University of Tennessee, Knoxville

Trace: Tennessee Research and Creative Exchange

Doctoral Dissertations

Graduate School

12-2008

Evaluation of statistical correlation and validation methods for construction of gene co-expression networks

Suman Duvvuru
University of Tennessee - Knoxville

Recommended Citation

Duvvuru, Suman, "Evaluation of statistical correlation and validation methods for construction of gene co-expression networks. " PhD diss., University of Tennessee, 2008.
https://trace.tennessee.edu/utk_graddiss/599

To the Graduate Council:

I am submitting herewith a dissertation written by Suman Duvvuru entitled "Evaluation of statistical correlation and validation methods for construction of gene co-expression networks." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Arnold M. Saxton, Major Professor

We have read this dissertation and recommend its acceptance:

Michael A. Langston, Brynn H. Voy, Russell Zaretzki, Elissa J. Chesler

Accepted for the Council:
Carolyn R. Hodges

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 dissertation written by Suman Duvvuru entitled "Evaluation of statistical correlation and validation methods for construction of gene co-expression networks". I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Arnold M. Saxton
Major Professor

We have read this dissertation and recommend its acceptance:

Michael A. Langston

Brynn H. Voy

Russell Zaretzki

Elissa J. Chesler

Accepted for the Council:

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

(Original Signatures are on file with official student records)

Evaluation of statistical correlation and validation methods for construction of gene co-expression networks

A Dissertation Presented for
the Doctoral Degree
The University of Tennessee, Knoxville

Suman Duvvuru
December 2009

This research is dedicated to the loving memory of my mother Late Shanthi Duvvuru,
my father Sudhakar Duvvuru and my sister Sirisha Duvvuru
for inspiring me all the way through.

Acknowledgements

First, I would like to thank my advisor, Professor Dr. Arnold Saxton, for giving me the wonderful opportunity to further my graduate education at the University of Tennessee. He has guided me through this journey with patience, intelligence, understanding, and caring. He has inspired me to take up my career in statistical genetics. This dissertation would not have been possible without his help and encouragement. Special thanks go to Dr. Elissa Chesler for guiding me through several challenges and for her helpfulness and suggestions on various matters. Her enthusiasm about systems biology has been always inspiring.

I would like to thank Dr. Brynn Voy for her suggestions on the issues relating to genetics. I would like to thank Dr. Mike Langston and his group for their support with the software for high-throughput computations. I would also like to thank Dr. Dan Goldwitz and his group for providing me with their experimental data and guiding me constantly through the biological issues. Thanks to Dr. Russell Zaretzki for taking the time to serve on my committee, his helpful advice and comments on the statistical issues.

Thanks to my sister, Sirisha Duvvuru and father, Sudhakar Duvvuru, for their love and moral support during the long years of my education. I would like to mention also my grandmother Audilakshmi Duvvuru whose life has been a constant source of solace and inspiration during the difficult times. I would like to thank my dear friends Haritha Kolli and Raghul Gunasekharan who have been very supportive during the setbacks in the long years of my education.

Abstract

High-throughput technologies such as microarrays have led to the rapid accumulation of large scale genomic data providing opportunities to systematically infer gene function and co-expression networks. Typical steps of co-expression network analysis using microarray data consist of estimation of pair-wise gene co-expression using some similarity measure, construction of co-expression networks, identification of clusters of co-expressed genes and post-cluster analyses such as cluster validation. This dissertation is primarily concerned with development and evaluation of approaches for the first and the last steps – estimation of gene co-expression matrices and validation of network clusters. Since clustering methods are not a focus, only a paraclique clustering algorithm will be used in this evaluation.

First, a novel Bayesian approach is presented for combining the Pearson correlation with prior biological information from Gene Ontology, yielding a biologically relevant estimate of gene co-expression. The addition of biological information by the Bayesian approach reduced noise in the paraclique gene clusters as indicated by high silhouette and increased homogeneity of clusters in terms of molecular function. Standard similarity measures including correlation coefficients from Pearson, Spearman, Kendall's Tau, Shrinkage, Partial, and Mutual information, and Euclidean and Manhattan distance measures were evaluated. Based on quality metrics such as cluster homogeneity and stability with respect to ontological categories, clusters resulting from partial correlation and mutual information were more biologically relevant than those from any other correlation measures.

Second, statistical quality of clusters was evaluated using approaches based on permutation tests and Mantel correlation to identify significant and informative clusters that capture most of the covariance in the dataset. Third, the utility of statistical contrasts was studied for classification of temporal patterns of gene expression. Specifically, polynomial and Helmert contrast analyses were shown to provide a means of labeling the co-expressed gene sets because they showed similar temporal profiles.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION AND BACKGROUND

Microarray gene expression arrays quantitatively and simultaneously monitor the expression of thousands of genes under different conditions. Genes with similar expression patterns under various conditions or time points may imply co-regulation or relationship in functional pathways [1]. Identification of groups of genes with similar expression patterns is usually achieved by exploratory techniques such as cluster analysis. Most algorithms used in the cluster analysis of large expression datasets fall into one of two categories: supervised methods (classification based on predictors of specific conditions using models constructed from prior information) and unsupervised methods (clustering data points without any prior information). Instead of learning the best way to predict a “correct answer,” unsupervised algorithms find useful or interesting patterns within a dataset. In a typical unsupervised cluster analysis, genes are assigned to clusters of similar expression patterns given a dissimilarity measure (usually correlation-based or distance-based) between any two genes. Similarly one can cluster samples to look for patients with similar expression signature in order to discover unknown subtypes of a disease [2]. Another approach known as bi-clustering or two-way clustering looks for groups of genes that have similar expression patterns only in a subset of samples or time periods [3]. These analyses involving transcriptional profiling are often used primarily to generate hypotheses for further investigation into specific pathways or genetic mechanisms.

Construction of coexpression networks from gene expression microarray datasets has recently become a popular alternative to the conventional analytic approaches, such as the detection of differential expression using statistical testing or coexpression analysis using unsupervised clustering. Network-based representation and analysis of microarray data is increasingly being used to both visualize and identify the components and their interactions involved in a given cellular system. Representing dependencies in the dataset as interaction networks allows the researcher to explore the whole spectrum of pairwise relationships among the genes as opposed to flat lists of genes from statistical tests or distinct groups of genes from clustering tools. Several approaches have been proposed for network construction including Boolean networks [4-6], Bayesian networks [7] and relevance networks [8]. The main focus of this dissertation is gene co-expression network construction using relevance networks.

1.1 Introduction to gene co-expression network construction

In general, a collection of nodes connected among each other represents a network or graph which thus provides a straightforward representation of interactions between the nodes. Network concepts such as node connectivity and cluster have been found useful for the analysis of complex interactions. Graph-theoretic methods have been found useful in many domains, e.g. gene co-expression networks [9], protein-protein interaction networks [10] and cell-cell interaction networks [11]

In this dissertation, the focus is on methods involved with construction of gene co-expression networks based on the transcriptional response of cells to changing conditions. Since the coordinated co-expression of genes encodes interacting proteins, studying co-expression patterns can provide insight into the underlying cellular processes [12]. It is standard to use the Pearson correlation coefficient as a co-expression measure, i.e. the absolute value of Pearson correlation is often used in a gene expression cluster analysis. Recently, several groups have suggested to threshold this Pearson correlation coefficient in order to arrive at gene co-expression networks, which are sometimes referred to as 'relevance' networks [9]. In these networks, a node corresponds to the gene expression profile of a given gene. Nodes are connected if they have a significant pairwise expression profile association across the conditions. There are several questions associated with thresholding a correlation to arrive at a network. On the simplest level, how to pick a threshold? Most of the strategies for picking a threshold are based on their definition of high enough correlation. Drawbacks of thresholding the network at a predetermined value include loss of information and sensitivity to the choice of the threshold [13].

A flowchart for constructing a gene co-expression networks is presented in Figure 1. It is assumed that the gene expression data have been suitably quantified and normalized. Each co-expression network corresponds to an adjacency matrix. The adjacency matrix encodes the correlation between each pair of genes. In unweighted networks, the adjacency matrix indicates whether or not a pair of nodes is connected, i.e.

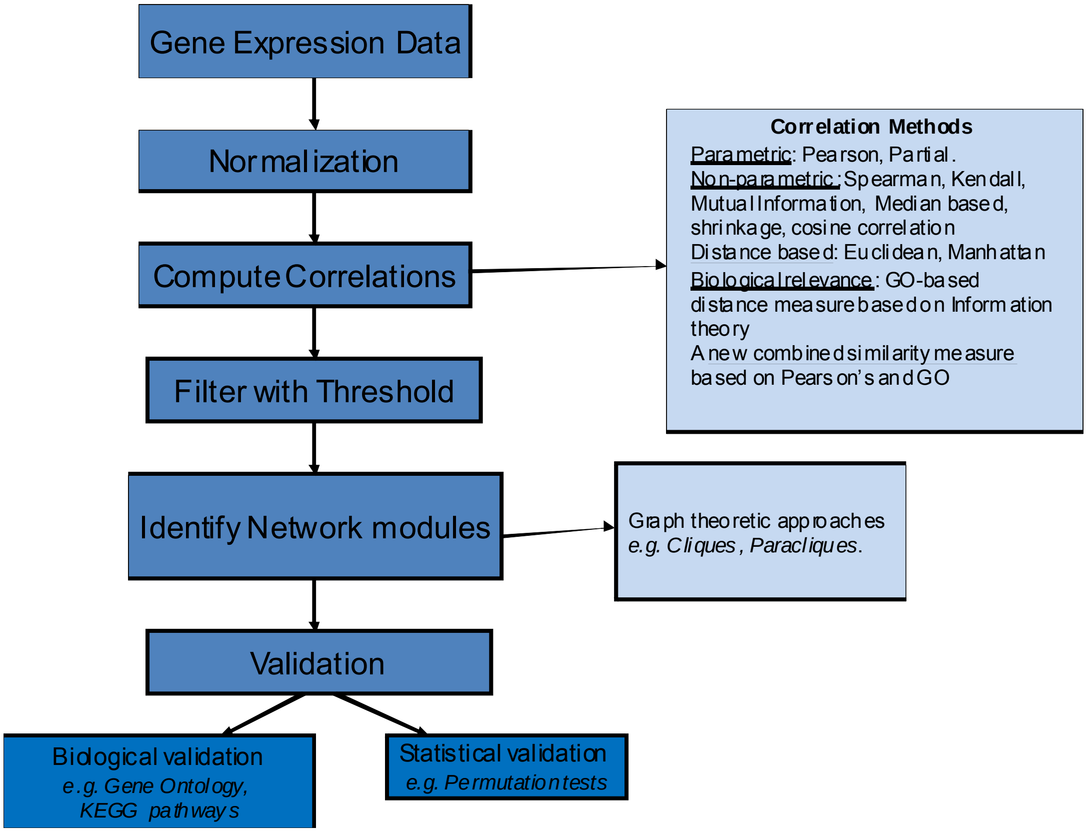

Gene Expression Data
Normalization
Compute Correlations
Filter with Threshold
Identify Network modules
Validation
Biological validation
e.g. Gene Ontology, KEGG pathways
Statistical validation
e.g. Permutation tests
Correlation Methods
Parametric: Pearson, Partial.
Non-parametric: Spearman, Kendall, Mutual Information, Median based, shrinkage, cosine correlation
Distance based: Euclidean, Manhattan
Biological relevance: GO-based distance measure based on Information theory
A new combined similarity measure based on Pearson's and GO
Graph theoretic approaches
e.g. Cliques, Paracliques.

Figure 1. Overview of gene co-expression network construction and validation

its entries are 1 or 0. To start, one needs to define a measure of similarity between the gene expression profiles. This similarity measures the level of concordance between gene expression profiles across the experiments. The n×n similarity matrix S = [s_{ij}] is transformed into an n × n adjacency matrix A = [a_{ij}], which encodes the correlation between pairs of nodes. Since the networks considered here are undirected, A is a symmetric matrix with nonnegative entries. It is commonly assumed that a_{ij} ε [0, 1] for weighted networks. The adjacency matrix is used to construct the co-expression network which is the foundation of all subsequent steps.

1.2 Clustering approaches for the identification of co-expression network modules

Many clustering algorithms have been proposed for gene expression data. Most methods use a correlation measure between expression levels to calculate a distance metric of similarity (or dissimilarity) of expression between each gene pair. Perhaps best known to biologists are the hierarchical clustering methods [12]. Spellman et al. [14] applied a variant of the hierarchical average-link clustering algorithm to identify groups of co-regulated yeast genes. In this family of techniques, all data instances start in their own clusters, and the two clusters most closely related by some similarity metric are merged. The process of merging the two closest clusters is repeated until a single cluster remains. This arranges the data into a tree structure that can be broken into the desired number of clusters by cutting across the tree at a particular height. Tree structures are easily viewed and understood, and the hierarchical structure provides potentially useful information about the relationships between clusters. Trees are known to reveal close relationships very well. However, as later merges often depend on aggregated measures of clusters containing many scattered elements, the broadest clusters can sometimes be hard to interpret.

Another common family of clustering methods is that of partition or centroid algorithms. These methods generally require specification of the number, *k*, of clusters, and start with *k* data points that may be chosen either randomly or deliberately. These *k* points are used as the 'centroids' which are the multidimensional center points of an initial set of clusters. The algorithm then partitions the samples into the *k* clusters, optimizing some objective function such as within-cluster similarity by iteratively assigning samples to the nearest centroid's cluster and adjusting the centroids to represent the center points of the new clusters. The *k*-means method [15] is a well-known centroid approach. A variation that allows samples to influence the location of neighboring clusters is known as the self-organizing map or Kohonen map. Such maps are particularly valuable for describing the relationships between clusters [16]. Tamayo et al. (1999) used self-organizing maps (SOM) to identify clusters in the yeast cell cycle and human hematopoietic differentiation data sets.

Some methods seek to optimize a measure of within-cluster similarity or separation between clusters, but avoid specifying the number of clusters ahead of time, instead specifying bounds on cluster membership using heuristic approaches [17, 18] . Model-based methods assume the data can be generated by a specified statistical model (such as a mixture of Gaussian distributions), and search for model parameters that best fit the data [19, 20]. So-called 'fuzzy' clustering finds groups, but may allow elements to appear in more than one cluster or in no clusters at all [21]. Most of these standard approaches do not allow for negative correlations which are quite meaningful from a biological point of view.

Another class of clustering techniques is based on graph-theoretical approaches. They have a major advantage over other approaches in network construction since the data when explicitly presented in terms of a graph convert the problem of clustering a dataset into such graph theoretical problems as finding minimum cut or cliques in the co-expression network. In the dissertation, the network construction is based on one of these graph theoretic approaches called clique. Some of the most popular graph theoretic approaches are outlined below:

Cliques and Paracliques: The clique-based clustering algorithm of [22] can applied to the co-expression network, to search for patterns of highly co-expressed genes or network motifs. A clique in the thresholded graph obtained from the previous step represents a set of genes with the property that every pair of its elements is highly correlated. This is widely interpreted as suggestive of putative co-regulation over the conditions in which the experiment was performed. Extracting cliques can be viewed as an especially stringent graph-theoretical form of clustering for gene co-expression data. Although clique is an exceedingly difficult computational problem, its advantages are many. It is important to note that finding cliques in a graph is a NP-hard problem. The main advantage that clique offers over other methods is that cliques need not be disjoint. A single vertex can be present in several cliques which accounts for when a gene might be involved in multiple regulatory networks.

Running clique analysis on high dimensional gene expression data however may yield very large numbers of highly-overlapping cliques, typically more than a million. To aggregate these data, a new algorithmic approach called paraclique has been introduced [23]. A paraclique is a clique augmented with vertices in a highly controlled manner to maintain density. It uses a “glom” factor to include new vertices, and an optional threshold to check the original weights of edges discarded by the high pass filter. Glom factor is the factor by which the degree constraint of the vertices is relaxed. Hence paraclique analysis gives rise to a very highly intercorrelated group of co-regulated genes whose transcript expression levels show highly significant but not necessarily pair-wise correlations above threshold. By using the computational power of tools such as fixed-parameter tractability, and then identifying paracliques, subgraphs much denser than are typically produced with traditional clustering algorithms are obtained [23]. The correlation matrix is then reduced to a select set of intercorrelated modules to simplify the discovery of functional significance that underlies gene expression variation.

CLICK. CLICK (CLuster Identification via Connectivity Kernels) [24] algorithm identifies highly connected components in the co-expression network as clusters. It makes the assumption that after standardization, pair-wise similarity values between elements are normally distributed. Under this assumption, the weight of an edge is defined as the probability that the corresponding vertices are in the same cluster. The clustering process of CLICK iteratively finds the minimum cut in the correlation graph and recursively splits the data set into a set of connected components from the minimum cut. CLICK also takes two post-pruning steps to refine the cluster results. The adoption step handles the remaining singletons and updates the current clusters, while the merging step iteratively merges two clusters with similarity exceeding a predefined threshold. The authors compared the clustering results of CLICK on two public gene expression data sets with those of a SOM approach and Eisen's hierarchical approach, respectively. In both cases, clusters obtained by CLICK demonstrated better quality in terms of homogeneity and separation. However, CLICK has little guarantee of not going astray and generating unbalanced partitions, e.g., a partition that only separates a few outliers from the remaining data objects.

CAST. Ben-Dor et al. [25] presented both a theoretical algorithm and a practical heuristic called *CAST* (Cluster Affinity Search Technique). They introduced the concept of a *corrupted clique graph* data model. The input data set is assumed to be from the underlying cluster structure by contamination with random errors caused by the complex process of gene expression measurement. Specifically, it is assumed that the true clusters of the data can be represented by a *clique graph*, which is a disjoint union of complete sub-graphs with each clique corresponding to a cluster. The similarity graph is derived from clique graph by flipping each edge or non-edge with a particular pre-defined probability. Therefore, clustering a dataset is equivalent to identifying the original clique graph from the corrupted version with as few flips (errors) as possible.

CAST takes as input a real, symmetric, n-by-n similarity matrix, and an *affinity threshold* which is actually the average of pairwise similarities within a cluster. The clusters are searched one at a time and the algorithm alternates between adding and removing elements to the current cluster based on their affinities to the cluster. When the process stabilizes, a cluster is finalized, and this process continues with each new cluster until all elements have been assigned to a cluster. CAST does not depend on a user-defined number of clusters and deals with outliers effectively.

1.3 Similarity measures for co-expression networks

When clustering genes based on microarray data in search for coordinated groups of coexpressed genes, the choice of the correlation metric has a great impact on the overall structure of overall co-expression network and thus on the clusters produced. Indeed, most clustering algorithms are based on pairwise distances between the expression profiles. Thus a crucial parameter for classification of genes is the choice of an appropriate metric to measure the similarity or dissimilarity between objects. Recent research in clustering analysis has been focused largely on two areas: estimating the number of clusters in data [26, 27] and the optimization of the clustering algorithms [28, 29]. In this dissertation, a different yet fundamental issue in clustering analysis was studied: to define an appropriate measure of similarity for gene expression patterns. Similarity measures can be based either on correlation or distance between the two vectors. Here is an overview of the different similarity measures available in literature.

Correlation measures:

1. Pearson's correlation:

   Pearson's correlation coefficient is widely used and has proven effective as a similarity measure for gene expression data [8, 18, 30]. Given two expression vectors x and y of dimension n, the Pearson's correlation *r* is defined as

   $$r = \frac{1}{n}\sum_{i=1}^{n} \frac{(x-\bar{x})(y-\bar{y})}{S_x\, S_y}$$

2. Spearman's correlation:

   Spearman's is simply a special case of the Pearson product-moment coefficient in which two vectors of expression profiles X_i and Y_i are converted to rankings before calculating the coefficient. Thus the classic Pearson's correlation coefficient between ranks is used to calculate the Spearman's correlation. As a consequence of ranking, a significant amount of information present in the data is lost which is a potential disadvantage of Spearman's correlation.

3. Kendall's Tau:

   Kendall's Tau has been applied to gene expression in a few studies [31]. Unlike the Pearson and Spearman correlations, there is an intuitive, graphical interpretation of Kendall's Tau. Given two genes, two ranked lists of the conditions are created based on the expression levels of each gene. In graph theory terminology, a bipartite graph is created with the conditions representing the two sets of vertices. Each condition from one ranked list is connected to the same condition in the other ranked list by an edge.

   Formally, given two genes x and y each with n expression values, Kendall's Tau is defined as

   $$\tau = \frac{1-2c}{m(m-1)/2} ,$$

   where c is the number of crossings in the bipartite graph and m is the number of conditions.

4. Partial correlation:

   The partial correlation coefficient of two genes measures the strength of the relation between these genes after the effect of other genes is removed or fixed, therefore indicating whether two genes are directly or indirectly linked. The partial correlations have been used in Gaussian Graphical Models (GGM) [32] to characterize strength of correlations between pairs of genes in the regulatory networks.

The partial correlation of genes x and y with respect to other genes whose effect is removed (fixed) is given by

$$r_{ij} = \frac{\boldsymbol{\omega}_{ij}}{\sqrt{\boldsymbol{\omega}_{ii}\boldsymbol{\omega}_{ij}}},$$

where $\boldsymbol{\omega} = P^{-1}$ is the inverse (or pseudo-inverse) of Pearson correlation matrix P. To overcome the degeneracy problem of the correlation matrix P for small samples, partial correlation estimators based on a shrinkage estimation of covariance matrix was introduced in [33].

5. Shrinkage based correlation:

The standard estimation of correlation matrix exhibits serious defects in the "small *n*, large *p*" data setting commonly encountered in functional genomics. Specifically, the empirical covariance matrix is not considered a good approximation of the true covariance matrix. For *n* smaller than *p,* covariance matrix loses its full rank as a growing number of eigenvalues become zero. This has several undesirable consequences. First the correlation matrix is not positive definite, and second, it cannot be inverted.

Schafer and Strimmer proposed an improved estimate of the correlation matrix called shrinkage estimate by shrinking the empirical correlations towards the identity matrix. In particular, they considered a recent analytic result from Ledoit and Wolf [34] that allows construction of an improved covariance estimator that is not only suitable for small sample size *n* and large numbers of variables *p* but at the same time is also inexpensive to compute.

The estimate of shrinkage correlation between two vectors x and y is given by

$$S_{xy} = r_{xy} * \max(0, 1 - \lambda) \,,$$

$$\text{where} = \frac{\sum var(r_{xy})}{\sum r_{xy}^2} \quad x \neq y \,,$$

r_{xy} is the standard Pearson's correlation coefficient between the two vectors and λ is the shrinkage intensity parameter. The details of the computation of Var(r_{xy}) and other variants of these shrinkage estimators are discussed in [33] .

6. Mutual information

Mutual information (MI) provides a general measure for dependencies in the data, in particular, positive, negative and nonlinear correlations [35]. It is a very well known measure in the field of information theory [36] that has been used to analyze gene-expression data [35, 37, 38]. The MI measure requires the expression patterns to be represented by discrete random variables. Given two random variables X and Y, and probability distribution functions $P(X = x_i) = p_i$, $P(Y = y_j) = p_j$, the Mutual information between two expression patterns, represented by random variables X and Y, is given by

$$I(x, y) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{p_i p_j} \,.$$

MI is always non-negative. It equals zero if and only if *X* and *Y* are statistically independent, meaning that *X* contains no information about *Y* and vice versa.

The use of the discrete form of the MI measure requires the discretization of the continuous expression values. The most straightforward and commonly used discretization technique is to use a histogram-based procedure [35] in which a two-dimensional histogram is used to approximate the joint probability density function of two expression patterns.

7. Euclidean distance

Euclidean distance is a measure of the difference between gene expression patterns. Euclidean distance between expression profiles x_i and y_i over n time points is a point in n-dimensional parameter space given by

$$d = \sqrt{\frac{\Sigma(x_i - y_i)^2}{n}} \ .$$

8. Manhattan distance

This is very similar to Euclidean distance and is given by

$$d = \frac{\Sigma\, |x_i - y_i|}{n} \ .$$

In two dimensional space, Manhattan distance is the distance between the data points on the first axis, plus the distance between them on second axis [39]. Manhattan distance is sometimes referred to as ‘city block distance’ as it measures the route one might have to travel between two points in a place such as Manhattan where the streets and avenues are at right angles to each other.

Merits and Demerits of Standard Correlation measures

Pearson's correlation coefficient is widely used and has proven effective as a similarity measure for gene expression data [8, 18, 30]. Some studies have shown that it is not robust with respect to outliers [18], thus potentially yielding false positives which assign a high similarity score to a pair of dissimilar patterns. The main drawback of Pearson's correlation coefficient is that it assumes an approximate Gaussian distribution of the points and may not be robust for non-Gaussian distributions [40]. To address this, Spearman's rank-order correlation coefficient has been suggested in literature as one of the alternative similarity measures [41]. Kendall's Tau has been applied to gene expression in a few studies [31].

In comparison with the standard empirical estimates, the shrinkage estimates exhibit a number of favorable properties [33]. For instance,

(i) They are typically much more efficient, i.e. they show better mean squared error.

(ii) The estimated covariance and correlation matrices are always positive definite and well conditioned so that there are no numerical problems when computing their inverse.

(iii) They are fully automatic and do not require any tuning parameters as the shrinkage intensity is analytically estimated from the data.

Mutual information (MI) provides a general measure for dependencies in the data, in particular, positive, negative and nonlinear correlations [35]. A zero MI indicates that the patterns do not follow *any kind* of dependence, an indication which is impossible to obtain from the Pearson correlation or the Euclidean distance. This property makes MI a generalized measure of correlation, which is advantageous in gene expression analysis. For instance, if a gene acts as a transcription factor only when it is expressed at a midrange level, then the scatter plot between this transcription factor and the other genes might closely resemble a normal distribution rather than a linear repsonse. The Pearson correlation coefficient in this case will give a low estimate, while the MI measure gives a high value [42]. Another important feature of the MI is its robustness with respect to missing expression values. In fact the MI can be estimated from datasets of different sizes. This is advantageous in analyzing expression datasets that often contain (up to 25%) missing values [43]. MI treats each expression level equally, regardless of the actual value, and thus is less biased by outliers.

Distance measures such as Euclidean and Manhattan measure the absolute level of gene regulation. Distance based measures may not be the most appropriate measure for gene expression profiles, as the absolute differences may not be meaningful if the gene expression data represent comparative expression measurements. For example, two genes whose expression levels were perfectly parallel to one another could still be far apart in Euclidean space if the absolute levels in each experiment were different. The Euclidean distance can also make genes that are uncorrelated appear close together. For example, if two genes had expression levels close to 0 but were otherwise randomly correlated they could still appear close in Euclidean space.

Novel correlation methods for gene co-expression

Apart from the standard measures of similarity, several new similarity metrics have been proposed to measure the coexpression of genes using gene expression data. Kim et al. [44] defined a new similarity metric called 'TransChisq' in a new feature space by modeling the shape and magnitude parameters separately in a gene expression profile. A new similarity metric was proposed for the analysis of microarray time course experiments that uses a local shape-based similarity measure based on Spearman's rank correlation [45]. Cherepinsky et al. [46] proposed a shrinkage based similarity metric for the cluster analysis of gene expression data. Son and Baek [47] proposed a modified correlation-based similarity measure for clustering time-course gene expression data. Li et al. [48] proposed a new algorithm based on B-spline approximation of coexpression between a pair of genes, followed by CoD (Coefficient of determination) estimation. Yona et al. [49] proposed a new measure that adjusts to the background distributions when measuring the similarity of two expression profiles. Each of these methods imposes its own criterion and generates clustering solutions with very different boundaries. Moreover none of the methods incorporate any biological information. In gene expression analysis, it is commonly assumed that genes with similar expression profiles are more likely to have similar biological function. However, clustering genes using gene expression data alone and then assigning biological function to the clusters may be suboptimal in a sense that it does not necessarily provide the best possible grouping by biological function. It is easy to find genes with mathematically similar expression profiles in the same cluster that do not share biological similarity and, vice versa, genes known to share similar functions which end up in different expression clusters. Similarity measures based solely on expression data may not handle such biological noise sufficiently.

Semantic similarity of genes using Gene Ontology

The Gene Ontology (GO) is one of the most important ontologies within the bioinformatics community and is developed by the Gene Ontology Consortium [50]. It is specifically intended for annotating gene products with a consistent, controlled and structured vocabulary. The GO is limited to the annotation of gene products and independent from any biological species. The GO represents terms within a Directed Acyclic Graph (DAG) covering three orthogonal taxonomies: molecular function, biological process and cellular component. The GO-graph consists of a number of terms, represented as nodes within the DAG, connected by relationships, represented as edges.

There are several semantic similarity measures that were proposed in literature that measure the functional relationship between the gene products based on the GO tree. Some of these measures are based on edge distances and consider the minimum number of edges that need to be traversed from one node to another. Shorter distances between the nodes imply high similarity and vice-versa. These edge-based methods were used in the lexical medical domain (Medline and Mesh) and have proved very useful in determining the relationships [51].

However most of these edge based methods assume that all the edges represent uniform distances which is not true in the case of the GO tree. Some GO branches may be very deep while others are not. Some terms may have many children terms while others have very few. Some edges may cover a large conceptual distance while others, at the same or even higher levels, cover only short conceptual distances. As a further drawback, higher sections of the taxonomy may seem too similar to each other. For instance, if two nodes high in the taxonomy (very general) are compared, the results that are equivalent to the comparison of two nodes far lower (very specific) might be obtained. This may lead to spurious similarity results as was shown by Richardson and Smeaton [52] when applying edge-based metrics to a broad domain such as the WordNet.

An alternative approach considers the information contained at the nodes applying concepts borrowed from information theory. When the probability of each node within the tree is known, this knowledge can be used to compute their information content. The lower the probability, the more information a node contains. These measures are based on the concept of information content that is defined as the frequency of each GO term, or any of its children, occurring in an annotated data set. Semantic similarity of gene products is estimated by the information content of specific GO annotations and their shared parents. The assumption is that the more information two terms share, the more similar they are. The shared information of two terms is indicated by the information content of the terms that subsume them in directed acyclic tree (DAG). Given the information content of each term, there are several ways to calculate similarity scores between annotated gene products. Similarity between two nodes can be seen as the information content that they share. This is indicated by the information contained in the set of their subsumers— common ancestor nodes.

Resnik, Lin, Jiang and Conrath [53-55] proposed GO semantic measures which are commonly used. Accordingly, Resnik defines semantic similarity between two nodes as the information content of their minimum subsumer [56]. When multiple inheritance is present, as happens in the GO, there might be minimum subsumers in several paths. In that case the most informative subsume is chosen. Similarity between two GO terms is defined by Resnik (1995) as:

$$\mathrm{Sim}(c1,c2) = -\log[p_{ms}(c1,c2)],$$

where c1 and c2 are GO terms; and $p_{ms}(c1,c2)$ is the probability of their minimum subsumer.

Jiang and Conrath [55] proposed a different approach for measuring semantic distance between GO categories. It is a mixed approach that inherits from the edge-based method and is enhanced by the information content calculation of node-based techniques. Lin [57] again presents another information-theoretic definition of similarity based on a defined probabilistic model.

Novel correlation methods using prior biological information

Few researchers have constructed combined measures with prior biological information. Hanisch et al. (2002) proposed a hybrid distance measure which combines biological network information with gene expression data. Their results confirmed that performing cluster analysis on the basis of network distances or expression distances alone is not sufficient to yield coregulated pathway-like clusters. Kustra and Zagdanski (2006) used Gene Ontology (GO) annotations to derive a GO-based dissimilarity measure, and constructed a combined measure by combining the GO-based dissimilarity measure with Pearson correlation. With the combined measure, their results revealed that combining comprehensive and reliable biological repository with expression data may improve performance of cluster analysis and yield biologically meaningful gene clusters. The wide application of combined measures is because of the general hypothesis that incorporating biological knowledge into statistical analysis of expression data is a reliable way to maximize statistical efficiency and enhance the interpretability of analysis results. A crucial point is the proper combination of individual similarity metrics. A linear or a non-linear function was used typically to combine the measures in the previous studies and finding the optimal function is still a challenge. If sufficiently many pathways and associated gene expression patterns are known to be relevant in advance, this knowledge might be utilized to learn an appropriate functional form by employing machine learning methods. However such comprehensive information is rarely available. In this dissertation, this issue is addressed and Bayesian setting is used for combining the two measures.

Different measures are likely to perform differently for a given gene expression dataset. If the effectiveness of pairwise measure can be simply evaluated before it is employed in clustering algorithm, it will save lots of time in correcting for errors in the clustering analysis. Priness et al. [58] performed a comparative study of MI and a few other correlation algorithms and observed that their best solutions are ranked almost oppositely when using different distance measures, despite the found correspondence between these measures when analyzing the averaged scores of groups of solutions. Their results show that it is very important to select a proper correlation measure for a given gene expression dataset.

1.4 Approaches for clustering validation

Interpreting the clustering results and validating the clusters found are as important as generating the clusters[15]. Given the same data set, different correlation metrics can potentially generate very different clusters of co-expressed genes. A biologist with a gene expression data set is faced with the problem of assessing the reliability of clustering results from an appropriate similarity measure for his or her data set. In much of the published clustering work on gene expression, the success of clustering algorithms is assessed by visual inspection using biological knowledge [59, 60]).

There are some studies that proposed measures that provide a quantitative data-driven framework to validate the clusters. Jain and Hubes [15] classified cluster validation procedures into external and internal criterion analyses. External criterion analysis validates a clustering result by comparing it to a given gold standard which is another partition of objects. The gold standard must be obtained by an independent process based on information other than the given data. There are many statistical measures that assess the agreement between an external criterion and a clustering result such as Biological Homogeneity Index proposed by Datta and Datta [61] and Rand index [62]. However reliable external criteria are rarely available when analyzing gene expression data. Internal criterion analysis uses information from within the given dataset to look at the goodness of fit between the input data and clustering results. Intra-cluster distances representing the homogeneity of the genes within clusters and inter-cluster distances representing separation between clusters are some of the possible measures of goodness of fit. Silhouette is one of the standard measures that has been used to evaluate the quality of clustering based on inter- and intra-cluster distances [63]. For validation of clustering results, external criterion analysis has the advantage of providing an independent and unbiased assessment of cluster quality. On the other hand, external criterion analysis has the strong disadvantage that an external gold standard is rarely available. Internal criterion analysis avoids the need for such a standard, but has the alternative problem that clusters are validated using the same information from which clusters are derived. Both these criterion can be used for a rigorous validation of the clustering results.

How best to compare clustering solutions again depends on the purpose of clustering. If clustering is to be used primarily for data reduction, one might evaluate it strictly from that point of view — the best clustering is the one that allows expression of the entire data set in minimal space. Based on this criterion, dimensionality reduction techniques such as Principal component analysis have been used for clustering gene expression data [64]. If clusters are to be used to predict classifications of other samples, one might choose to evaluate each clustering by its predictive power. A measure such as Figure of Merit (FOM) proposed by Yeung et al. [65] uses a jackknife approach in which the clustering algorithm is applied to all but one experimental condition in a data set, and uses the left-out condition to assess the predictive power of the clustering algorithm.
Another desirable property of clustering is stability, i.e. if the experiment were repeated again and again one would hope to obtain similar clusters. A standard technique for testing cluster reliability involves adding a small amount of noise to the data and re-clustering. Several microarray studies have incorporated these techniques, either using simple but reasonable noise models [66], or by sampling the noise distribution directly from the data [67].

The issue of statistical validation of clustering solutions has been poorly studied. How likely is it that the clustering solution that was obtained is seen by chance? Randomization approaches such as permutation tests and bootstrapping can be used to assess the significance of the clusters [67, 68].

1.5 Overview of the dissertation

Many different techniques have been used in the context of finding co-expression network clusters and instances of success from many different methods have been reported in specific applications. It is evident from the literature that graph theoretic approaches such as clique and paraclique are well suited for co-expression network construction. As is seen, the choice of a correlation or distance metric is the starting and the crucial step that determines the network structure. There is little or no systematic comparison of different correlation metrics in the context of clustering co-expressed genes. The main goal of the dissertation is the evaluation of correlation measures and investigation of approaches for statistical validation of co-expression network structures.

In chapter 2, a new combined similarity metric is proposed using a Bayesian methodology that incorporates prior biological information based on the general hypothesis that incorporating biological knowledge into statistical analysis of expression data is a reliable way to maximize statistical efficiency and enhance the interpretability of analysis results. This metric is based on a strong statistical foundation which is lacking in many of the measures proposed in the literature. Secondly the incorporation of functional annotation adds more confidence to the clustering results.

In chapter 3, the focus is on the statistical assessment of the clusters of co-expressed genes obtained from gene expression data. Though most of the literature on cluster validation is focused on proposing new validation indices that can be used to compare different clustering results, this comparison will not reveal the reliability of the resulting clusters, i.e. the probability that the clusters are not formed by chance. Also, many previous studies optimized the cluster attributes such as number and size of clusters based on specific criteria such as silhouette. However, none of them dealt with the evaluation of their statistical significance. Chapter 3 is primarily concerned with developing approaches for evaluating the statistical significance of clusters obtained from a co-expression based clustering algorithm. Firstly, permutation tests are used to evaluate the significance of the cluster attributes. Secondly, Mantel correlation is used to evaluate the information content of a cluster based on how well the correlation matrix in the cluster space correlates with that in the original space, and permutation tests are used to compute the p-value associated with the Mantel correlation of a cluster. These tools will help biologists eliminate the non-significant and non-informative clusters before proceeding with biological validation.

The clusters obtained from the gene clustering algorithms are usually labeled using external information such from GO and KEGG pathway databases. But insufficient annotation in some organisms makes it harder to assign meaningful labels to all the clusters of genes. Not much attention was paid in the literature to interpreting clusters of coexpressed genes using the internal information such as shapes of gene expression patterns. Chapter 4 focuses on Helmert and polynomial contrast analysis for the differential profiling of genes in time course microarray data and the use of these contrast patterns as labels to the co-expression network modules.

CHAPTER 2 COMPARISON OF A NOVEL BAYESIAN AND OTHER METRICS IN CO-EXPRESSION NETWORK CONSTRUCTION

2.1 Abstract

The choice of correlation metric used to measure the pairwise coexpression of genes from microarray data has a great impact on the structure of gene coexpression networks. The main objective of this study is the development of a combined correlation metric which is driven by both data and prior biological information, used for the construction of gene co-expression networks and also the comparison of different correlation metrics in gene co-expression networks. This study provided evidence that this Bayesian metric produces clusters of genes that are highly correlated with biologically relevant external standards. The results confirm that incorporation of biological information increases the homogeneity of clusters both in terms of biological functional categories and intra-cluster distances. Based on the analysis, incorporation of biological information decreases the noise level in the correlations. A second objective was a comparative survey of standard correlation methods, which revealed that all the metrics except for partial correlation produced similar degree distributions of vertices (genes), number and size distributions of co-expression network modules. Furthermore it was shown that all the correlation metrics revealed a poor correlation between gene

expression, protein-protein interaction and pathway membership using the chosen datasets.

2.2 Introduction

An increasing number of methodologies are available for finding clusters of co-expressed genes using gene expression data. The initial step before implementing any clustering algorithm is to construct a similarity matrix based on a chosen similarity measure. The choice of similarity/distance measures between genes may significantly affect the clustering results. Though there is a lot of literature on clustering methods with random or designed sets of conditions and different definitions of similarity, there is much less attention paid to derive a statistically robust definition for the similarity of genes. There are several mathematical approaches for measuring the co-expression of genes using microarray data including Pearson, Spearman, Kendall's Tau, mutual information and the distance based measures such as Euclidean and Manhattan. Although the results of all these approaches are useful, one basic problem remains: none of these methods incorporates known biological information. Therefore, biologists are still forced to do a sequential analysis of their data by first clustering the expression data alone and afterwards annotating the genes of each cluster by hand and thus incorporating biological information into their models. Such an approach is slow and exhausting and may also result in a suboptimal clustering since information from other resources could often help in resolving ambiguities or avoiding errors caused by linkages based on noisy data or spurious similarities. Another problem with clustering methods is that cluster boundaries may be very close and arbitrary to some degree.

On the other hand there are several semantic similarity measures proposed in literature based on genomic annotation that give a measure of functional relationship between the genes. GO (Gene Ontology) is one of the most organized and comprehensive ontologies for annotations of genes and gene products. It provides a structured controlled vocabulary of gene and protein biological roles describing the following aspects: function, process and component. GO is organized as a DAG (Directed Acyclic Graph), one for each aspect. Many semantic similarity measures applied to ontologies have been proposed such as Resnik, Lin and Jiang's distances [53-55]. These measures are dependent on the annotation similarity of genes and are not based on any particular microarray dataset. Resnik defines semantic similarity between two nodes (GO terms) in the graph as the information content of their minimum subsumer [56]. When multiple inheritance is present, as happens in the GO, there might be minimum subsumers in several paths. The most informative subsumer is then chosen. Jiang and Conrath [55] proposed a different approach for measuring semantic distance between GO categories. It is a mixed approach that inherits from the edge-based method and is enhanced by the information content calculation of node-based techniques. Lin [57] again presents another information-theoretic definition of similarity based on a defined probabilistic model.

Mathematical correlations give a good measure of correlation between the gene expression levels whereas semantic similarity indicates biological relevance. For this reason, a combined measure which incorporates prior biological information in determining the relationship between genes would be very useful in the elimination of false relationships that would result from using the data correlations alone. The clusters that result from such biologically valid co-expression network will be more meaningful in terms of functional relationships and be representative of pathways.

Except for designing new measures for gene expression data, few attempts have been made to construct combined measures with prior biological information. Hanisch et al. [69] constructed a combined distance measure which combines biological network information with gene expression data. Their results confirmed that performing cluster analysis on the basis of network distances or expression distances alone is not sufficient to end up with coregulated pathway-like clusters. Kustra and Zagdanski [70] used Gene Ontology (GO) annotations to derive a GO-based dissimilarity measure, and developed a combined measure by combining the GO-based dissimilarity measure with Pearson correlation. With the combined measure, their results revealed that combining comprehensive and reliable biological repository with expression data may improve performance of cluster analysis and yield biologically meaningful gene clusters. However the function used to combine the measures was empirically determined based on a few experiments and finding an optimal function for combining the measures is still an issue.

There is no general consensus regarding which distance measure is optimal for capturing similarities between GO categories. Lord et al. [71] investigated the three measures to compare GO semantic similarity and its correlation to protein sequences. It was shown that the Resnik measure may be the most discriminatory while the Jiang distance shows the weakest correlation to protein sequences. Seivilla et al. [72] showed that Resnik outperforms the other measures by showing that semantic gene similarities obtained using Resnik measure are the most correlated with the gene expression data correlations. Hence Resnik measure was chosen as the semantic similarity measure in the current study.

Though a few similarity measures that combine semantic similarity and correlation measures have been suggested, a statistical foundation for combining the measures has not been established before. Schisterman et al. [73] suggested a Bayesian approach for combining correlation coefficients in which knowledge from previous studies was incorporated to improve estimation in epidemiological studies. A Bayesian approach provides a valid statistical methodology for combining the prior biological information and statistical correlation. In this study a new similarity metric called *BaySim* is proposed that uses this approach to combine Resnik GO similarity and the Pearson correlation.

It is possible that different correlation or distance measures might work differently for the same datasets and yield different clustering solutions. How well the clustering solutions resulting from these different measures agree with each other is an objective of the current study. Though many different metrics have been used in gene expression studies, a single study comparing all the metrics and outlining the relative merits and demerits of the metrics in gene expression studies has not been carried out.

Also the effect of these metrics in producing gene networks and clustering solutions has not been studied. Thus there is a need for comparison of the correlation methods to see which algorithms unveil the true networking of the genes and determine the factors which cause the differences in the results. In the present chapter, a comprehensive comparison of many different correlation metrics is performed and their effects on co-expression networks are studied. This study proposes to evaluate different correlation measures for gene expression data using several public gene expression datasets. The comparison includes a biological validation and quality examination of clustering solutions from different metrics.

2.3 Methods

2.3.1 A Bayesian approach to infer correlation using prior biological information

Methodology

A Bayesian methodology is used to integrate prior biological similarity based on GO with the correlation derived from the data to arrive at a posterior distribution of gene similarity.

Bayes' theorem [74] is a theorem of probability theory originally stated by the Reverend Thomas Bayes. It can be seen as a way of understanding how the probability that a theory is true is affected by a new piece of evidence. In other words, it provides a means of adding new information to the existing information thereby updating the prior knowledge. It is used to calculate the posterior probability of a hypothesis. The major difference of the Bayesian approach, compared with a standard likelihood (data-driven) approach, is that it modifies the likelihood into a posterior distribution. According to Bayes' theorem, the posterior probability of Pearson correlation ρ is given by:

$$P(\rho/data) = \frac{P(data/\rho)*P(\rho)}{P(data)}$$

P(ρ) is the prior probability and it needs to be determined before the analysis. A motivation for this approach is that the prior distribution summarizes the prior information on ρ, i.e. the knowledge that is available on ρ prior to the observation of the sample data. The denominator P(data), referred to as the normalizing constant, is the sum or integral of the numerator over all ρ's.

Bayes' theorem [8] can be rewritten as

$$\text{Posterior Probability} \propto \text{Likelihood} \times \text{Prior Probability} ,$$

where ∝ stands for “proportional to”.

The two variables of interest, X and Y, are supposed to follow a bivariate normal distribution with a population correlation coefficient ρ(x,y) = ρ. Let the population means be μx and μy and variances be σx and σy, respectively.

Pearson correlation is given by the following formula:

$$\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{S_x\,S_y} ,$$

where $\bar{x}$ and $\bar{y}$ represent the sample means of X and Y respectively, and S_x and S_y represent the standard deviation of X and Y respectively.

Using the standard reference priors for μ_x, μ_y, σ_x, and σ_y, a reasonable approximation to the posterior density of ρ is given by Schisterman.et al. [73] as

$$P(\rho/x, y) \propto \frac{P(\rho)\,(1-\rho^2)^{(n-1)/2}}{(1-\rho * r)^{n-(\frac{3}{2})}}$$

Using the substitution ρ = tanh ξ and r = tanh z, ξ is found to be approximately normal with mean z and variance 1/n. These results were derived in a series of complicated substitutions by Fisher [75, 76]. The hyperbolic tangent transformation (ρ = tanh ξ and r = tanh z) allows taking full advantage of the conjugate properties of the normal distribution, which is accomplished by combining correlation coefficients from different studies. The prior and likelihood functions are combined together to form the posterior density, which will follow a normal distribution with mean

$$\mu_{Posterior} = \sigma^2_{posterior} * (n_{prior} * \tanh^{-1} r_{prior} + n_{Likelihood} * \tanh^{-1} r_{Likelihood}) \quad (1)$$

and variance

$$\sigma^2_{posterior} = \frac{1}{n_{prior} + n_{likelihood}} \quad , \qquad (2)$$

where r_{prior} and $r_{Likelihood}$ are the correlations based on the prior and the dataset respectively. $\mu_{Posterior}$ and $\sigma^2_{posterior}$ are the posterior mean and the variance respectively. n_{prior} and $n_{likelihood}$ are the sample sizes based on the prior and likelihood respectively.

Though many priors can be used for ρ, the same prior as in Schisterman et al. [73] is chosen based on the idea that the inference becomes easier if a prior is in the following form for c:

$$P(\rho) \propto (1-\rho^2)^c$$

The choice of c will determine the weight the prior will have in estimation and is crucial in estimating the posterior. If there is no information from previous studies, a common choice for c will be 0, that is, p (ρ) $\propto$ 1. Other choices for c, such as -3/2 using multiple parameter Jeffreys' rule [74] can be used.

Since the posterior distribution is normally distributed, the 95% posterior confidence interval is defined by

$$\mu_{Posterior} \pm 1.96 * \sqrt{\sigma^2_{posterior}} \quad . \tag{3}$$

Application to gene expression data

The standard mathematical correlation functions such as Pearson’s r quantify the degree of similarity of gene expression profiles. However, a perfect correlation of expression in a pathway is not observed for several reasons. First gene expression measurements reflect the amount of mRNA in the sample. Second measurements from current high throughput technology such as microarrays are very noisy. The combined correlation function should assign a high correlation to genes that are in the same or closely related pathway and show similar expression patterns. Genes which are far apart in terms of pathways are considered to be in different biological context and should be far apart according to the new similarity function.

If normality is assumed on the distribution of Resnik similarities which are treated as prior information and Pearson’s correlation from the data is used to determine the likelihood, the equation (3) in the previous section can be used to calculate a point estimate and Confidence Interval estimates of the new Bayesian coefficient.

Since the normally distributed prior and likelihood functions are conjugate functions, the posterior distribution then is also normally distributed with mean and variance as defined in equations (1) and (2). Since the semantic similarity is not based on microarray data, the sample size is not available. An equal sample size is assumed for prior as in the data to avoid any bias.

It is assumed that $n_{prior} = n_{Likelihood}$ and equation (2) becomes

$$\sigma^2_{posterior} = \frac{1}{2 * n_{likelihood}} \quad . \qquad (4)$$

Equation (1) can be then rewritten as

$$\mu_{Posterior} = \sigma^2_{posterior} * (n_{prior} * \tanh^{-1} r_{Resnik} + n_{Likelihood} * \tanh^{-1} r_{Likelihood})$$

$$= \sigma^2_{posterior} * n_{Likelihood} \, (tanh^{-1} r_{Resnik} + tanh^{-1} r_{Likelihood}) \ . \qquad (5)$$

For example, a gene pair with high Pearson's correlation of 0.80 and low semantic similarity of 0.20 with a sample size of 78 results in a mean estimate of 0.65. It can be written as Normal (Mean=0.65, Variance=0.0001) resulting in a point estimate of the correlation coefficient of tanh(0.65) =0.57 on the original scale. The 95% confidence interval of the mean is $0.65 \pm 1.96 * \sqrt{0.0001}$ = [0.6304, 0.6696]. The corresponding 95% confidence interval for the posterior ρ is obtained using the hyperbolic tangent transformation as [tanh(0.6304), tanh(0.6696)] = [0.56,0.58] on the original scale.

2.3.2 Correlation measures

One of the objectives of this study was to compare different correlation measures to determine the best correlation method. Nine co-expression similarity measures were compared: Pearson's correlation, Spearman's correlation, Kendall's correlation, Shrinkage based correlation, Partial correlation, Mutual Information, Euclidean distance, Manhattan distance, and *BaySim* (the proposed new Bayesian measure).

The standard Pearson correlation is essentially a measure as to how similar the directions in which two expression vectors are. The Pearson correlation treats the vectors as if they were the same (unit) length, and is thus insensitive to the amplitude of changes that may be seen in the expression profiles. A mathematically rigorous correlation coefficient of two data vectors is considered based on James-Stein shrinkage estimators. These estimators are obtained by introducing a "shrinkage coefficient" taking a value between 0 and 1. The classical Pearson estimator is a special case of this family of estimators when the shrinkage coefficients are 1 and 0 respectively. The rank-based metrics Spearman's and Kendall's correlation metrics are also considered. The Spearman's correlation uses ranks rather than raw expression levels. This makes it less sensitive to extreme values in the data. The rank correlation methods are suited for data that are far from normal. The standard Euclidean and Manhattan distances measure the geometric distance between two vectors. They consider difference between two gene expression levels directly for comparison and hence take the magnitude of changes in the gene expression levels into account. It therefore preserves more information about the gene expression levels compared to rank based methods. As opposed to these measures, it is well known that *mutual information* (MI) provides a general measurement for dependencies in the data, in particular positive, negative and nonlinear correlations [35]. It is a generalized measure of statistical dependence in the data, and it is reasonably robust against missing data and outliers. Michaels et al. [59] indicate that the Euclidean distance and the MI measure have a high degree of correspondence.

2.3.3 Co-expression network analysis

The microarray data is converted to a graph structure by representing genes as nodes and the correlation between the genes as the edges. A correlation threshold is used to filter the graph in order to retain the high correlated edges. The threshold used in the current study is based on the top 1% of the correlation distribution. The 99th percentile of correlation distribution is chosen as the threshold.

Clusters of co-expressed genes are generated using the graph theory based paraclique algorithm described in detail at [23]. Paraclique is a modified version of clique algorithm [77] proposed to mitigate the effects of noise as well as to view correlation structures at a more interpretable level of granularity. Paraclique is very similar to clique in that it is an extremely densely-connected subgraph, but one that may be missing a small number of edges. In our context, this corresponds to a very highly correlated group of genes whose representational levels show highly significant but not necessarily perfect pair-wise correlations. A maximum clique is the largest clique in a given graph. A paraclique consists of the maximum clique and all vertices with at least some proportion of edges to the maximum clique. The proportion is called the *glom factor*.

Briefly the paraclique algorithm is described as follows: Beginning with a clique, C, of size k, each non-clique vertex, v is considered. Vertex v is marked if and only if it is adjacent to at least k-1 vertices in C. After each vertex has been considered, a paraclique, P is defined to be the union of C and the set of all marked vertices. P is removed from the graph and the process is reiterated.

2.3.4 Silhouette validation

A good clustering algorithm is expected to produce groups with distinct non-overlapping boundaries, although a perfect separation cannot typically be achieved in practice. Clustering validity measures such as the silhouette width [63] can be used to evaluate the separation of groups obtained from a clustering algorithm. The Silhouette validation technique is a way to assess the strength of clusters: Is the data set clustered well based on intra-cluster and inter-cluster distances? Cluster silhouettes are a classical way of depicting the quality of a given clustering of objects. The silhouette value for each point in a cluster is a measure of how similar that point is to points in its own cluster vs. points in other clusters, and ranges from -1 to +1. The average silhouette width for each cluster and overall average silhouette width for a total data set can be calculated. Using this approach each cluster could be represented by so-called silhouette that is based on the comparison of its tightness and separation. The average silhouette width could be applied for evaluation of clustering validity and can also be used to decide how good the clustering solution is.

To construct the silhouettes S(i) the following formula [63] is used:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

where a(i) = average dissimilarity of ith-object to all other objects in the same cluster b(i) = minimum of average dissimilarity of ith-object to all objects in other cluster (in the closest cluster).

2.3.5 Metrics for biological validation of paracliques

Two metrics proposed by Datta and Datta [61], Biological Homogeneity index (BHI) and Biological Stability Index (BSI), are used to validate the clustering solutions obtained from different correlation measures.

Suppose that G is the set of all genes in a given microarray experiment. Let C_1 …..C_F be *F* functional classes, not necessarily disjoint. Public databases (e.g., Gene Ontology, Entrez Gene, Unigene cluster) can be used to annotate and organize the expression values from a microarray experiment into families related by the biological characteristics of the genes or of their encoded proteins. In this study Gene Ontology database was used to assign each gene to the molecular function GO category with highest information content.

Biological homogeneity index (BHI)

Consider two annotated genes *x, y* that belong to the same statistical cluster D. Let us say that C(*x*) is a functional class containing gene *x*. Similarly C(*y*) contains gene *y*. The indicator function *I*(C (*x*) = C(*y*)) will be assigned the value 1 if C(*x*) and C(*y*) match (in case of membership to multiple functional classes, any one match will be sufficient). As genes *x* and *y* are in the same statistical cluster, it is expected that the two functional classes match. Thus, the following measure [61] evaluates the biological similarity of the statistical clusters:

$$BHI = \frac{1}{k} \sum_{j=1}^{k} \frac{1}{n_j(n_j-1)} \sum_{x \neq y \in D_j} I(\mathrm{C}\,(x) = \mathrm{C}(y)) \,,$$

where k is the number of statistical clusters and for cluster D_j, $n_j = n(D_j \cap C)$ is the number of annotated genes in D_j, and where for a set *A*, *n*(*A*) denotes its size or cardinality.

This is a simple measure that is easy to interpret and implement once the reference collection of functional classes are in place. This also works with overlapping functional classes. This measure can be interpreted an average proportion of gene pairs with matched functional classes that are statistically clustered together based on their expression profiles.

Biological stability index

Next the stability of a clustering algorithm is captured by inspecting the consistency of the biological results produced when the expression profile is reduced by one observational unit.

In a microarray study, each gene has an expression profile that can be represented as a multivariate data value in R^p, for some $p > 1$ where p is the number of samples. For example, in a time course microarray study, p could be the number of time points at which expression readouts were taken. In a two sample comparison, p could be the total sample size, and so on. For each $i = 1, 2, ..., p$, the clustering algorithm is repeated for each of the p data sets in R^{p-1} obtained by deleting the observations at the ith position of the expression profile vectors. For each gene g, let $D^{g,i}$ denote the cluster containing gene g in the clustering based on the reduced expression profile. Let $D^{g,0}$ be the cluster containing gene g using the full expression profile. For each pair of genes x and y in a biological class, the statistical clusters containing x based on the original and the statistical cluster containing y based on the reduced profile are compared. A stable clustering algorithm would produce similar answers, as judged biologically, based on the original and the reduced data. Thus, the clusters using full and reduced data, respectively, containing two functionally similar genes should have substantial overlaps. This is captured by the stability measure and larger values of this index indicate more consistent answers. BSI is given by the following formula

$$BSI = \frac{1}{F}\sum_{i=1}^{F}\frac{1}{n(C_i)(n(C_i)-1)p}\sum_{j=1}^{p}\sum_{x\neq y\in C_i}\frac{n(D^{x,0}\cap D^{y,j})}{n(D^{x,0})}.$$

A biologically valid clustering is characterized by high values of both of these indices.

2.3.6 *Datasets*

Three datasets, two from Yeast and one from Human will be used for the study. The yeast datasets were chosen for several reasons. First, Yeast is a model organism for which extensive experimental protein-protein interaction information and GO annotations are available. Other factors such as sample size and number of replicates are taken in to consideration while choosing the datasets.

1) The Cho et al. [78] cell-cycle yeast dataset consists of 6601 genes comprising all the genes in yeast at 17 different time points past the cell cycle arrest. The dataset represents the complete characterization of mRNA transcript levels during the cell cycle of the budding yeast Saccharomyces cerevisiae. Ge et al. [79] assembled gene expression data from the Cho et al. [78] yeast cell-cycle experiment, literature protein-protein interaction (PPI) data and yeast two-hybrid data. The reduced data consisting of 2885 genes that were common to all experiments was used for examining the biological relevance of the correlations. The PPI data consisted of 315 protein interactions among the 2885 genes.
2) The Spellman.et al. [14] microarray data originally contained the gene expression of 4289 genes at 24 time points during the cell-cycle. The data-set was taken as

provided by the public download site for Spellman et al. [14] paper material (http://genome-www.stanford.edu/clustering). Signals represent log (ratio) where ratio is calculated between the absolute signals of two dyes (spotted microarray technology). Data from separate time courses of gene expression in the yeast S. cerevisiae were combined and then used for the correlation analysis. Data were drawn from time courses during the following processes: the cell division cycle after synchronization by alpha factor arrest (18 time points); centrifugal elutriation (14 time points), and with a temperature-sensitive cdc15 mutant (15 time points); sporulation (7 time points plus four additional samples); shock by high temperature (6 time points); reducing agents (4 time points) and low temperature (4 time points) and the diauxic shift (7 time points). All data were collected by using DNA microarrays with elements representing nearly all of the ORFs from the fully sequenced S. cerevisiae genome. All measurements were made against a time 0 reference sample except for the cell-cycle experiments, where an unsynchronized sample was used. About 2467 genes which were well annotated in the Saccharomyces Genome Database were included.

3) In Tian et al. [80], microarray data from human patients with diabetes, inflammatory myopathies and Alzheimers data sets were analyzed. The inflammatory myopathies data consisted of 7 normal and 8 inclusion body myositis (IBM) samples. The 5000 probe sets in this matrix represent the most variable probe sets (by expression value) in the 15 arrays.

K-nearest neighbor imputation [81] was used to treat missing data since it was found in the previous studies to be more robust to missing value estimation as compared to the standard row average estimation.

2.3.7 Outline of analysis

Correlation matrices are generated using the different correlation methods – *BaySim*, Pearson's, Spearman's, Kendall's, shrinkage, partial, mutual information, Euclidean and Manhattan distances - which are then input to the paraclique algorithm run using a threshold based on top 1% of the correlation distribution and a glom factor of 1. This results in the generation of paraclique solutions corresponding to the chosen correlation metrics. Then the quality metrics such as silhouette and biological validity indices - BHI and BSI – of the paraclique solutions from different correlation metrics were calculated to check which metric produces valid and biologically relevant clusters.

2.3.8 Software

The correlation values were computed in R/Bioconductor [82, 83]. The *cor* function in the BASE library was used for computing Pearson, Spearman and Kendall correlations, *corpcor* package for partial and shrinkage estimates, *bioDist* package for distance functions, *GOSim* for semantic similarity, *clValid* for BHI and BSI biological validation measures. Paraclique analysis was done using a C software package developed by M.A Langston's group at the University of Tennessee. This software employed principles of fixed parameter tractability [84] to find vertex covers [85] which were then used to extract cliques. The protein-protein interaction data and pathway annotations for

Yeast have been obtained from the bioconductor packages *YeastExpData* and *org.Sc.sgd.db* respectively.

2.4 Results

2.4.1 Results from BaySim.

Biological example

The co-expression between every pair of genes was estimated using the Spellman et al. [14] data consisting of 2467 genes using *BaySim*. Figure 2 shows the distribution of BaySim using Spellman et al. [14] data which is approximately normal as derived by the posterior distribution of the correlation (see *Methods*).

By adding prior biological information, it is highly likely that the number of false negatives is reduced. As an example, two histone proteins HHT1 and HTB2 with high functional similarity had a Pearson's correlation of 0.32 conveying a low co-expression using the data alone. However the new measure yields a reasonably high correlation of 0.76 by incorporation of biological similarity. These two genes were found to belong to the same pathway based on annotation. On the other hand, if a gene pair shows high Pearson's correlation and a low biological similarity, the new measure shrinks the Pearson's estimate towards the GO similarity which serves as a means of reducing false positives.

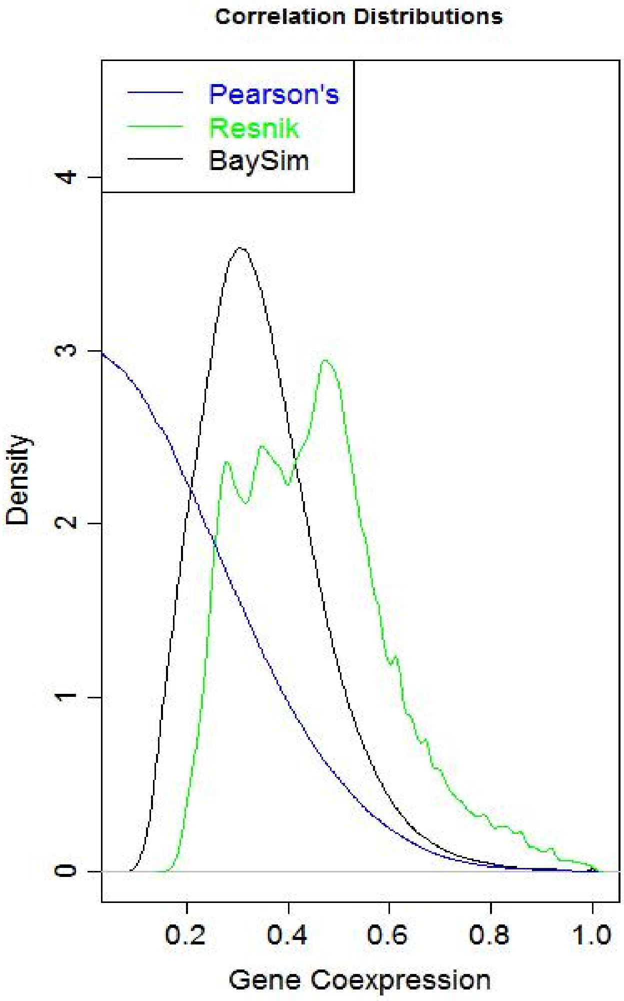


Figure 2. Density plot of Baysim and its components -Pearson's correlation and Resnik's GO similarity.

Microarray data is very noisy and relying on the data alone might lead to erroneous conclusions. Hence addition of prior biological information deals with the reduction of false positives and false negatives.

Biological validation

Genes from each paraclique were submitted to the functional analysis tools developed at DAVID Bioinformatics Resources http://david.abcc.ncifcrf.gov/ [86] to cluster the genes based on annotation. It was seen that *Baysim* yielded more meaningful and homogenous clusters compared to that resulted from the Pearson measure alone. It is observed that each cluster of genes represented distinct GO categories and pathways. For example cluster 1 consisted of 138 genes mostly consisting ribosomal genes. Most of the genes were highly enriched for the ribosomal categories in GO ontology and mapped to ribosome pathway using KEGG. Cluster 2 consisted of 40 genes which are mainly involved in metabolic processes and contained some transcription factors as evident from the enriched GO categories represented by the cluster. Cluster 3 and 4 mainly represented the mRNA processing and mRNA splicing GO categories. This method performed very well in producing gene clusters with distinct functional categories. Though there are multiple clusters representing similar functions, most of the genes in a single cluster represented a common function. In other words, the clusters obtained using *BaySim* were very homogeneous in terms of gene function.

Evaluation of Clustering quality using Silhouette width

Does adding biological information to the data decrease the noise level in terms of the quality of clustering produced? To address this question the average silhouettes of the paraclique sets resulting from Pearson's and *BaySim* similarity measures were compared using the Spellman et al. [14] and Cho et al. [78] datasets. If silhouette value is close to 1, it means that sample is "well-clustered" and it was assigned to a very appropriate cluster. If silhouette value is about zero, it means that that sample could be assigned to another closest cluster as well, and the sample lies equally far away from both clusters. If silhouette value is close to –1, it means that sample is "misclassified" and is merely somewhere in between the clusters. The overall average silhouette width is simply the average of the S(i) for all objects in the whole dataset.

The largest overall average silhouette indicates the best clustering. As seen in the Table 1, *BaySim* performs better in the case of the two Yeast datasets. A difference in silhouettes of 0.1328 and 0.329 respectively is observed in the two cases. However, *BaySim* did not show any difference in the case of Tian et al. [80] dataset. This is might be due to the lack of sufficient annotation of the human genes in the GO database. It is important to note that Spellman and Cho datasets are from yeast which is well annotated in the GO database and hence the incorporation of biological information is very reliable.

Table 1. Silhouettes of clustering solutions obtained from Pearson's and BaySim on the three datasets.

Silhouette	Spellman et al. [14]	Cho et al.[78]	Tian et al. [80]
Pearson	0.5997	0.0116	0.6553
BaySim	0.7325	0.3406	0.6012
Spearman’s	0.1946	0.2923	0.2011
Kendall’s	0.2090	0.2177	0.1964
Partial correlation	0.1571	0.1695	0.5932
Mutual information	0.2376	0.3211	0.3195
Euclidean	0.0597	0.1654	0.0219
Manhattan	0.0739	0.1046	0.0551

On the other hand, Tian et al. [80] human dataset has generally incomplete gene annotation hence prior information might not be as informative as in the case of yeast. This could be one of the possible reasons for the lack of reduction of noise in the Tian et al. dataset.

2.4.2 Comparison of correlation metrics for gene co-expression networks.

Agreement of correlation metrics

A non-parametric correlation, in this case – Spearman's correlation, is used to calculate the correlation of the correlation matrices. From Table 2, it is evident that there is no good agreement between all the different correlation measures. As expected, distance measures Manhattan and Euclidean are well correlated and the non parametric measures Spearman and Kendall are also well correlated. Since shrinkage is a linear combination of Pearson correlation, there is a perfect correlation between Shrinkage and Pearson measures.

Paraclique comparison

Genes are clustered using paraclique algorithm [22] applied to the Spellman et.al [14] dataset based on the various definitions of correlation coefficients i.e **the chosen clustering algorithm (paraclique) is run using each of the chosen correlation measures.** The threshold parameter is set using top 1% of the correlation distribution for all the methods**.** Then the proximity of the clustering solutions from each of these metrics is estimated using a similarity coefficient. This allows us to see which metrics are similar in terms of co-expression networks produced. A similarity coefficient S is used to judge if two clustering solutions A and B are close to each other. The similarity $\mathrm{Sim}(C_i,C_j)$ between two clusters C_i and C_j with $n(C_i)$ and $n(C_j)$ number of genes respectively can be found by

$$\mathrm{Sim}\left(C_i, C_j\right) = \max\left\{\frac{n\left(C_i \cap C_j\right)}{\sqrt{n(C_i) * n\left(C_j\right)}}\right\}$$

$$\mathrm{Sim}(A, B) = \overline{\mathrm{S\imath m}\left(C_\imath, C_\jmath\right)}\,, i \epsilon A, j \epsilon B\ .$$

In Table 3, most of the values of similarity coefficients are less than 0.5, implying that different correlation methods yield paraclique solutions that are quite different from each other. Comparing Tables 2 and 3, it is observed that paraclique similarity across any pairs of methods is not high as corresponding correlation metric similarity. *Baysim* as expected yielded paraclique structures that was most similar to those from Pearson's. The same trend is observed as in the comparison of correlation values in that the non-parametric correlation metrics – Spearman's and Kendall's – and the distance measures – Euclidean and Manhattan yielded clustering solutions that agree well among each other with a similarity of greater than 0.8. Since shrinkage is a linear combination of Pearson correlation, the paracliques resulting from the both measures are the same and hence a perfect correlation is observed.

Table 2. Spearman's correlation between the different similarity metrics using Spellman et al. data [14].

	Pea	**Spe**	**Ken**	**Shr**	**Euc**	**Man**	**MI**	***BaySim***
Pea		0.7428	0.7293	1	0.0060	0.0086	0.2778	0.5356
Spe			0.9944	0.7428	0.0745	0.0806	0.2387	0.3391
Ken				0.7293	0.0761	0.0854	0.2461	0.2384
Shr					0.0060	0.0086	0.2778	0.2256
Euc						0.9712	0.0103	0.2054
Man							0.0075	0.1078
MI								0.2420

Abbreviations used: Pea-Pearson, Spe-Spearman's, Ken-Kendall's Tau, Shr – Shrinkage, Man-Manhattan, MI- Mutual information

Table 3. Agreement of paraclique solutions from different similarity metrics using a cluster similarity coefficient (Spellman et al. data [14])

	Pea	Spe	Ken	Shr	Euc	Man	MI	*BaySim*
Pea		0.4051	0.4117	1	0.2845	0.2984	0.2765	0.5123
Spe			0.8432	0.6477	0.1660	0.1925	0.1885	0.2485
Ken				0.2374	0.1626	0.1874	0.1869	0.2369
Shr					0.1220	0.1372	0.1214	0.5123
Euc						0.8319	0.1779	0.0418
Man							0.1720	0.0724
MI								0.3062

Abbreviations used: Pea-Pearson, Spe-Spearman's, Ken-Kendall's Tau, Shr – Shrinkage, Man-Manhattan, MI- Mutual information

Effect of correlation metrics on Co-expression Network features

Degree profiles:

Correlation graphs obtained using the different correlation metrics on Spellman et al. dataset at 1% edge threshold slightly differ from each other in the number of vertices (Table 4), but showed very similar degree distributions. All the methods yielded graphs with similar degree distributions except for the graph from the partial correlation estimates as shown in Figure 3.

The degree distribution of the graph from partial correlations was very different from that of the other measures. First the range of the degrees was much smaller compared to that from the other methods. Second, the shape of the distribution was smoother and concave shaped compared to L-shaped distribution that was seen in the other ones. As expected, the degree distribution of the graph using *BaySim* was closest to that from Pearson's correlation. The methods that has yielded graphs with the highest maximum degrees are the distance measures such as Manhattan and Euclidean followed by Mutual information. This is in accordance with the notion that mutual information captures any general dependency present in the data as compared to other measures which only aim to extract specific kind of relationship (linear etc). While most of correlations yielded graphs that had a maximum degree of less than 250, graphs from distances measures were extremely dense and had very high degrees of greater than 400. However the average degree per vertex was highest in the graph using Pearson correlation (Table 4).

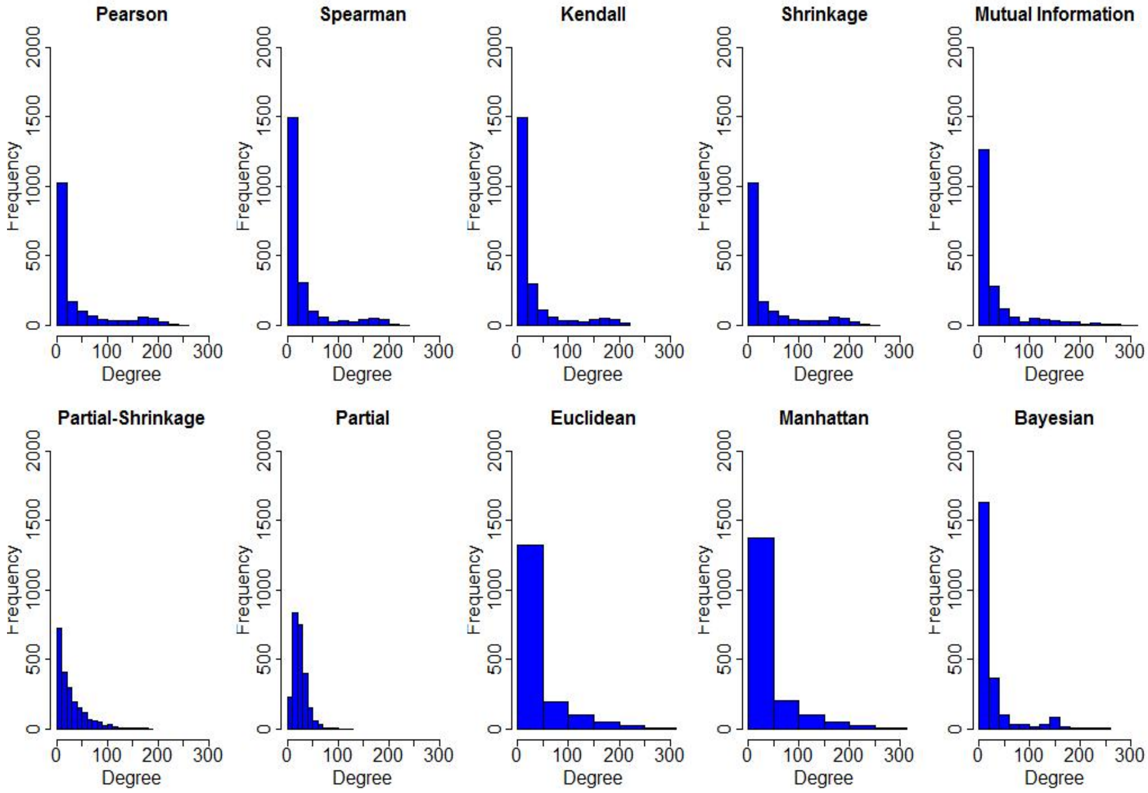


Figure 3. Degree distributions of graphs from different correlation metrics - using Spellman et al. [14] data.

Table 4. Correlation graph features using different correlation metrics using Spellman et al. [14] data.

Correlation Metric	**Number of Vertices**	**Average Degree**	**Max Degree**
Pearson	1625	37.43754	248
Spearman’s	2182	27.88084	221
Kendall’s Tau	2175	27.97057	217
Shrinkage	1625	37.43754	248
Mutual Information	1954	31.13408	309
Partial-Shrinkage	2150	28.29581	182
Partial	2466	24.66991	121
Euclidean	1694	35.91263	408
Manhattan	1749	34.78559	434
Baysim	2321	25.95519	255

Graphs from most of the correlation methods exhibited the scale-free behavior that is typically expected in gene co-expression networks as evident from the degree distributions. Only the graph resulting from the standard estimates of partial correlation did not exhibit the scale free behavior. This could be due to the improper estimation of the covariance matrices due to problems arising from matrix inversion. Shrinkage estimation of covariance matrix retained the scale-free behavior of the network as shown in Figure 3.

Paraclique sizes: Different correlation methods yield different distributions of paraclique sizes as shown in Figures 4 and 5 for Spellman et al. [14] and Tian et al. [80] datasets respectively. The size distributions are very similar for Pearson's, Spearman's and Kendall Tau correlation methods. They all have large number of modest sized paracliques and small number of large sized paracliques. The rank-based methods such as Spearman's and Kendall's correlation display similar patterns as expected. The partial correlations produced with and without shrinkage estimation produce distributions of paraclique sizes which are quite different from those of the other methods. The size distributions are identical for Pearson and shrinkage since they have the same paraclique structure. Also the number of vertices (genes) and number of paracliques resulting from different metrics are reasonably similar among each other.

.

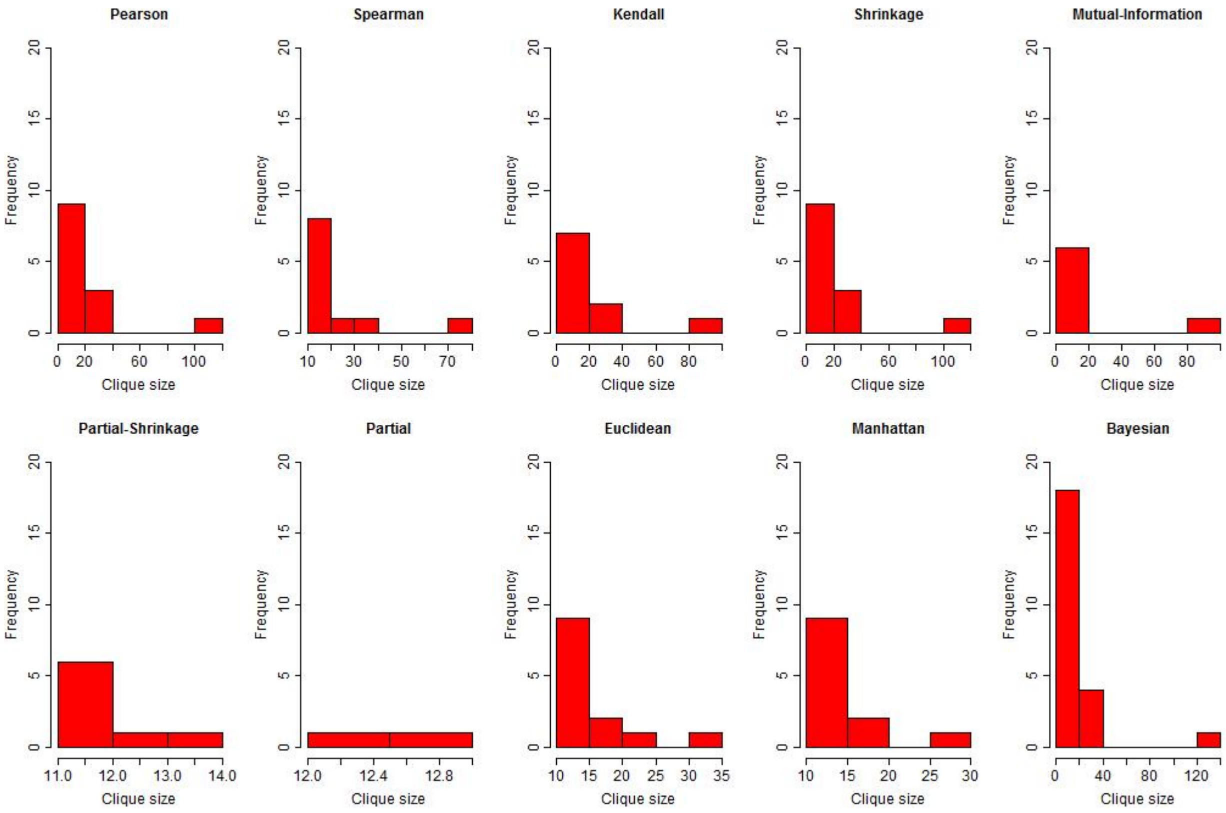


Figure 4. Histogram of sizes of paracliques from different correlation metrics using Spellman et al. [14] data

Table 5. Number of paracliques and mean paraclique sizes from different correlation metrics

	Spellman et al. [14]		Tian et al. [80]	
	Number	Average size	Number	Average size
Pearson’s	13	24	39	23
Spearman’s	11	22	52	19
Kendall’s	10	23	51	17
Shrinkage	13	24	39	23
Mutual Information	7	23	17	23
Partial	2	12	47	15
Partial-Shrinkage	8	11	43	15
Euclidean	13	15	38	24
Manhattan	12	15	34	25
BaySim	23	22	13	20

Biological Validation: Correlation between gene expression, protein-protein interactions and pathways

Cho et al. [78] dataset was used to study the agreement of different correlation metrics in the degree of relationship between gene co-expression and protein level interaction. The 315 gene pairs from the Cho et al dataset that had the corresponding protein-protein interaction data were used for analysis. Using Pearson correlation, it was shown that given a high correlation such as 0.8 between two genes, the probability that the corresponding proteins would interact was quite low. Out of the 315 pairs of interacting proteins, only **3%** were found to have a Pearson correlation of above 0.8. Only about **20%** of the interaction pairs were found to have a correlation above 0.5. From Figure 6, it is evident that Pearson and Partial shows decreasing trend in frequency of interaction pairs as correlation increases and it is interesting to note that mutual information shows the opposite trend. Bayesian correlation metric however showed a distribution which is approximately normal and very different from the others. Incorporation of GO information tends to increase the mean similarity of the gene pairs representing Protein-Protein interactions. Mutual information shows a very high correlation for most of the gene pairs in general and hence shows a mean correlation of above 0.9 for the PPI gene pairs.

If the two genes had a high correlation using a specific correlation metric, are they most likely to be part of the same pathway? The Cho et al. dataset was again used to look at the biological relevance of all the chosen correlation metrics in terms of pathways.

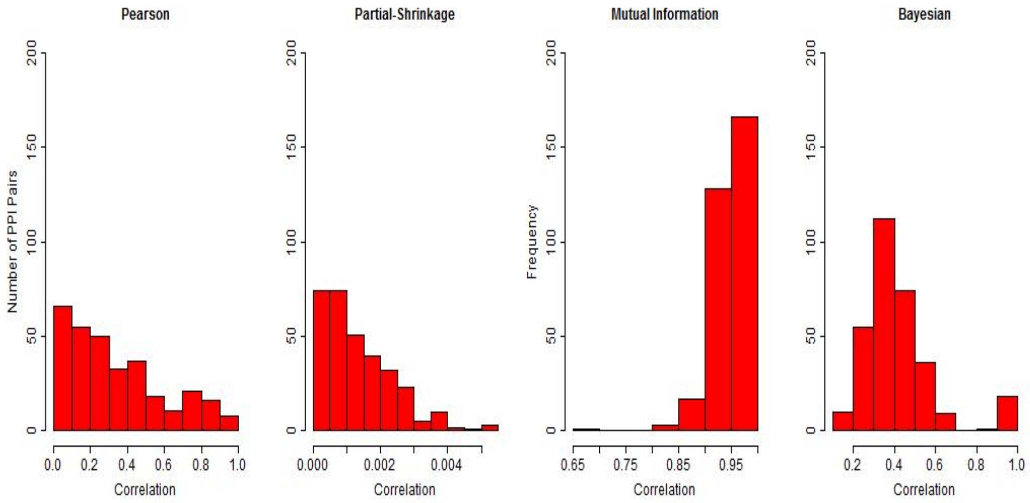


Figure 5. Correlation distributions of the gene pairs with protein-protein interactions using Cho et al. [78] data.

About **51%** of the total pairs of genes that belong to the same pathway have a pearson correlation above 0.5 and **29%** of them have a correlation above 0.8. Shown in the Table 6 are the means of correlation distributions from all the other correlation metrics. Figures 7,8,9 and 10 show the distribution of correlations of the genes belonging to the same pathway using the all the correlation methods (using Cho et al. [78] dataset). In the case of Pearson's, a large number of highly correlated genes belong to the same pathway. For gene pairs with correlation higher than 0.6, an increasing trend in the number of genes in the same pathway is observed with an increase in correlation. We see an increased signal at the right end of the distribution in the case of Pearson's whereas addition of the biological information shaves off the high signals at the either ends of the Pearson's distribution yielding an approximately normal distribution. However *Baysim* yielded a mean correlation of 0.60 which is higher than that in the case of Pearson's which is 0.51. Partial correlations of the gene pairs belonging to the same pathway follow a similar distribution to that of Pearson's, but on a compressed scale (Figure 9).

Non-parametric measures such as Spearman's and Kendall's correlations yielded similar distributions which are a bit noisier compared to the other ones since there is no good separation between the low, moderate and high correlations as shown in Figure 8. In the case of mutual information, most of the correlations of gene pairs belonging to the same pathway are greater than 0.8. However this distribution is not any different from the general distribution of mutual information which makes it difficult to predict pathway membership using the correlation distribution.

Table 6. Mean of correlation distribution of gene pairs belonging to a common pathway-using Cho et al. [65] data.

	Mean correlation
Pearson	0.51
Spearman’s	0.45
Kendall’s	0.34
Mutual Information	0.92
Partial (Shrinkage estimation)	0.20
Euclidean	0.88
Manhattan	0.87
Baysim	0.60

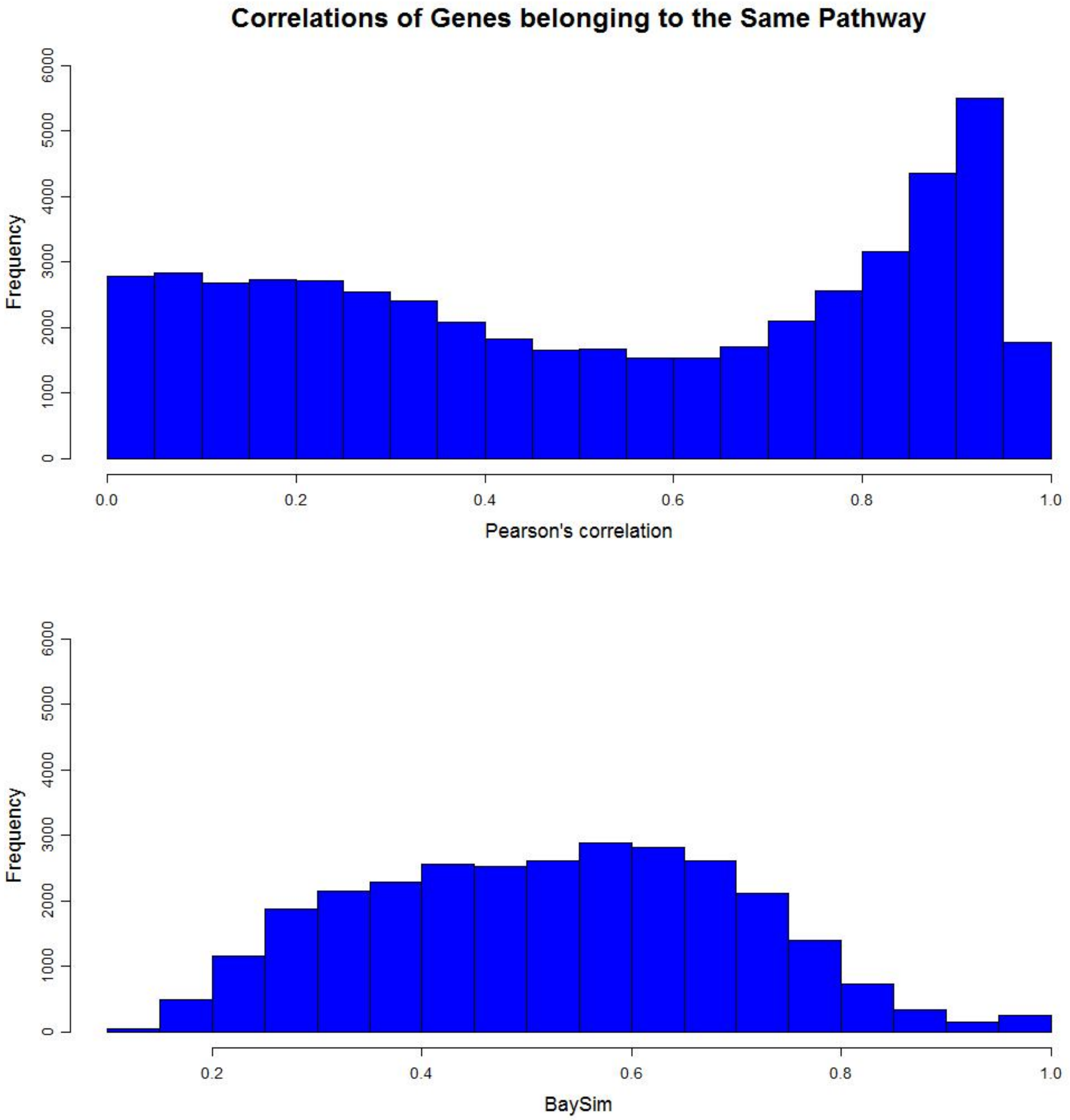


Figure 6. Histogram of correlations of the gene pairs belonging to the same pathway using Pearson's and Baysim (Cho et al. [65] dataset).

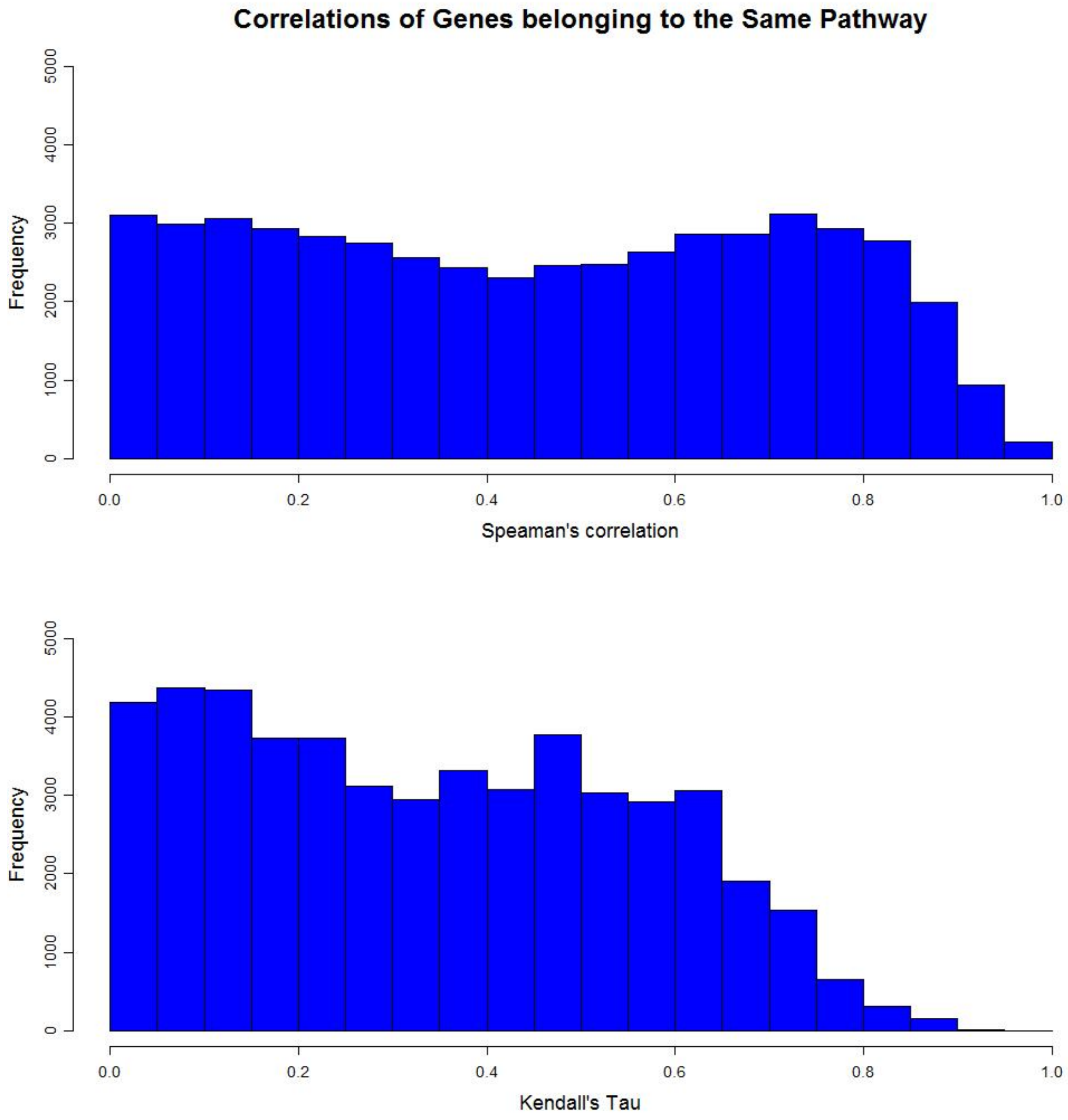


Figure 7. Histogram of correlations of the gene pairs belonging to the same pathway using Spearman's and Kendall's measures (Cho et al. [65] dataset).

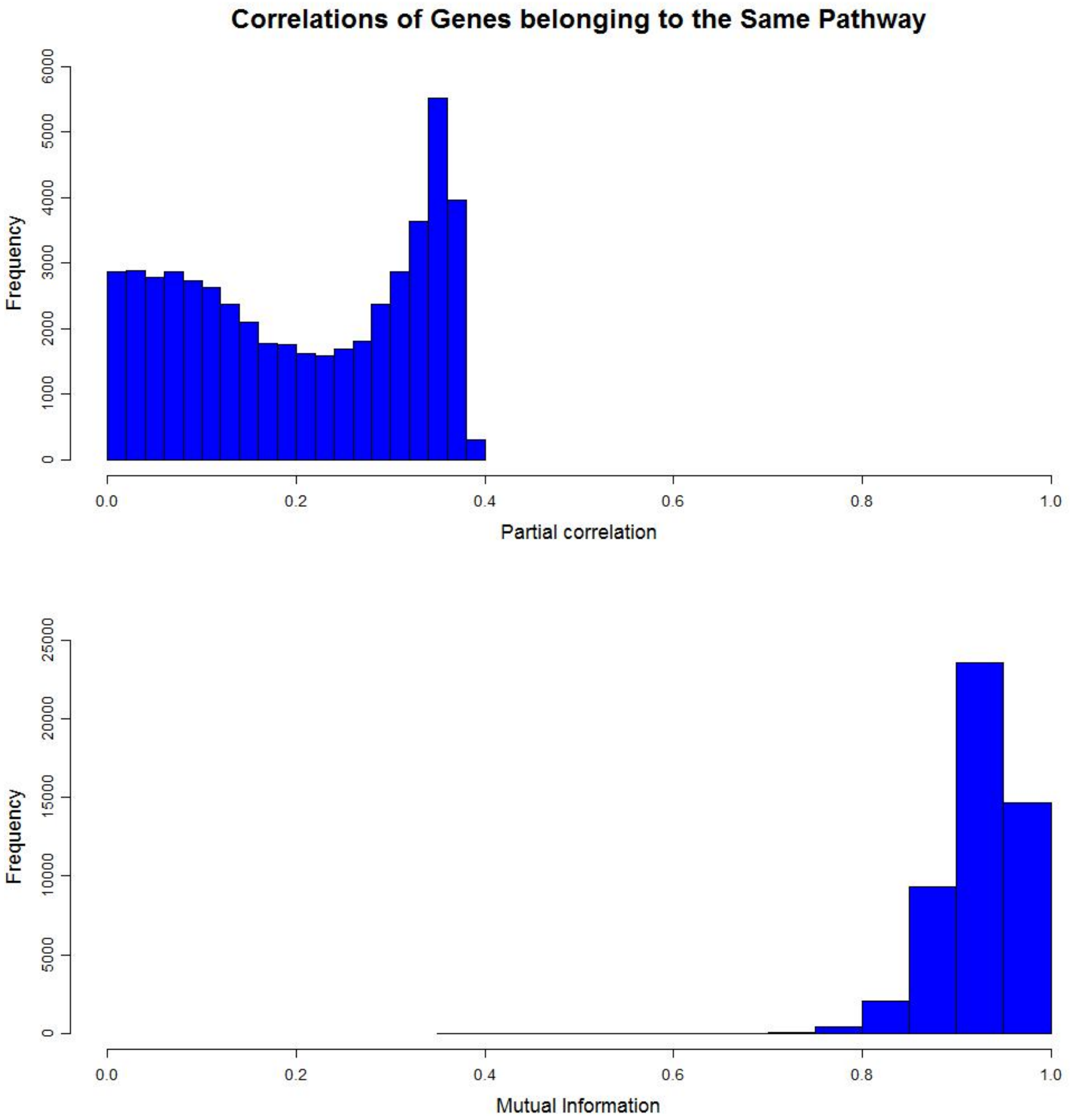


Figure 8. Histogram of correlations of the gene pairs belonging to the same pathway using Partial correlation and Mutual Information (Cho et al. [65] dataset).

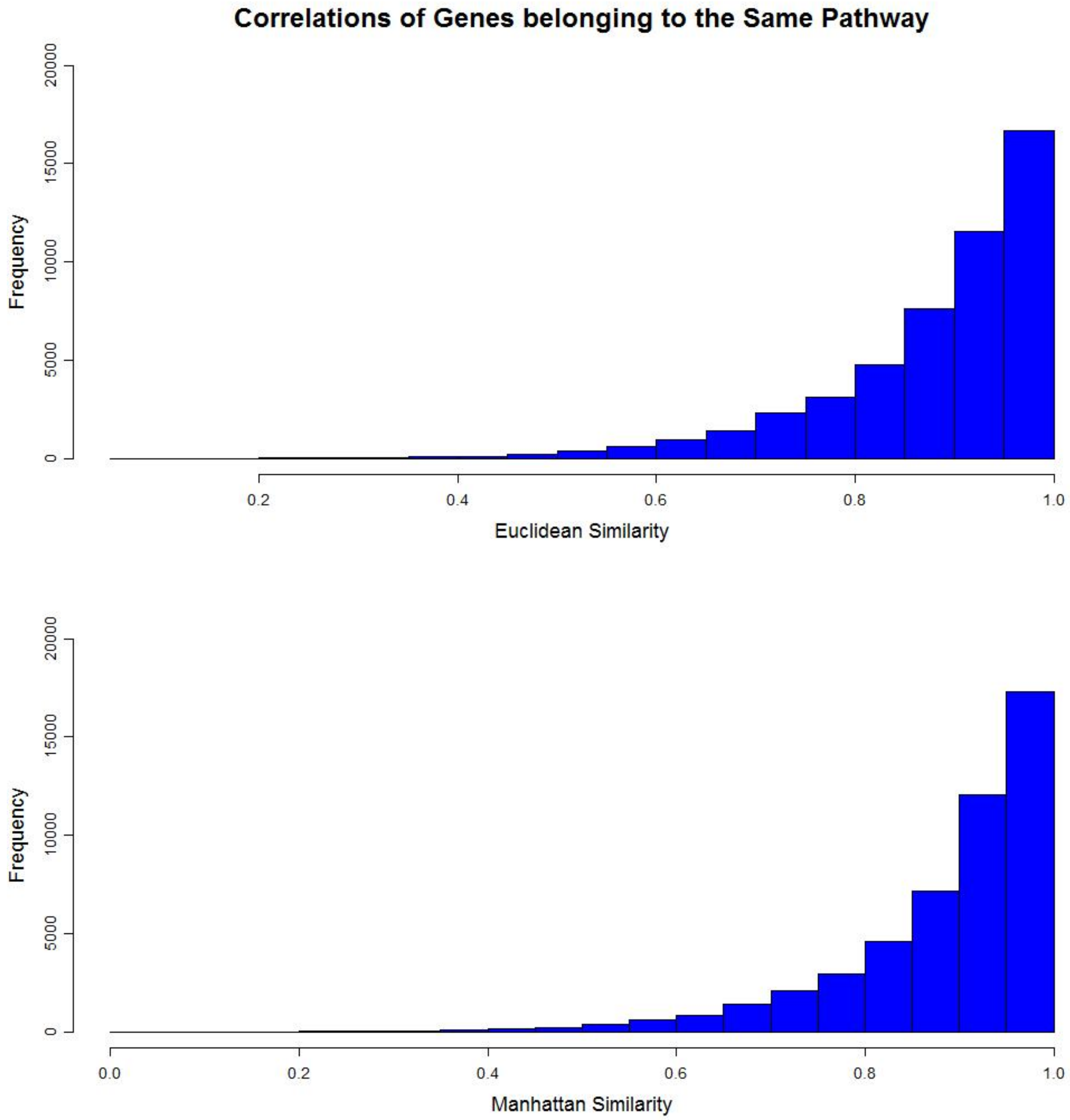


Figure 9. Histogram of correlations of the gene pairs belonging to the same pathway using Euclidean and Manhattan distance measures (Cho et al. [65] dataset).

From this analysis it is implied that high correlation does not necessarily imply pathway membership and that the modest correlations might be quite informative and should be given good consideration in the co-expression analysis.

Biological validation of Paracliques

Validation indices introduced by Datta et al. [61] - Biological homogeneity index (BHI) and biological stability index (BSI) measuring statistical stability and biological congruence, respectively are used for the validation of paraclique solutions from different correlation metrics. For each dataset, the BHI and BSI were computed for each clustering solution (paraclique set) obtained from each of the correlation methods. The genes are assigned to functional classes based on GO categories. Since *BaySim* is partly based on the ontological similarties, BaySim gives highest values of BHI and BSI. From the Table 7, it is seen that apart from *BaySim*, partial correlation using shrinkage estimation and mutual information consistently gives the best values for both BHI and BSI for all the datasets. Based on the ranking of the BHI's and BSI's of the paracliques from the different measures, it is concluded that the paracliques resulting from these matrices are more biologically relevant than the clustering solutions using the other correlation measures. However it has to be noted that the differences in the BHI resulting from the different measures are not large and more datasets need to be tested in order to confirm this conclusion.

Table 7. Biological validation of paracliques from different correlation metrics using Cho et al. [65] data

	Spellman et al. [14]		Cho et al. [78]		Tian et al.[80]	
	BHI	BSI	BHI	BSI	BHI	BSI
Pearson	0.5656	0.2323	0.4032	0.4333	0.5432	0.2968
Spearman	03452	0.3219	0.2849	0.4677	01634	0.2664
Kendall	0.2634	0.3491	0.4729	0.3466	0.1433	0.4322
Mutual Information	0.5643	0.4933	0.5034	0.4944	0.7543	0.3491
Partial	0.3442	0.3789	0.3201	0.4334	0.3645	0.4581
Partial (Shrinkage estimation)	0.5764	0.5323	0.6344	0.3426	0.5377	0.6314
Euclidean	0.3792	0.2691	0.3211	0.2663	0.4421	0.6389
Manhattan	0.3831	0.2943	0.3432	0.2943	0.4270	0.4234
BaySim	0.8143	0.6213	0.8719	0.3125	0.7941	0.6421

2.5 Discussion

Most studies that looked at co-expression networks used one measure or another to quantify the similarity of expression profiles without objectively assessing their merit, and without an underlying statistical justification. This is important and needs further attention as microarray data is noisy, and it is often difficult to separate real signals from random fluctuations. Therefore, the choice of the metric can greatly affect the microarray analysis results when looking for clusters of co-expressed genes.

In this study, the quality of different similarity measures for expression profiles was evaluated and a new measure called *BaySim* was proposed that is the most effective for detecting functional links. In terms of the network topology, all correlation metrics except for partial correlation produced very similar features such as degree distribution and cluster sizes. The similarity between different metrics is, however, confined only to the network structures. The correlation metrics do not agree in terms of the elements of the clusters produced. Each correlation metric imposes its own criterion in the quantification of the relationship between two profiles and hence the genesets produced from the different metrics vary. It is important to note that *BaySim* which incorporates functional similarity follows a normal distribution unlike the other methods.

We used silhouette as a metric for the assessment of the quality of clustering. Quality metrics based on intra-cluster and inter-cluster distances are most suitable for centroid based clustering approaches such as k-means and SOM. In networking based clustering approaches, however, metrics based on connectivity could be used to assess the co-expression network. However we did not use connectivity as a quality metric since the connectivity of cliques and paracliques are predefined and hence does not vary across the clusters.

Noise due to random measurement or error attenuates co-expression measure towards the null (i.e. toward no association). Strategies for correcting measurement error require knowledge about the reliability of the gene expression measurements which is not usually available, or increasing the sample size, which is not always possible. However, when there is knowledge of the association from previous studies, it can be coupled with the data collected and inference can be improved. Gene similarity measures from biological databases such as GO and KEGG pathways can be used in this way to deattenuate the effects of measurement error. Furthermore, the Bayesian approach can be used to combine as many correlation coefficients as necessary to achieve improved point estimates with narrower confidence intervals.

One of the advantages of the Bayesian method is that the confidence intervals can be interpreted as probabilities as they are based on a true probability function. This enables the investigator to assess the nature of the relation between two variables (genes) more intuitively. It is recognized that special attention should be given to the choice of prior when using Bayesian estimation procedures, since differences in the correlation estimates between the sampled population and the prior may reflect population heterogeneity. Evaluation of different prior distributions still needs to be performed in order to select the best distribution for the prior information.

BaySim is very reliable on annotation and is not appropriate for use in datasets from organisms with poor genomic annotation. In such cases the prior is not of any value and *BaySim* is simply equivalent to the Pearson's correlation. As the semantic similarity changes with the updates in annotations of the GO database, *BaySim* needs to be kept updated with such changes. Further investigation is required to determine which semantic similarity measure is most appropriate for use.

For this study, all the analyses were limited to graph theoretic approaches such as clique and paraclique. It would be interesting to see if *BaySim* and other correlation metrics exhibit similar effects on the clustering results obtained using standard clustering approaches such as K-means, hierarchical clustering etc. Since all the conclusions in this study have been based on few datasets, it is of importance to test several independent microarray datasets in order to further validate the robustness of *BaySim* measure.

Though *BaySim* produced clusters which are homogenous in terms of gene function, it did not show improved (higher) estimates of correlation for gene pairs belonging to a common pathway. This could be due to poor correlation between annotation similarity and pathway membership. As of today, Gene ontology is one of the most organized databases to look for annotation information of whole genome and hence *BaySim* was entirely based on the GO information. However the method is by no means limited to Ontological similarity. It sets up a standard platform to include any kind of appropriate biological distances based on pathway relatedness, annotations of specific and relevant tissue types or diseases if available in the future.

The current work also investigated the relationship of protein interactions with gene co-expression using all the chosen correlation metrics. Interacting proteins are more likely to be involved in similar biological functions and processes and thus they are more likely to be co-expressed. Earlier, Grigoriev [87] analyzed physical interactions in yeast and observed that proteins encoded by co-expressed genes interact with each other more frequently than with random pairs. Ge et al. [79] showed that interacting protein pairs are more likely to be in the same expression cluster than random pairs for yeast. On a genomic scale, they attempted to relate the absolute mRNA expression levels and the expression profiles in yeast to protein–protein interactions. In this study, it was seen that there is no correlation between gene co-expression and protein-protein interaction using any of the correlation metrics. However, several datasets need to be tested in order to further confirm this conclusion.

CHAPTER 3 STATISTICAL VALIDATION OF CO-EXPRESSION NETWORK MODULES

3.1 Abstract

An approach using Mantel statistics and permutation tests is presented to evaluate the significance of co-expression network modules. It was illustrated how this measure can be used to rank gene clusters likely to have important characteristics. An example using human myopathy data was used to illustrate this method only, and is not meant to be viewed as a definitive analysis of myopathy data. The statistical significance of cluster features such as paraclique size, number of paracliques and silhouette were evaluated using the standard permutation approaches. Several other network features such as connectivity and edge threshold needs to be evaluated further for significance in order to validate all aspects of the co-expression network.

3.2 Introduction

Clustering, the process of grouping genes based on their co-expression is a crucial step in the analysis of gene expression data. Some of the most commonly used clustering techniques applied to gene expression data include hierarchical clustering algorithms [12] , k-means [88], fuzzy c-means [21], mixture models [19] and SOMs [89]. Many improved clustering techniques such as biclustering [3] and gene shaving [90] have been developed to deal with the challenges posed by the high dimensional gene expression data. However traditional clustering techniques remain as the most predominant methods in post-genomics due to their conceptual simplicity, ease of representation and their widespread availability in standard software packages. Another class of clustering techniques is based on graph-theoretical approaches. They have a major advantage over other approaches in that the data when explicitly presented in terms of a graph convert the problem of clustering a dataset into such graph theoretical problems as finding minimum cut or cliques in the co-expression network. Moreover, graphical representations such as clique and paraclique provide displays of gene expression based information that may be explored to generate insights about pathways.

There is hardly any consensus on the best correlation measure or clustering method to be used for microarray data. As a consequence, it is common practice among researchers to employ a particular clustering algorithm that best suits their needs to analyze a dataset, and then to use visual inspection and prior biological knowledge to select what is considered the most appropriate result. In such inspection there is a high possibility that the researchers overrate clusters that reinforce their own assumptions and ignore results from other clusters that might be informative which potentially hinders the process of identification of surprising or unexpected patterns in the data that might then serve for hypothesis generation. Thus a cluster validation step in which the quality and significance of individual clusters are evaluated, is needed before the use of prior biological knowledge and assumptions in the final interpretation of a cluster analysis.

Cluster-validation provides an assessment of the quality and type of structure captured by clustering, and is therefore be a key tool in the interpretation of clustering results. The literature provides a range of different validation techniques broadly divided in to external and internal validation measures. External validation measures refer to all those methods that evaluate a clustering result based on the knowledge of the true clustering solution. In cases where no prior information on the clustering is available, an evaluation based on internal validation measures is appropriate. Internal validation techniques estimate the quality of clustering solution based on the information intrinsic to the data alone. Several internal validation measures have been proposed in literature based on compactness, connectedness, and separation of the cluster partitions. Connectedness relates to what extent observations are placed in the same cluster as their nearest neighbors in the data space, and is here measured by the connectivity [91]. Compactness assesses cluster homogeneity, usually by looking at the intra-cluster variance, while separation quantifies the degree of separation between clusters (usually by measuring the distance between cluster centroids). Since compactness and separation demonstrate opposing trends (compactness increases with the number of clusters but separation decreases), popular methods combine the two measures into a single score. The Dunn index [92] and silhouette width [63] are both examples of non-linear combinations of the compactness and separation. The details of each measure and a good overview of internal measures in general are presented in Handl et al. [91].

All these approaches validate the quality of clustering, but do not address the statistical significance of clustering solution. The issue of determining the statistical significance of clustering has been poorly studied. What is the probability that a particular clustering solution occurs just by chance? A statistical validation step is needed due to following two issues that arise when clustering the gene expression data. First, correlation and clustering algorithms are biased towards partitions that are in accordance with their own criterion and properties. Secondly, though clustering relies on the existence of a distinct structure within the data, most algorithms return a clustering even in the absence of actual structure and it is the responsibility of the user to detect the lack of significance of the results. It would be misleading if a clustering solution that is non-significant is used for the subsequent biological validation such as Gene set enrichment analysis (GSEA) and pathway analysis. It is very critical to evaluate the significance of the paracliques solutions to make sure that they are not just random clusters, but are strongly driven by the observed gene expression data. Kerr and Churchill [67] applied bootstrapping to assess the stability of results from cluster analysis. However the analysis was based on pre-defined target profiles and stability was evaluated by matching the actual and bootstrap clusters to predefined target profiles. This method will not be applicable when the knowledge of target clustering profiles is not available. The first objective of this study is to develop a general permutation based approach for the assessment of statistical significance of clustering solutions.

As part of the cluster validation process, it is also essential to determine the clusters likely to have the most information. Mantel correlation was originally developed to evaluate spatial and temporal clustering of diseases like leukemia [93]. The Mantel test is an alternative to regressing one set of variables against another. Mantel statistics have been applied with success to correlate gene expression levels with clinical covariates [94]. Mantel correlation can be used to evaluate the information content of a gene cluster based on how well the correlation matrix in the cluster space correlates with that in the original space and significance associated with the Mantel correlation of a cluster can be determined using permutation tests. The second objective of this study was to assess the 'informativeness' of individual paracliques using a permutation test based Mantel correlation approach. These tools will help the biologist to eliminate the non-significant and non-informative clusters before proceeding with biological validation.

3.3 Methods

3.3.1 Dataset

Tian et al. [80] microarray data from human patients with inflammatory myopathy consisted of 7 normal and 8 inclusion body myositis (IBM) samples. The 5000 probe sets used represent the most variable probe sets (by expression value) in all the arrays. This dataset was chosen because of the high variability in the gene expression across the samples.

3.3.2 Randomization strategy

The raw expression data within each gene is randomly permuted 1000 times and paracliques were generated using 1% edge threshold. The total number of paracliques, paraclique size distribution and silhouette are the test statistics that are computed at each permutation run and thus random distribution of these parameters is obtained. P-value is computed as the proportion of values from the random distribution that are as extreme as the observed test statistic.

3.3.3 Silhouette width

Silhouette width has been one of the most widely accepted standards to measure the quality of clustering based on inter- and intra-cluster distances. The average silhouette width of a cluster is the average of each observation's silhouette value within the cluster. The silhouette value measures the degree of confidence in the clustering assignment of a particular observation, with well-clustered observations having values near 1 and poorly clustered observations having values near -1. For the ith observation, it is defined as

$$\mathrm{S(i)} = \frac{b(i) - a(i)}{\max\,(b(i), a(i)},$$

where a_i is the average distance between i and all other observations in the same cluster, and b_i is the average distance between i and the observations in the nearest neighboring cluster.

Silhouette width which has been the most widely used metric to measure the internal quality of clustering is used as the test statistic in the permutation tests for assessment of quality of paracliques. The permutation procedure uses the permutation distribution of average silhouette width to determine whether a paraclique structure has a nonrandom distribution.

3.3.4 Mantel correlation

The Mantel test is used to evaluate the congruence between two distance matrices of the same dimensions. The two matrices must have the same set of sample units in the same order. Mantel correlation seeks linear relationships between two matrices. Because the cells of distance matrices are not independent of each other, the p-values from standard techniques that assume independence of the observations are not acceptable. A standardized Mantel statistic (r) is calculated as the Pearson correlation coefficient between the two matrices.

Let D_G and D_X be the sample distance matrices calculated using the gene expression data from the full dataset and the subset dataset corresponding to each paraclique respectively. The Mantel correlation is calculated on the (i, j) elements of the two distance matrices using the Mantel correlation statistic:

$$\rho(D^G, D^X) = \frac{\Sigma_{i<j}\left(d^G_{i,j} - \overline{d^G}\right)\left(d^X_{i,j} - \overline{d^X}\right)}{\sqrt{\Sigma_{i<j}\left(d^G_{i,j} - \overline{d^G}\right)^2}\sqrt{\Sigma_{i<j}\left(d^X_{i,j} - \overline{d^X}\right)^2}} \ ,$$

where d^Gi,j and d^Xi,j are the distances between samples (i, j) measured on the gene expressions from the full and paraclique subset data respectively, and $\overline{d^G}$ and $\overline{d^X}$ are the average of the distances for all pairs (i, j) in the distance matrices calculated for the full and paraclique subset data respectively.

After the paraclique method partitioned the gene space into k non-overlapping clusters, the Mantel correlation was used to assess the significance of individual paracliques. First, two types of sample correlation matrices are computed, one based on the original dataset containing all the genes and the others based on the genes from each paraclique. The correlation matrices are then converted to distance matrices by subtracting the correlation values from 1. This results in two types of dissimilarity matrices, one based on the original data D-full, and one for each resultant cluster, D-subset (k). The two dissimilarity matrices are then correlated using the Mantel correlation statistic described before. A high cluster Mantel correlation indicates that the cluster captures most of sample correlation structure in the dataset. The Mantel correlation is a measure of proportion of sample covariance captured by the cluster.

In order to destroy the distance dependent nature of D-full and to obtain an empirical null distribution of distance independence, a permutation test is done. The significance of the correlation between matrices was tested by evaluating results from repeated randomization. Strong correlation structure between matrices will rarely be preserved or enhanced if one matrix is shuffled. Specifically, the significance level provides the criterion value (p-value) at which a given paraclique is considered significant or non-significant. A test statistic, the standardized Mantel statistic (r), was calculated for each run. A p-value is calculated from the number of randomizations that yield a test statistic equal to or more extreme than the observed value.

3.3.5 Software

Paraclique analysis was done using a C software package developed by M.A Langston's group at the University of Tennessee. This software employed principles of fixed parameter tractability [84] to find vertex covers [85] which were then used to extract cliques. Permutation tests of the paraclique features were done using a custom scripts written in R programming language. Mantel correlations of the clusters were computed using the bioconductor package *MantelCorr.*

3.4 Results

3.4.1 Permutation test results for Co-expression network features

In this study, a permutation based approach is used for the assessment of statistical significance of paracliques. Based on 1000 permutations of the expression values within each gene, the random distributions of attributes of paracliques such as number of paracliques, mean paraclique sizes and average silhouette widths are obtained which are then used to compute a permutation p-value.

(i) **Number of Paracliques**: It can be seen from Figure 11 that the distribution of number of paracliques is approximately normal with a mean of about 55. The number of paracliques using the actual dataset is 39 which yields a one-sided p-value of 0.044 based on the random distribution. Since the p-value is below the significance level of 0.05, the number of observed paracliques of 39 is much higher than observed at random and is not likely to be obtained by chance. It is interesting to note that the number of paracliques observed is to the left tail of random distribution of paracliques. So whenever a large number of paracliques are obtained, it is important to determine its significance to make sure that they do occur by chance.

(ii) Mean Paraclique sizes: From Figure 12, it is evident that most of the random paraclique sets had a mean paraclique size of 15. The mean paraclique size for the actual dataset is 22 and is associated with a p-value of 0 based on the random distribution. Hence the observed mean paraclique size of 22 is highly significant and not likely to be obtained by chance. These results showed that large number of small-sized paracliques is likely to be observed at random.

(iii) Quality of clusters (Silhouette): Next the statistical significance of clustering quality was evaluated using the random distribution of silhouette. Is the clustering quality that is observed using the dataset likely seen to be at random? Shown in Figure 13 is the random distribution of average Silhouette width based on 1000 permutations. For the actual dataset, the observed silhouette was 0.60 and the associated permutation p-value is 0 which implies that the clustering quality observed from the paraclique analysis is not likely to be obtained at random.

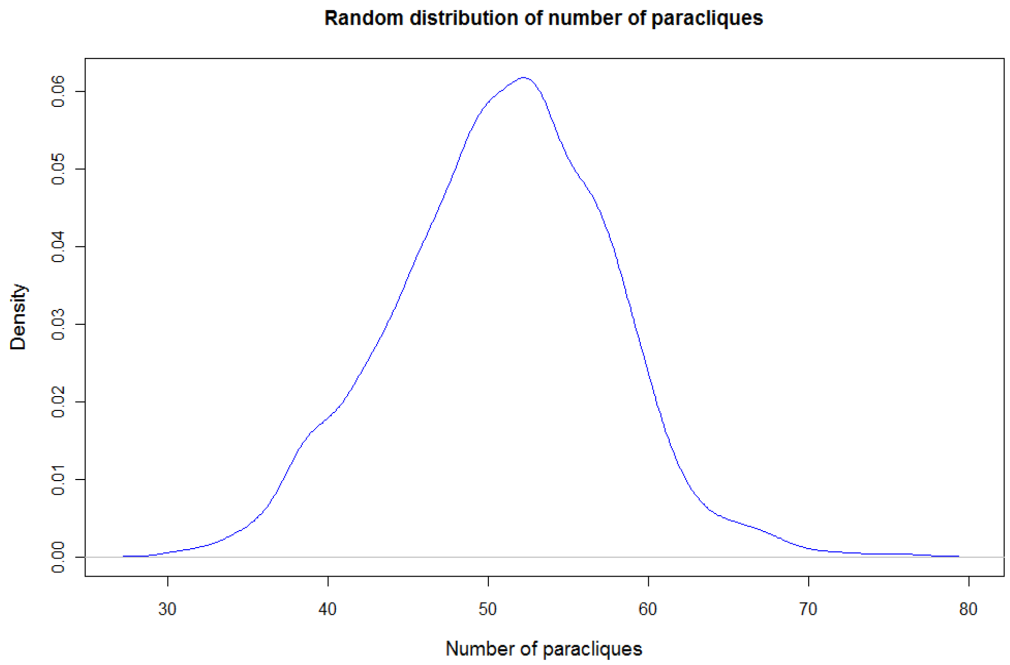


Figure 10. Density plot of number of paracliques using 1000 random permutations

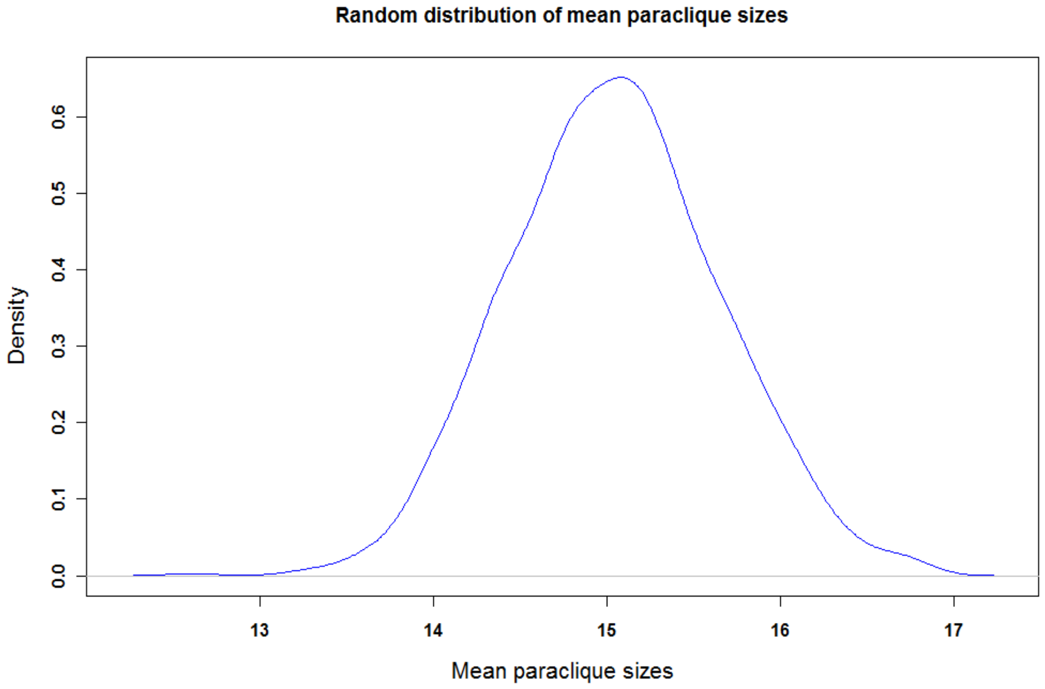


Figure 11. Random distribution of mean paraclique sizes using 1000 permutations.

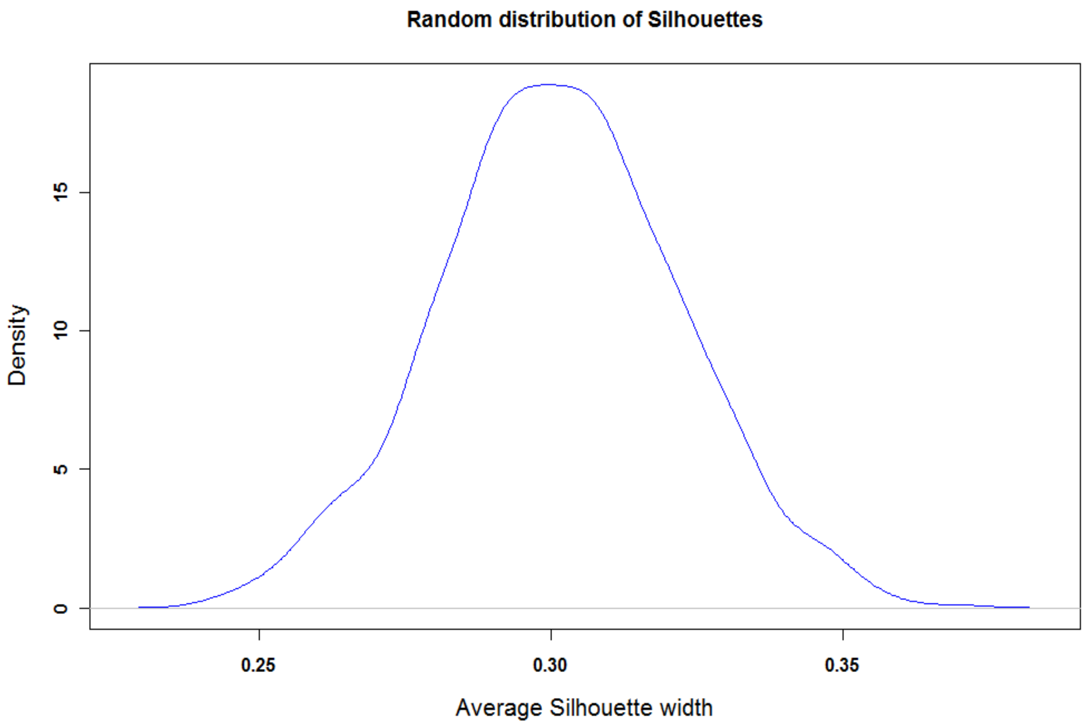


Figure 12. Random distribution of average silhouette width using 1000 permutations

3.4.2 Cluster significance using Mantel correlation

Mantel correlation was used to evaluate the significance of paracliques. Based on 1000 permutations, 18 out of 39 paracliques had significant Mantel correlation as listed in Table 8. Mantel correlations associated with significant and non-significant paracliques were further correlated with the corresponding GSEA enrichment p-value of the most significant category for the corresponding paracliques. The significant paracliques showed a correlation of -0.64 as shown in Figure 14 whereas the non-significant ones showed a correlation of -0.25 which implies that the clusters with high Mantel correlation are more likely to belong to a biological grouping than those with non-significant correlation.

3.5 Discussion

The goal of permutation tests in this study is to make statistical inference about the clustering solution obtained using a particular clustering algorithm on a given dataset. The ‘‘stability” of a clustering structure evaluated by the comparison of the silhouettes of random and the actual clustering solutions is a reasonable first approximation to the confidence of the clustering quality. The significance of the co-expression network features gives more confidence to the results at the level of co-expression network obtained using a particular threshold. Threshold is a crucial parameter in the paraclique algorithm that affects the structure of the clustering solution. Higher threshold lowers the number of edges and thereby the number of paracliques and vice-versa.

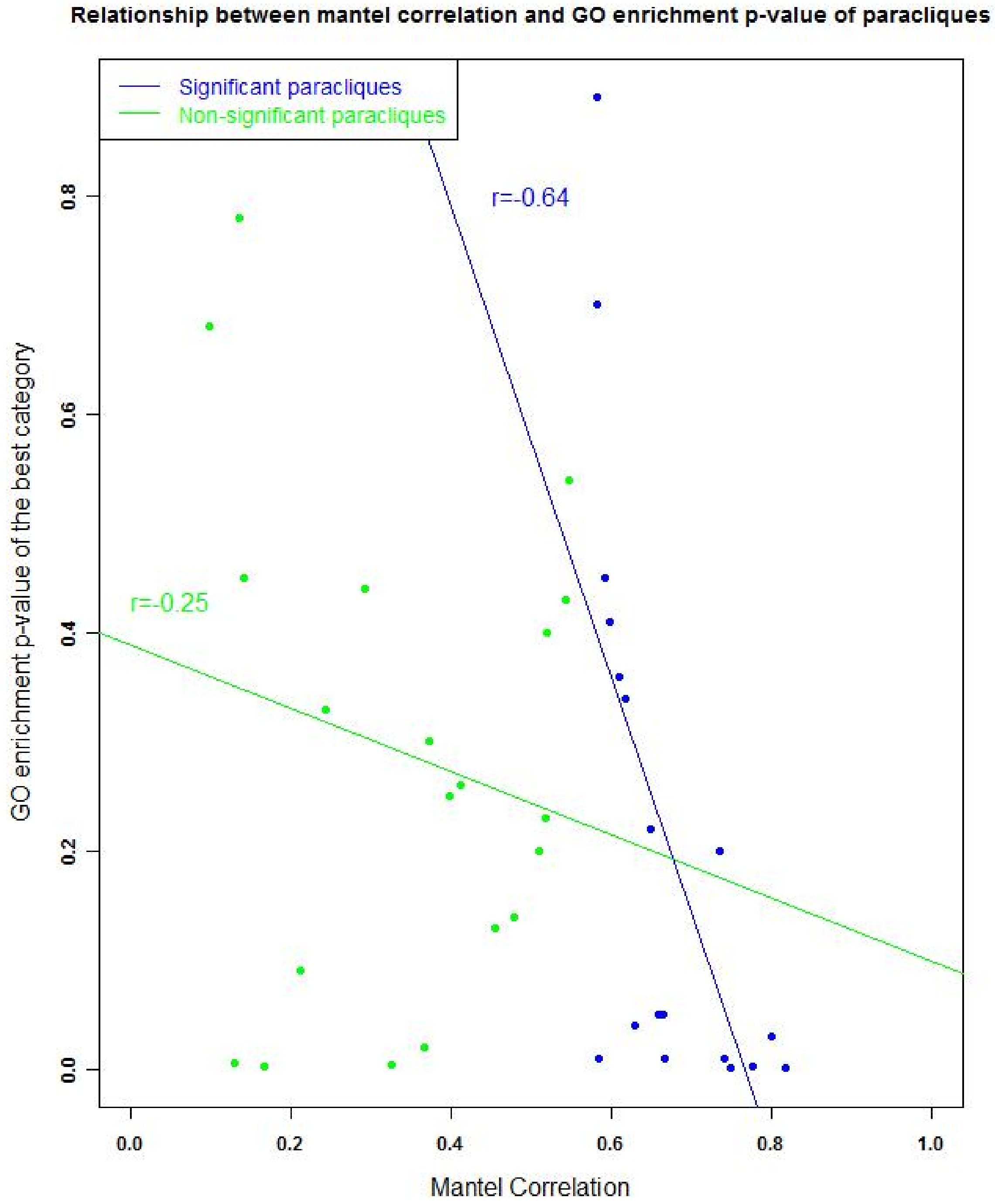


Figure 13. Relationship between Mantel correlation and the GO enrichment p-value of the most significant category. (Tian et al. [80] data)

Table 8. Paracliques with significant Mantel correlation (Tian et al. [80] data)

Paraclique ID	**Mantel correlation**	**Size**
1	0.74	75
2	0.74	70
4	0.59	53
5	0.63	41
6	0.62	33
7	0.75	28
11	0.58	26
12	0.82	21
13	0.78	22
15	0.58	22
16	0.60	20
17	0.67	19
18	0.66	19
19	0.80	15
20	0.67	16
36	0.65	11
37	0.58	10
38	0.61	11

Table 9. Paracliques with non-significant Mantel correlation (Tian et al. [80] data)

Paraclique ID	Mantel correlation	Size
3	0.54	66
8	0.51	26
9	0.13	24
10	0.24	23
14	0.10	22
21	0.33	15
22	0.55	16
23	0.48	16
24	0.17	14
25	0.52	15
26	0.37	13
27	0.40	13
28	0.41	13
29	0.14	12
30	0.29	12
31	0.52	11
32	0.21	12
33	0.37	12
34	-0.25	12
35	0.46	10
39	0.14	11

The current study is based on the paraclique analysis using 1% edge threshold and it still needs to be investigated how different choices of threshold affects the significance and Mantel correlation of paracliques.

There are several ways in which permutation tests can be applied. We used a naïve and straightforward permutation approach by randomizing the expression values within each gene independently and then running the co-expression analysis using the paraclique algorithm. More sophisticated permutation approaches use a reference distribution or a model from which random datasets are generated which are then compared to the original data through some statistic, or by seeking repeated occurrences of same elements in a cluster. The simplest method, for instance, may be to sample from a uniform distribution for each variable, from the range of that variable found in the original data. A more sophisticated but computationally intensive method is to sample uniformly from the convex hull computed from the data. Advantages of using such uniform reference distributions are not clear, however, particularly in high dimensional situations. Other null distributions include randomizing the dissimilarity matrix [95] and adding normally distributed errors to the data [66, 68, 96]. Perturbing the data with noise can be reasonable when one has a good idea of errors associated with each variable. For gene expression, however, the quantities that are needed are gene-specific variances, which cannot be obtained except in relatively large studies with enough replicates. Hence we adopt a within-gene permutation approach in this study that accounts for the gene-specific variances.

Mantel statistics can become an important post-processing aid to the clustering of gene expression data. However, it remains to investigate the statistical properties of Mantel statistics and modeling approaches for analyzing gene chip data. Standard statistics that can be estimated using pairwise distances (e.g., Pearson correlation) are used for calculating Mantel correlation. However, many other models (e.g., nonlinear, multivariate regression with interactions) can be fit using pairwise distances, and need to be investigated as better fitting models to gene expression data. This will require appropriate diagnostics such as goodness-of-fit statistics and graphical analyses (e.g., scatter plot of pairwise distances to assess appropriateness of the Pearson correlation).

Due to the many sources of noise and the high dimensionality of the data, the above statistical validation techniques on their own may often be insufficient in biological data analysis. Frequently, the most obvious cluster structure in the data may be artifacts due to experimental factors. The artifacts will ultimately have to be removed if a researcher is interested in biologically meaningful results. Towards this goal, external validation measures can be applied to assess the degree of agreement with prior biological knowledge. This information can also provide additional feedback on the quality of the data and of previous pre-processing steps. However finding a golden standard for a biological validation is a difficult task. A good final clustering solution will ideally combine validity under both internal and external measures and exhibit a distinct underlying cluster structure revealed by statistical validation while being consistent with prior biological knowledge.

CHAPTER 4 STATISTICAL ANALYSIS OF TIME COURSE DATA USING CONTRAST ANALYSIS TO REVEAL CO-EXPRESSION NETWORK SIGNATURES

4.1 Abstract

In gene co-expression networks, the pattern determined by timing of significant changes in the expression level of each gene may be the most critical information in developmental time course expression profiles. In this study, applied linear modeling approach called planned linear contrasts was implemented to analyze time-course microarray data from developing mouse cerebellum. Helmert contrast analysis identified 7644 and 9336 genes in DBA/2J and C57BL/6 strains respectively with significant changes in expression in a microarray study of early cerebellum development. Polynomial contrast analysis identified 13066 and 14982 genes in DBA/2J and C57BL/6 strains respectively with significant changes in expression. A contingency table analysis was then used to identify genes that are differentially patterned in the two strains of mice. This step yielded 2015 and 5758 differentially patterned genes from Helmert and polynomial contrast analyses respectively. Criteria such as a fold change cut-off and low expression cutoff were further used for filtering the genes and identified 28 and 200 genes from Helmert and polynomial contrast analyses respectively that are differentially patterned genes across strains with large changes in expression over time. The validity of the resulting gene sets was demonstrated by biological enrichment using Gene ontology

and pathway databases which identified several key genes involved in the brain developmental process. Finally, these contrast patterns were used as a means of labeling clusters of co-expressed genes.

4.2 Introduction

Microarray time course experiments typically involve gene expression measurements of genes over relatively few time points under one or more biological conditions. The time points at which the RNA samples are taken are usually determined by the investigator's judgment concerning the biological events of interest and are therefore frequently irregularly placed although for many other time course experiments, equally spaced times are conventional. A major advantage of time course microarray studies is that they give us the ability to monitor temporal behavior of a biological process of interest through the expression levels of thousands of genes simultaneously. Hence this can be a very good experimental design for identifying patterns of gene expression across the different units of interest.

Time course experiments fall in to three main categories: periodic, developmental and time-to-event types. Periodic time courses include natural biological processes whose temporal profiles follow regular patterns. Cell cycle [78] and circadian rhythms [97] are examples and the genes in these processes are expected to have periodic expression patterns. In the developmental time course experiments, gene expression levels are measured at successive times during a developing process. In these cases, there are usually few prior expectations concerning the form of temporal profiles. A third type (which developmental may be a class of) is time locked to an event, e.g. injury or drug injection. This is a common design, and is challenging because the experiment will occur with circadian and other effects confounding it although replication gives some randomization with respect to these other events. Another unique issue in these designs is that these are not repeated measures, but rather time point sampling from independent individuals.

The most critical information in time course expression profiles is the timing of the changes in expression level for each gene, and secondarily is the general shape of its expression pattern. In addition, different genes will be activated or inactivated at each level of a gene network. Therefore it may not be reasonable to expect that the expression levels of those co-expressed genes will go up and down concordantly all the way through the entire sampling period. With the same timing of initial change, genes which share similar pattern of expression for any number of sampling intervals from the beginning might be considered co-expressed at certain levels in the gene network.

A simple but powerful tool for extracting temporal patterns is found in contrasts: linear combinations of gene expression over time. Contrast analysis methods are a general linear model technique generally suitable for time-course experiments based on the most widely used kinds of microarray platforms including one-color and two-color arrays in order to identify genes associated with temporal differences between groups, i.e., the point(s) in time in which the groups show big differences [98]. An example of the use of contrasts can be seen in Lonnstedt et al. [99] where samples were taken from cells at 0.5, 1, 4 and 24 h after stimulation with a growth factor and contrast patterns were used to categorize genes into late and early responders. Smyth et al. [100] used contrasts in the univariate linear model setting and used F-statistic for testing whether there is any change in gene expression levels over time. This approach assumes that the samples are independent and so would be appropriate for cross-sectional data. Li et al. [98] applied linear planned contrast analysis to categorize the genes with specific expression patterns. However statistical methods to analyze these temporal patterns across multiple biological conditions have not yet been reported.

In this study the focus is on the statistical analysis of microarray time course data using Helmert contrast analysis and polynomial regression with a focus on developmental time course experiments. Two different strategies based on Helmert contrast analysis and polynomial regression followed by a contingency table analysis were used for the differential profiling of genes across multiple biological conditions. Both these approaches take into consideration the temporal order in the data. Helmert contrast approach focuses on the timing of a gene's initial response and the regression approach is useful to look at the general shapes of gene expression patterns along the subsequent sampling time points. These methods are particularly suitable for analysis of microarray experiments in which it is often difficult to take sufficiently frequent measurements and/or the sampling intervals are non-uniform. These methods were implemented on the microarray data from mouse cerebellum at eleven different time points from embryonic and postnatal stages and different biological conditions (DBA/2J and C57BL/6 strains of mice). These methods were performed on each strain dataset independently. A contingency analysis based approach is then used to identify the genes that are differentially profiled or "patterned" across the two conditions. Though a typical ANOVA analysis looking at significant interaction effects between condition and time helps to identify the genes that show different temporal effects in both the conditions, it does not allow us to characterize the differences in specific patterns as in contingency table analysis of contrast patterns which enables us to look at the differences in the shapes of overall time courses across the conditions.

Systems approaches to developmental biology and genetics often describe complex relationships using networks. A co-expression network consists of a set of nodes representing genes and a set of edges that connect those nodes defined by co-expression between genes. This network is then used to extract clusters of co-expressed genes using a graph theoretic approach such as clique or paracliques [23, 101]. In the context of developmental time course microarray data from cerebellum, paracliques in graphs constructed from time series data have the property that most genes in the paraclique are very highly correlated across time with most other genes in the paraclique, which suggests coregulation over time in the developing cerebellum. An approach for deriving time profile "signature" of each paraclique, by labeling the paraclique with the contrast design associated with most of the genes in it, is then presented.

4.3 Methods

4.3.1 Data Preprocessing

Illumina chip raw data files were preprocessed using BeadStudio software (Illumina Systems, San Diego, CA). The rank invariant normalization [102] without background subtraction was used to normalize the data. This method was chosen based on a comparison analysis of various normalization methods in which the rank invariant normalization yielded the highest signal to noise ratio based on intraclass correlation analysis in large multi-group designs. Quality control analysis was performed on the arrays using arrayQualityMetrics—a bioconductor package for quality assessment of microarray data [103] and the data from all the arrays was retained based on the analysis.

4.3.2 Analysis of variance

For each transcript, a two-way (11 X 2) general linear model was fit using factors: age, strain and their interaction. The first factor has 11 levels starting from E12 to P9 time points. The second factor has two levels: DBA/2J and C57BL/6 strains of mice. P-values are calculated for the main and interaction effects. This analysis is useful in identifying genes for which there is a significant strain differences and interaction between development stage and strain effect indicating that the strain alters the time course of gene expression.

4.3.3 Development classifications using Helmert contrasts

Post-hoc contrast analysis is widely used for small time series experiments (those in which a few time points were sampled). A set of orthogonal contrast vectors is applied to the data matrix to test specific hypotheses regarding the pattern of group differences. Our goal is to characterize the time patterns so the Helmert contrasts was identified which test for changes across time by comparing expression at each time point to all preceding time points. Table 10 shows the contrast vectors used for generating the 10 Helmert designs. The designs have been labeled as “D-X” where X represented the time point which is compared to average of the preceding time points. They measure the rate of change in expression between the time point X and all the preceding time points.

4.3.4 Development classification using polynomial regression

A step-down polynomial (cubic) regression model was used for characterization of genes based on the overall shapes of the expression profiles. The first step is to fit the following quadratic regression model to the each gene:

$$Y_{ij} = \beta 0_j + \beta 1_j * x + \beta 2_j * x^2 + \beta 3_j * x^3 + \varepsilon_{ij}$$

where Y_{ij} denotes the expression of the jth gene at the ith replication, x denotes time, $\beta 0_j$ is the mean expression of the jth gene at $x = 0$, $\beta 1_j$ is the linear effect parameter of the jth gene, $\beta 2_j$ is the quadratic effect parameter of the jth gene, and, ε_{ij} is the random error associated with the expression of the jth gene at the ith replication and is assumed to be independently distributed normal with mean 0 and variance.

Table 10. Helmert contrast coefficient matrix.

Time/Design	D-E13	D-E14	D-E15	D-E16	D-E17	D-E18	D-P0	D-P3	D-P6	D-P9
E12	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
E13	1	-1	-1	-1	-1	-1	-1	-1	-1	-1
E14		2	-1	-1	-1	-1	-1	-1	-1	-1
E15			3	-1	-1	-1	-1	-1	-1	-1
E16				4	-1	-1	-1	-1	-1	-1
E17					5	-1	-1	-1	-1	-1
E18						6	-1	-1	-1	-1
P0							7	-1	-1	-1
P3								8	-1	-1
P6									9	-1
P9										10

"D-X" represents a Helmert pattern where X is the time point which is compared to average of the preceding time points.

If overall model p-value >α_0, the jth gene is considered to have no significant differential expression over time. The expression pattern of the gene is "flat". If overall model p-value ≤ α_0, the jth gene will be considered to have significant differential expression over time. The patterns are then determined based on the p-values obtained from F tests. All the p-values have been adjusted for False discovery rate (FDR) using the Benjamini-Hochberg (BH) algorithm.

If only the p-value of linear effect is ≤0.05 and p-values of quadratic and cubic effects >0.05, the j^{th} gene is considered to be significant in linear term and is uniquely characterized by a “linear” pattern. If p-value of quadratic effect ≤0.05 and p-value of linear and cubic effects *>0.05*, the jth gene is considered to be significant only in the quadratic term. The expression pattern of the gene is uniquely "quadratic". If p-value of cubic effect ≤0.05 and p-value of linear and cubic effects *>0.05*, the jth gene is considered to be significant only in the quadratic term. The expression pattern of the gene is then uniquely "cubic”.

4.3.5 Contingency analysis of contrast patterns

Association between contrast designs and strain was evaluated using standard contingency table analysis.

4.3.6 Literature-based Gene set enrichment

GCAT is a web-based tool (Ramin Homayouni, University of Memphis) that lets the researchers evaluate the cohesion of sets of genes according to information derived from PUBMED literature (http://motif.memphis.edu/gcat). It determines the functional coherence of gene sets by performing latent semantic analysis of Medline abstracts. It generates an enrichment p-value for the geneset using a fisher's exact test. GCAT currently holds pair-wise literature correlation information for the mouse and human genes.

4.3.7 Software

Contrast and contingency analyses were done using custom scripts written in R programming language [82]. The linear model function "lm" was used for model fitting for contrast analysis in R. Bowker's test of agreement was performed using the JMP 7 software (SAS Institute). Paraclique analysis was done using a C software package developed by M.A Langston's group at the University of Tennessee. This software employed principles of fixed parameter tractability [84] to find vertex covers [85] which were then used to extract cliques.

4.3.8 Outline of the analysis

A brief overview of the analysis of the time course data is as follows:

(1) Two different strategies, Helmert contrast analysis and Polynomial regression were applied to the gene expression data from DBA/2J and C57BL/6 strains

independently as described in the sections 4.3.3 and 4.3.4. Genes with significant fit to only a single pattern (unique significant p-value) were assigned corresponding patterns. Genes with significant fit to multiple patterns were not considered.

(2) Contingency table analysis of contrast patterns was applied to the common significant genes from DBA/2J and C57BL/6 strains (from step 1).

(3) Biological validation using GO enrichment and GCAT was then performed on the genesets from the diagonal and off-diagonal cells in the contingency table.

(4) Genes with significant contrast designs were then filtered for fold change greater than 2 and mean expression level greater than 8 in order to identify the genes that are differential patterned across strains with large changes in expression over time.

4.4 Results

4.4.1 Helmert contrast analysis

Helmert analysis was performed on the data from DBA/2J and C57BL/6 separately. The genes are then binned in to 10 classes corresponding to the 10 Helmert designs. Figure 15 shows the histogram of genes with specific Helmert design patterns. Clearly the shapes of distributions of designs in both the strains are different. There are many genes (49%) that have a significant initial spike at E18 and P0 in DBA/2J whereas in C57BL/6, E15 change seems to characterize many genes (32%). This could also imply that most of the genes in DBA/2J are late responders as compared to C57BL/6 in which there are a large number of genes that have an initial spike at early embryonic time points.

Contingency table analysis was performed using the design information of the 2105 genes that had a unique significant Helmert fit in both the strains. In the contingency table (Table 11), diagonal cells consisting of 436 genes represent the gene sets that have the same developmental Helmert patterns across both the strains. The off-diagonals consisting of 1579 genes correspond to the genesets with shifts in initial responses between the strains. Extreme off-diagonal cells representing genes with huge shifts in intial responses between the strains are very sparsely populated. Bowker’s test [104] is used to test for the differences in the proportions of Helmert designs across both the strains for the same set of genes. Bowker's test is a generalization of McNemar's test [105] which is in general used to test the hypothesis of symmetry. The test yielded a p-value of <0.0001 which clearly indicates that the DBA/2J and C57BL/6 strains differ in their design categories of the genes with significant Helmert fit and thus are characterized by different initial gene responses. This conclusion is further supported by *kappa* statistic value of 0.14 that indicates a very low level of agreement in the design profiles of the genes in DBA/2J and C57BL/6.

The scatterplot in Figure 16 shows the major transition points where a lot of genes are different in the timing of initial responses in expression. For instance, there is a smear at the region corresponding to E18 in DBA/2J and E15 in C57BL/6 which implies that the initial significant response for the corresponding genes is at E15 in C57BL/6 but is delayed till E18 in the case of DBA/2J.

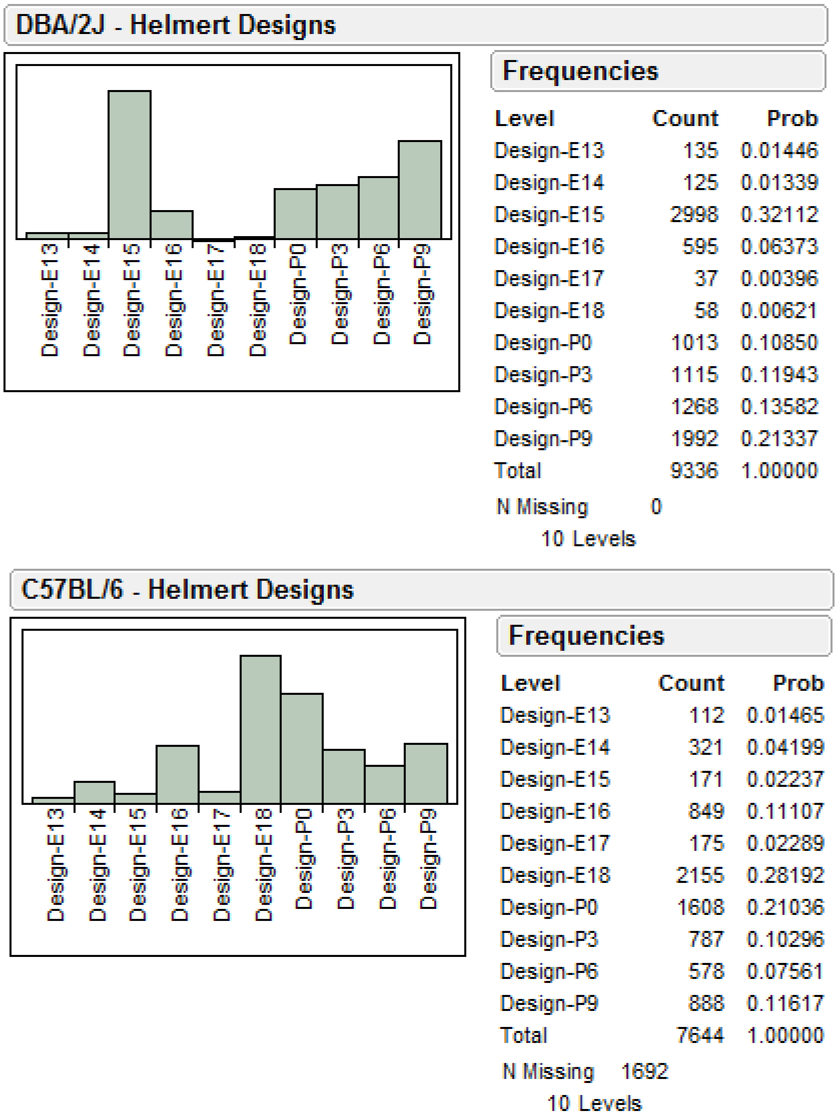
For each gene, an ANOVA F test was then performed with Strain in the model, and the corresponding P-value was obtained. To consider the genes that are significantly different across the strains, the genes that were differentially expressed across strains are then retained in the contingency table which is displayed in Table 12.

The literature-based geneset enrichment tool GCAT was used to validate the 131 genesets which showed pattern change from E18-design in DBA/2J to E15-design in C57BL/6. The low literature association p-value indicates that these genes that are differentially patterned in the two strains are highly related and networked to each other as evident from the literature (Figure 17).

GO analysis of selected cells.

Genesets corresponding to the diagonal cells in the contingency table were enriched for categories such as metabolic process, cell motility, apoptosis, cell proliferation and cell differentiation. The genes in these categories are expected to be the housekeeping genes which are necessary for cerebellum development. Off-diagonal cells correspond to the genes which are differentially patterned. Most of the off-diagonal cells are sparsely populated indicating that there are only a few genes which have a big time shift in the initial response. There are several genes in the category corresponding to E15 in C57BL/6 and E18 in DBA/2J which were enriched for development, particularly in embryonic development. Some of the notable genes are AATF, SBDS, POFUT1, FOXI1 and MYST3. It was found in the previous studies that Protein o-fucosyltransferase 1 (POTUF1) plays a crucial role in Notch signaling pathway and a striking effect of the *Pofut1* mutation was the marked up-regulation of several Notch pathway genes in neural tube and brain [106]. Apoptosis antagonizing transcription factor (AATF) is another essential gene in embryonic development which functions as a general inhibitor of the histone deacetylase HDAC1, leading to the activation of E2F target genes and cell cycle progression [107]. Figure 18A shows the time course profile of AATF in the two strains. Synaptic proteins, such as SNAP-25, are considered to form a core complex that coordinates vesicle docking and fusion for neurotransmitter release [108]. This gene belonged to one of the off-diagonal cells and had a first significant initial response at P3 in DBA/2J where as it had a negative response at E15 in C57BL/6 as shown in Figure 18B.

Thus many genes that were profiled based on Helmert contrast pattern differences in DBA/2J and C57BL/6 were shown to be involved in biological processes during early cerebellum development such as cell proliferation, apoptosis, synaptogenesis and developmental pathways such as Notch signaling pathway.



```
DBA/2J - Helmert Designs

Frequencies
Level        Count     Prob
Design-E13     135  0.01446
Design-E14     125  0.01339
Design-E15    2998  0.32112
Design-E16     595  0.06373
Design-E17      37  0.00396
Design-E18      58  0.00621
Design-P0     1013  0.10850
Design-P3     1115  0.11943
Design-P6     1268  0.13582
Design-P9     1992  0.21337
Total         9336  1.00000
N Missing        0
     10 Levels

C57BL/6 - Helmert Designs

Frequencies
Level        Count     Prob
Design-E13     112  0.01465
Design-E14     321  0.04199
Design-E15     171  0.02237
Design-E16     849  0.11107
Design-E17     175  0.02289
Design-E18    2155  0.28192
Design-P0     1608  0.21036
Design-P3      787  0.10296
Design-P6      578  0.07561
Design-P9      888  0.11617
Total         7644  1.00000
N Missing     1692
     10 Levels
```

Figure 14. Distribution of Helmert designs of genes with significant fit in DBA/2J and C57BL/6

Table 11. 10X10 Contingency table of Helmert designs with significant fit in DBA/2J and C57BL/6

Count C57BL/6 x DBA/2J	D-E13	D-E14	D-E15	D-E16	D-E17	D-E18	D-P0	D-P3	D-P6	D-P9	
D-E13	1	3	0	0	1	3	4	1	3	1	17
D-E14	0	1	2	5	0	5	3	1	0	0	17
D-E15	10	22	15	77	6	441	72	22	16	34	715
D-E16	2	3	3	30	6	32	31	6	12	15	140
D-E17	0	0	0	1	1	2	0	0	1	0	5
D-E18	1	2	0	2	0	10	2	0	1	1	19
D-P0	2	4	1	21	5	18	114	16	20	40	241
D-P3	4	2	1	32	7	18	17	96	16	28	221
D-P6	3	8	4	18	3	18	110	28	40	23	255
D-P9	6	12	5	52	4	77	58	24	19	128	385
	29	57	31	238	33	624	411	194	128	270	2015

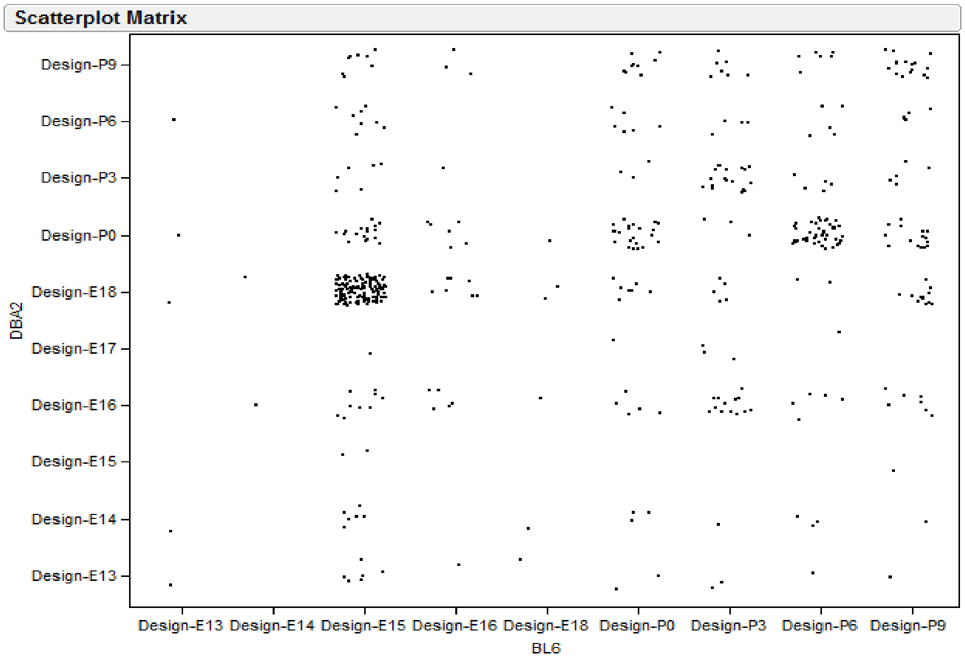


Figure 15. Scatterplot of Helmert designs in DBA/2J and C57BL/6

Table 12. Contingency table of Helmert designs of genes with significant contrast fit and strain differences.

Count C57BL/6 by DBA/2J	D-E13	D-E14	D-E15	D-E16	D-E17	D-E18	D-P0	D-P3	D-P6	D-P9	
D-E13	1	1	0	0	0	1	1	0	1	0	5
D-E14	0	0	0	1	0	1	0	0	0	0	2
D-E15	6	6	2	10	1	131	17	6	8	8	195
D-E16	1	0	0	5	0	7	6	1	0	3	23
D-E18	1	1	0	1	0	2	1	0	0	0	6
D-P0	2	3	0	5	1	7	22	3	6	9	58
D-P3	2	1	0	13	3	5	3	18	4	7	56
D-P6	1	3	0	5	1	2	44	5	5	6	72
D-P9	1	1	1	7	0	12	15	5	5	16	63
	15	16	3	47	6	168	109	38	29	49	480

Literature cohesion p-value = 3.423350e-11

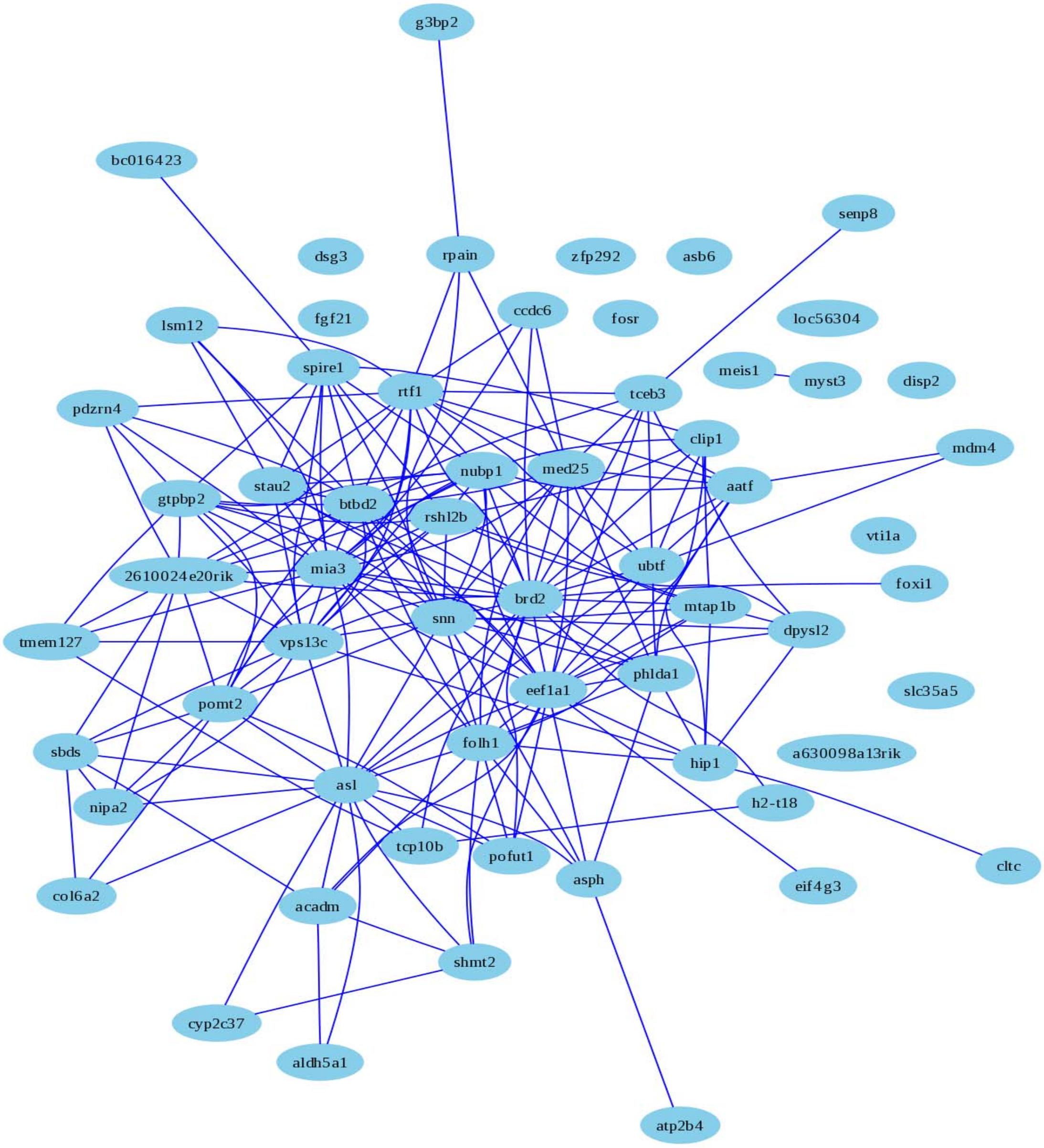


Figure 16. GCAT literature association of the gene cluster that showed change from E18-design in DBA/2J to E15-design in C57BL/6

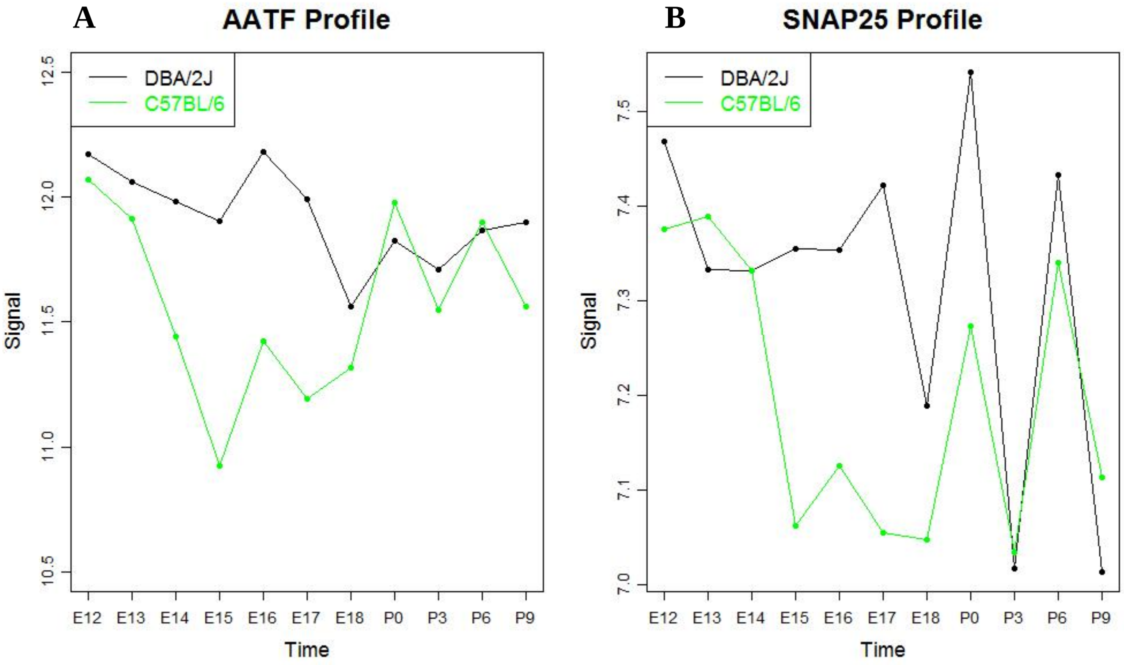


Figure 17. Timecourse profile of AATF (Panel A) and SNAP25 (Panel B) genes in DBA/2J and C57BL/6.

(A) AATF showing E18-design in DBA/2J and E15-design in C57BL/6 strains. (B) SNAP25 showing P3-design in DBA/2J and E15-design in C57BL/6.

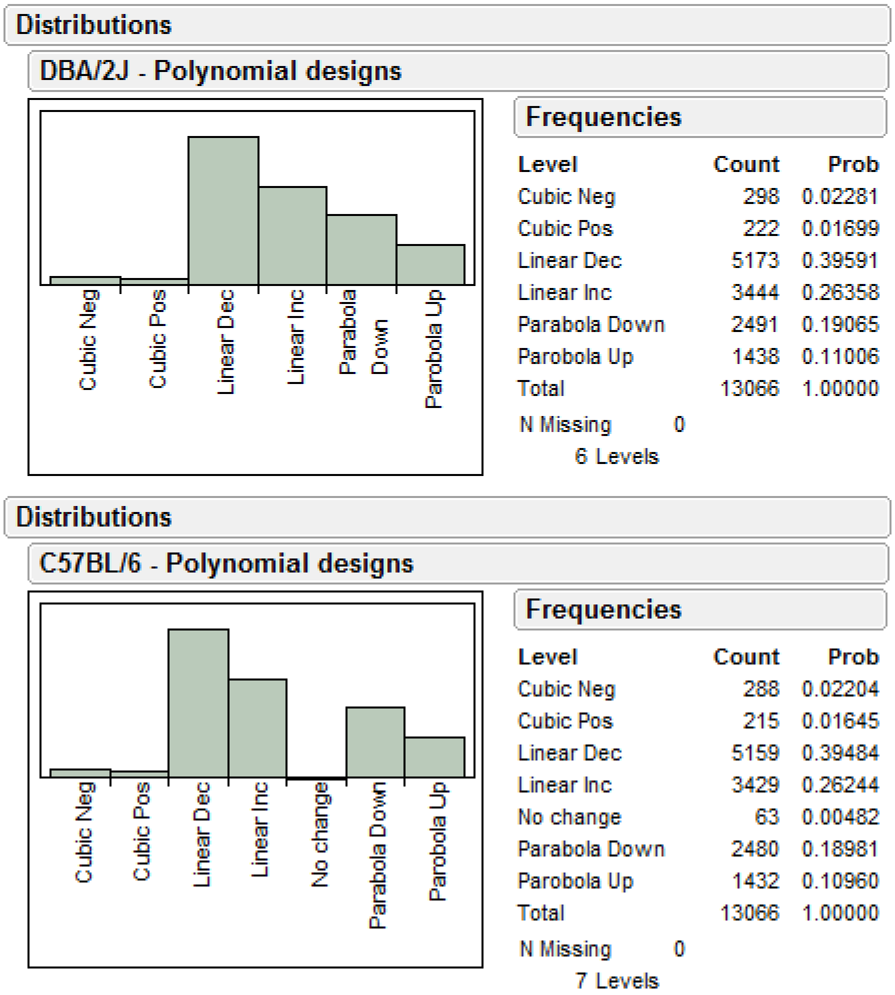
4.4.2 Polynomial contrast analysis

Figure 19 shows the difference in the shapes of distributions of polynomial designs in both the strains. About 56% of the C57BL/6 genes in the dataset are characterized by non linear patterns (parabola up and down), whereas only 30% of DBA/2J genes show non-linearity across the time points. The nonlinear patterns are characterized by increase and decrease of expression levels at certain time points. About 65% of the genes in DBA/2J are characterized by linear patterns. Since many DBA/2J genes are late responders as seen from the Helmert analysis, it is possible that they might have an increase or decrease till the P9 and might change at later time points. This could be one reason for not being able to detect complex non-linear patterns in DBA/2J.

Contingency analysis of the results from the polynomial regression analysis

Contingency table analysis was performed using the polynomial design information of the 5878 genes that had a unique significant polynomial fit in both the strains. The 7x7 contingency table (Table 13) shows the number of genes corresponding to the all combinations of polynomial design categories in DBA/2J and C57BL/6 strains of mice. The kappa statistic of 0.48 indicates good agreement of polynomial designs between the strains.

The diagonals are heavily populated (3654 genes) compared to the off-diagonals (2224 genes). So a major portion of the genes with significant polynomial fit are not changing the overall expression pattern. The highest count in the off-diagonals corresponds to the genes with linear decrease pattern in DBA/2J with upward parabola in C57BL/6.



Distributions

DBA/2J - Polynomial designs

Frequencies

Level	Count	Prob
Cubic Neg	298	0.02281
Cubic Pos	222	0.01699
Linear Dec	5173	0.39591
Linear Inc	3444	0.26358
Parabola Down	2491	0.19065
Parobola Up	1438	0.11006
Total	13066	1.00000

N Missing 0
6 Levels

Distributions

C57BL/6 - Polynomial designs

Frequencies

Level	Count	Prob
Cubic Neg	288	0.02204
Cubic Pos	215	0.01645
Linear Dec	5159	0.39484
Linear Inc	3429	0.26244
No change	63	0.00482
Parabola Down	2480	0.18981
Parobola Up	1432	0.10960
Total	13066	1.00000

N Missing 0
7 Levels

Figure 18. Distribution of polynomial contrast designs of genes with significant fit in DBA/2J and C57BL/6 (Using JMP software)

Table 13. Contingency table of polynomial designs with significant contrast fit.

Count C57BL/6 By DBA/2J	Cubic Neg	Cubic Pos	Linear Dec	Linear Inc	Parabola Down	Parobola Up	
Cubic Neg	64	0	26	40	22	12	164
Cubic Pos	0	27	28	10	3	20	88
Linear Dec	12	18	1304	7	90	76	1507
Linear Inc	12	2	7	998	41	21	1081
Parabola Down	19	17	33	313	707	5	1094
Parobola Up	6	10	1351	18	5	554	1944
	113	74	2749	1386	868	688	5878

They represent genes whose expression levels are decreasing over time in DBA/2J, but in C57BL/6 they are going down till a particular time point after which the expression levels start to increase.

Notable genes in this category are NRG1 and SYN3 which are involved in synapsogenesis. NRG1 is a neuronal signal that promotes the proliferation and survival of the oligodendrocyte, the myelinating cell of the central nervous system [109]. SYN3 belong to the family of Synapsins which are neuron-specific synaptic vesicle-associated phosphoproteins that have been implicated in synaptogenesis and in the modulation of neurotransmitter release [110]. SYN3 is associated with synaptic vesicles, and its expression appears to be neuron-specific and highly expressed in the brain [111]. The difference in the time course profiles of these genes across the strains could have a significant impact on the differences in the developmental phenotypes. Fin15 (fibroblast growth factor inducible 15) belongs to a group of genes that are stimulated by fibroblast growth factors [112]. Expression of FIN15 was characterized a linear decrease in DBA/2J whereas it was found to have linear increase in C57BL/6 (Figure 20). It was found that most of the FIN genes are in involved in cell proliferation and apoptosis [113]. Thus several genes were identified based on polynomial contrast pattern differences in DBA/2J and C57BL/6 to be involved in biological processes during early cerebellum development such as cell proliferation, apoptosis and synaptogenesis.

After considering only the genes that are differentially expressed between strains (p-value < 0.05) in the contingency table (Table 14), the profile agreement between the strains increases slightly with a kappa value of 0.51. It is interesting to note that although the genes are differentially expressed by strain, the overall pattern of the expression remains the same for the majority of the genes.

Gene selection using filtering analysis

The following filters were applied to genesets obtained from DBA/2J and C57BL/6 Helmert and polynomial contrast analysis:

(i) Fold change filter: Maximum Fold change between time points > 2

(ii) Low expression filter: Mean expression level of each gene > 8

Filtering Helmert contrast results

This criterion was first applied to the 2015 genes that have a unique significant Helmert design fit in both the strains. 66 and 51 genes with significant Helmert designs were found to pass both the filtering criteria in DBA/2J and C57BL/6 strains respectively. Comparison of these two gene sets yielded 28 common genes with significant Helmert fit that passed all the filtering criteria in both the strains. These results are represented in a Venn diagram in Figure 21.

Shown in the Figure 22 are few examples of genes from the filtered list (MAGEH1, CREBBP, ZC3H13, MTAP1B) showing E15-design in C57BL/6 and E18-design in DBA/2J. CREBBP, a creb-binding protein, plays a role in transcriptional activation by binding specifically to phosphorylated CREB and enhances its transcriptional activity

Table 14.7x7 Contingency table of polynomial designs of genes with significant contrast fit and strain differences.

Count	Cubic Neg	Cubic Pos	Linear Dec	Linear Inc	Parabola Down	Parobola Up	
Cubic Neg	21	0	13	10	6	3	53
Cubic Pos	0	2	5	1	1	0	9
Linear Dec	3	4	254	4	11	18	294
Linear Inc	3	1	7	233	15	11	270
Parabola Down	2	2	20	64	121	4	213
Parobola Up	3	4	263	10	1	99	380
	32	13	562	322	155	135	1219

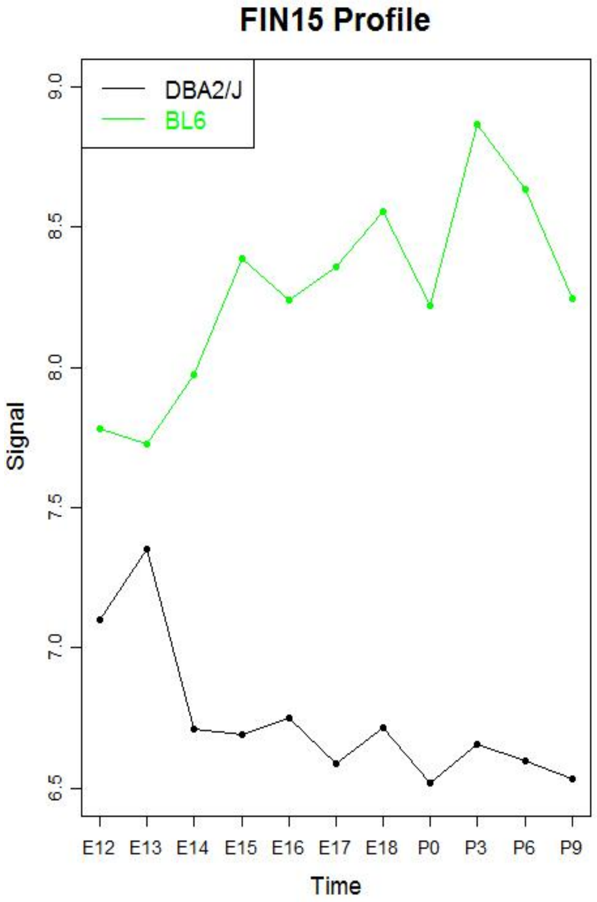


Figure 19. Timecourse profile of FIN15 gene showing linear decrease and linear increase patterns in DBA/2J and C57BL/6 respectively.

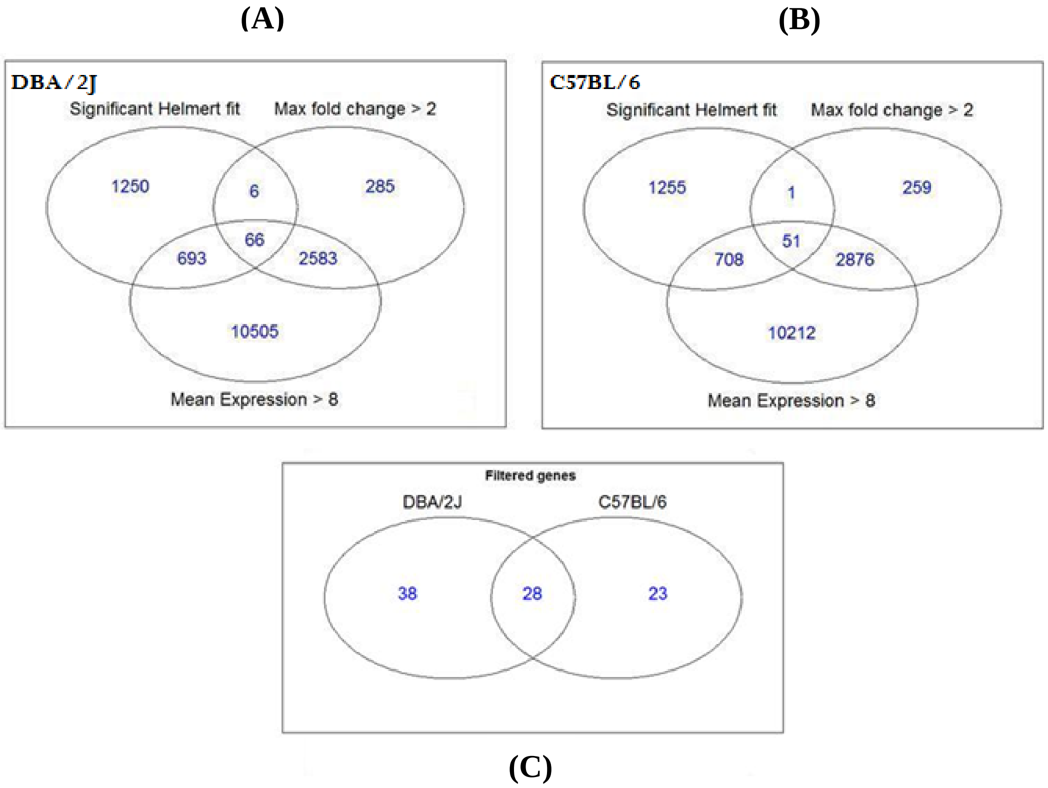


Figure 20. Gene selection using fold change and low expression cutoffs as filters.

(A) and (B) show Venn diagrams representing genesets from different filters in DBA/2J and C57BL/6 strains respectively (c) Venn diagram representing the common filtered genes in DBA/2J and C57BL/6 strains.

toward cAMP-responsive genes [114]. MTAP1B (Microtubule associated protein 1b) was shown to be involved in the cytoskeletal organization that accompany neurite extension [115]. Other notable genes in this category are KIF1B that codes for a Kinesin protein which is a microtubule-dependent motor protein that transports organelles [116] and ACTL6A which is involved in transcriptional activation and repression of select genes by chromatin remodeling [117].

Filtering polynomial contrast results

Genesets obtained from polynomial contrast analysis are also filtered using the same criteria. 303 and 331 genes with significant polynomial designs were found to pass both the filtering criteria satisfied in DBA/2J and C57BL/6 strains respectively. Comparison of these two genesets yielded 200 common genes with significant polynomial fit that passed all the filtering criteria in both the strains. These results are represented in a Venn diagram in Figure 23. Out the 200 genes, 166 were found to have the same designs in both the strains. The other 34 genes were found to be differentially patterned by strain. Many of the genes in this list are microtubule associated and involved in nervous system development. Notable genes in this list are MTAP2, MTAP1B, DBN1 and GNAI1.

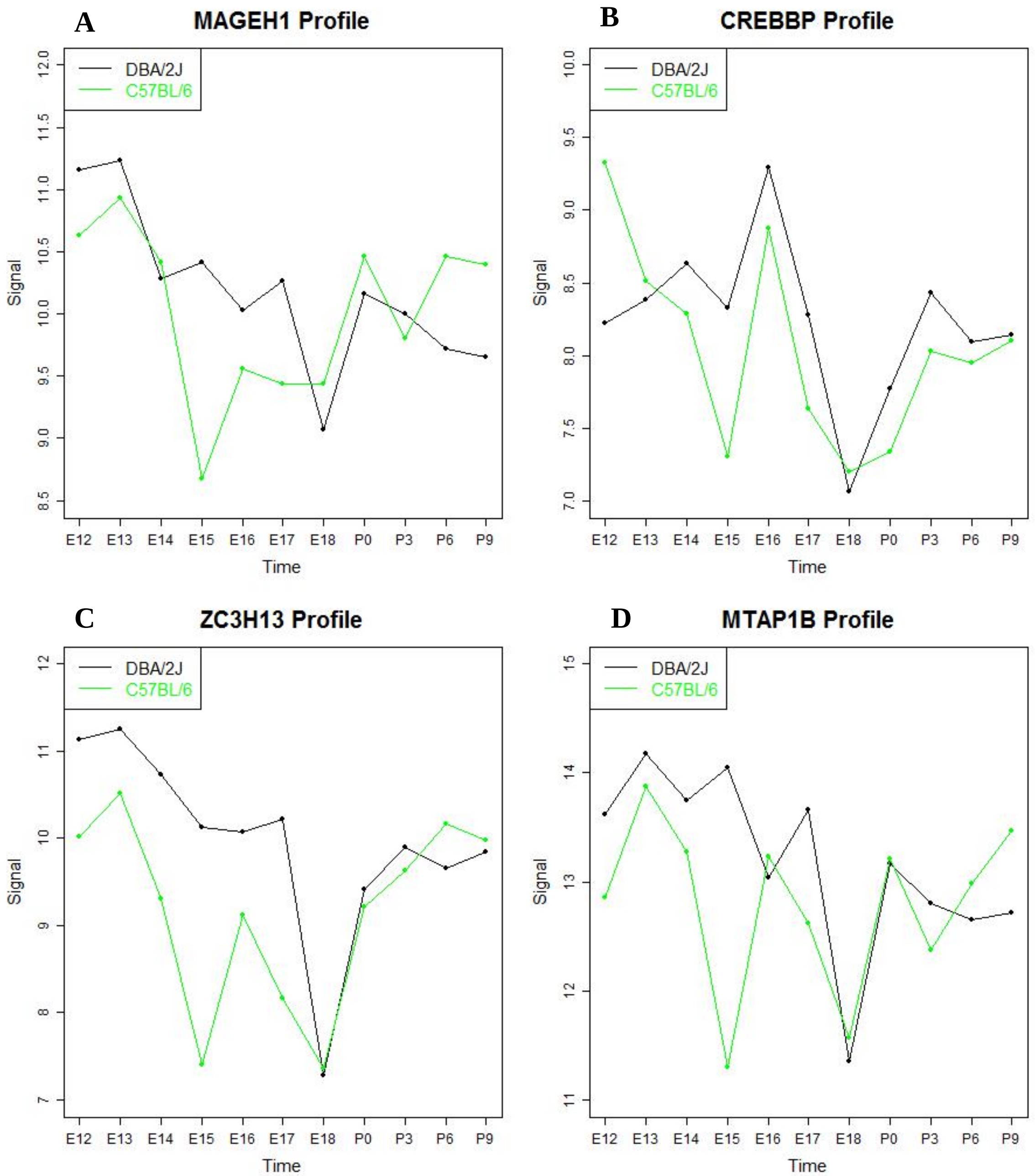


Figure 21. Timecourse profiles of MEGEH1 (Panel A), CREBBP (Panel B), ZC3H13 (Panel C) and MTAP1B (Panel D) genes in DBA/2J and C57BL/6.

All genes show E18-design in DBA/2J and E15-design in C57BL/6.

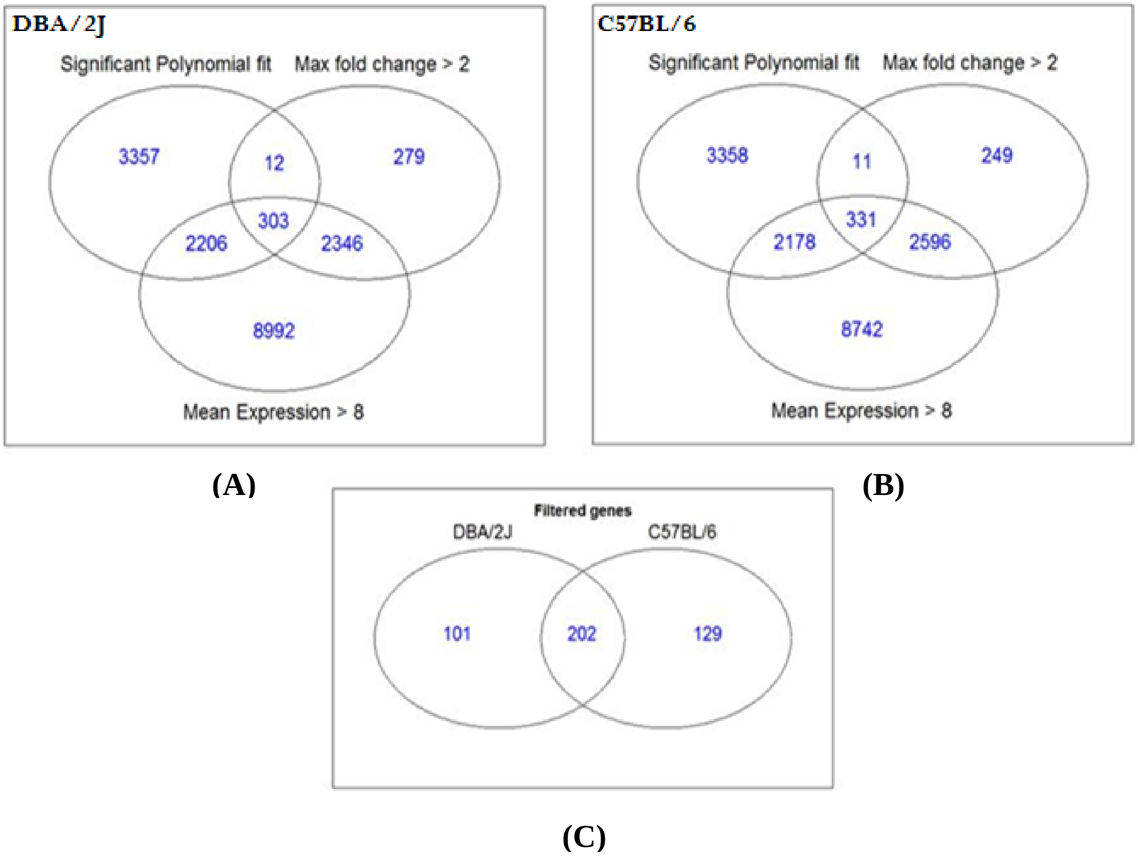


Figure 22. Filtering genesets from polynomial regression using fold change and low expression cutoffs as filters.

(A) and (B) show Venn diagrams representing genesets from different filters in DBA/2J and C57BL/6 strains (c) Venn diagram representing the common filtered genes in DBA/2J and C57BL/6.

GNAI1 (guanine nucleotide binding protein, alpha inhibiting 1), a gene involved in the axon guidance pathway showed cubic pattern in DBA/2J and linear increase in C57BL/6 as shown in Figure 24B. It was shown in a study that loss of GNAI1 amplifies the responsiveness of postsynaptic neurons to stimuli that strengthen synaptic efficacy, thereby diminishing synapse-specific plasticity required for new memory formation [118]. Since this gene shows variation of expression patterns in both the strains, it might be of interest to further investigate if it differentially regulates the memory formation in the two strains. DBN1 (Debrin1) is high expressed in brain and might play some role in cell migration, extension of neuronal processes and plasticity of dendrites [119]. It shows linear decrease pattern in DBA/2J and upward parabolic pattern in C57BL/6 as shown in Figure 24A. MTAP2 and MTAP1B are involved in neuronal migration, dendritic outgrowth, and microtubule organization [120]. Variation of expression patterns of the genes across the strains motivates further investigation in to the differential regulation of the processes controlled by these genes in the strains.

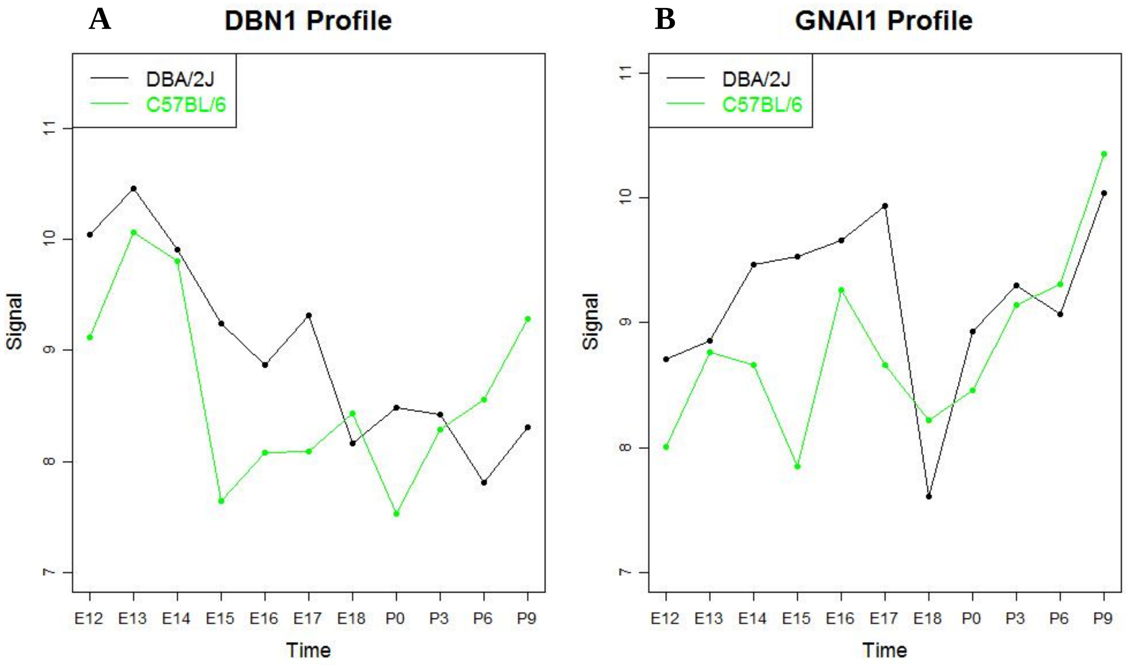


Figure 23. Timecourse profiles of DBN1 (Panel A) and GNAI1 (Panel B) genes in DBA/2J and C57BL/6.

(A) DBN1 shows linear decrease pattern in DBA/2J and upward parabolic pattern in C57BL/6. (B) GNAI1 shows cubic pattern in DBA/2J and linear increase pattern in C57BL/6.

4.4.3 Co-expression network signatures using contrast patterns

It is expected that co-expressed genes in a cluster tend to have the same developmental time course patterns. This would enable us to label the paracliques using the associated contrast pattern. Thus the contrast patterns will be useful for labeling the clusters. The 2015 genes that showed significant Helmert contrast patterns in both DBA/2J and C57BL/6 were used for cluster analysis. Paracliques were generated using both the DBA/2J and C57BL/6 datasets separately. A high threshold of 0.80 is used for generating the networks in both the datasets. In the case of DBA/2J, the network graph consisted of 1224 genes and 22794 edges which resulted in 10 paracliques of varying sizes. In the case of C57BL/6, the network graph consisted of 1127 genes and 24360 edges which resulted in 7 paracliques.

Shown in the Tables 15 and 16 are the frequencies of different Helmert patterns associated with all genes within each paraclique. Almost all the genes in each paraclique were found to be characterized by the same Helmert pattern. The homogeneity index (HI) defined in Datta et al. [61] which is a measure of how homogenous the clusters are in terms of the design categories, is used to assess all the paracliques. It was found the paracliques from both C57BL/6 and DBA/2J yielded very high HI values of 0.92 and 0.90 respectively. Hence Helmert contrast signatures are very useful in labeling the clusters of genes from the paracliques.

Table 15. Distribution of Helmert patterns in paracliques for C57BL/6 dataset

Number of Genes	PC-1	PC-2	PC-3	PC-4	PC-5	PC-6	PC-7	PC-8	PC-9	PC-10
D-E13	0	0	0	0	0	0	0	0	0	0
D-E14	0	0	0	0	0	0	0	0	0	0
D-E15	0	28	17	18	0	0	1	12	11	11
D-E16	0	0	0	0	0	0	0	0	0	0
D-E17	0	0	0	0	0	0	0	0	0	0
D-E18	0	0	0	0	0	0	0	0	0	0
D-P0	3	0	0	0	16	1	0	0	0	0
D-P3	0	0	0	0	0	0	0	0	0	0
D-P6	26	0	0	0	0	4	12	0	0	0
D-P9	1	0	0	0	0	8	0	0	0	0

Table 16. Distribution of Helmert patterns in paracliques for DBA/2J dataset

Number of Genes	PC-1	PC-2	PC-3	PC-4	PC-5	PC-6	PC-7
D-E13	0	0	0	0	0	0	0
D-E14	0	0	0	0	0	0	0
D-E15	0	0	0	0	0	0	0
D-E16	0	25	0	0	2	11	11
D-E17	0	0	0	0	0	0	0
D-E18	0	4	0	0	13	0	0
D-P0	39	0	16	15	0	0	0
D-P3	0	0	0	0	0	0	0
D-P6	0	0	0	0	0	0	0
D-P9	0	0	0	0	0	0	0

4.5 Discussion

The contrasts patterns enabled the classification of genes based on specific patterns of gene expression and further provided insight in to genetic regulation of cerebellum development. Helmert patterns are concerned about the initial responses in gene expression. The first contrast design, for example, measures the mean differences between the first and second embryonic age, the second measures the mean differences between the third and average of first and second ages and so on. A larger number of transcripts had initial responses at E18 in DBA/2J and at E15 in C57BL/6. This suggests that most of the changes happen early in the embryonic development in C57BL/6 and at later stages of embryonic development in DBA/2J. Polynomial patterns, on the other hand, give information on the overall pattern of gene expression across all the time points. A linear pattern, for example indicates that there is a linear increase or decrease in the expression across the time points. Developmental events which require constant increase or decrease of expression levels across the embryonic to postnatal time points could be regulated by genes that belong to this category. Some events require gene expression characterized by increase across the embryonic stages and decrease across the postnatal stages and vice versa. Genes characterized by quadratic patterns belong to this category. Many complex patterns are also possible, but this analysis is confined to the linear, quadratic and cubic patterns which are easily interpretable in terms of the shapes of the expression profiles. More complex patterns can be fit using spline-fitting approaches that allow for fitting flexible models when identifying genes that are temporally differentially expressed. These methods are yet to be explored.

Statistical methods based on linear contrasts are very suitable in cases with experimental designs with few number of time points, typically less than 15. As the number of time points increase, the number of patterns increases exponentially and hence it might be computationally very expensive to fit all possible contrast patterns.

Genes that are co-regulated over time could be characterized by specific contrast pattern. In this way a cluster of co-expressed genes or a paraclique characterized by a particular pattern could be involved in the regulation of specific developmental events. Therefore, correlating the developmental events to the pattern of gene expression leads to identification of the key players involved in gene regulation associated with specific events.

It is important to note the distinction between ANOVA F-test and a specific contrast test such as Helmert or polynomial contrasts. A significant ANOVA F test among a group of means indicates that the largest contrast among all possible contrasts is significant. Therefore, a gene with a significant F test does not necessarily have a significant selected contrast. Therefore the expression patterns of these genes should be interpreted carefully.

Our methods emphasized the relative differences between adjacent sampling time points and the direction of the differences. The information about exact magnitudes of gene expressed at each time point was not included in our methods. A maximum fold change of two between any pairs of time points was used as one of the filtering criteria for the selection of genes. However it does not take in to consideration the magnitude of changes at all the individual time points. For example, two genes may have the same pattern but the magnitude of changes for the two genes may be dramatically different. So, even for genes belonging to the same pattern, their expression patterns should be examined with care.

A contingency table based method to identify differentially patterned genes in time course microarray experiments. The method may also be applied to more complicated situations, where three or more groups are compared, for example. Bowker's test, a generalization of McNemar's test which is in general used to test the hypothesis of symmetry was performed. A traditional chi-square test, used to test differences in the proportions, is not appropriate in this case since 20% of the cells had expected counts less than 5. This method focuses on differential profiling based on pattern differences, but do not assess the significance of the differences using p-values. However, if desired, generating p-values from a bootstrap analysis should be successful in this context.

There is a need for further annotation of all the genes that are expressed in different time patterns across the two strains by integrating these findings with available biological information. Further extensions to the factorial modeling of time patterns can be made to include allelic variation across the BXD RI lines [121] for QTL mapping of genes which are expressed under specific temporal patterns.

REFERENCES

1. Bonner AE, Lemon WJ, You M: **Gene expression signatures identify novel regulatory pathways during murine lung development: implications for lung tumorigenesis.** *J Med Genet* 2003, **40:**408-417.
2. Sarlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, et al: **Repeated observation of breast tumor subtypes in independent gene expression data sets.** *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100:**8418-8423.
3. Cheng Y, Church GM: **Biclustering of expression data.** *Proc Int Conf Intell Syst Mol Biol* 2000, **8:**93-103.
4. Liang S, Fuhrman S, Somogyi R: **Reveal, a general reverse engineering algorithm for inference of genetic network architectures.** *Pac Symp Biocomput* 1998**:**18-29.
5. Szallasi Z, Liang S: **Modeling the normal and neoplastic cell cycle with "realistic Boolean genetic networks": their application for understanding carcinogenesis and assessing therapeutic strategies.** *Pac Symp Biocomput* 1998**:**66-76.
6. Wuensche A: **Genomic regulation modeled as a network with basins of attraction.** *Pac Symp Biocomput* 1998**:**89-102.
7. Friedman N, Linial M, Nachman I, Pe'er D: **Using Bayesian networks to analyze expression data.** *J Comput Biol* 2000, **7:**601-620.
8. Butte A, Kohane I: **Relevance Networks: A First Step Toward Finding Genetic Regulatory Networks Within Microarray Data.** In *The Analysis of Gene Expression Data.* New York: Springer-Verlag; 2003: 428-446
9. Carter SL, Brechbuhler CM, Griffin M, Bond AT: **Gene co-expression network topology provides a framework for molecular characterization of cellular state.** *Bioinformatics* 2004, **20:**2242-2250.
10. Jeong H, Mason SP, Barabasi AL, Oltvai ZN: **Lethality and centrality in protein networks.** *Nature* 2001, **411:**41-42.
11. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology.** *Nature* 1999, **402:**C47-52.
12. Eisen M, Spellman P, Brown P, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proceedings of the National Academy of Sciences of the United States of America* 1998, **95:**14863 - 14868.
13. Zhang B, Horvath S: **A general framework for weighted gene co-expression network analysis.** *Stat Appl Genet Mol Biol* 2005, **4:**Article17.
14. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization.** *Mol Biol Cell* 1998, **9:**3273-3297.
15. Jain A, Dubes R: *Algorithms for clustering data.* Prentice-Hall, Inc.; 1988.

16. Kohonen T: **The self-organizing map.** *Neurocomputing* 1998, **21:**1-6.

17. De Smet F, Mathys J, Marchal K, Thijs G, De Moor B, Moreau Y: **Adaptive quality-based clustering of gene expression profiles.** *Bioinformatics* 2002, **18:**735-746.

18. Heyer LJ, Kruglyak S, Yooseph S: **Exploring Expression Data: Identification and Analysis of Coexpressed Genes.** *Genome Research* 1999, **9:**1106-1115.

19. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL: **Model-based clustering and data transformations for gene expression data.** *Bioinformatics* 2001, **17:**977-987.

20. Fraley C, Raftery AE: **Model-Based Clustering, Discriminant Analysis, and Density Estimation.** *Journal of the American Statistical Association* 2002, **97:**611-631.

21. Gasch AP, Eisen MB: **Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering.** *Genome Biol* 2002, **3:**RESEARCH0059.

22. Langston M, Lin L, Peng X, Baldwin N, Symons C, Zhang B, Snoddy J: **A Combinatorial Approach to the Analysis of Differential Gene Expression Data.** In *Methods of Microarray Data Analysis.* New York: Springer-Verlag; 2005: 223-238

23. Chesler E, Langston M: **Combinatorial Genetic Regulatory Network Analysis Tools for High Throughput Transcriptomic Data.** In *Systems Biology and Regulatory Genomics.* 2006: 150-165

24. Sharan R, Shamir R: **CLICK: a clustering algorithm with applications to gene expression analysis.** *Proc Int Conf Intell Syst Mol Biol* 2000, **8:**307-316.

25. Ben-Dor A, Shamir R, Yakhini Z: **Clustering gene expression patterns.** *J Comput Biol* 1999, **6:**281-297.

26. Tibshirani R, Walther G, Hastie T: **Estimating the number of clusters in a data set via the gap statistic.** *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2001, **63:**411-423.

27. Okada Y, Sahara T, Mitsubayashi H, Ohgiya S, Nagashima T: **Knowledge-assisted recognition of cluster boundaries in gene expression data.** *Artif Intell Med* 2005, **35:**171-183.

28. Famili AF, Liu G, Liu Z: **Evaluation and optimization of clustering in gene expression data analysis.** *Bioinformatics* 2004, **20:**1535-1545.

29. Schachter AD, Kohane IS: **An unsupervised self-optimizing gene clustering algorithm.** *Proc AMIA Symp* 2002**:**682-686.

30. Chun T, Li Z, Aidong Z, Ramanathan M: **Interrelated two-way clustering: an unsupervised approach for gene expression data analysis.** In *Bioinformatics and Bioengineering Conference, 2001 Proceedings of the IEEE 2nd International Symposium on*. 2001: 41-48.

31. Ye C, Eskin E: **Discovering tightly regulated and differentially expressed gene sets in whole genome expression data.** *Bioinformatics* 2007, **23:**e84-90.

32. Dobra A, Hans C, Jones B, Nevins JRJR, Yao G, West M: **Sparse graphical models for exploring gene expression data.** *Journal of Multivariate Analysis* 2004, **90:**196-212.

33. Schafer J, Strimmer K: **A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics.** *Stat Appl Genet Mol Biol* 2005, **4:**Article32.

34. Ledoit O, Wolf M: **Improved estimation of the covariance matrix of stock returns with an application to portfolio selection.** *Journal of Empirical Finance* 2003, **10:**603-621.

35. Steuer R, Kurths J, Daub CO, Weise J, Selbig J: **The mutual information: Detecting and evaluating dependencies between variables.** *Bioinformatics* 2002, **18:**S231-240.

36. Shannon CE: **A mathematical theory of communication.** *SIGMOBILE Mob Comput Commun Rev* 2001, **5:**3-55.

37. D'haeseleer P, Liang S, Somogyi R: **Genetic network inference: from co-expression clustering to reverse engineering.** *Bioinformatics* 2000, **16:**707-726.

38. Priness I, Maimon O, Ben-Gal I: **Evaluation of gene-expression clustering via mutual information distance measure.** *BMC Bioinformatics* 2007, **8:**111.

39. Morrison DA, Ellis J, Johnson AM: **An empirical comparison of distance matrix techniques for estimating codon usage divergence.** *Journal of Molecular Evolution* 1994, **39:**533-536.

40. Luo F, Khan L: **Data Complexity in Clustering Analysis of Gene Microarray Expression Profiles.** In *Data Complexity in Pattern Recognition.* 2006: 217-239

41. Kuo WP, Mendez E, Chen C, Whipple ME, Farell G, Agoff N, Park PJ: **Functional relationships between gene pairs in oral squamous cell carcinoma.** *AMIA Annu Symp Proc* 2003**:**371-375.

42. Butte AJ, Kohane IS: **Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements.** *Pac Symp Biocomput* 2000**:**418-429.

43. Herwig R, Poustka AJ, Muller C, Bull C, Lehrach H, O'Brien J: **Large-scale clustering of cDNA-fingerprinting data.** *Genome Res* 1999, **9:**1093-1105.

44. Kim K, Zhang S, Jiang K, Cai L, Lee I-B, Feldman L, Huang H: **Measuring similarities between gene expression profiles through new data transformations.** *BMC Bioinformatics* 2007, **8:**29.

45. Balasubramaniyan R, Hullermeier E, Weskamp N, Kamper J: **Clustering of gene expression data using a local shape-based similarity measure.** *Bioinformatics* 2005, **21:**1069-1077.

46. Cherepinsky V Fau - Feng J, Feng J Fau - Rejali M, Rejali M Fau - Mishra B, Mishra B: **Shrinkage-based similarity metric for cluster analysis of microarray data.**

47. Son YS, Baek J: **A modified correlation coefficient based similarity measure for clustering time-course gene expression data.** *Pattern Recogn Lett* 2008, **29:**232-242.

48. Li H, Sun Y, Zhan M: **Analysis of gene coexpression by B-spline based CoD estimation.** *EURASIP J Bioinformatics Syst Biol* 2007, **2007:**6-6.

49. Yona G, Dirks W, Rahman S, Lin DM: **Effective similarity measures for expression profiles.** *Bioinformatics* 2006, **22:**1616-1622.

50. **Gene Ontology Consortium** [http://www.geneontology.org.]

51. Rada R, Mili H, Bicknell E, Blettner M: **Development and application of a metric on semantic nets.** *Systems, Man and Cybernetics, IEEE Transactions on* 1989, **19:**17-30.

52. Varelas G, Voutsakis E, Raftopoulou P, Petrakis EGM, Milios EE: **Semantic similarity methods in wordNet and their application to information retrieval on the web.** In *Proceedings of the 7th annual ACM international workshop on Web information and data management*. Bremen, Germany: ACM; 2005.

53. Resnik P: **Using Information Content to Evaluate Semantic Similarity in a Taxonomy.** *Proceedings of the 14th International Joint Conference on Artificial Intelligence* 1995**:**448 - 453.

54. Lin D: **An Information-Theoretic Definition of Similarity.** *Proceedings of the Fifteenth International Conference on Machine Learning* 1998.

55. Jiang J, Conrath D: **Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy.** 1997.

56. Resnik P: **Using Information Content to Evaluate Semantic Similarity in a Taxonomy.** In *IJCAI*. 1995: 448-453.

57. Lin D: **An Information-Theoretic Definition of Similarity.** In *Proceedings of the Fifteenth International Conference on Machine Learning*: Morgan Kaufmann Publishers Inc.; 1998.

58. Priness I, Maimon O, Ben-Gal I: **Evaluation of gene-expression clustering via mutual information distance measure.** *BMC Bioinformatics* 2007, **8:**111.

59. Michaels GS, Carr DB, Askenazi M, Fuhrman S, Wen X, Somogyi R: **Cluster analysis and data visualization of large-scale gene expression data.** *Pac Symp Biocomput* 1998**:**42-53.

60. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci U S A* 1998, **95:**14863-14868.

61. Datta S, Datta S: **Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes.** *BMC Bioinformatics* 2006, **7:**397.

62. Rand W: **Objective Criteria for the Evaluation of Clustering Methods.** *Journal of the American Statistical Association* 1971, **66:**846-850.

63. Rousseeuw P: **Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.** *J Comput Appl Math* 1987, **20:**53-65.

64. Yeung KY, Ruzzo WL: **Principal component analysis for clustering gene expression data.** *Bioinformatics* 2001, **17:**763-774.

65. Yeung KY, Haynor DR, Ruzzo WL: **Validating clustering for gene expression data.** *Bioinformatics* 2001, **17:**309-318.

66. McShane LM, Radmacher MD, Freidlin B, Yu R, Li M-C, Simon R: **Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data.** *Bioinformatics* 2002, **18:**1462-1469.

67. Kerr MK, Churchill GA: **Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments.** *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98:**8961-8965.

68. Park PJ, Manjourides J, Bonetti M, Pagano M: **A permutation test for determining significance of clusters with applications to spatial and gene expression data.** *Computational Statistics & Data Analysis* 2009, **53:**4290-4300.

69. Hanisch D, Zien A, Zimmer R, Lengauer T: **Co-clustering of biological networks and gene expression data.** *Bioinformatics* 2002, **18:**S145-154.

70. Kustra R, Zagdanski A: **Incorporating Gene Ontology in Clustering Gene Expression Data.** In *Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems*: IEEE Computer Society; 2006.

71. Lord PW, Stevens RD, Brass A, Goble CA: **Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation.** *Bioinformatics* 2003, **19:**1275-1283.

72. Sevilla JL, Segura V, Podhorski A, Guruceaga E, Mato JM, Martinez-Cruz LA, Corrales FJ, Rubio A: **Correlation between Gene Expression and GO Semantic Similarity.** *IEEE/ACM Trans Comput Biol Bioinformatics* 2005, **2:**330-338.

73. Schisterman E, Moysich K, England L, Rao M: **Estimation of the correlation coefficient using the Bayesian Approach and its applications for epidemiologic research.** *BMC Medical Research Methodology* 2003, **3:**5.

74. Box G, Tao D: **Bayesian Inference in Statistical Analysis.** *Reading, MA, Addison-Wesley* 1973.

75. Bashir SA, Duffy SW: **The correction of risk estimates for measurement error.** *Annals of Epidemiology* 1997, **7:**154-164.

76. Fisher R: **Frequency distributions of the values of the correlation coefficient in samples from an indefinitely large population.** *Biometrika* 1915, **10:**507 - 521.

77. Langston MA, Perkins AD, Saxton AM, Scharff JA, Voy BH: **Innovative Computational Methods for Transcriptomic Data Analysis: A Case Study in the Use of FPT for Practical Algorithm Design and Implementation.** *The Computer Journal* 2008, **51:**26-38.

78. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW: **A**

genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* 1998, **2:**65-73.

79. Ge H, Liu Z, Church GM, Vidal M: **Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae.** *Nat Genet* 2001, **29:**482-486.
80. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ: **Discovering statistically significant pathways in expression profiling studies.** *Proc Natl Acad Sci U S A* 2005, **102:**13544-13549.
81. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB: **Missing value estimation methods for DNA microarrays.** *Bioinformatics* 2001, **17:**520-525.
82. **R Project for Statistical Computing** [http://www.r-project.org]
83. Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biology* 2004, **5:**R80.
84. Downey R, Fellows MR: *Parameterized Complexity.* Springer; 1999.
85. Faisal NA-k, Michael AL, Suters WH: **Fast, effective vertex cover kernelization: a tale of two algorithms.** 2005.
86. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protoc* 2009, **4:**44-57.
87. Grigoriev A: **A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae.** *Nucl Acids Res* 2001, **29:**3513-3519.
88. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture.** *Nat Genet* 1999, **22:**281-285.
89. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR: **Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.** *Proc Natl Acad Sci U S A* 1999, **96:**2907-2912.
90. Hastie T, Tibshirani R, Eisen M, Alizadeh A, Levy R, Staudt L, Chan W, Botstein D, Brown P: **'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns.** *Genome Biology* 2000, **1:**research0003.0001 - research0003.0021.
91. Handl J, Knowles J, Kell DB: **Computational cluster validation in post-genomic data analysis.** *Bioinformatics* 2005, **21:**3201-3212.
92. Dunn JC: **Well separated clusters and optimal fuzzy-partitions.** *Journal of Cybernetics* 1974, **4:**95-104.
93. Mantel N: **The Detection of Disease Clustering and a Generalized Regression Approach.** *Cancer Res* 1967, **27:**209-220.

94. Shannon WD, Watson MA, Perry A, Rich K: **Mantel statistics to correlate gene expression levels from microarrays with clinical covariates.** *Genetic Epidemiology* 2002, **23:**87-96.

95. Ling RF: **A Probability Theory of Cluster Analysis.** *Journal of the American Statistical Association* 1973, **68:**159-164.

96. Wald A, Wolfowitz J: **Statistical Tests Based on Permutations of the Observations.** *The Annals of Mathematical Statistics* 1944, **15:**358-372.

97. Storch KF, Lipan O, Leykin I, Viswanathan N, Davis FC, Wong WH, Weitz CJ: **Extensive and divergent circadian gene expression in liver and heart.** *Nature* 2002, **417:**78-83.

98. Li H, Wood CL, Liu Y, Getchell TV, Getchell ML, Stromberg AJ: **Identification of gene expression patterns using planned linear contrasts.** *BMC Bioinformatics* 2006, **7:**245.

99. Lonnstedt I. GS, Begley G., Speed TP.: **Microarray analysis of two interacting treatments: a linear model and trends in expression over time.** Sweden: Department of Mathematics, Uppsala University; 2001.

100. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol* 2004, **3:**Article3.

101. Voy BH, Scharff JA, Perkins AD, Saxton AM, Borate B, Chesler EJ, Branstetter LK, Langston MA: **Extracting Gene Networks for Low-Dose Radiation Using Graph Theoretical Algorithms.** *PLoS Comput Biol* 2006, **2:**e89.

102. Pelz C, Martin M, Bagby G, Sears R: **Global Rank-invariant Set Normalization (GRSN) to reduce systematic distortions in microarray data.** *BMC Bioinformatics* 2008, **9**.

103. Kauffmann A, Gentleman R, Huber W: **arrayQualityMetrics--a bioconductor package for quality assessment of microarray data.** *Bioinformatics* 2009, **25:**415-416.

104. Krampe A, Kuhnt S: **Bowker's test for symmetry and modifications within the algebraic framework.** *Computational Statistics & Data Analysis* 2007, **51:**4124-4142.

105. Agresti A: *Categorical Data Analysis (Wiley Series in Probability and Statistics).* Wiley-Interscience; 2002.

106. Shi S, Stanley P: **Protein O-fucosyltransferase 1 is an essential component of Notch signaling pathways.** *Proc Natl Acad Sci U S A* 2003, **100:**5234-5239.

107. Di Padova M, Bruno T, De Nicola F, Iezzi S, D'Angelo C, Gallo R, Nicosia D, Corbi N, Biroccio A, Floridi A, et al: **Che-1 arrests human colon carcinoma cell proliferation by displacing HDAC1 from the p21WAF1/CIP1 promoter.** *J Biol Chem* 2003, **278:**36496-36504.

108. Ramírez MJ, Honer WG, Minger SL, Francis PT: **Changes in hippocampal SNAP-25 expression following afferent lesions.** *Brain Research* 2004, **997:**133-135.

109. Canoll PD, Musacchio JM, Hardy R, Reynolds R, Marchionni MA, Salzer JL: **GGF/Neuregulin Is a Neuronal Signal That Promotes the Proliferation and Survival and Inhibits the Differentiation of Oligodendrocyte Progenitors.** 1996, **17:**229-243.

110. Kao HT, Porton B, Czernik AJ, Feng J, Yiu G, Haring M, Benfenati F, Greengard P: **A third member of the synapsin gene family.** *Proc Natl Acad Sci U S A* 1998, **95:**4667-4672.

111. Bahler M, Benfenati F, Valtorta F, Greengard P: **The synapsins and the regulation of synaptic function.** *Bioessays* 1990, **12:**259-263.

112. Johnson-Anuna LN, Eckert GP, Keller JH, Igbavboa U, Franke C, Fechner T, Schubert-Zsilavecz M, Karas M, Muller WE, Wood WG: **Chronic Administration of Statins Alters Multiple Gene Expression Patterns in Mouse Cerebral Cortex.** *J Pharmacol Exp Ther* 2005, **312:**786-793.

113. Yan XD, Hanson AJ, Nahreini P, Koustas WT, Andreatta C, Prasad KN: **Altered expression of genes regulating cell growth, proliferation, and apoptosis during adenosine 3',5'-cyclic monophosphate-induced differentiation of neuroblastoma cells in culture.** *In Vitro Cell Dev Biol Anim* 2002, **38:**529-537.

114. Elliott AM, de Miguel MP, Rebel VI, Donovan PJ: **Identifying genes differentially expressed between PGCs and ES cells reveals a role for CREB-binding protein in germ cell survival.** *Dev Biol* 2007, **311:**347-358.

115. Del Rio JA, Gonzalez-Billault C, Urena JM, Jimenez EM, Barallobre MJ, Pascual M, Pujadas L, Simo S, La Torre A, Wandosell F, et al: **MAP1B is required for Netrin 1 signaling in neuronal migration and axonal guidance.** *Curr Biol* 2004, **14:**840-850.

116. Gong TW, Winnicki RS, Kohrman DC, Lomax MI: **A novel mouse kinesin of the UNC-104/KIF1 subfamily encoded by the Kif1b gene.** *Gene* 1999, **239:**117-127.

117. Zhao K, Wang W, Rando OJ, Xue Y, Swiderek K, Kuo A, Crabtree GR: **Rapid and phosphoinositol-dependent binding of the SWI/SNF-like BAF complex to chromatin after T lymphocyte receptor signaling.** *Cell* 1998, **95:**625-636.

118. Pineda VV, Athos JI, Wang H, Celver J, Ippolito D, Boulay G, Birnbaumer L, Storm DR: **Removal of G(ialpha1) constraints on adenylyl cyclase in the hippocampus enhances LTP and impairs memory formation.** *Neuron* 2004, **41:**153-163.

119. Peitsch WK, Bulkescher J, Spring H, Hofmann I, Goerdt S, Franke WW: **Dynamics of the actin-binding protein drebrin in motile cells and definition of a juxtanuclear drebrin-enriched zone.** *Exp Cell Res* 2006, **312:**2605-2618.

120. Teng J, Takei Y, Harada A, Nakata T, Chen J, Hirokawa N: **Synergistic effects of MAP2 and MAP1B knockout in neuronal migration, dendritic outgrowth, and microtubule organization.** *J Cell Biol* 2001, **155:**65-76.

121. Peirce J, Lu L, Gu J, Silver L, Williams R: **A new set of BXD recombinant inbred lines from advanced intercross populations in mice.** *BMC Genetics* 2004, **5:**7.

VITA

Suman Duvvuru was born in city of Hyderabad in India. After graduating from high school in April 2000, he moved to United States to pursue his bachelor's degree in computer science at University of Missouri- Kansas city. During the senior year of his undergraduate studies, he worked on a research project in the area of bioinformatics involving development of an XML standard for proteomic data. He graduated from college in May 2004 and then moved to Knoxville, Tennessee to pursue graduate work in Genomic Sciences and Statistics at the University of Tennessee (Ph.D. 2009). During his graduate career at University of Tennessee, he was involved in research rotations in the fields of genetics, proteomics and computational biology. It was during the second year of graduate study that he became interested in statistical genomics and got involved in the analysis of high dimensional genomic data. His statistical interests led to earning a Master of Science in Statistics simultaneously with his Ph.D. studies. He worked with the Systems Genetics group at Oak Ridge National Laboratory and gained research experience in the design and analysis of several research studies in the areas of brain and behavioral genetics. In addition, he also gained experience in industry working in the field of pharmacogenomics at two major pharmaceutical companies, Pfizer and GlaxoSmithKline. As a graduate student at University of Tennessee, he was mentored by professors Arnold Saxton and Elissa Chesler. He will work with Eli Lilly and Company starting November 2009 as research scientist in the pharmacogenomics division.