method 1: Threshold based on number of Maximal Cliques in graph

The algorithm is based on graph properties, specifically the distribution of the number of *maximal* cliques in the network as the threshold is lowered step-wise from a very high (0.99) to low correlation values. By definition, a *maximal* clique is a clique that cannot be expanded by the inclusion of any other vertex in the graph. This needs to be distinguished from a *maximum* clique, which is the largest clique in a graph [Zhang et al. 2005, Baldwin et al. 2005]. Tomita et al. (2004) showed that a network with *n* nodes could at the most have $3^{n/3}$ *maximal* cliques. Thus, as the threshold is lowered, the number of *maximal* cliques may grow exponentially with number of nodes included in the network. Two important issues come up here. First, as threshold is lowered, computational complexity of enumeration of maximal cliques increases. Also, at lower thresholds, noise or the number of false positives may increase. To circumvent both these issues, the algorithm looks for an inflection point where the number of maximal cliques grows to more than two times the previous value (Maximal Clique-2). Also, a simple modification of the algorithm looks for an inflection point where the number of maximal cliques grows to more than three times the previous value (Maximal Clique-3). The use of an inflection point makes the algorithm capable of adapting to the properties of the correlation matrix (distribution of correlation values), thus evaluating a different threshold for each dataset. However, the algorithm is dependent on selection of parametric value: selecting the inflection point where the number of maximal cliques grows to more than 2.5 or 3.5 should give a different threshold. We plan to eliminate such arbitrariness by applying data dependent techniques such as inflection point obtained on the basis of slope of the curve.

A major advantage of this methodology is that it depends on the occurrence of clique in the graph, which by itself is a very stringent criterion to guard against false positives. Moreover, since cliques represent putatively co-regulated sets of genes, using this information to arrive at a threshold seems biologically reasonable. To prevent the algorithm from halting at very high thresholds, another condition applied is that the number of maximal cliques be greater than a particular minimum value. We selected the value to be 50000 based on our experience with various microarray datasets.

Method 2: Threshold based on information extracted from Control Spots

Control spots are spots distributed throughout the microarray chip in a defined pattern containing either just the buffer or labeled with gene sequences from a distant, unrelated species (e.g. *Arabidopsis thaliana* genes are selected as control spots for studies involving mammalian genes). For an Affymetrix microarray chip, control spots are designated by identifiers that begin with ‘AFFX’. Ideally, the control spots should not hybridize any RNA and thus should not display any signal. However, due to nonspecific binding, signal intensities above the background are routinely observed from these spots. Voy et al. (2006) used correlations of all genes with control spots on the array as a guide to consider the most specific correlations for analysis. Using Fisher's z-transform in reverse and Bonferroni correction for multiple testing, they showed that such correlations represented statistically significant correlations (p-value < 0.01).

In this study, we evaluate correlation with control spots to estimate the level of noise in the correlation matrix. By considering top 1% of the control-spot correlation distribution (absolute value) to represent a threshold, the algorithm filters insignificant correlations arising due to non-specific binding from further analysis.

The algorithm however, heavily relies on availability of control spot information for each dataset. For the Anoxia and Reoxygenation datasets, information from 20 *Arabidopsis* oligonucleotide spike controls [Lai et al. 2006] available only for glucose arrays (about half of total arrays) was utilized, while for the Alpha dataset, information from 8 salmon sperm DNA and 300 buffer spots available for all arrays was used towards the algorithm.

Method 3: Threshold based on Top 1% of Correlations

This is a simplistic way of picking up the most significant correlations. A microarray experiment with *n* genes has *n*(n-1)/2* correlations. The top 1% of these correlations (absolute value) is chosen to represent the threshold as done by Lee et al. (2004). One advantage of such a threshold is that there is no assumption (normality) made about the distribution. However, there is no statistical justification for picking the top 1% of correlations.

Method 4: Threshold based on Spectral Graph Clustering

Spectral graph theory is the study of graphs with respect to eigen values and eigen vectors derived from the adjacency matrix. Eigen values represent an important methodology to realize the principal properties of graphs [Chung 1994].

Gene interaction networks are well known to display modularity [Hartwell et al. 1999, Hintze and Adami 2008]. The modules, which represent clusters of genes working in synchrony, are not to be considered as completely disconnected components but components with high intra-component connectivity and low inter-component connectivity [Albert 2005]. We use spectral graph theory to identify the modules in a graph and select the correlation at which the best modular separation is possible as the threshold. One of the advantages of spectral clustering is that it is fast and unsupervised. Thus, it requires no prior knowledge of the number of clusters in the graph.

The algorithm thresholds the correlation matrix at a random, low correlation value to create the Laplacian matrix based on the binary adjacency matrix *A* and Degree matrix *D*. This is illustrated with a simple graph consisting of four vertices as shown in Figure 2.

The adjacency (A) and the Degree (D) matrices for the above graph are as follows:

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \qquad D = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The Laplacian matrix (L) becomes

$$L = D - A = \begin{bmatrix} 3 & -1 & -1 & -1 \\ -1 & 2 & -1 & 0 \\ -1 & -1 & 2 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix}$$

The eigen problem is solved for the Laplacian matrix. In the context of full-scale biological graphs, we solve the eigen problem for the largest cluster in the graph.

Eigen values: $\lambda_0 < \lambda_1 < \lambda_2 < \lambda_3 < \lambda_4 \ldots\ldots\ldots\ldots < \lambda_{n-1}$

Eigen vectors: $v_0, v_1, v_2, v_3, v_4 \ldots\ldots\ldots\ldots v_{n-1}$

where n = number of vertices in the graph/largest cluster.

The algebraic connectivity of the graph is represented by the lowest non-zero eigen value (λ_1) [Chung 1994]. The eigen vector, v1, associated with λ_1 is chosen to segregate the graph into spectral clusters [Ding et al. 2001]. Our algorithm uses a sliding window of 10 to identify the clusters. A tolerance level of (median + (0.5 * stdev)) needs to be exceeded by the difference between the highest and the lowest value in the sliding window for a new cluster to be formed [Perkins 2008].

The total number of clusters for the correlation *X* is noted. The procedure is reiterated in increments of 0.01. The correlation value with the maximum number of clusters, which represents the best modular separation of the graph, is chosen as the threshold.

Method 5: Threshold based on Bonferroni correction of p-values

For every correlation value a corresponding p-value is obtained by computing the *t* statistic:

$$t = corr * \sqrt{\frac{n-2}{1-corr^2}} \qquad \textit{(Equation 1)}$$

where n is the number of arrays in the experiment (thus, n-2 is the degrees of freedom). The cutoff p-value (α), which is used to determine the threshold, is based on the Bonferroni correction. We also evaluated statistical FDR and q-value as measures to identify threshold but found them to provide little protection in presence of large number of significant p-values; their distributions were almost the same as the raw p-values.

Method 6: Threshold based on Statistical Power

This algorithm identifies threshold based on statistical power. Depending on the number of arrays *N* in the experiment, statistical power to differentiate correlation between genes against a baseline correlation of 0 is evaluated. The alpha level was Bonferroni-adjusted to correct for multiple testing. Statistical standard of 80% power was chosen to represent the threshold.

Two-tailed hypothesis test was constructed as follows:

H_o: $\rho = \rho_o$ (null hypothesis that true correlation is a specific value ρ_o and $\rho_o = 0$)

H_A: $\rho = \rho_1$ (alternative hypothesis that true correlation is a specific value ρ_1 and $\rho_1 <> \rho_o$)

Thus, the hypothesis was constructed such that statistical power measures the probability that the test will reject H_o when it is truly null, i.e. gene relationships that are not statistically significant. The algorithm first finds out the critical value r_α, such that the probability of rejecting H_o when H_o is true is equal to α (calculated as in Equation 2). Mathematically, we find r_α such that

$$1 - R(r > r_\alpha \mid N, \rho_o) = \alpha \qquad \textit{(Equation 2)}$$

where N = sample size or the number or arrays and $R(r \mid N, \rho)$ represents area under correlation density curve to the left of r. Statistical power is then calculated as the probability of rejecting H_o when H_A is true. Mathematically,

$$Power = 1 - R(r > r_\alpha \mid N, \rho_1) \qquad \textit{(Equation 3)}$$

Analysis of performance of thresholding algorithms

Performance analysis for the above thresholding algorithms was carried out by: i) Bootstrapping over the arrays in the original gene expression datasets to evaluate the stability or reliability of the derived thresholds, and ii) Comparison against threshold based on underlying biological information as quantified using gene ontology to evaluate the validity of the derived thresholds.

Bootstrapping

Bootstrap datasets (n=10000) were created from the expression data files. These datasets were used to obtain a bootstrap distribution of thresholds with each of the thresholding algorithms. Comparing this distribution to the estimated threshold obtained for the original or real gene expression dataset gives an idea of the robustness of the thresholding algorithm.

All datasets used in this study are time-series data. Bootstrapping for time-series data has been a challenging topic of research as the underlying assumption of independency of samples is violated. A good review of the problem and the various approaches employed to address it is presented by Hardle et al. (2001).

Block bootstrapping strategy, which remains the oldest and best non-parametric method to capture the dependence structure of neighboring observations in time-series data [Hardle et al. 2001], was used in this study. Non-overlapping blocks of 3 consecutive arrays were formed and the blocks were randomly sampled with replacement.

Perl scripts were written to perform the bootstrap analysis.

Comparison with threshold estimated from Gene Ontology

Many current algorithms utilize Gene Ontology (GO) [Ashburner et al. 2000, Harris et al. 2004] to understand the biological relevance of relationships derived from gene expression data [Khatri et al. 2002, Zeeberg et al. 2003, Doniger et al. 2003, Zhang et al. 2004]. It is well known that the biological meaning decreases while the noise increases as correlation is lowered. The biological meaning for each correlation bin (1 - 0.99, 0.989 - 0.98, 0.979 - 0.97……… 0.769 - 0.76, 0.759 - 0.75, 0.749 - 0.74 …….) is evaluated as the average of the functional similarity scores for all gene pairs (average functional similarity or GO Score) whose correlations fall within that correlation bin. To calculate the functional similarity for a pair of genes, say gene A and gene B, the algorithm searches for a GO category *X* that covers both gene A and gene B and has the minimum number of genes (*n*). Normalization of *n* to a range of 0 to 1 is done using the following formula:

$$Functional\ \ Similarity = 1 - \left(\frac{-\log(2/N) - (-\log(n/N))}{-\log(2/N)} \right) \qquad \textit{(Equation 4)}$$

where *N* represents the total number of genes annotated for the particular organism under study.

The rationale behind the algorithm is based on the “guilt-by-association” concept [Wolfe et al. 2005]. Pairs of genes with similar expression patterns (high correlation values) tend to be involved in the same biological processes or perform similar cellular functions and are found in deeper, more specific levels of GO tree hierarchy. Thus, they occur under GO categories with a comparatively lower n and feature a high functional similarity score (close to 1).

We consider the GO Score only for positive correlations as we show later in our analysis that negative correlations fail to display any biological significance.

Threshold is identified as the correlation at which the change in GO Score exceeds the (median + (0.5 * stdev)) tolerance for all positive correlations. The median – as against the mean – of the GO Score and half of overall standard deviation help guard against extreme values in the data. Also, such a threshold is completely dependent on the inherent biological characteristics of the data as reflected through gene ontology and thus, automatically adapts to different datasets.

In order to measure the performance of each method, we define a difference metric d_{TM} for a method as the difference between estimated threshold based on gene ontology (τ_{GO}) and threshold derived from that method (τ_{TM}). The metric is calculated for each dataset.

$$d_{TM} = \tau_{GO} - \tau_{TM} \qquad \textit{(Equation 5)}$$

To evaluate the overall performance of each thresholding method (across all datasets) we define another metric S_{TM}, as the summation of d_{TM}'s for the particular thresholding method over the three datasets.

$$S_{TM} = d_{TM}(anoxia) + d_{TM}(reoxygenat\,ion) + d_{TM}(alpha)$$

(Equation 6)

The gene2go.gz file available at NCBI ftp location (*ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/)* was used to map GO annotation for genes on the respective arrays. GO Pairwise Similarity analysis code version 1.0 (software written by Dr.Langston's research group) was used for the method.

CHAPTER IV
RESULTS

The study analyzes six completely different methods for thresholding: four of them are evaluated for robustness by creating bootstrap datasets from original (real) datasets while all six are evaluated against threshold obtained from biological information using Gene Ontology. Bootstrap analysis was carried out only for the first four methods. Bootstrap analysis was carried out only for the first four methods. Bootstrapping on Bonferroni correction of p-values was not carried out due to time. Statistical power being solely dependent on the number of arrays and number of genes in the microarray experiment was also not considered for bootstrapping.

Derivation of the estimated threshold for the original datasets with each of the methods is discussed below.

Derivation of estimated threshold for each method

Method 1: Threshold based on number of Maximal Cliques in graph

Table 1, 2 and 3 give the parameters of the graph (number of vertices, edges, maximal cliques and size of maximum clique) at each correlation for Anoxia, Reoxygenation and Alpha datasets respectively.

For Anoxia dataset, the number of maximal cliques grows to more than 50000 and the first instance of doubling to that at the previous correlation of 0.91 occurs at correlation threshold of 0.9. Thus, 0.9 is chosen as the threshold for the anoxia dataset with the Maximal Clique-2 method. For the Maximal Clique-3 method, the threshold becomes 0.87, when the first instance of tripling of the number of maximal cliques to that at the previous correlation occurs (Table 1).

Similarly, the Maximal Clique-2 method found threshold of 0.91 and 0.74 for the Reoxygenation and Alpha datasets, respectively. The Maximal Clique-3 method found a threshold of 0.89 for the Reoxygenation dataset.

For the Alpha dataset, the tripling of the number of maximal cliques does not occur until the threshold falls very low. This is because of the relative sparseness of edges in the graph for the Alpha dataset at high correlations. Figure 3 depicts the growth of number of maximal cliques in the graph for the three datasets as correlation threshold is lowered. From the distribution of maximal cliques for the three datasets (Table 1, 2 and 3 and Figure 3), it is clear that the Alpha dataset produces comparatively sparse graphs at high correlations. At correlation of 0.61, the number of maximal cliques falls to almost half to that at correlation of 0.62. At correlation threshold of 0.6, however, the number of maximal cliques recovers to almost 2.8 times at correlation threshold of 0.61, which is very close to tripling. Due to increase in computational time at correlation threshold below 0.6, we have assigned 0.6 as the estimated threshold for the Alpha dataset with the Maximal Clique-3 method.

Thus, the Maximal Clique algorithm inherently adjusts the threshold according to the graph characteristics for the respective datasets: a higher threshold is identified for Anoxia and Reoxygenation datasets that display a greater number of high correlations and subsequently produce comparatively denser graphs at high correlation thresholds, while a considerably lower threshold is identified for the Alpha dataset.

Method 2: Threshold based on information extracted from Control Spots

We found the control spots to display a high degree of correlation – many even to the extent of 0.98 – with rest of the genes on the array (Figure 4). One reason for this could be the presence of high correlation within the control spots themselves (Figure 5). The distribution of negative correlations was very much similar to positive correlations.

The distribution appears very close to normal for the Alpha dataset, which is not so for the other datasets. The reason for this could be the far greater number of control spots considered for analysis in the case of Alpha dataset (8 salmon sperm DNA spots and 300 3XSSC buffer spots). For the Anoxia and

Reoxygenation datasets, information from only 20 *Arabidopsis* oligonucleotide spike controls was available. Also, this information was restricted to only the glucose arrays, which were about half of the total arrays in the datasets.

Also, the relative low variance of normality leaves very low number of control spot correlations at the extremes of the distribution. Considering the top 1% of the control spot correlations correspondingly offers a lower threshold (τ = 0.7) for the Alpha dataset. As for the Anoxia and Reoxygenation datasets, which do not display a normal distribution, a greater number of total correlations fall in the extremes of the distribution, much more so for Anoxia as compared to Reoxygenation (Figure 4). The thresholds identified for the two datasets correspondingly reveal this difference in distribution: a lower threshold is identified for Reoxygenation dataset (τ = 0.83) as compared to Anoxia dataset (τ = 0.93).

Method 3: Threshold based on Top 1% of Correlations

The distribution of correlations of all genes (excluding the control spots) on the array follows a normal or near-normal distribution (Figure 6).

Comparison of this distribution for the three datasets reveals a lower variance for the Alpha dataset; the distribution for this dataset is also much closer to normality. The Anoxia and Reoxygenation datasets display a similar distribution of gene correlations. The estimated threshold reflects this apparent difference: a lower threshold if obtained for the Alpha dataset (τ = 0.72), while a comparatively higher threshold is obtained for Anoxia (τ = 0.81) and Reoxygenation datasets (τ = 0.81).

Method 4: Threshold based on Spectral Graph Clustering

Figure 7 shows the number of spectral clusters obtained at every correlation for the three datasets. For the Anoxia dataset, the maximum number of clusters (7) is seen at estimated threshold of 0.93. For the Reoxygenation and Alpha datasets, the maximum number of clusters (6 for Reoxygenation and 5 for Alpha dataset) is seen at estimated threshold of 0.97 and 0.89 respectively.

Method 5: Threshold based on Bonferroni correction of p-values

For Anoxia dataset, with 5525 genes and 31 arrays, the threshold obtained was 0.85. For Reoxygenation dataset, with 5525 genes and 21 arrays, the estimated threshold was: τ = 0.93, while for Alpha dataset, with 5466 and 18 arrays, it was τ = 0.95.

Method 6: Threshold based on Statistical Power

The threshold based on statistical power depends on the number of conditions used in the microarray experiment. More conditions give higher statistical power and correspondingly lower threshold.

The Anoxia dataset with 31 conditions displays a very high statistical power (Figure 9) at the various correlation thresholds as compared to Reoxygenation dataset with 21 conditions (Figure 10) and Alpha dataset with 18 conditions (Figure 11). Using 80% as the standard cut-off for statistical power, Anoxia dataset was assigned τ = 0.88, Reoxygenation dataset τ = 0.94 and Alpha dataset τ = 0.96. Figures 9, 10 and 11 represent the output from PASS analysis software [*http://www.ncss.com/pass.html*].

Results of Bootstrapping

Estimated threshold (τ) was compared to bootstrap distribution of thresholds obtained from 10000 bootstrap datasets generated for each of the original (real) datasets. Bootstrap analysis was carried out only for the first four methods. Bootstrapping on the Bonferroni correction of p-values method was not done due to time constraints. Threshold based on statistical power, on the other hand, is derived only on the basis of the number of arrays and number of genes considered in the experiment and thus would not be affected by bootstrapping.

The block bootstrapping methodology we adopted, made us able to create datasets from a huge sample space. For example, the Reoxygenation dataset contains 21 arrays. 7 blocks of 3 arrays were made across the time series and arrays within each block were randomly sampled 3 times with replacement to build a bootstrap dataset with 21 arrays. Since each block had 10 different possibilities, overall there were 10^7 different possibilities of creating bootstrap datasets.

Similarly, the 18 arrays in Alpha dataset were grouped into 6 blocks of 3 arrays. The Anoxia dataset, which has 31 arrays, was grouped into 9 blocks of 3 arrays and one block of 4 arrays.

Some general comments can be made over results from the overall bootstrapping procedure (Table 4). The bootstrap threshold distribution for all methods is pushed higher. The bootstrap mean and mode of threshold distribution are always greater than the estimated threshold τ. Even the 95% confidence intervals for bootstrap mean do not encompass the estimated threshold τ.

Maximal Clique algorithm

The Maximal Clique-2 method performs well with the Anoxia and Alpha datasets: the bootstrap frequency of τ is very close to the bootstrap frequency of mode (Figure 12). However, for the Reoxygenation dataset, this is not so.

The algorithm's performance is enhanced with the Maximal Clique-3 method: the bootstrap frequency of τ is pushed closer to the bootstrap frequency of mode for the Reoxygenation dataset. Similar conclusion can be drawn for the Anoxia dataset (Figure 13). However, the variance of bootstrap distribution for both datasets is increased as compared to the Maximal Clique-2 method.

The extremely dense nature of the graph at low estimated threshold (τ = 0.6) for the Alpha dataset precluded generating bootstrap results for the Maximal Clique-3 method.

Control-Spot verification algorithm

The Control spot method performs poorly with all three datasets (Figure 14). The bootstrap frequency of *τ* is very low as compared to the bootstrap frequency of mode. Even with the comparatively high variance for the Reoxygenation and Alpha datasets, there is no improvement in the method's robustness.

Top 1% Correlations algorithm

This algorithm also performs poorly with all three datasets (Figure 15). The bootstrap frequency of *τ* is very low compared to the bootstrap frequency of mode.

Spectral graph clustering algorithm

The Spectral Clustering method performs exceptionally well – even better than the Maximal Clique-3 method – for the Reoxygenation dataset (Figure 16): the bootstrap frequency of *τ* (34.9%) is very close to the bootstrap frequency of mode (39.87%). Importantly, the bootstrap standard deviation for all three datasets is comparatively low only for this algorithm (Table 4).

We also analyzed the bootstrapping results with respect to each dataset. For the Anoxia dataset, the estimated threshold *τ* lies very close to bootstrap mean and mode for all methods. However, the bootstrap frequency of *τ* is very close to the bootstrap frequency of mode only for the Maximal Clique-2 method: 19.87% to 24.55%. *τ* is also very close to the 95% confidence interval for this method. The performance repeats in Maximal Clique-3 method.

In case of the Reoxygenation dataset, the bootstrap frequency of *τ* is very close to the bootstrap frequency of mode for the Spectral Clustering method: 34.9% to 39.87%. Maximal Clique method performs poorly when we take the correlation at which the number of maximal cliques grows to more than twice (Maximal Clique-2 method) as the threshold. However in Maximal Clique-3 method - a much more stringent algorithm - the bootstrap frequency of *τ* gets closer to bootstrap frequency of mode: 9.72% to 13.45%.

For the Alpha dataset, τ is far away from the bootstrap mean and mode for all methods. However, it is comparatively closest for the Maximal clique-2 method. Also, the bootstrap frequency of τ is closest to bootstrap frequency of mode for Maximal clique-2 method: 6.56% to 6.75%. Results on bootstrapping at Maximal Clique-3 level for this dataset were not derivable since below threshold of 0.7 the graphs became too dense and computational time became unreasonable.

For all three datasets, the Spectral Clustering algorithm displays a comparatively low variance for the bootstrap threshold distribution.

Comparison with threshold estimated from Gene Ontology

Figure 17 shows the distribution of functional similarity score against correlation for each of the three datasets. The score is high at very high positive correlations and displays a sharp drop early on. At high negative correlations the score falls almost to 0, except for Alpha dataset in which the score shows some rise. However, this rise is not as high as at the positive correlation end.

Figure 19 depicts the change in GO Score occurring at each correlation value for the three datasets. Although the graphs are more so flat at low correlations, fluctuations in GO Score begin to arise around 0.7-0.8 correlation values and become huge at higher correlations.

Estimated threshold τ obtained from each of the algorithms for all three datasets are listed in Table 5 for comparison against the estimated threshold derived on the basis of gene ontology. A good thresholding method is one that maximizes the proportion of true positives and true negatives against the number of false negatives and false positives.

d_{TM} values for each thresholding method and dataset are shown in brackets in Table 5. Negative values for d_{TM} indicate the method provides a threshold higher than the biological threshold thus incorporating a high number of false negatives. While positive values for d_{TM} indicate a threshold below the biological threshold and incorporate a high number of false positives. However, in contrast to false negatives, biologists work under discovery-based settings and so can tolerate some amount of false positives, provided they are very few. Thus, a thresholding method with a low positive d_{TM} indicates a desirable performance.

Similarly, a negative S_{TM} – just like a negative d_{TM} – indicates a threshold accommodating a high number of false negatives. Higher the S_{TM}, more the number of false positives accommodated by the thresholding method. Thus, a low and positive S_{TM} is preferred.

As outlined in Table 5, for the Anoxia dataset, Maximal Clique-2 (d_{TM} = 0.07), Control-Spot verification (d_{TM} = 0.04) and Spectral Clustering (d_{TM} = 0.04) methods give thresholds that are lower and close to the biological threshold. Thus, thresholds from these methods do not miss the underlying information in the dataset. And being close to the biological threshold assures that these methods also limit the noise factor very well.

Similarly, for the Reoxygenation dataset, Maximal Clique-2 (d_{TM} = 0.01), Maximal Clique-3 (d_{TM} = 0.03) and statistical p-value (d_{TM} = 0.02) methods identify a threshold lower and close to the biological threshold.

For the Alpha dataset, only the Maximal Clique-2 method provides a relatively low d_{TM} value of 0.11, and thus performs better than the rest.

Thus, Maximal Clique-2 method performs the best in comparison with the other methods. This is indicated by the positive d_{TM} evaluated for the three datasets while S_{TM} for the method has the lowest positive value (0.19). Although threshold based on Bonferroni correction of p-values has S_{TM} of 0.01, which is lower than Maximal Clique-2 method, it gives a negative d_{TM} for Reoxygenation and Alpha dataset.

CHAPTER V
DISCUSSION AND CONCLUSIONS

Thresholding of data to pick information-rich sections is an important research problem that has significant application for large volumes of data. In the context of transcriptomic research, there have only been studies that mention and handle the thresholding issue in passing [Bredel et al. 2005, Sanoudou et al. 2003]. Many researchers have based their choice of threshold either on one or the other method elucidating the validity of their approach [Moriyama et al. 2003, Lee et al. 2004, Voy et al. 2006]. Although Allocco et al. (2004) report an interesting study, their results are confined more so to the datasets they analyzed.

This study compares and analyzes different approaches for thresholding the gene correlation matrix on the basis of robustness and underlying biological information. Two of the methods are based on graph theory, two on statistical theory, while the other two based on correlation distribution.

Correlation as a measure of association between genes is very much susceptible to fluctuations in expression values that occur as a result of high variability and noise associated with microarrays. Although at high correlations the effect of such susceptibility is very low, it is difficult to ascertain or measure such an effect. Since the threshold is a pivotal resolution to such a binary decision problem, it is important that the threshold not be sensitive to the high variability and noise that affect values in the gene correlation matrix. Thus, a high level of robustness is a desirable property for a threshold.

Bootstrapping

The bootstrap methodology helps to estimate a method’s robustness by deriving distribution information obtained by resampling the data. Two important issues come up when performing bootstrap analysis on dependent data: the bias and the variance.

The persistent bias in the bootstrap distribution has been documented to be a classical drawback of bootstrapping on time-series data [Hardle et al. 2001].

Various researchers have modified the block bootstrapping approach with matched blocks [Carlstein et al. 1998], overlapping blocks [Hall 1985] or stochastic block sizes [Politis and Romano 1993]. Papers highlighting the influence of bootstrap block size and block assignment on the outcome of the bootstrap procedure abound in literature [Lahiri 1999, Hardle et al. 2001]. We have employed non-overlapping blocks of size 3 in our bootstrap approach. Subtle differences in results may arise upon adopting a different bootstrap procedure but we expect to see a similar comparative performance of the thresholding algorithms.

The bias, however, does not invalidate the purpose or the results of this study. Since the same bootstrapping strategy was employed for all datasets, the bias can be concluded to affect each thresholding method in a similar fashion. Thus, although presence of such bias makes it difficult to identify methods that are comparatively more robust, it does not hinder us from proceeding with the analysis. In fact, the bias can be considered to act similar to the presence of outlier arrays – there weren't any for the datasets in this study – and robustness to bootstrap bias can be conceived as robustness against outlier arrays.

The bias-variance issue regarding the bootstrap distribution could be compared to the precision-accuracy problem. In a much-cited paper, Lahiri (1999) reports a thorough analysis of the issue with different bootstrapping methods on dependent data.

A detailed observation of the bootstrap results reveals that at higher estimated threshold, the bootstrap distribution for threshold displays low variance. Instead, when the estimated threshold is low, the bootstrap distribution shows high variance. Moreover, the estimated threshold at which such changes are observed is dependent on the dataset.

Considering the proximity of estimated threshold to the mode of the bootstrap distribution as a metric, the Maximal Clique algorithm comes out to be more robust. It performs well at Maximal Clique-2 level for Anoxia and Alpha datasets. For the Reoxygenation dataset, it performs better at Maximal Clique-3 level, which is a far more exacting algorithm. Importantly, the Anoxia dataset also performs well at this level. This further demonstrates the algorithm's robustness. On the other hand, considering the bootstrap variance as a metric, the Spectral Clustering method proves to be more robust as it generates threshold distributions with comparatively low variance for all three datasets. Thus, interpretation of bootstrapping results is dependent on the metric used to evaluate stability.

We propose the modular basis of these algorithms to be responsible for their robust performances. Even though the bootstrap bias (or equivalently, the presence of outlier arrays) tends to skew the correlation distribution and subsequently add or remove a significant number of edges from the graph, such a phenomenon does not affect the existing number of clusters or the formation of new gene clusters (maximal cliques or spectral clusters) as much.

We analyze the performance of other methods and identify reasons for their lack of robustness. The Top 1% Correlation method is tightly linked to the distribution of gene correlations, which is easily perturbed by the bias of bootstrapping (or by the presence of outlier arrays). This is illustrated by the huge disparity in the bootstrap frequency of Mode and τ for this method with all datasets. Even the Control-Spot verification method fails to perform in bootstrapping. It is worthwhile to note here that the Anoxia and Reoxygenation datasets have control-spot data only for about half the arrays. However, even with control-spot data from all the arrays – as is the case for the Alpha dataset – we do not have any improvement on the method's performance with the bootstrap datasets. The method's lack of robustness is a reflection of the immoderately high degree of correlation displayed by control spots with rest of the genes on the array. The exaggeration of such high correlations upon bootstrapping leads to a higher bootstrap mean and mode and 95% Confidence Intervals that do not encompass the estimated threshold. Besides, the immoderately high degree of control-spot correlations is likely to be very much susceptible to the presence of outlier arrays.

Comparison with threshold estimated from Gene Ontology

The second half of results compares the thresholds from different algorithms against the scale of biological threshold. The scale we have used, however, is by no means standard and is just one way of quantifying biological information. In spite of the tremendous popularity of Gene Ontology – as indicated by numerous bioinformatics servers dedicated to the subject, a string of which could be found at *www.gene-ontology.org* – to represent biologically true relationships, the utility of such a controlled vocabulary system suffers from various limitations that hinder it from being an accurate reflection of the inherent biological information in the data. Khatri and Draghici (2005) have enlisted these limitations in detail. The more important of these limitations are: *Incompleteness*, the ontology is far from being complete and many more genes from sequenced genomes are yet to find their way through a formal annotation, *Exclusion* of known biological information either due to human error or time lag between discovery and data processing and subsequent inclusion, *Incorrect annotations* resulting from inferences made from automatic data parsing and curation, *Annotation bias* towards genes that are studied more extensively, *Discrepancy* in known information arising from absence of one-to-one mapping between various gene identifiers used by autonomous data collecting organizations.

Figure 17 displays the inadequately low GO score at high negative correlations as against the high GO score associated with high positive correlations for all three datasets. The drop in GO score at high negative correlations could be linked to various reasons. First, there exist experimental and analytical limitations to detect biologically negative correlations amongst genes [Lee et al. 2004] even in the face of today's highly developed microarray technology. Second, active gene-specific transcriptional repression is not as common in eukaryotes as in prokaryotes [Struhl 1999]. Lastly, such a drop in GO Score at negative correlations could also be a drawback of limited gene annotations [Lee et al. 2004]. So, in the use of GO Similarity to identify a biologically relevant threshold, we have considered only the GO Score for positive correlations.

The limitations of gene ontology, however, bring up future challenges to innovate and/or improvise ways to quantify biological information [Khatri and Draghici 2005]. Various regulatory pathway-dependent analyses like MAPPFinder [Doniger et al. 2003], Pathway-Express [Khatri et al. 2005], Cytoscape [Shannon et al. 2003] have already opened up interesting avenues to do this. As the quantification of biological information in data gets more precise, the validation of choice of a particular threshold should become easier and undebatable.

The comparison of estimated thresholds from different methods to estimated threshold from gene ontology points towards certain general conclusions. Methods like Power and P-value that are completely based on statistical properties of data are not able to represent the underlying biological information any better than the other methods. These statistical methods have significant impact on the success of microarray experiments if they are used towards designing and planning them [Wei et al. 2004, Page et al. 2006].

Methods like Control-Spot and Top 1% of Correlations are directly dependent on the correlation distribution and so fail to demonstrate a satisfactory amount of robustness or to represent biological relationships. Although the Control-Spot verification method is based on a sound biological reasoning, the very high correlation of control spots with rest of the genes on arrays weakens the method’s validity. The Top 1% Correlations, on the other hand, seems a random approach to pick up information from correlation matrix data, thus failing to conform to the biological aspect of it.

Spectral Clustering and Maximal clique algorithms reflect the modular nature of biological networks. However, spectral clustering lacks the stringency associated with Maximal clique algorithm. The comparative ease of formation of spectral clusters leads to higher threshold as compared to the Maximal clique algorithm.

For all three datasets, the Maximal Clique-2 algorithm works the best when compared to the biological threshold: it has the lowest positive d_{TM} value for all datasets and lowest positive S_{TM} value, thus indicating that the thresholds offered by the Maximal Clique-2 method are lower than the biological threshold and conveniently close to it. Maximal Clique-3 method, as a result of a higher stringency, pushes the threshold lower down thus accommodating for a higher level of noise.

Conclusions and Future Work

A threshold derived on the basis of aggregate gene relationships is much more robust than one derived on the basis of pair-wise relationships. The study carried out by Zhang and Horvath (2005) also resulted in a similar conclusion: threshold based on the scale-free topology criterion – which relies on the formation of hubs and densely-connected sub-graphs – was shown to produce more robust results.

The Maximal Clique algorithm performs very well in terms of stability (as indicated by results from bootstrapping in Table 4) as well as validity (as indicated from comparison with biological threshold in Table 5). Though Maximal Clique-2 method pushes the threshold very close to the biological threshold, the method does not display robustness in case of the Reoxygenation dataset. The Maximal Clique-3 method, on the other hand, seems more robust (and more stringent) but pushes the threshold lower down and accommodates a large amount of noise. Thus, a balance between robustness and noise accommodation needs to be reached for the algorithm to perform to its optimum. It is well known that the chance of random occurrence for a clique is inversely related to the size of the clique. Thus, the robustness of the Maximal Clique-2 algorithm would easily be enhanced by exclusion of smaller cliques in the graph, for e.g. cliques of size 3.

Spectral clustering seems promising as an approach to thresholding. It performs very well for the Reoxygenation dataset and generates threshold distributions with comparatively low variance for all three datasets. Though the modular basis of the algorithm resembles the nature of gene networks, it fails to generate a biologically valid threshold. Further analysis of the spectral clustering algorithm by tweaking the various parameters in the algorithm (size of sliding window, different tolerance levels for cluster formation) should be required to harness the method's robustness as well as its validity. In a recent paper, Almendral and Díaz-Guilera (2007) have documented the sensitivity of the non-zero eigen value to network changes. Alterations to the algorithm towards lowering such sensitivity need to be explored.

Although this study has implications beyond transcriptomic research, important limitations need to be mentioned. The analysis for robustness in this study was carried out upon preliminary exploratory analysis of the datasets, which concluded that none of them had any outlier arrays. Though we anticipate that robustness of a thresholding algorithm against bootstrap bias is tantamount to robustness against outlier arrays, our observation needs to be validated. Further studies will involve a thorough analysis of the bias of threshold distribution upon bootstrapping of transcriptomic data. The influence of block size and block assignment on threshold distribution will be investigated to identify ways to reduce the bias. Using the least bias bootstrap methodology, robustness of thresholding algorithms will be tested upon introduction of one or two outlier arrays for each of the datasets. Besides this, all limitations of gene ontology apply to the present study. With availability of metabolic and pathway databases, we plan to replace gene ontology with more accurate ways to quantify biological information and overcome this limitation.

The results of our analysis help to assess the relative performance of thresholding algorithms. The bootstrap experiment affords identification of robust methods towards thresholding the data. Comparison to a threshold based on informational aspect of data identifies methods that yield thresholds that allow for maximum information and minimum noise. Future work will involve research on elimination of the dependency of thresholding algorithms to chosen parametric values. We hope to achieve this by using inflection points derived on the basis of inherent properties of each dataset. Also, development of 'soft' algorithms for thresholding and ‘combinatorial’ strategies that bring in strong attributes of all algorithms remains an open-ended problem and a further challenge to ongoing research.

LIST OF REFERENCES

LIST OF REFERENCES

Abu-Khzam F. N., Langston M. A., and Suters W. H. (2005) Fast effective vertex cover kernelization: A tale of two algorithms. Proceedings of the ACS/IEEE International Conference on Computer Systems and Applications, Cairo, Egypt.

Albert R. (2005) Scale-free networks in cell biology. Journal of Cell Science, 118: 4947-57.

Allison D. B., Cui X., Page G. P. and Sabripour M. (2006) Microarray data analysis: from disarray to consolidation and consensus. Nature Reviews Genetics, 7: 55-65.

Allocco D. J., Kohane I. S., and Butte A. J. (2004) Quantifying the relationship between co-expression, co-regulation and gene function. BMC Bioinformatics, 5:18.

Almendral J. A. and Díaz-Guilera A. (2007) Dynamical and spectral properties of complex networks. New J. Phys. 9, 187.

Alon U. (2003) Biological Networks: The Tinkerer as an Engineer; Science, 301(5641): 1866-7.

Aloy P., and Russell R. B. (2004) Taking the mystery out of biological networks. EMBO Rep., 5(4): 349-50.

Alter O., Brown P. O., and Botstein D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci U S A., 97(18): 10101-6.

Anisimov S. V., Christophersen N. S., Correia A. S., Li J. Y., and Brundin P. (2007) “NeuroStem Chip”: a novel highly specialized tool to study neural differentiation pathways in human stem cells. *BMC Genomics*, 8:46.

Ashburner M., Ball C. A., Blake J. A., Botstein D., Butler H., Cherry J. M., Davis A. P., Dolinski K., Dwight S. S., Eppig J. T., Harris M. A., Hill D. P., Issel-Tarver L., Kasarskis A., Lewis S., Matese J. C., Richardson J. E., Ringwald M., Rubin G. M., and Sherlock G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet., 25(1):25-9.

Bader G. D., and Enright A. J. (2005) Intermolecular interactions and biological pathways. In: Baxevanis A. D., and Ouellette B. F. F., editors. Bioinformatics: A practical analysis of genes and proteins. Third edition. New York: John Wiley & Sons, Inc.; 280 p.

Baldwin N. E., Chesler E. J., Kirov S., Langston M. A., Snoddy J. R., Williams R. W., and Zhang B. (2005) Computational, Integrative, and Comparative Methods for the Elucidation of Genetic Coexpression Networks. J Biomed Biotechnol., 2005(2):172-80.

Barabási A. L., and Oltvai Z. N. (2004) Network Biology: Understanding the Cell's Functional Organization. Nature Reviews Genetics, 5(2):101-13.

Bellaachia A., Portnoy D., Chen Y., and Elkahloun A.G., editors. (2002) ECAST: a data-mining algorithm for gene expression data. In: 2nd Workshop on Data Mining in Bioinformatics (BIOKDD 2002); Alberta, Canada. 49–54.

Ben-Dor A., Shamir R., and Yakhini Z. (1999) Clustering gene expression patterns. J Comput Biol. 6(3-4):281–297.

Ben-Dor A., Bruhn L., Friedman N., Nachman I., Schummer M., and Yakhini Z. (2000) Tissue classification with gene expression profiles. J Comput Biol., 7(3-4):559–583.

Benjamini Y., and Hochberg Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc B., 57:289–300.

Bergmann S., Ihmels J., and Barkai N. (2004) Similarities and differences in genome-wide expression data of six organisms. PLOS Biology, 2(1), 85-93.

Bomze I. M., Budinich M., Pardalos P. M., and Pelillo M. (1999) The maximum clique problem. Handbook of Combinatorial Optimization (Supplement Volume A), Du D.-Z. and Pardalos P. M. (Eds.), Kluwer Academic Publishers, Boston, MA, 1-74.

Bray D. (2003) Molecular networks: The top-down view. Science, 301(5641):1864-5.

Bredel M., Bredel C., Juric D., Harsh G. R., Vogel H., Recht L. D. and Sikic B. I. (2005) Functional network analysis reveals extended gliomagenesis pathway maps and three novel MYC-interacting genes in human gliomas. Cancer Res., 65(19): 8679-89.

Butte A. J., and Kohane I. S. (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput. 418-429.

Butte A. J., Tamayo P., Slonim D., Golub T. R., and Kohane I. S. (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks; Proc Natl Acad Sci U S A., 97(22):12182-6.

Carlstein E., Do K. -A., Hall P., Hesterberg T. and Künsch H. R. (1998) Matched block-bootstrap for dependent data. Bernoulli 4, 305-328.

Carter S. L., Brechbühler C. M., Griffin M., Bond A. T. (2004) Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics, 20(14), 2242-50.

Chesler E. J., Lu L., Shou S., Qu Y., Gu J., Wang J., Hsu H. C., Mountz J. D., Baldwin N. E., Langston M. A., Threadgill D. W., Manly K. F., and Williams R. W. (2005) Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nat Genet., 37(3):233-42.

Chesler E. J., and Langston M. A. (2005) Combinatorial Genetic Regulatory Network Analysis Tools for High Throughput Transcriptomic Data. RECOMB Satellite Workshop on Systems Biology and Regulatory Genomics; 2-4 December; San Diego, California.

Cho R.J., Campbell M. J., Winzeler E. A., Steinmetz L., Conway A., Wodicka L., Wolfsberg T. G., Gabrielian A. E., Landsman D., Lockhart D. J., and Davis R. W. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. Mol. Cell, 2, 65-73.

Chung Fan R. K. (1994) Spectral Graph Theory, Volume 92 of Regional Conference Series in Mathematics. American Mathematical Society.

Davidson G. S., Wylie B. N., and Boyack K. W. (2001) Cluster stability and the use of noise in interpretation of clustering. IEEE Information Visualization, 23-30.

Dehmer M. and Emmert-Streib F. (2008) Analysis of Microarray Data: A network-based approach. Publisher Wiley-VCH, ISBN:3527318224.

Ding C. H. Q., He X. and Zha H. (2001) A spectral method to separate disconnected and nearly disconnected Web graph components. Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California. August. Pages: 275 – 280.

Doniger S. W., Salomonis N., Dahlquist K. D., Vranizan K., Lawlor S. C., and Conklin B. R. (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol*, 4:R7.

Dougherty E. R., Barrera J., Brun M., Kim S., Cesar R. M., Chen Y., Bittner M., and Trent J. M. (2002) Inference from clustering with application to gene-expression microarrays. J Comput Biol, 9(1): 105-126.

Dudoit S., Shaffer J. P., and Boldrick J. C. (2003) Multiple Hypothesis Testing in Microarray Experiments. *Statistical Science*, 18: 71-103.

Eisen M. B., Spellman P. T., Brown P. O., and Botstein D. (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA, 95: 14863–14868.

Elkin P. L. (2003) Primer on medical genomics. Part V: Bioinformatics. Mayo Clinic Proc. 78:57–64.

Elowitz M. B., Levine A. J., Siggia E. D., and Swain P. S. (2002) Stochastic gene expression in a single cell. Science. 297(5584):1183-6.

Feige U., Peleg D., and Kortsarz G. (2001) The dense k-subgraph problem. Algorithmica. 29:410–421.

Fellows M. R., and Langston M. A. (1994) On search, decision, and the efficiency of polynomial-time algorithms. J Comp Sys Sci. 49:769–779.

Freeman T. C., Goldovsky L., Brosch M., van Dongen S., Mazière P., Grocock R. J., Freilich S., Thornton J., and Enright A. J. (2007) Construction, Visualisation, and Clustering of Transcription Networks from Microarray Expression Data; PLoS Comput Biol., 3(10): 2032-42.

Garey M. R. and Johnson D. S. (1979) “Computers and Intractability: A guide to the theory of *NP*-completeness.” W. H. Freeman, ISBN 0-7167-1044-7.

Girolami M., and Breitling R. (2004) Biologically valid linear factor models of gene expression. Bioinformatics, 20(17):3021-33.

Hall P. (1985) Resampling a coverage process. Stochastic process applications 19, 259-269.

Hansen P., and Jaumard B. (1997) Cluster analysis and mathematical programming. Math Program. 79(1-3):191–215.

Härdle W., Horowitz J. and Kreiss J. P. (2001) Bootstrap Methods For Time Series. *Sonderforschungsbereich 373*, 2001-59, Humboldt Universitaet Berlin.

Harris M. A., Clark J., Ireland A., Lomax J., Ashburner M., Foulger R., Eilbeck K., Lewis S., Marshall B., Mungall C., Richter J., Rubin G. M., Blake J. A., Bult C., Dolan M., Drabkin H., Eppig J. T., Hill D. P., Ni L., Ringwald M., Balakrishnan R., Cherry J. M., Christie K. R., Costanzo M. C., Dwight S. S., Engel S., Fisk D. G., Hirschman J. E., Hong E. L., Nash R. S., Sethuraman A., Theesfeld C. L., Botstein D., Dolinski K., Feierbach B., Berardini T., Mundodi S., Rhee S. Y., Apweiler R., Barrell D., Camon E., Dimmer E., Lee V., Chisholm R., Gaudet P., Kibbe W., Kishore R., Schwarz E. M., Sternberg P., Gwinn M., Hannick L., Wortman J., Berriman M., Wood V., de la Cruz N., Tonellato P., Jaiswal P., Seigfried T., and White R. (2004) Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res., 32 (Database issue): D258-61.

Hartuv E., Schmitt A., Lange J., Meier-Ewert S., Lehrachs H., and Shamir R., editors. (1999) An algorithm for clustering cDNAs for gene expression analysis. In: Proceedings of the 3rd Annual International Conference on Computational Molecular Biology (RECOMB '99); Lyon, France. 188–197.

Hartwell L. H., Hopfield J. J., Leibler S., and Murray A. W. (1999) From molecular to modular cell biology. Nature, 402 (6761 Suppl):C47-52.

Hintze A. and Adami C. (2008) Evolution of Complex Modular Biological Networks. PLoS Comput Biol 4(2): e23. doi:10.1371/journal.pcbi.0040023.

Jeong H., Tombor B., Albert R., Oltvai Z. N., and Barabasi A. L. (2000) The large-scale organization of metabolic networks. Nature, 407: 651–654.

Khatri P., Draghici S., Ostermeier G. C., and Krawetz S. A. (2002) Profiling gene expression using Onto-Express. *Genomics*, 79:266-270.

Khatri P. and Drăghici S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics, 21(18):3587-95.

Khatri P., Sellamuthu S., Malhotra P., Amin K., Done A. and Draghici S. (2005) Recent additions and improvements to the Onto-Tools. Nucleic Acids Research, 33(Web Server issue):W762-5.

Kim D., and Chung J. (2002) Akt: Versatile mediator of cell survival and beyond. J Biochem Mol Biol., 35:106–115.

Lahiri S. N. (1999) Theoretical comparisons of block bootstrap methods. Ann. Statist., 27(1), 386-404.

Lai L. C., Kosorukoff A. L., Burke P. V., and Kwast K. E. (2006) Metabolic-State-Dependent Remodeling of the Transcriptome in Response to Anoxia and Subsequent Reoxygenation in Saccharomyces cerevisiae. Eukaryot Cell, 5(9): 1468-89.

Langston M. A. (2004) Practical FPT Implementations and Applications. Proceedings of Parameterized and Exact Computation: First International Workshop, IWPEC 2004 Bergen, Norway. DOI: 10.1007/b100584.

Langston M. A., Perkins A. D., Saxton A. M., Scharff J. A. and Voy B. H. (2006) Innovative Computational Methods For Transcriptomic Data Analysis: A Case Study in the Use Of FPT For Practical Algorithm Design and Implementation; Proceedings of the 2006 ACM symposium on Applied computing; Dijon, France.

Lee H. K., Hsu A. K., Sajdak J., Qin J., and Pavlidis P. (2004) Coexpression analysis of human genes across many microarray data sets. Genome Res., 14(6):1085-94.

Lopez J., and Martinez A. (2002) Cell and molecular biology of the multifunctional peptide, adrenomedullin. Int Rev Cytol., 221:1–92.

Lorentz C. P., Wieben E. D., Tefferi A., Whiteman D. A., and Dewald G. W. (2002) Primer on medical genomics part I: History of genetics and sequencing of the human genome. Mayo Clin. Proc., 77:773–782.

Manfield I. W., Jen C. H., Pinney J. W., Michalopoulos I., Bradford J. R., Gilmartin P. M., and Westhead D. R. (2006) Arabidopsis Co-expression Tool (ACT): web server tools for microarray-based gene expression analysis; Nucleic Acids Res. July 1; 34(Web Server issue): W504–W509.

Mayo M. S., Gajewski B. J., and Morris J. S. (2006) Some statistical issues in microarray gene expression data. Radiat. Res. 165, 745–748.

Moriyama M., Hoshida Y., Otsuka M., Nishimura S., Kato N., Goto T., Taniguchi H., Shiratori Y., Seki N., and Omata M. (2003) Relevance network between chemosensitivity and transcriptome in human hepatoma cells. Mol Cancer Ther., 2(2):199-205.

New J. R., Kendall W., Huang J., and Chesler E. (2008). Dynamic visualization of co-expression in systems genetics data. In press. IEEE Transactions on Visualization and Computer Graphics. 14(5).

Oltvai Z. N., and Barabási A. L. (2002) Systems Biology. Life's complexity pyramid. Science, 298(5594):763-4.

Ozbudak E. M., Thattai M., Kurtser I., Grossman A. D. and van Oudenaarden A. (2002) Regulation of noise in the expression of a single gene. Nat. Genet., 31(1): 69-73.

Page G. P., Edwards J. W., Gadbury G. L., Yelisetti P., Wang J., Trivedi P. and Allison D. B. (2006) The PowerAtlas: a power and sample size atlas for microarray experimental design and research. BMC Bioinformatics, 7:84.

Perkins A. D. (2008) Addressing Challenges in a Graph-Based Analysis of High Throughput Microarray Data. PhD thesis, University of Tennessee, Knoxville.

Politis D. N. and Romano J. P. (1993) The stationary bootstrap. Journal of the American Statistical Association, 89: 1303-1313.

Press W. H., Flannery B. P., Teukolsky S. A. and Vetterling W. T. (1992) Numerical Recipes in C: The Art of Scientific Computing. Second edition. ISBN: 0521437202.

Quackenbush J. (2001) Computational analysis of microarray data. Nature reviews. 2(6):418-427.

Quackenbush J. (2003) Genomics. Microarrays--guilt by association. Science. 302(5643):240-1.

Quackenbush J. (2007) Extracting biology from high-dimensional biological data. J Exp Biol., 210(9):1507-17.

Rajasekaran S. A., Barwe S. P., and Rajasekaran A. K. (2005) Multiple functions of Na,K-ATPase in epithelial cells. Semin Nephrol., 25:328–334.

Rougemont J., and Hingamp P. (2003) DNA microarray data and contextual analysis of correlation graphs. BMC Bioinformatics, 4:15.

Sanoudou D., Haslett J. N., Kho A. T., Guo S., Gazda H. T., Greenberg S. A., Lidov H. G., Kohane I. S., Kunkel L. M., and Beggs A. H. (2003) Expression profiling reveals altered satellite cell numbers and glycolytic enzyme transcription in nemaline myopathy muscle. Proc Natl Acad Sci U S A., 100(8):4666-71.

Schena M., Shalon D., Davis R. W., and Brown P. O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 270: 467–470.

Setubal J. C., and Meidanis J. (1997) Introduction to Computational Molecular Biology, Boston, Mass: PWS Publishing Company.

Shannon P., Markiel A., Ozier O., Baliga N.S., Wang J.T., Ramage D., Amin N., Schwikowski B. and Ideker T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research, 13(11):2498-504.

Simon R., Radmacher M. D., and Dobbin K. (2002) Design of studies using DNA microarrays. Genet. Epidemiol., 23:21–36.

Slonim D. K. (2002) From patterns to pathways: gene expression data analysis comes of age. Nat Genet. 32 Suppl:502-8.

Smyth G. K., Yang Y. H., and Speed T. (2003) Statistical issues in cDNA microarray data analysis. Methods Mol Biol. 224:111-36.

Spellman P. T., Sherlock G., Zhang M. Q., Iyer V. R., Anders K., Eisen M. B., Brown P. O., Botstein D., and Futcher B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Molecular Biology of the Cell 9, 3273-3297.

Storey J. D. (2002) A direct approach to false discovery rates. Journal of Royal Statistical Society, Series B (Statistical Methodology). 64:3, 479-498.

Storey J. D. (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. The Annals of Statistics, 31:6, 2013-2035.

Storey J. D., and Tibshirani R. (2003) Statistical significance for genome wide studies. Proc Natl Acad Sci U S A. 100:9440–9445.

Struhl K. (1999) Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. Cell. 98(1):1-4.

Stuart J. M., Segal E., Koller D., and Kim S. K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science. 302(5643): 249-55.

Szodoray P., Alex P., Frank M. B., Turner M., Turner S., Knowlton N., Cadwell C., Dozmorov I., Tang Y., Wilson P. C., Jonsson R., and Centola M. (2006) A genome-scale assessment of peripheral blood B-cell molecular homeostasis in patients with rheumatoid arthritis. Rheumatology 45(12): 1466-1476.

Tefferi A., Wieben E. D., Dewald G. W., Whiteman D. A. H., Bernard M. E., and Spelsberg T. C. (2002) Primer on medical genomics part II: Background principles and methods in molecular genetics. Mayo Clinic Proc. 77:785–808.

Tomita E., Tanaka A., and Takahashi H. (2004) The worst-case time complexity for generating all Maximal Cliques, Proceedings, Computing and Combinatorics Conference, Jeju Island, Korea.

Tusher V. G., Tibshirani R., and Chu G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A. 98:5116–21.

van Noort V., Snel B., and Huynen M. A. (2003) Predicting gene function by conserved co-expression. Trends Genet. 19:238–242.

Voy B. H., Scharff J. A., Perkins A. D., Saxton A. M., Borate B., Chesler E. J., Branstetter L. K., and Langston M. A. (2006) Extracting gene networks for low-dose radiation using graph theoretical algorithms; PLoS Comput Biol. 2(7):e89.

Watts D.J, and Strogatz S.H. (1998) Collective dynamics of “small-world” networks. Nature. 393(6684):440–442.

Wei C., Li J. and Bumgarner R. E. (2004) Sample size for detecting differentially expressed genes in microarray experiments. BMC Genomics, 5: 87.

Wolfe C. J., Kohane I. S., and Butte A. J. (2005) Systematic survey reveals general applicability of “guilt-by-association” within gene coexpression networks. BMC Bioinformatics, 6: 227.

Wu L. F., Hughes T. R., Davierwala A. P., Robinson M. D., Stoughton R., and Altschuler S. J. (2002) Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. Nat Genet, 31: 255–265.

Wu S., and Li J. (2007) Comparative Analysis of Gene-Coexpression Networks Across Species. Book: Bioinformatics Research and Applications. Măndoiu I. and Zelikovsky A. (Eds.): ISBRA 2007, LNBI 4463, 615–626.

Yan X., Mehan M. R., Huang Y., Waterman M. S., Yu P. S., and Zhou X. J. (2007) A graph-based approach to systematically reconstruct human transcriptional regulatory modules; Bioinformatics, 23(13):i577-86.

Zeeberg B. R., Feng W., Wang G., Wang M. D., Fojo A. T., Sunshine M., Narasimhan S., Kane D. W., Reinhold W. C., Lababidi S., Bussey K. J., Riss J., Barrett J. C., and Weinstein J. N. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol, 4:R28.

Zhang B., Schmoyer D., Kirov S., and Snoddy J. (2004) GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. BMC Bioinformatics. 5:16.

Zhang Y., Abu-Khzam F. N., Baldwin N. E., Chesler E. J., Langston M. A., and Samatova N. F. (2005) Genome-scale computational approaches to memoryintensive applications in systems biology. Supercomputing 2005. Proceedings of the ACM/IEEE SC 2005 Conference, 12-18 Nov. Page(s):12 - 12.

Zhang B. and Horvath S. (2005) A general framework for weighted gene co-expression network analysis", Statistical Applications in Genetics and Molecular Biology. 4:1, Article 17.

Zhou X., Kao M. C., and Wong W.H. (2002) Transitive functional annotation by shortest-path analysis of gene expression data. Proc Natl Acad Sci U S A. 99(20):12783-8.

APPENDIX

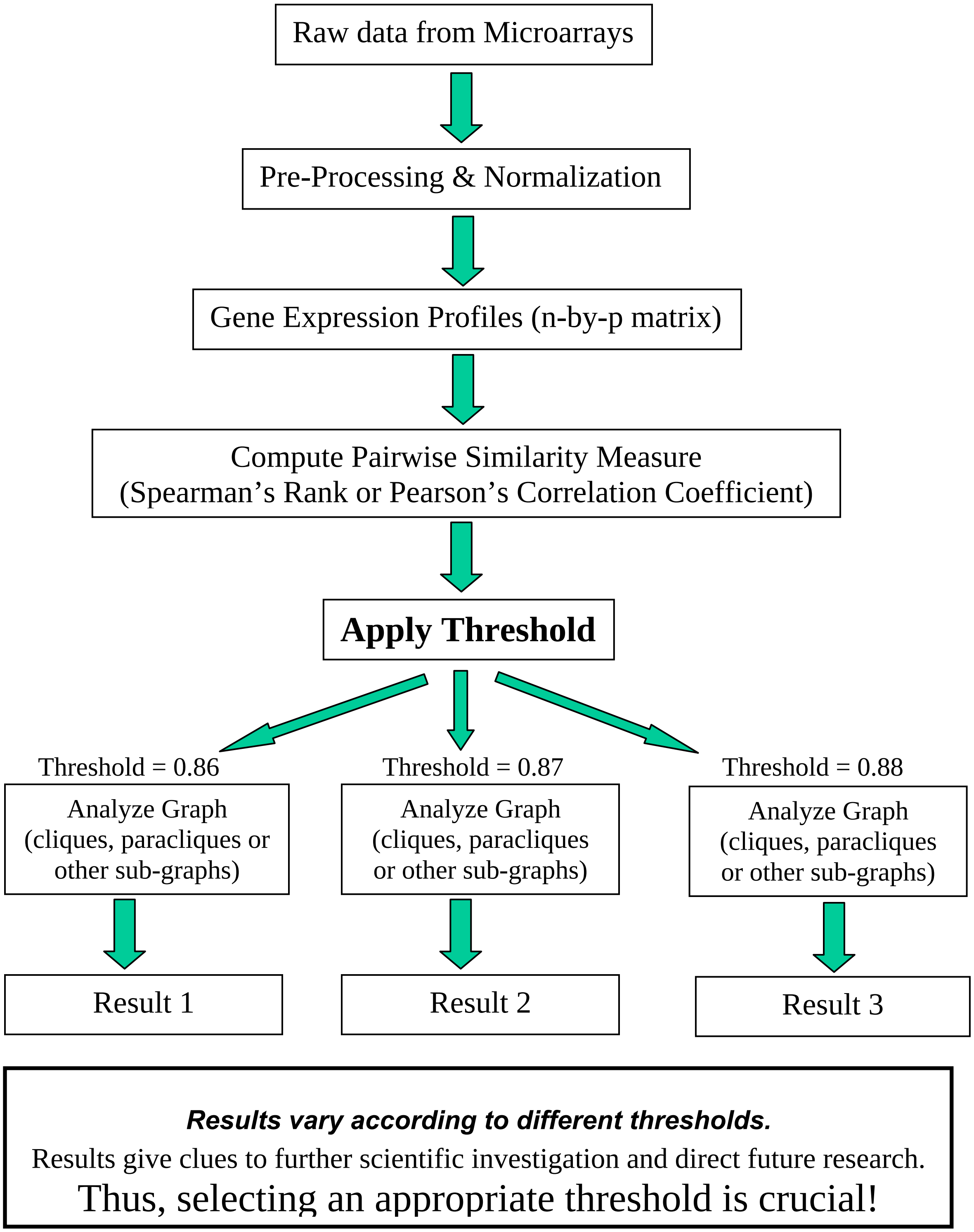

Raw data from Microarrays
Pre-Processing & Normalization
Gene Expression Profiles (n-by-p matrix)
Compute Pairwise Similarity Measure
(Spearman’s Rank or Pearson’s Correlation Coefficient)
Apply Threshold
Threshold = 0.86
Threshold = 0.87
Threshold = 0.88
Analyze Graph (cliques, paracliques or other sub-graphs)
Analyze Graph (cliques, paracliques or other sub-graphs)
Analyze Graph (cliques, paracliques or other sub-graphs)
Result 1
Result 2
Result 3
Results vary according to different thresholds.
Results give clues to further scientific investigation and direct future research.
Thus, selecting an appropriate threshold is crucial!

Figure 1. Flowchart routine in graph-based microarray data analysis [Voy et al. 2006].

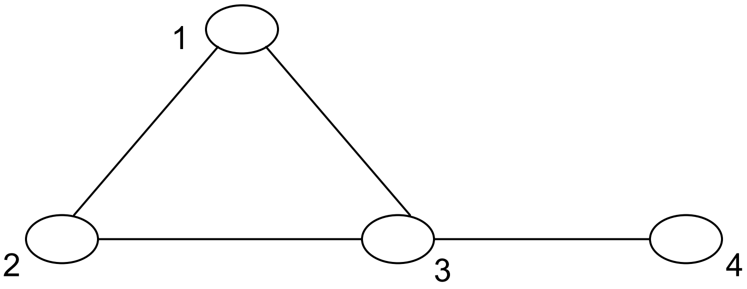


Figure 2. A simple graph

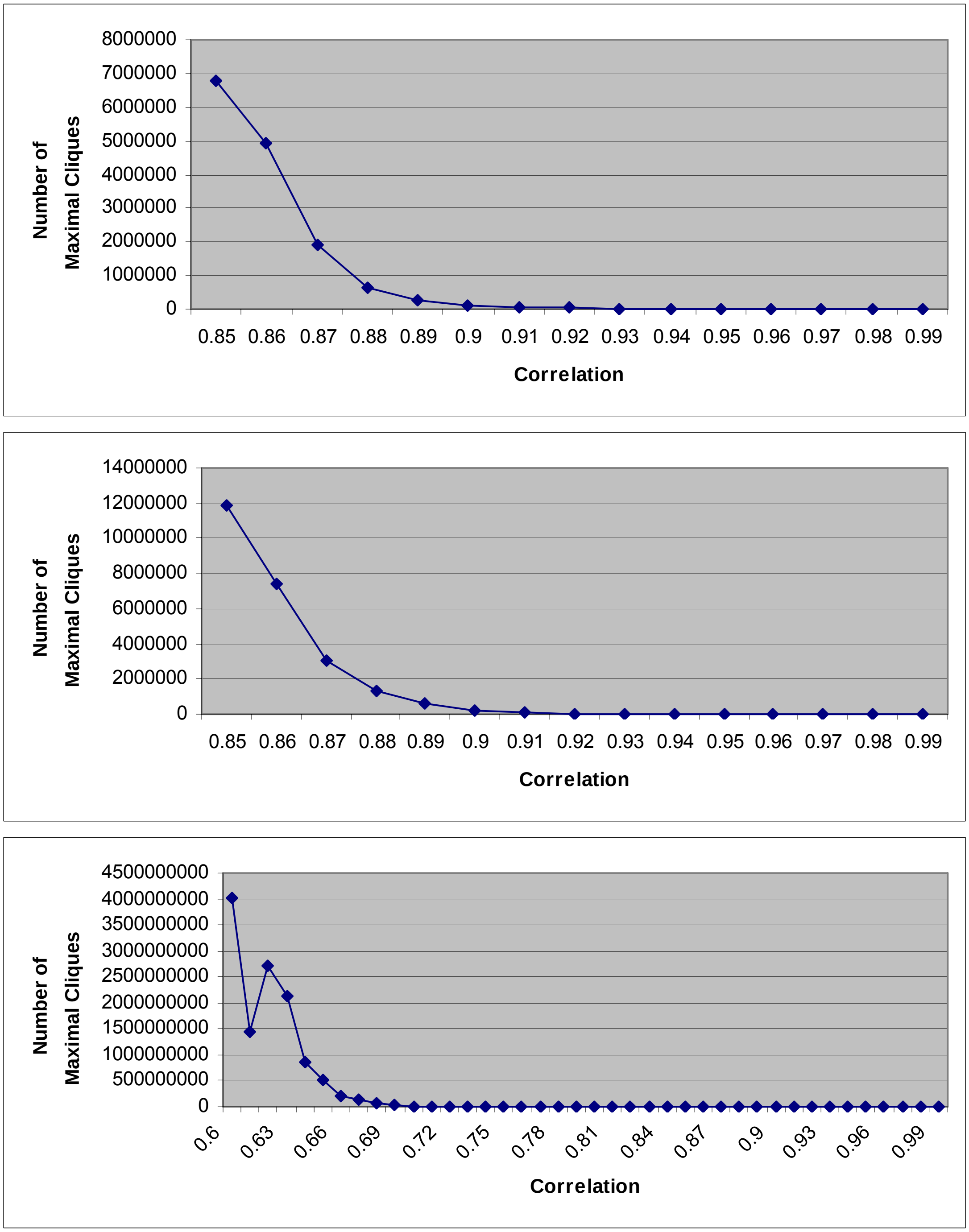


Figure 3. Distribution of Maximal Cliques. *Top: Anoxia data (τ = 0.9), Middle: Reoxygenation data (τ = 0.91), Bottom: Alpha data (τ = 0.74). Graphs generated from Anoxia and Reoxygenation datasets are very dense as compared to Alpha dataset. This is reflected in the thresholds: τ for Alpha dataset is very low.*

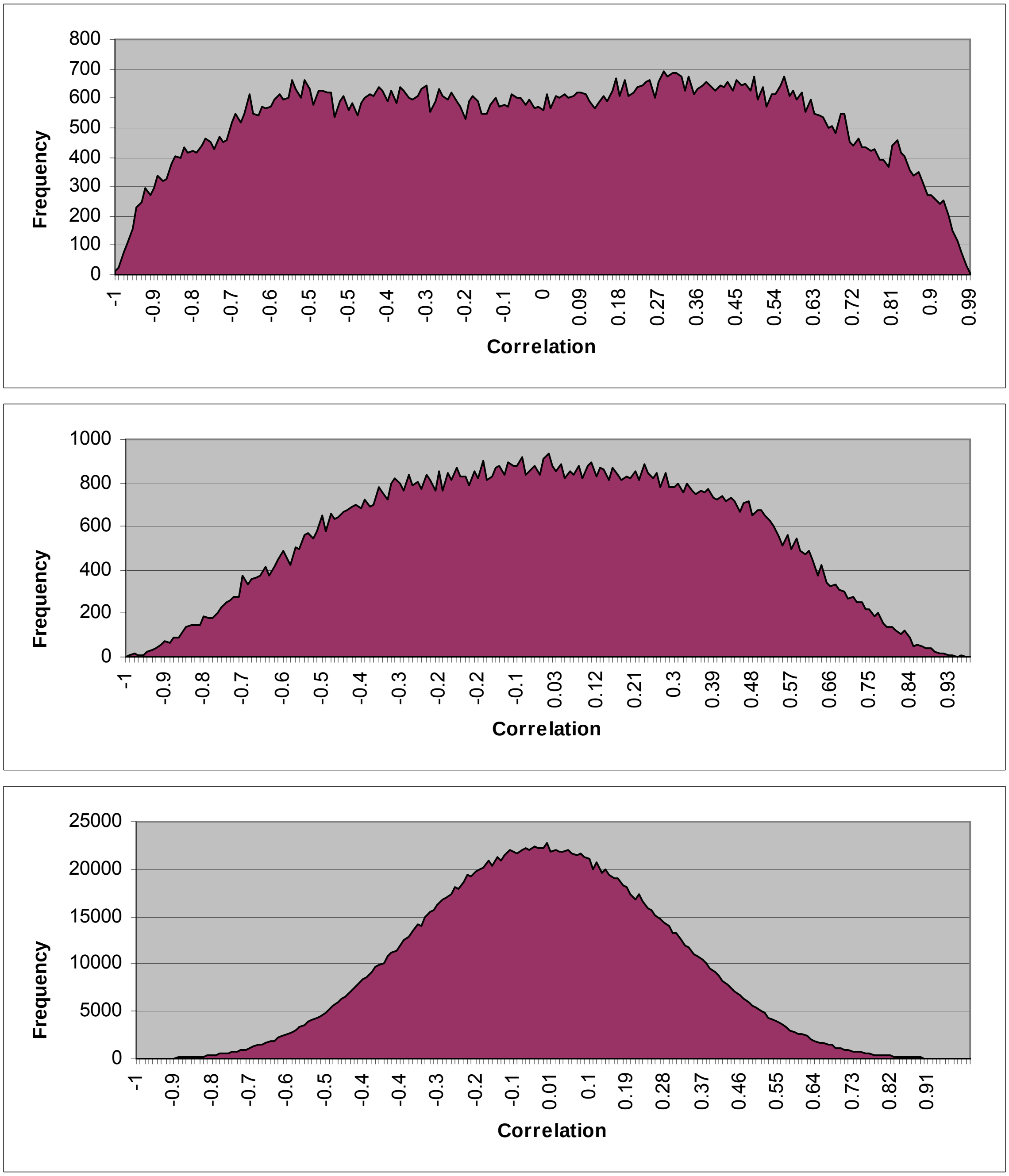


Figure 4. Distribution of Control Spot Correlations. *Top: Anoxia data (τ = 0.93), Middle: Reoxygenation data (τ = 0.83), Bottom: Alpha data (τ = 0.7). The distribution for Alpha dataset is close to normal, unlike Anoxia and Reoxygenation datasets, from a higher number of control spot data available for analysis.*

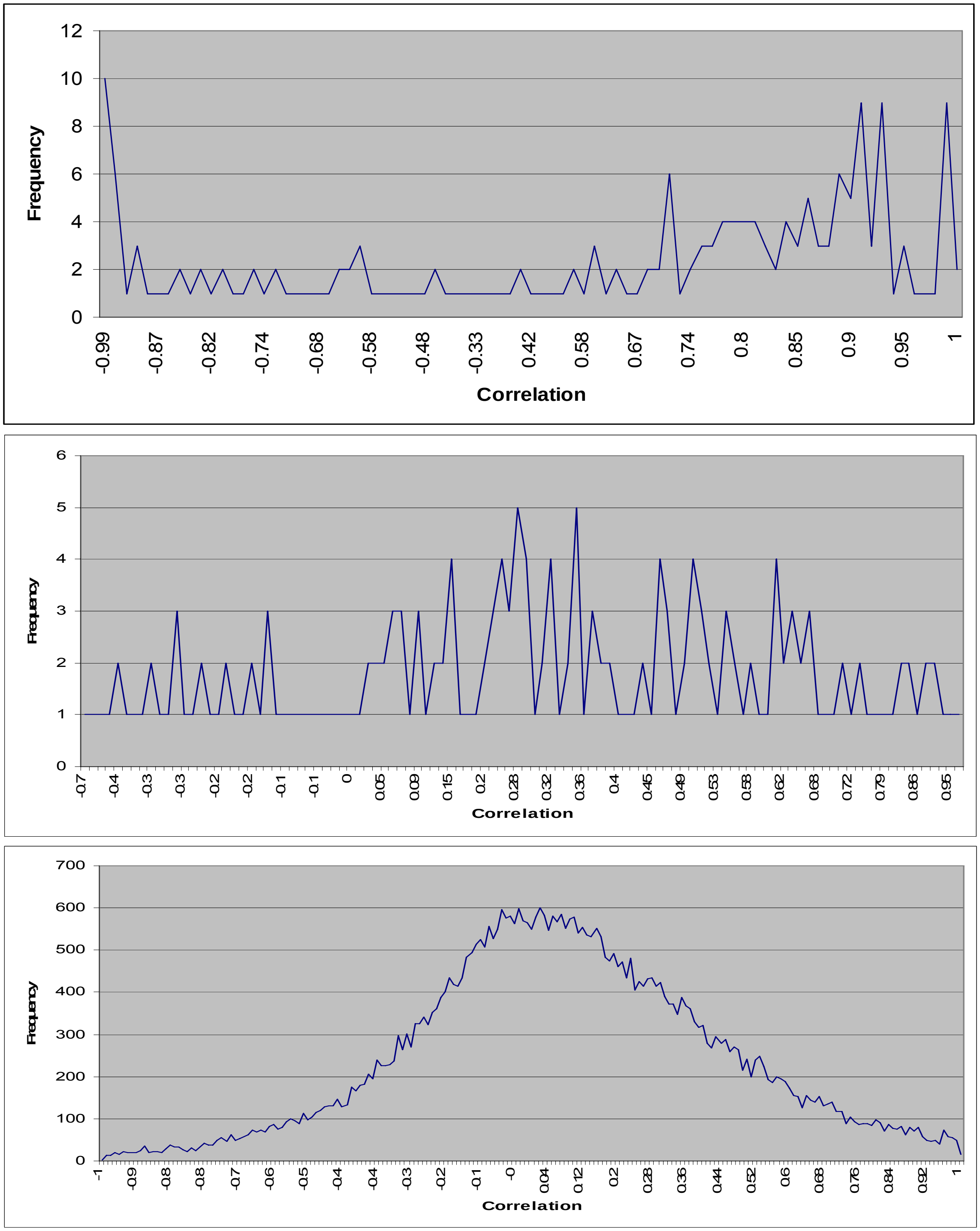


Figure 5. Correlations within Control spots. *Top: Anoxia data (τ = 0.93) with information from 20 control spots for 15 arrays; Middle: Reoxygenation data (τ = 0.83) with information from 20 control spots for 10 arrays; Bottom: Alpha data (τ = 0.7) with information from 308 control spots for 18 arrays.*

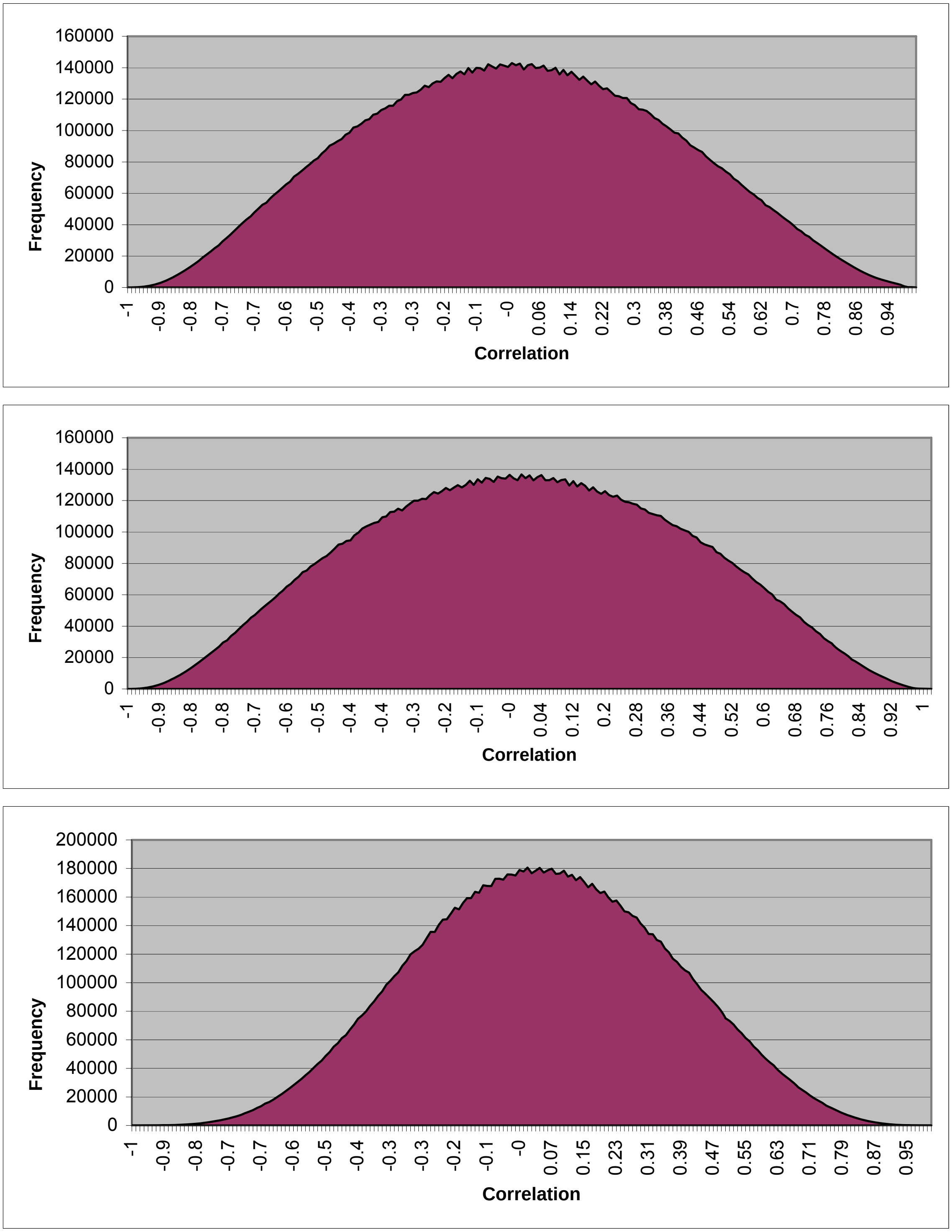


Figure 6. Distribution of correlations for Top 1% Method. *Top: Anoxia data (τ = 0.81), Middle: Reoxygenation data (τ = 0.81), Bottom: Alpha data (τ = 0.72). The distribution is closer to normal for the Alpha dataset.*

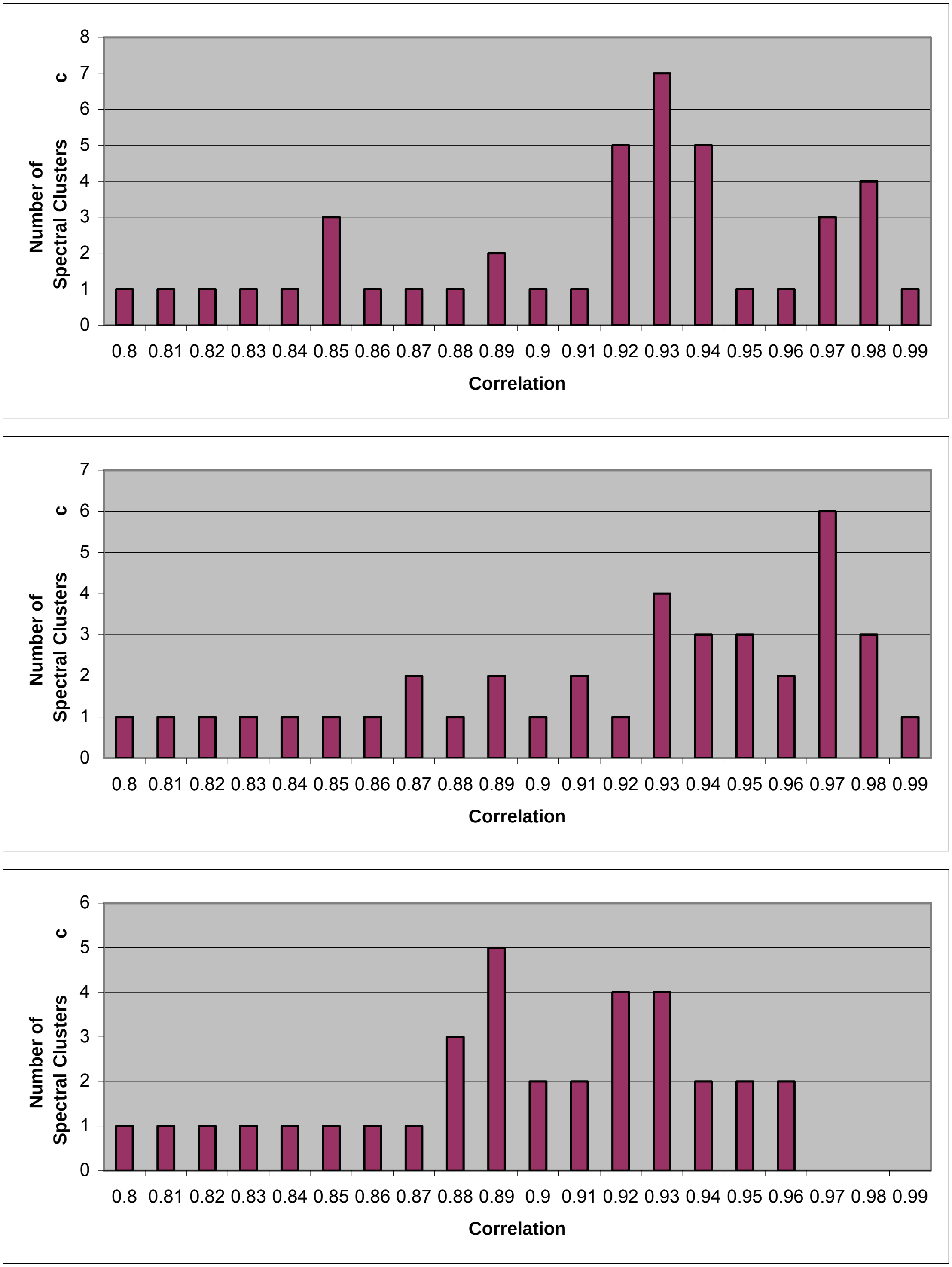


Figure 7. Spectral Clusters. *Top: Anoxia data (τ = 0.93), Middle: Reoxygenation data (τ = 0.97), Bottom: Alpha data (τ = 0.89). The correlation value with the highest number of spectral clusters represents the threshold.*

Numeric Results when Ha: R0<>R1

Power	N	Alpha	Beta	R0	R1
0.65833	31	0.00000	0.34167	0	0.85
0.72854	31	0.00000	0.27146	0	0.86
0.79452	31	0.00000	0.20548	0	0.87
0.85343	**31**	**0.00000**	**0.14657**	**0**	**0.88**
0.90285	31	0.00000	0.09715	0	0.89
0.94125	31	0.00000	0.05875	0	0.9
0.96837	31	0.00000	0.03163	0	0.91
0.98534	31	0.00000	0.01466	0	0.92
0.99442	31	0.00000	0.00558	0	0.93
0.99838	31	0.00000	0.00162	0	0.94
0.99968	31	0.00000	0.00032	0	0.95
0.99996	31	0.00000	0.00004	0	0.96
1.00000	31	0.00000	0.00000	0	0.97
1.00000	31	0.00000	0.00000	0	0.98
1.00000	31	0.00000	0.00000	0	0.99

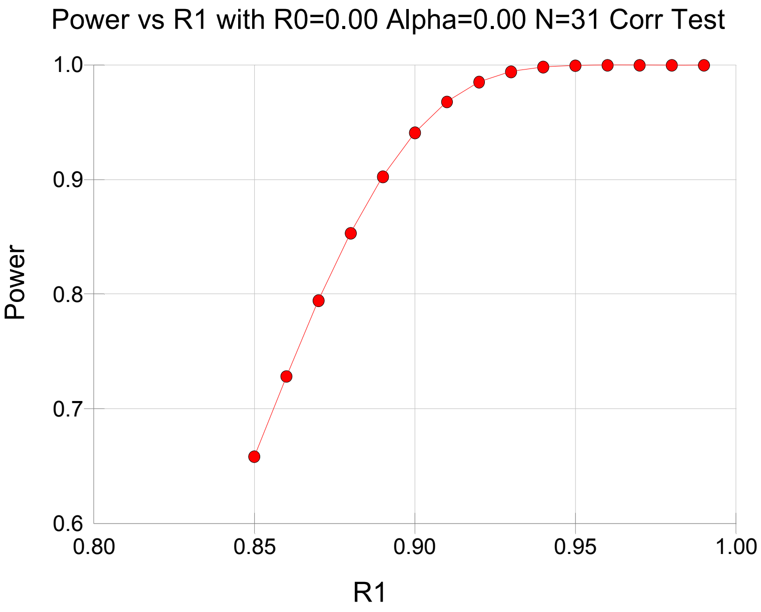


Figure 8. Output from PASS statistical software (*http://www.ncss.com/pass.html*) depicting Power versus Correlation for Anoxia data with 31 arrays (τ = 0.88).

Numeric Results when Ha: R0<>R1

Power	**N**	**Alpha**	**Beta**	**R0**	**R1**
0.11655	21	0.00000	0.88345	0	0.85
0.1508	21	0.00000	0.8492	0	0.86
0.19437	21	0.00000	0.80563	0	0.87
0.24912	21	0.00000	0.75088	0	0.88
0.31674	21	0.00000	0.68326	0	0.89
0.39829	21	0.00000	0.60171	0	0.9
0.49347	21	0.00000	0.50653	0	0.91
0.5996	21	0.00000	0.4004	0	0.92
0.71056	21	0.00000	0.28944	0	0.93
0.8163	**21**	**0.00000**	**0.1837**	**0**	**0.94**
0.90418	21	0.00000	0.09582	0	0.95
0.96339	21	0.00000	0.03661	0	0.96
0.9918	21	0.00000	0.0082	0	0.97
0.99936	21	0.00000	0.00064	0	0.98
1.00000	21	0.00000	0.00000	0	0.99

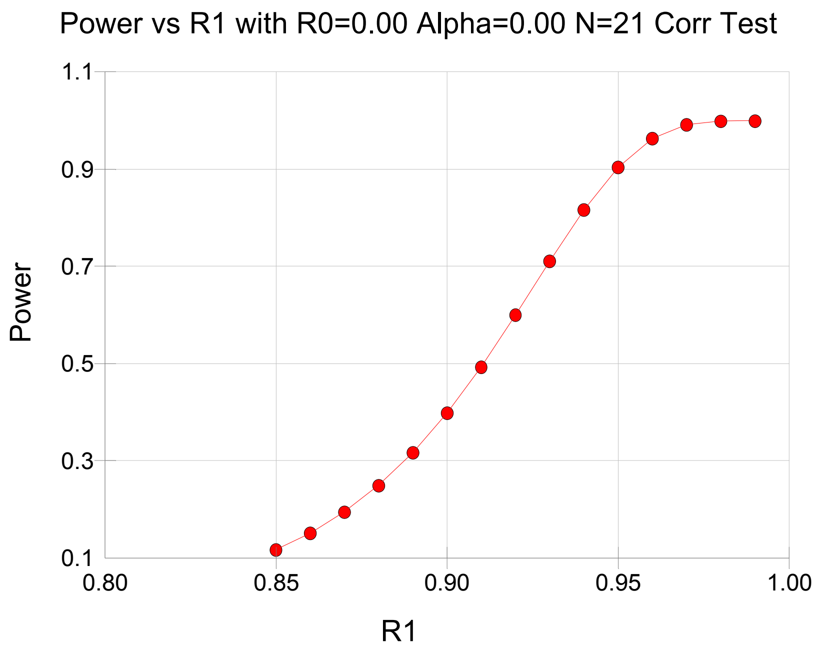


Figure 9. Output from PASS statistical software (*http://www.ncss.com/pass.html*) depicting Power versus Correlation for Reoxygenation dataset with 21 arrays (τ = 0.94).

Numeric Results when Ha: R0<>R1

Power	**N**	**Alpha**	**Beta**	**R0**	**R1**
0.03689	18	0.00000	0.96311	0	0.85
0.04984	18	0.00000	0.95016	0	0.86
0.0676	18	0.00000	0.9324	0	0.87
0.09195	18	0.00000	0.90805	0	0.88
0.12527	18	0.00000	0.87473	0	0.89
0.17062	18	0.00000	0.82938	0	0.9
0.2317	18	0.00000	0.7683	0	0.91
0.31247	18	0.00000	0.68753	0	0.92
0.4161	18	0.00000	0.5839	0	0.93
0.54268	18	0.00000	0.45732	0	0.94
0.68533	18	0.00000	0.31467	0	0.95
0.82581	**18**	**0.00000**	**0.17419**	**0**	**0.96**
0.93497	18	0.00000	0.06503	0	0.97
0.98923	18	0.00000	0.01077	0	0.98
0.9998	18	0.00000	0.0002	0	0.99

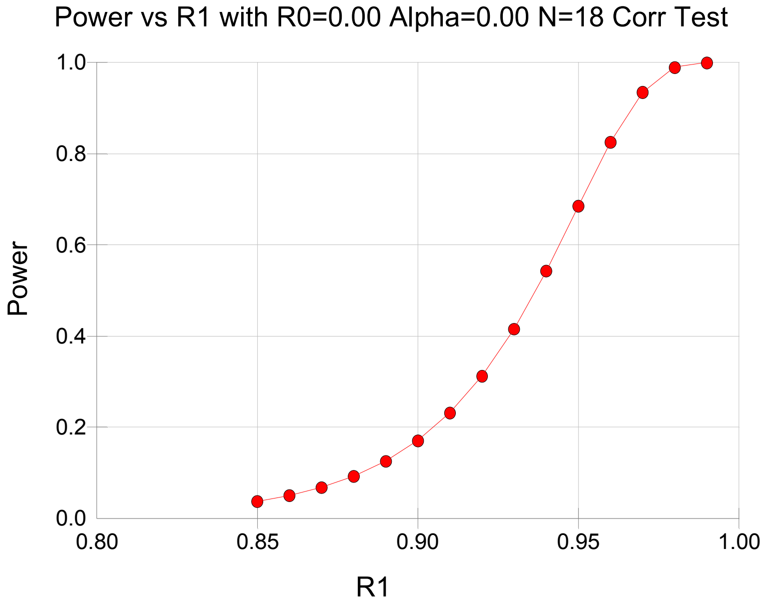


Figure 10.Output from PASS statistical software (*http://www.ncss.com/pass.html*) depicting Power versus Correlation for Alpha dataset with 18 arrays (τ = 0.96).

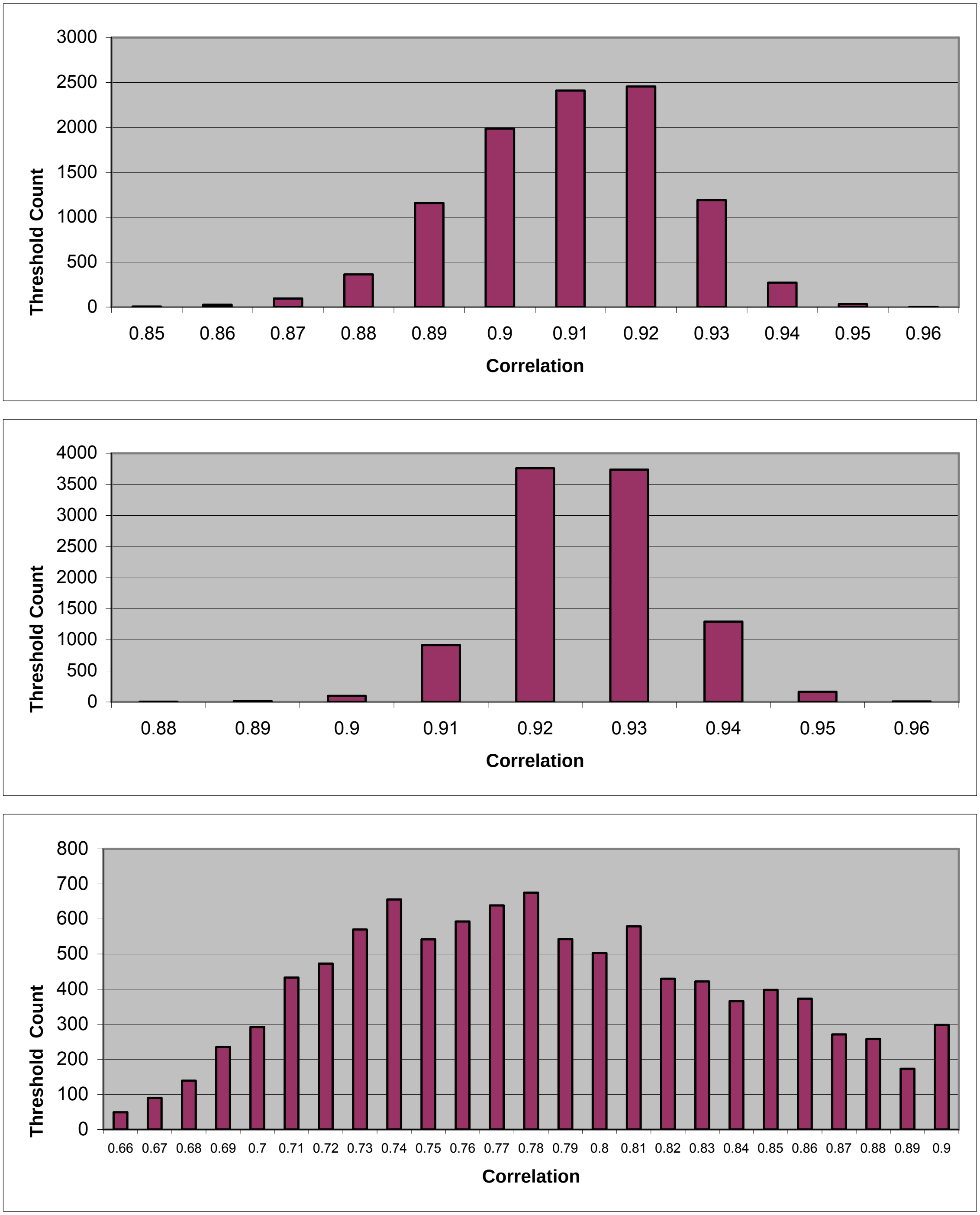
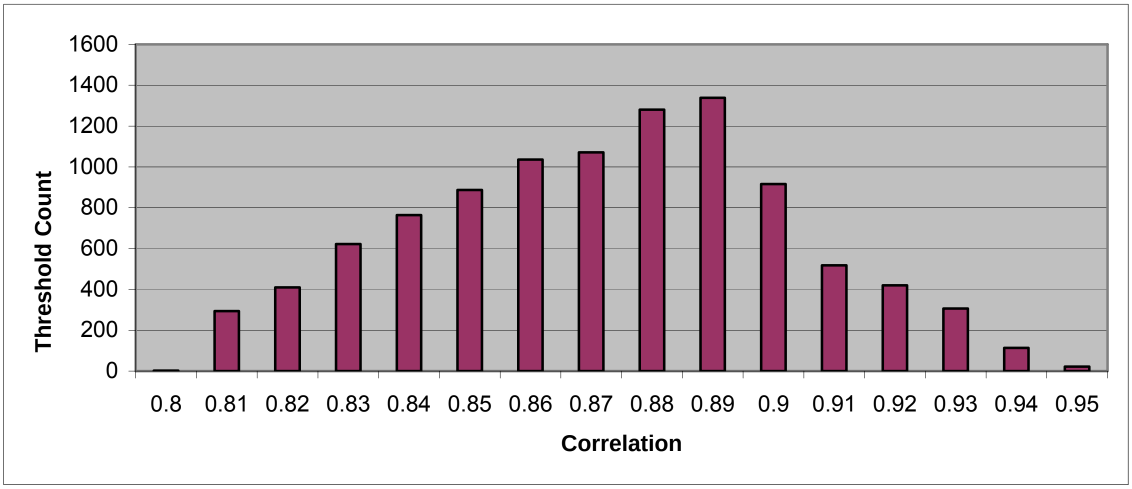


Figure 11. Bootstrap Results for Maximal Clique-2 method. Top: Anoxia data (τ = 0.9), Middle: Reoxygenation data (τ = 0.91), Bottom: Alpha data (τ = 0.74). τ is close to the mode of threshold distribution for Anoxia and Alpha datasets.



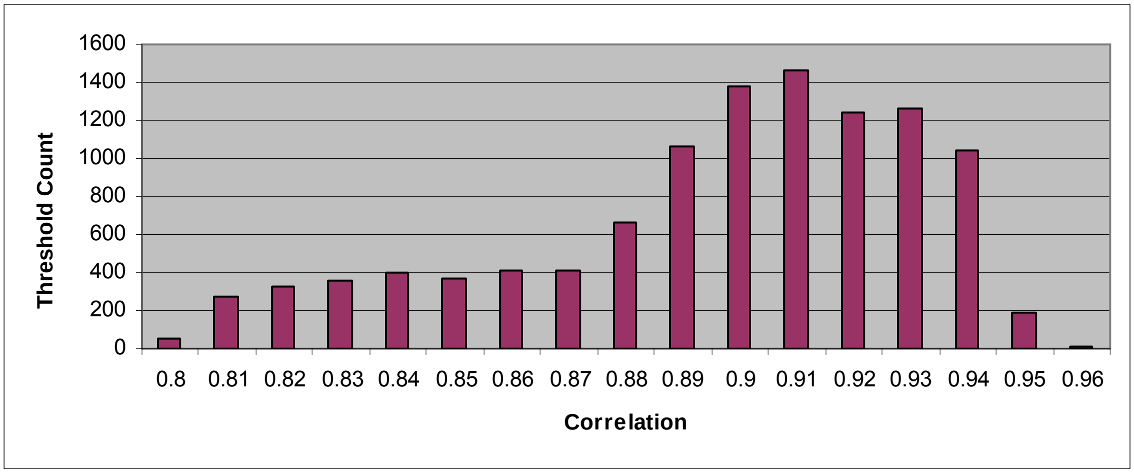


Figure 12. Bootstrap Results for Maximal Clique-3 method. *Top: Anoxia data (τ = 0.87). Bottom: Reoxygenation data (τ = 0.89). τ is close to the mode of threshold distribution for both Anoxia and Reoxygenation datasets.*

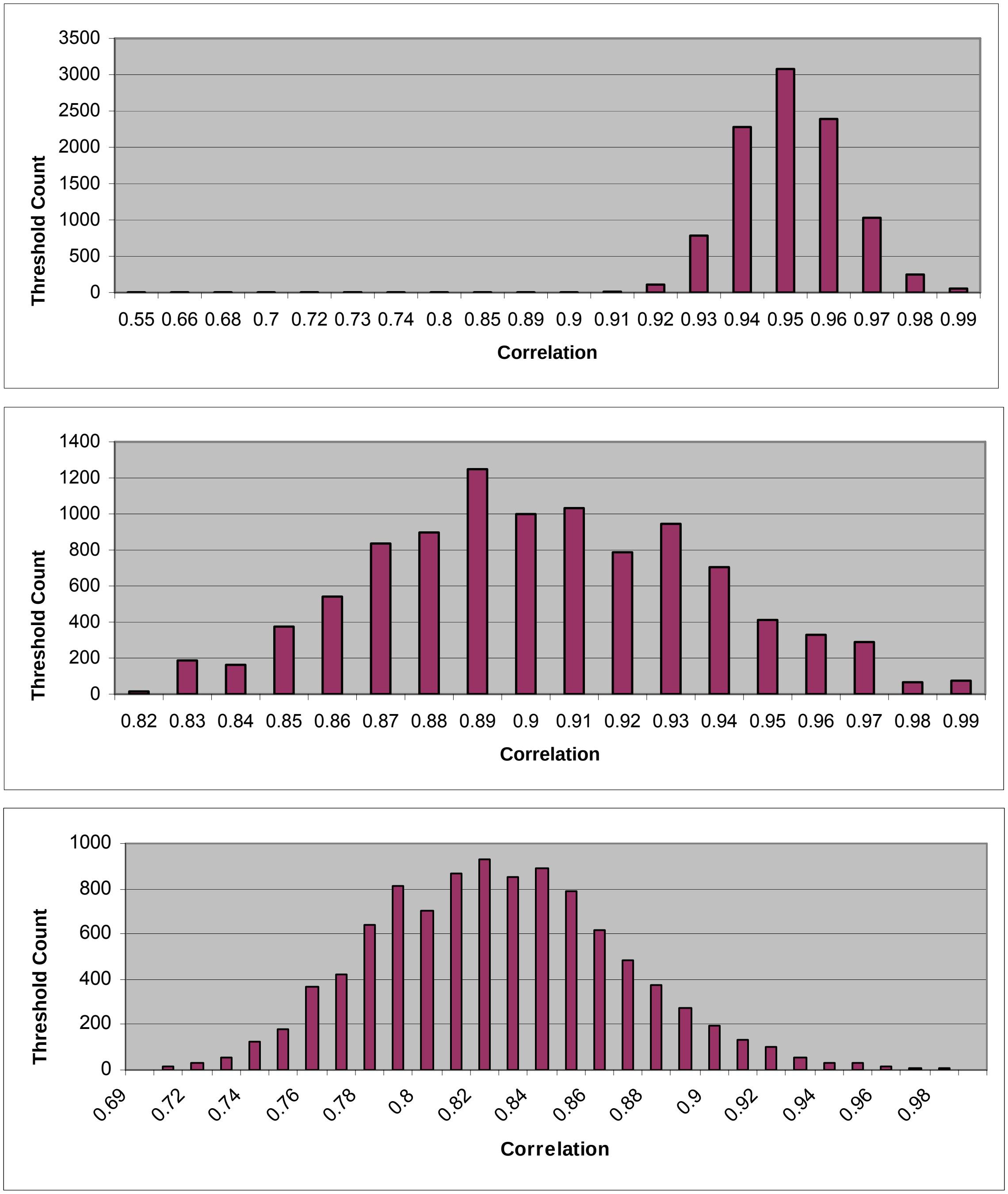


Figure 13. Bootstrap Results for Control Spot Verification method. *Top: Anoxia data (τ = 0.93). Middle: Reoxygenation data (τ = 0.83). Bottom: Alpha data (τ = 0.7). τ is far away from the mode of the threshold distribution for all three datasets.*

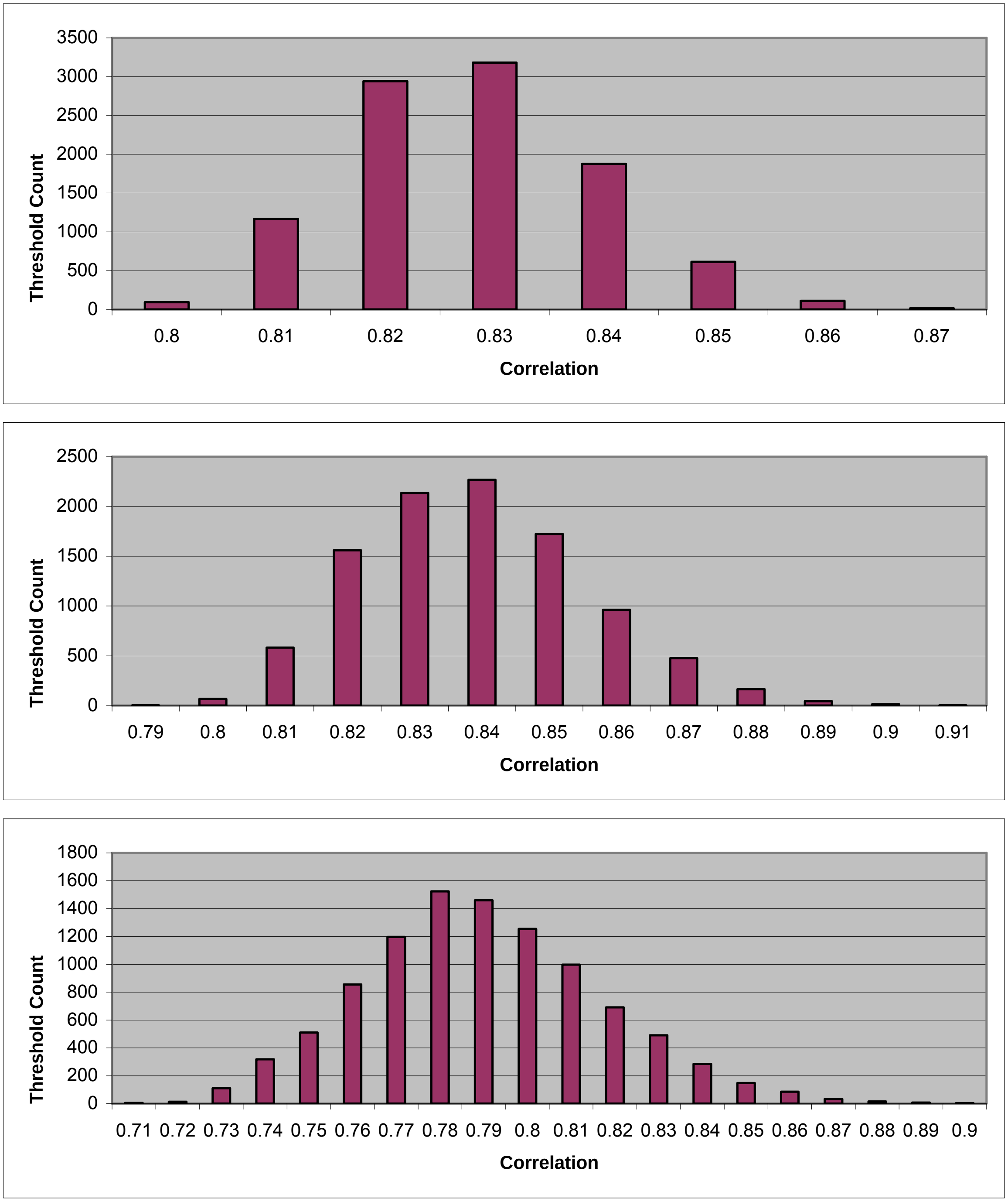


Figure 14. Bootstrap Results for Top 1% of Correlations method. *Top: Anoxia data (τ = 0.81). Middle: Reoxygenation data (τ = 0.81). Bottom: Alpha data (τ = 0.72). τ is far away from the mode of the threshold distribution for all three datasets.*

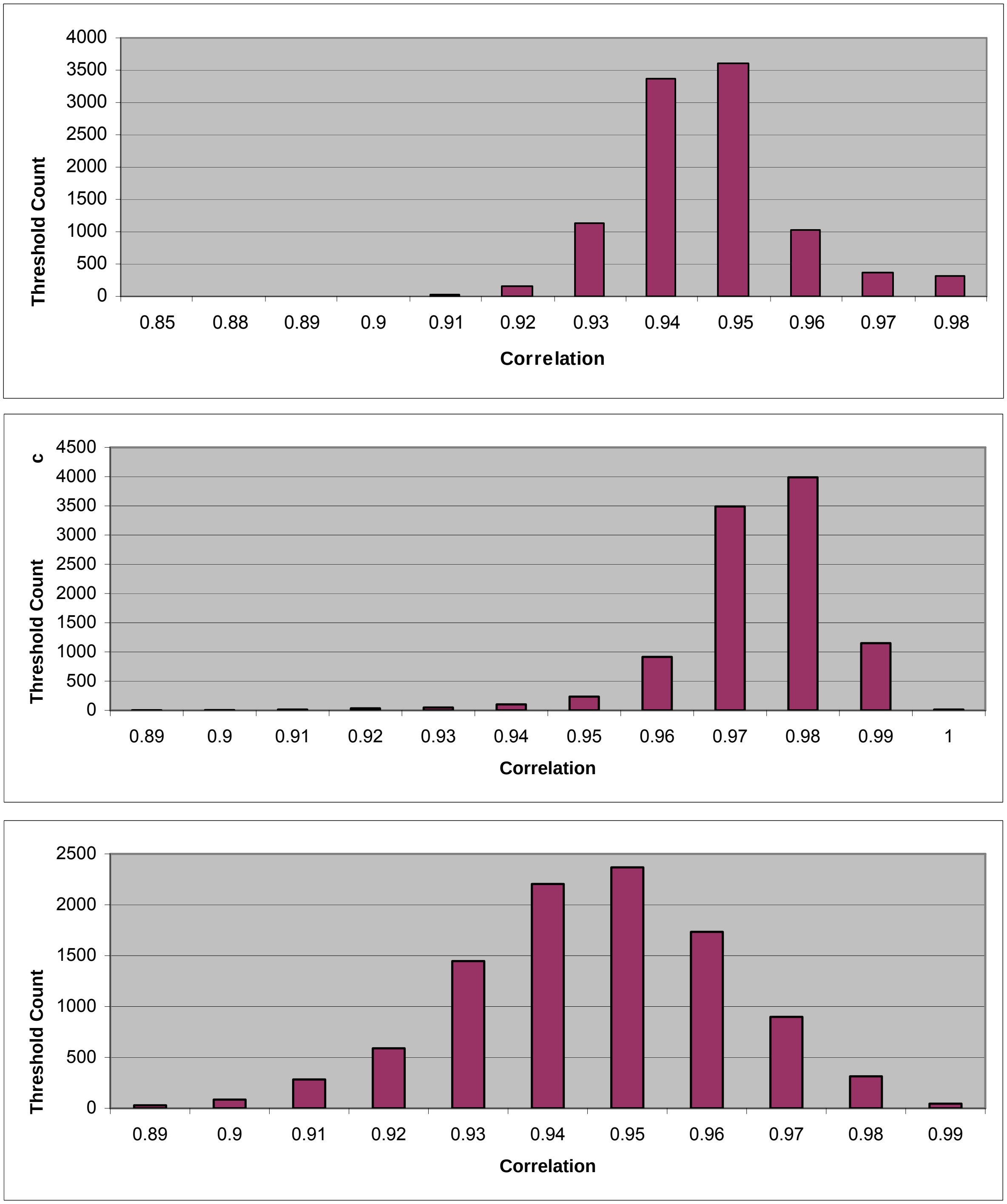
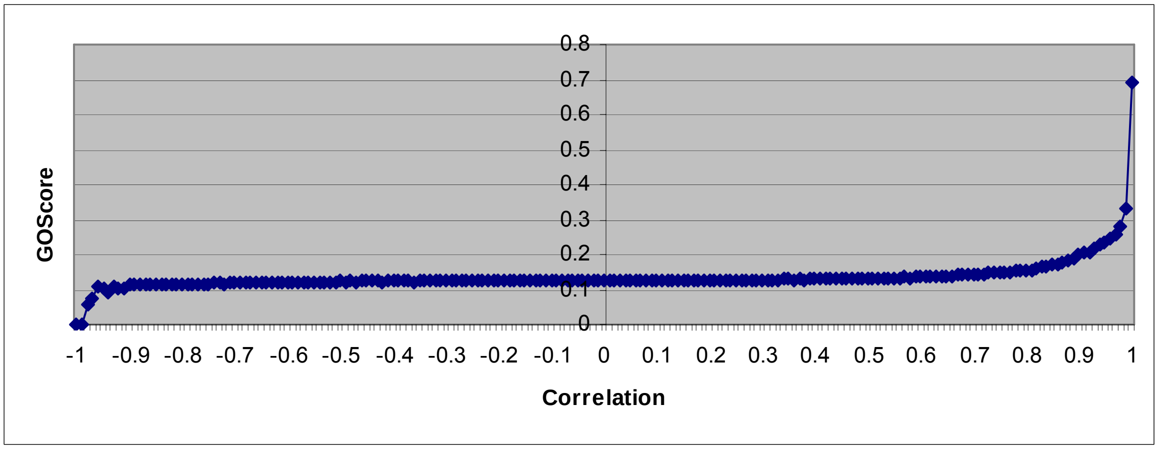
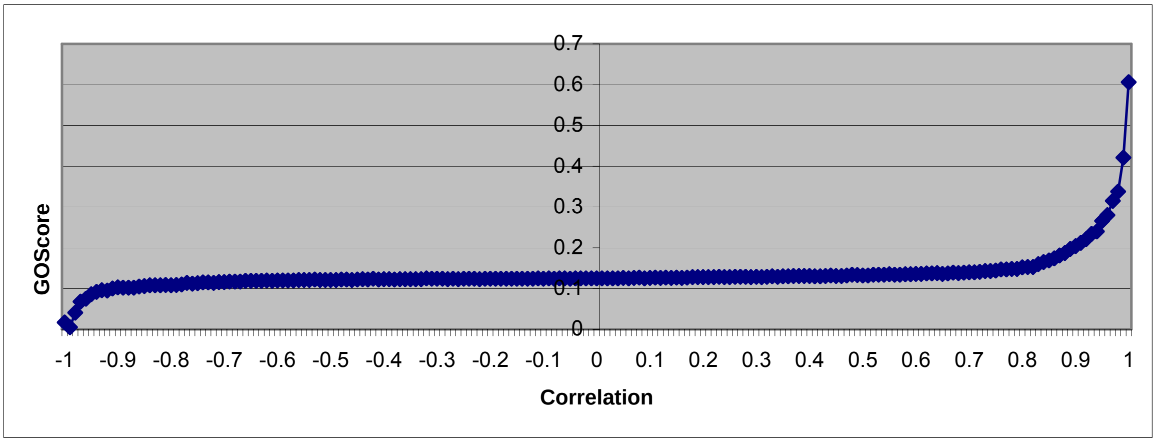


Figure 15. Bootstrap Results for Spectral Graph Clustering method. *Top: Anoxia data (τ = 0.93). Middle: Reoxygenation data (τ = 0.97). Bottom: Alpha data (τ = 0.89). τ is very close to mode of threshold distribution only for Reoxygenation dataset.*





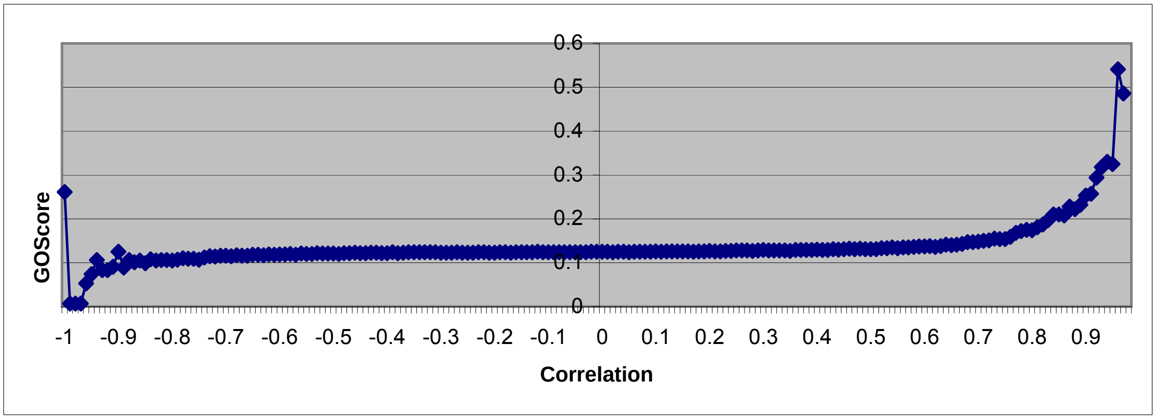


Figure 16. Distribution of Functional Similarity Score against correlations. *Top: Anoxia data. Middle: Reoxygenation data. Bottom: Alpha data. The score is high at very high positive correlations. At high negative correlations, the score falls almost to 0, except for Alpha dataset in which the score shows a rise. However, the rise is not as high as at the positive correlation end.*

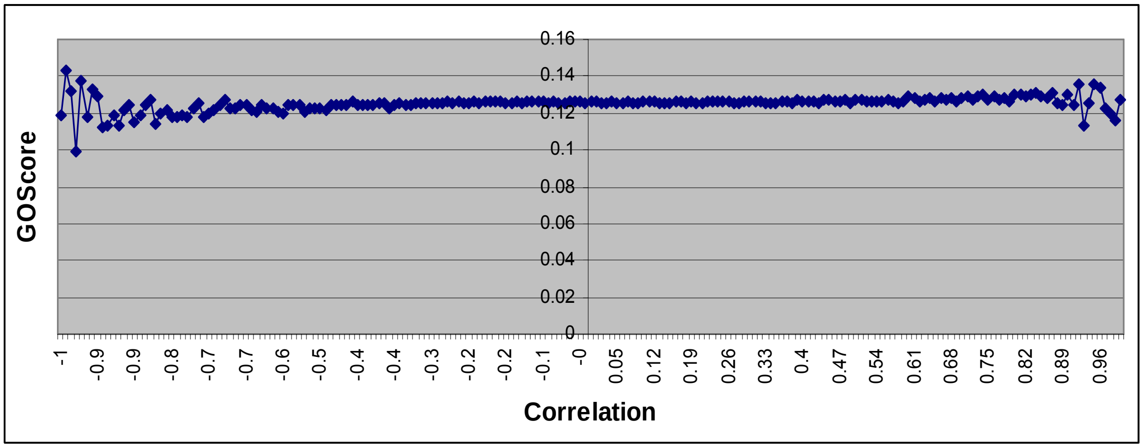


Figure 17. Distribution of Functional Similarity Score against correlations for cdc15 dataset from Yeast Cell Cycle Project [Spellman et al. 1998]. *The score does not show a rise at high positive correlations.*

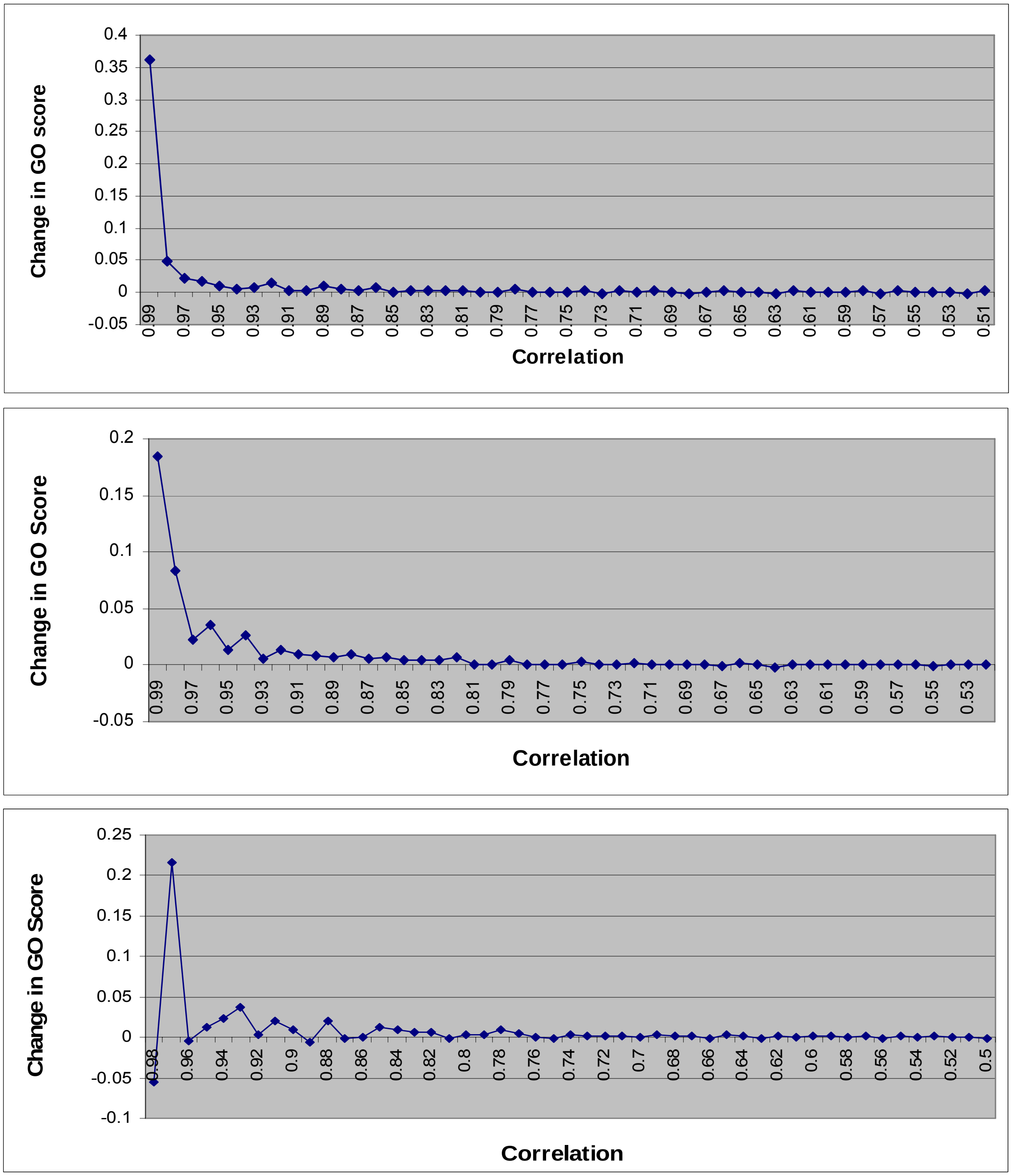


Figure 18. Change in GO Score versus Correlation. *Top: Anoxia data (τ = 0.97). Middle: Reoxygenation data (τ = 0.92). Bottom: Alpha data (τ = 0.85).*

Table 1. Graph properties for Anoxia data. *The number of maximal cliques grows to more than 50000 and the first instance of doubling over the previous correlation of 0.91 occurs at correlation threshold of 0.9, which is thus chosen as the threshold for the Maximal Clique-2 method. For the Maximal Clique-3 method, the threshold becomes 0.87, when the first instance of tripling of the number of maximal cliques occurs.*

Threshold	Vertices	Edges	Density	Number of maximal cliques	Maximum clique size
0.99	202	325	0.016	32	13
0.98	392	1073	0.014	312	15
0.97	600	2896	0.016	1675	20
0.96	825	5477	0.016	8077	32
0.95	1108	8896	0.015	11037	49
0.94	1385	13129	0.014	15730	60
0.93	1708	18413	0.013	16320	73
0.92	2033	25079	0.012	30257	82
0.91	2345	33109	0.012	46283	91
0.9	**2609**	**42841**	**0.013**	**114907**	**98**
0.89	2873	54669	0.013	278115	105
0.88	3119	68762	0.014	624074	112
0.87	**3359**	**85074**	**0.015**	**1887870**	**119**
0.86	3594	104168	0.016	4936760	127
0.85	3703	114963	0.017	6766028	132

Table 2. Graph properties for Reoxygenation data. *The number of maximal cliques grows to more than 50000 and the first instance of doubling over the previous correlation of 0.92 occurs at correlation threshold of 0.91, which is thus chosen as the threshold for the Maximal Clique-2 method. For the Maximal Clique-3 method, the threshold becomes 0.89, when the first instance of tripling of the number of maximal cliques occurs.*

Threshold	**Vertices**	**Edges**	**Density**	**Number of maximal cliques**	**Maximum clique size**
0.99	223	321	0.013	29	11
0.98	485	787	0.007	110	14
0.97	811	1894	0.006	433	17
0.96	1202	3927	0.005	1304	19
0.95	1619	7099	0.005	3005	23
0.94	2041	11687	0.006	6956	32
0.93	2398	17766	0.006	16616	37
0.92	2731	25589	0.007	31988	45
0.91	**3036**	**35563**	**0.008**	**78070**	**52**
0.9	3335	47784	0.009	206786	61
0.89	**3626**	**62394**	**0.009**	**637051**	**67**
0.88	3892	79522	0.011	1323852	79
0.87	4169	99227	0.011	3041128	88
0.86	4381	121972	0.013	7361883	95
0.85	4478	134884	0.013	11858152	100

Table 3. Graph properties for Alpha data. *The number of maximal cliques grows to more than 50000 and the first instance of doubling over the previous correlation of 0.75 occurs at correlation threshold of 0.74, which is thus chosen as the threshold for the Maximal Clique-2 method. For the Maximal Clique-3 method, the threshold becomes 0.6, when the first instance of tripling of the number of maximal cliques occurs.*

Threshold	**Vertices**	**Edges**	**Density**	**Number of maximal cliques**	**Maximum clique size**
0.99	8	4	0.143	0	0
0.98	39	31	0.042	5	3
0.97	97	89	0.019	11	5
0.96	167	181	0.013	28	7
0.95	284	349	0.009	53	8
0.94	464	608	0.006	94	8
0.93	706	1026	0.004	197	10
0.92	1006	1655	0.003	349	10
0.91	1380	2500	0.003	574	13
0.9	1788	3646	0.002	954	13
0.89	2238	5207	0.002	1467	15
0.88	2671	7219	0.002	2303	17
0.87	3082	9701	0.002	3550	19
0.86	3482	12818	0.002	5562	21
0.85	3843	16593	0.002	8579	23
0.84	4186	21126	0.002	13344	25
0.83	4480	26688	0.003	20970	27
0.82	4718	33322	0.003	32927	30
0.81	4926	41145	0.003	50746	32
0.8	5077	50209	0.004	72477	35
0.79	5188	60697	0.005	115990	39
0.78	5266	72963	0.005	207441	42
0.77	5329	86802	0.006	283811	47
0.76	5373	102750	0.007	511424	50
0.75	5401	120346	0.008	819951	53
0.74	**5427**	**140513**	**0.01**	**1664203**	**59**
0.73	5436	163034	0.011	2869894	60
0.72	5450	187756	0.013	4755801	64
0.71	5459	215112	0.014	8707605	68
0.7	5460	245579	0.016	15105804	74
0.69	5463	279162	0.019	36879521	76
0.68	5465	315590	0.021	77793385	79
0.67	5466	356055	0.024	137292075	83
0.66	5466	400035	0.027	216063925	89
0.65	5466	447757	0.03	505219484	94
0.64	5466	499670	0.033	868420486	99
0.63	5466	554607	0.037	2122778657	102
0.62	5466	614720	0.041	2702356249	110
0.61	5466	678883	0.045	1425759180	114
0.6	**5466**	**713226**	**0.048**	**4023958621**	**119**

Table 4. Summary of bootstrap results. *The estimated threshold is compared with various parameters of the bootstrap distribution. The bootstrap frequency of* $\boldsymbol{\tau}$ *is the percentage of times the estimated threshold was selected as the threshold for the bootstrap datasets. Significant results are highlighted in bold.*

Method	Datasets	Estimated Threshold *(τ)*	Bootstrap Mean	Bootstrap Mode	Bootstrap Standard Deviation	95% Confidence Interval for Bootstrap Mean	Bootstrap Frequency of Mode (%)	Bootstrap Frequency of $\boldsymbol{\tau}$ (%)
Maximal Clique-2	Anoxia	0.9	0.91	0.92	0.015	0.9095 – 0.9101	24.55	19.87
	Reoxy	0.91	0.9257	0.92	0.009	0.9254 – 0.9258	37.6	9.16
	Alpha	0.74	0.7833	0.78	0.057	0.7822 – 0.7844	6.75	6.56
Maximal Clique-3	Anoxia	0.87	0.8722	0.89	0.03	0.8716 – 0.8728	13.38	10.71
	Reoxy	0.89	0.8958	0.91	0.036	0.8951 – 0.8965	13.45	9.72
	Alpha	0.6	-	-	-	-	-	-
Control-Spot	Anoxia	0.93	0.9509	0.95	0.015	0.9506 – 0.9512	30.79	7.8
	Reoxy	0.83	0.9035	0.89	0.034	0.9028 – 0.9042	12.48	1.87
	Alpha	0.7	0.8248	0.82	0.043	0.8239 – 0.8256	9.33	0.12
Top1%	Anoxia	0.81	0.8279	0.83	0.011	0.8277 – 0.8281	31.81	11.67
	Reoxy	0.81	0.8387	0.84	0.016	0.8384 – 0.8391	22.69	5.82
	Alpha	0.72	0.7898	0.78	0.027	0.7892 – 0.7903	15.24	0.13
Spectral Clustering	Anoxia	0.93	0.9464	0.95	0.012	0.9461 – 0.9466	35.99	11.21
	Reoxy	0.97	0.9741	0.98	0.011	0.9739 – 0.9743	39.87	34.9
	Alpha	0.89	0.946	0.95	0.017	0.9457 – 0.9463	23.67	0.29

Table 5. Estimated threshold (*τ*) for each dataset with different methods. *The bracketed values represent d_{TM} values. Thresholding methods with low positive d_{TM} and S_{TM} values are preferred. Significant results are highlighted in bold.*

Method	**Anoxia**	**Reoxygenation**	**Alpha**	**S_{TM}**
1. Maximal Clique-2 **Maximal Clique-3**	0.9 (0.07) 0.87(0.1)	0.91 (0.01) 0.89 (0.03)	0.74 (0.11) 0.6 (0.25)	**0.19** 0.38
2. Control-Spot	0.93 (0.04)	0.83 (0.09)	0.70 (0.15)	0.28
3. Top1Percent	0.81 (0.16)	0.81(0.11)	0.72 (0.13)	0.4
4. Spectral Clustering	0.93 (0.04)	0.97 (-0.05)	0.89 (-0.04)	-0.05
5. Bonferroni-adjusted p-value	0.85 (0.12)	0.93 (-0.01)	0.95 (-0.1)	0.01
6. Power	0.88 (0.09)	0.94 (-0.02)	0.96 (-0.11)	-0.04
GO-Functional Similarity (median + (0.5*stdev))	0.97	0.92	0.85	

VITA

Bhavesh R. Borate was born on August 16, 1978 in Bombay, India. He did his Bachelors in Medicine and Surgery from Grant Medical College, Bombay and later got an Advanced Diploma in Bioinformatics from University of Pune, India. In July 2008, he graduated from University of Tennessee, Knoxville with a Double Master's in Genome Science & Technology and Statistics.