Published in final edited form as:

Biometrics. 2016 December ; 72(4): 1216–1225. doi:10.1111/biom.12508.

CCor: a whole genome network-based similarity measure between two genes

Yiming Hu and

Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, U.S.A.
yiming.hu@yale.edu

Hongyu Zhao

Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, U.S.A.;
Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT, U.S.A.
hongyu.zhao@yale.edu

Summary

Measuring the similarity between genes is often the starting point for building gene regulatory networks. Most similarity measures used in practice only consider pairwise information with a few also consider network structure. Although theoretical properties of pairwise measures are well understood in the statistics literature, little is known about their statistical properties of those similarity measures based on network structure. In this article, we consider a new whole genome network-based similarity measure, called *CCor*, that makes use of information of all the genes in the network. We derive a concentration inequality of *CCor* and compare it with the commonly used Pearson correlation coefficient for inferring network modules. Both theoretical analysis and real data example demonstrate the advantages of *CCor* over existing measures for inferring gene modules.

Keywords

similarity measure; concentration inequality; co-expression; Pearson correlation; Gaussian graphical model; topological overlap measure; clustering; gene module

1. Introduction

Understanding the interactions and functional relationships between genes is a major goal in systems biology. Various statistical and computational methods have been developed to infer network modules and gene regulatory networks using gene expression data, which are usually generated by microarrays or RNA-sequencing. A gene expression microarray typically contains tens of thousands of probes fixed on a solid surface, where each probe corresponds to fragment of a gene to infer the expression level of the target gene (Hoheisel, 2006; Schulze and Downward, 2001). In contrast, RNA-sequencing directly quantifies the

Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 3, 4, 5 are available with this paper at the Biometrics website on Wiley Online Library.

expression levels of genes by applying next generation sequencing to RNA in cells. Although gene expression microarrays were routinely used in biological research, RNA-sequencing is becoming more commonly used in recent years (Wang et al., 2009). Grouping genes with similar expression patterns is usually the first step in these analyses. The genes grouped together (clusters or modules) can reveal important biological insights. Typically, a similarity measure is the first and key step to perform such analyses (Song et al., 2012).

Most similarity measures fall into one of two categories: (1) measures using only *pairwise* information of the two genes studied, such as correlation coefficients (Eisen et al., 1998; Zhou et al., 2002; Stuart et al., 2003) and mutual information measures (Butte et al., 2000; Daub et al., 2004; Basso et al., 2005; Margolin et al., 2006; Priness et al., 2007; Meyer et al., 2008; Cadeiras et al., 2011); (2) measures incorporating information of all the genes in the network, such as the topological overlap measure (*TOM*) (Zhang and Horvath, 2005; Langfelder and Horvath, 2008). Methods in the first category define similarities through measuring co-expression of two genes. This is accomplished by formulating the problem as detecting bivariate association between two vectors representing the expression levels for two genes across samples (Song et al., 2012; Wang et al., 2014; Kumari et al., 2012).

Although conceptually intuitive and computationally straightforward, these methods do not incorporate the biological structure between the two genes of interest and other genes in the network. It is likely that similarity may be better defined between two genes using the information of all the other genes related to this pair of genes in the network. We illustrate this using *TOM* as an example. *TOM* first obtains an adjacency matrix defined by pairwise similarity measures, e.g. Pearson correlation coefficients between each pair of genes in the network. It then defines the similarity between two genes by the number of shared neighbors of this pair of genes in the network (with standardization). Compared with those *pairwise* measures, *TOM* has been found to perform better in both simulations and real applications (Horvath et al., 2006) in reconstructing gene networks (Allen et al., 2012) and detecting modules (Song et al., 2012).

Although the literature suggests that borrowing information from other genes in the network may better characterize the similarity between genes, theoretical analysis of these measures, such as *TOM*, is challenging due to complexities in these statistics. Ideally, we would like to show that a good similarity measure can better distinguish within-module gene pairs from between-module gene pairs. That is, we expect that there is a better separation of the distribution of similarities for gene pairs from the same module from that for gene pairs from different modules.

In this article, we propose a new similarity measure, correlation of correlations (*CCor*), to define similarity between a pair of genes that can incorporate expression profiles from all the other genes. We establish the concentration inequalities of this measure for gene pairs in the same module (to be defined later) and those in different modules. We show that *CCor* can better separate gene pairs in the same module from those in different modules than the *pairwise* measures such as the most commonly used Pearson correlation coefficient (*PCor*). Simulations and real data applications substantiate the theoretical results. The article is organized as follows: in Section 2, we introduce the proposed measure *CCor* and its

motivation. Statistical properties of $CCor$ are derived in Section 3 together with simulation results. Comparisons with other similarity measure by simulating more complicated cases when the number of genes is much larger than sample size and noise exists are shown in Section 4. Real data analysis is performed in Section 5 with discussion and conclusions in Section 6.

2. Model

In this section, we introduce $CCor$, its motivation, and biological interpretation. A modified version of $CCor$ is also introduced.

2.1 $CCor$

Assume that we have the expression profiles for p genes of n individuals summarized in an $n \times p$ matrix $X_{n \times p}$ with rows $(X_k, k = 1, \dots, n)$ representing individuals and columns $(X_i, i = 1, \dots, p)$ representing genes. We can calculate the sample Pearson correlation coefficient,

$$r_{ij} = \frac{\sum_k (X_{ki} - \bar{X}_{\cdot i})(X_{kj} - \bar{X}_{\cdot j})}{\sqrt{\sum_k (X_{ki} - \bar{X}_{\cdot i})^2} \sqrt{\sum_k (X_{kj} - \bar{X}_{\cdot j})^2}}, \text{ between gene } i \text{ and gene } j. \text{ } CCor \text{ between genes } i \text{ and } j, i < j, \text{ is defined as the Pearson correlation between coefficients of } X_i \text{ vs genes except } i, j \text{ and } X_j \text{ vs genes except } i, j. \text{ To be more specific, let}$$

$$\begin{aligned} \mathbf{v}_i &\triangleq (r_{i1}, r_{i2}, \dots, r_{i(i-1)}, r_{i(i+1)}, \dots, r_{i(j-1)}, r_{i(j+1)}, \dots, r_{ip}) \\ \mathbf{v}_j &\triangleq (r_{j1}, r_{j2}, \dots, r_{j(i-1)}, r_{j(i+1)}, \dots, r_{j(j-1)}, r_{j(j+1)}, \dots, r_{jp}) \end{aligned} \quad (1)$$

$CCor$ is defined as:

$$CCor_{ij} \triangleq \widehat{cor}(\mathbf{v}_i, \mathbf{v}_j)$$

To study the statistical properties of $CCor$, we first assume that the expression profiles from the samples are independent and have a multivariate normal distribution. For simplicity, we assume that the genes are all standardized with mean 0 and variance 1. Therefore, the

covariance matrix is the same as the correlation matrix. That is $X_j \stackrel{i.i.d.}{\sim} N(0, W)$, in which the diagonal elements of W are 1. In network analysis, a module is defined as a group of nodes with high interconnections, i.e. the within-module adjacencies are high. Therefore, we define a gene co-expression module as a group of genes with high within-module correlations. We assume that there are M gene modules. If we sort the genes so that genes in the same module are grouped together, the covariance matrix W can be written with a blockwise structure:

$$W = \begin{bmatrix} W_{11} & W_{12} & \dots & W_{1M} \\ & W_{22} & \dots & W_{2M} \\ & & \dots & \dots \\ & & & W_{MM} \end{bmatrix} \quad (2)$$

The diagonal blocks correspond to the covariance matrices of genes in the same module and the off-diagonal blocks represent the correlation structures of genes in different modules. According to the definition of module, we expect that the values in the diagonal blocks to be higher than those in the off-diagonal blocks since a gene should be more correlated with genes from the same module than those from different modules.

Under these assumptions, we will show in later sections that $CCor$ is high for gene pairs in the same module and low for pairs from different modules. Furthermore, we will show that although the module is defined by the covariance matrix, $CCor$ can better distinguish gene pairs from the same module and those from different modules than Pearson correlation. We use the following example to illustrate the idea of $CCor$. In this example, suppose we have two modules of sizes p_1 and p_2 , respectively, where $p_1 = 30$, $p_2 = 20$ and the correlation matrix is

$$W = \begin{bmatrix} 1 & w_1 & w_1 & \dots & w_1 & w_2 & w_2 & \dots & \dots & w_2 \\ & 1 & w_1 & \dots & w_1 & w_2 & w_2 & \dots & \dots & w_2 \\ & & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ & & & & 1 & w_2 & w_2 & \dots & \dots & w_2 \\ & & & & & 1 & w_1 & w_1 & \dots & w_1 \\ & & & & & & 1 & w_1 & \dots & w_1 \\ & & & & & & & \dots & \dots & \dots \\ & & & & & & & & & 1 \end{bmatrix} \quad (3)$$

with $w_1 = 0.5$ and $w_2 = 0.3$. The two diagonal blocks represent the within-module covariance matrix of two modules. In Figure 1, when $n = 100$, we plot the observations of genes 1 and 2 (genes from the same module) and genes 1 and $p_1 + 1$ (genes from different modules). The corresponding v_1 , v_2 and v_1 , v_{p_1+1} (see (1)) are also plotted in Figure 1. Note that in this simulation, the sample size is larger than the total number of genes, where the sample size is likely to be smaller than the number of genes in microarray data sets. Here, our purpose is to illustrate the idea of $CCor$ and the performance of $CCor$ when $n \ll p$ is investigated in Sections 4 and 5.

A biological interpretation of $CCor$ is that it quantifies similarity between genes i and j defined through their relationships with other genes. Since correlation coefficients characterize the similarity of each gene pair, $CCor$ is the correlation of these pairwise similarities. That is, if two genes are both highly correlated with the same set of genes, $CCor$ is high. Therefore, $CCor$ may better separate gene pairs in the same module from those in different modules than Pearson correlation through the incorporation of information from other genes.

2.2 Modified CCor

The example in Figure 1 illustrates the intuition of *CCor*; however, it is a rather simplified and ideal case (only two modules and the sizes of the modules are balanced). In Figure 2, we show a more complex case with three modules and unbalanced module sizes. The choice of $\mathbf{W}$ is similar to (3) except for having one more diagonal block. The values and associated scatter plots for *CCor* of $(1, 2)$, $(1, p_1 + 1)$, $(p_1 + p_2 + 1, p_1 + p_2 + 2)$, $(1, p_1 + p_2 + 1)$, where $p_1 = p_2 = 10$ and $p_3 = 70$, for $n = 100$ are shown in Figure 2, representing the *CCor* within the first module, between the first and second modules, within the third module, and between the first and third modules, respectively.

The two right panels show that the between-module similarity measured by *CCor* is not as low as that in the balanced module size case since they are affected by points in the bottom left part of the plots, which are the correlations with genes that are not highly correlated with both genes in the gene pair. An intuitive way to tackle this problem is to not consider genes that are not highly correlated with both genes of the gene pair. For example, if we want to measure the similarity between genes A and B , we will not use the correlations of A and B with genes that are not highly correlated with both A and B . A more sophisticated way would be to put weights on the correlation coefficients based on their magnitudes. Therefore, we propose a modified version of *CCor* (*mCCor*) with hard thresholding and weights as follow. Assume we have a pre-specified threshold t and weight coefficient w , then *mCCor* for genes i and j is defined as the weighted Pearson correlation coefficient

$$\frac{\sum_k w_k \left(X_{ki} - \frac{\sum_k w_k X_{ki}}{\sum_k w_k} \right) \left(X_{kj} - \frac{\sum_k w_k X_{kj}}{\sum_k w_k} \right)}{\left(\sqrt{\sum_k w_k \left(X_{ki} - \frac{\sum_k w_k X_{ki}}{\sum_k w_k} \right)^2} \sqrt{\sum_k w_k \left(X_{kj} - \frac{\sum_k w_k X_{kj}}{\sum_k w_k} \right)^2} \right)} \quad \text{of (1), where the weight for the } k\text{th pair in (1) is defined as}$$

$$w_k = \begin{cases} w^2, & \mathbf{1}_{\{|r_{ik}| > t\}} + \mathbf{1}_{\{|r_{jk}| > t\}} = 2 \\ w, & \mathbf{1}_{\{|r_{ik}| > t\}} + \mathbf{1}_{\{|r_{jk}| > t\}} = 1 \\ 1, & \mathbf{1}_{\{|r_{ik}| > t\}} + \mathbf{1}_{\{|r_{jk}| > t\}} = 0 \end{cases} \quad (4)$$

Besides hard-thresholding, we can also use soft-thresholding and other methods to weigh the connections between genes. The biological interpretation of *mCCor* is that we can better measure the similarity between two genes by the set of genes that are ‘strongly’ connected to them.

3. Theoretical Results

As mentioned in the introduction, a good similarity measure should have higher scores for pairs from the same module and lower scores for pairs from different modules. Therefore, the distribution of the similarity measure can reflect its properties. In this section, we will show the statistical properties of *CCor* under moderate assumptions. We will derive a concentration inequality to characterize the distribution of *CCor*.

THEOREM 1: Assume we have $X_{n \times p}, X_j \stackrel{i.i.d.}{\sim} N(\mathbf{0}, \mathbf{W})$, in which $\mathbf{W}$ was defined as in (2)

and with all diagonal terms equal to 1. $\mathbf{v}_i, \mathbf{v}_j$ are defined as in (1) with $r_{kt} \triangleq \frac{1}{n} \sum_{z=1}^n X_{kz} X_{tz}$

and $K \triangleq p-2$. $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix} \triangleq \text{var} \left(\begin{bmatrix} \mathbf{v}_i \\ \mathbf{v}_j \end{bmatrix} \right)$ and $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \triangleq \mathbb{E} \left[\begin{bmatrix} \mathbf{v}_i \\ \mathbf{v}_j \end{bmatrix} \right]$. Denote

$A \triangleq I_K - \frac{1}{K} \mathbf{1}_K \mathbf{1}_K^T$. Then for a vector $\begin{bmatrix} U \\ V \end{bmatrix} \sim \text{MVN}(\mu, \Sigma)$,

$$\widehat{CCor}_{ij} \triangleq \frac{\sum_k (U_k - \bar{U})(V_k - \bar{V})}{\sqrt{\sum_k (U_k - \bar{U})^2 \sum_k (V_k - \bar{V})^2}}$$

$$\left| \widehat{CCor}_{ij} - \frac{\frac{1}{2} \sum_{i=1}^{2K} \lambda_{1i} (1 + C_{1i}^2)}{\sqrt{\sum_{i=1}^K \lambda_{2i} (1 + C_{2i}^2) \sum_{i=1}^K \lambda_{3i} (1 + C_{3i}^2)}} \right| < \epsilon \quad (5)$$

with probability no less than

$$1 - 2 \exp \left(-K c_1 \min \left(\frac{\epsilon^2}{\mathcal{K}_1^2}, \frac{\epsilon}{\mathcal{K}_1} \right) \right) - 2 \exp \left(-K c_2 \min \left(\frac{\epsilon^2}{\mathcal{K}_2^2}, \frac{\epsilon}{\mathcal{K}_2} \right) \right) - 2 \exp \left(-K c_3 \min \left(\frac{\epsilon^2}{\mathcal{K}_3^2}, \frac{\epsilon}{\mathcal{K}_3} \right) \right)$$

. In all the formulas above, $\lambda_{1i}, \lambda_{2i}, \lambda_{3i}, C_{1i}, C_{2i}, C_{3i}, \mathcal{K}_1, \mathcal{K}_2, \mathcal{K}_3$ are functions of Σ, μ and A . $c_1, c_2, c_3 > 0$ are absolute constants.

The proof of this theorem is shown in the Supplementary Materials. We argue that the

$$CCor_{ij} = \frac{\sum_k (\mathbf{v}_{ik} - \bar{\mathbf{v}}_i)(\mathbf{v}_{jk} - \bar{\mathbf{v}}_j)}{\sqrt{\sum_k (\mathbf{v}_{ik} - \bar{\mathbf{v}}_i)^2 \sum_k (\mathbf{v}_{jk} - \bar{\mathbf{v}}_j)^2}}$$

distribution of $\frac{\sum_k (\mathbf{v}_{ik} - \bar{\mathbf{v}}_i)(\mathbf{v}_{jk} - \bar{\mathbf{v}}_j)}{\sqrt{\sum_k (\mathbf{v}_{ik} - \bar{\mathbf{v}}_i)^2 \sum_k (\mathbf{v}_{jk} - \bar{\mathbf{v}}_j)^2}}$ is very similar to $\widehat{CCor}_{ij}$ defined in Theorem 1 since $\text{MVM}(\mu, \Sigma)$ is the asymptotic approximation of the distribution of

$\begin{bmatrix} \mathbf{v}_i \\ \mathbf{v}_j \end{bmatrix}$. Simulation results (see Supplementary Materials) also show that the approximation is very accurate even when n is rather small. Therefore, we could use this concentration inequality to study the theoretical performance of $CCor$. Theorem 1 provides a way to study the statistical properties of $CCor$ since when p is large, $CCor$ will concentrate in a small neighborhood with a high probability, and the concentrating value in (5) is a function of $\mathbf{W}$. Although we do not have the closed form expression of this function, we can calculate the concentrating value and study how they behave with different choices of $\mathbf{W}$.

We continue to use the two-module case here to illustrate how $CCor$ performs theoretically. We choose $p_1 = p_2 = 100$, $n = 100$ and $\mathbf{W}$ is the same as (3). We fix w_2 and calculate the concentrating values of $CCor_{12}$ and $CCor_{1(p_1+1)}$ with different choices of w_1 . Here, we use $CCor_{12}$ and $CCor_{1(p_1+1)}$ to represent within-module and between-module similarity measures, respectively. Due to the structure of $\mathbf{W}$ in (3), the $CCor$ of any within-module pair

or between-module pair will have the same distribution, therefore it is representative to study $CCor_{12}$ and $CCor_{1(p_1+1)}$ only. For convenience, we only use within-module and between-module in the following discussion. The theoretical and empirical values of $CCor$ of within- and between-module pairs are shown in the left panels of Fig. 3, whereas the corresponding Pearson correlations are shown in the right panels of this figure. Empirical values of $CCor$ are the means of $CCor$ calculated from 100 simulated datasets. When the difference between w_1 and w_2 increases, the difference between within-module and between-module $CCor$ also increases. Furthermore, the difference between $CCors$ is much larger than that between Pearson correlations under different choices of w_1 and w_2 . These results show that $CCor$ can better distinguish within-module pairs from between-module pairs than Pearson correlation under different settings both theoretically and empirically in this simple scenario.

Besides the simple case discussed above, we can also study the theoretical performance of $CCor$ in more complicated cases. Here, we use a 5-module unbalanced population as an example. We choose $p_1 = 50, p_2 = 10, p_3 = 70, p_4 = 70, p_5 = 60$ and the structure of $\mathbf{W}$ is similar to (3). That is, for diagonal blocks, all elements except the diagonal terms are chosen to be the same and the diagonal terms are 1. Therefore, we can represent $\mathbf{W}$ as follows:

$$\mathbf{W} = \begin{bmatrix} 0.5 & 0.35 & 0.35 & 0.2 & 0.3 \\ & 0.45 & 0.3 & 0.2 & 0.3 \\ & & 0.55 & 0.3 & 0.2 \\ & & & 0.6 & 0.25 \\ & & & & 0.5 \end{bmatrix}, \quad \mathbf{W}_{est}^{PCor} = \begin{bmatrix} 0.50 & 0.35 & 0.35 & 0.19 & 0.29 \\ & 0.45 & 0.30 & 0.19 & 0.30 \\ & & 0.54 & 0.29 & 0.20 \\ & & & 0.58 & 0.24 \\ & & & & 0.49 \end{bmatrix} \quad (6)$$

$$\mathbf{W}_{ccen}^{CCor} = \begin{bmatrix} 0.80 & 0.56 & 0.31 & -0.52 & 0.10 \\ & 0.65 & 0.20 & -0.42 & 0.24 \\ & & 0.87 & -0.01 & -0.49 \\ & & & 0.92 & -0.2 \\ & & & & 0.83 \end{bmatrix}, \quad \mathbf{W}_{est}^{CCor} = \begin{bmatrix} 0.80 & 0.55 & 0.31 & -0.53 & 0.09 \\ & 0.64 & 0.18 & -0.42 & 0.25 \\ & & 0.87 & -0.03 & -0.49 \\ & & & 0.92 & -0.20 \\ & & & & 0.81 \end{bmatrix} \quad (7)$$

We calculate the theoretical values of $CCor$ of $(1, 2),$

$(1, p_1+1), (1, p_1+p_2+1), \dots, (\sum_{i=1}^4 p_i+1, \sum_{i=1}^4 p_i+2)$ representing within-module and between-module similarity measures. They represent the similarity measure of pairs from all the blocks respectively. In fact, all the diagonal blocks represent the similarity structure within each module and the off-diagonal blocks show the similarity structure between modules. For example, $CCor_{p_1+1, p_1+2}$ represents the similarity measure of pairs within the second module and $CCor_{p_1+1, p_1+p_2+1}$ represents the similarity measure of a gene from the second module and another gene from the third module. The estimated means of $PCor$ of each block are denoted as $\mathbf{W}_{est}^{PCor}$ and shown in (6). The concentrating values of the $CCor$ of the representing pairs of each block are denoted as $\mathbf{W}_{ccen}^{CCor}$ and the estimated means using

simulations are denoted as W_{est}^{CCor} (7). By comparing (6) with (7), we can see that $CCor$ s are higher for within-module pairs and lower for between-module pairs and the difference is larger than that of using $PCor$, which suggests that $CCor$ could amplify the difference between the within-module similarity and between-module similarity.

It is worth noting that the performance of $CCor$ on distinguishing the second module with other modules is not as good as others. The within-module $CCor$ is lower and the between-module $CCor$ is higher. We can observe that the size of the second module is much smaller than that of the first module ($p_1 = 50, p_2 = 10$) and the correlation within the second module is not as high as with the other modules, which makes sense since it is indeed hard to distinguish a small module with lower within-module similarity from modules with larger sizes and within-module similarities. This was also observed in Section 2.2 indicating that the performance of $CCor$ could be compromised by the imbalance of module sizes.

We therefore calculate $mCCor$ using simulated datasets and the averages of 200 replicates are shown below. Threshold t in (4) is chosen to be 0.45 and weight is 5. Results using different thresholds and weights are shown in the Supplementary Materials. Overall, the results are not sensitive to the values of weights. Within-module $mCCor$ changes little across all choices of weights and thresholds. However, between-module $mCCor$ s are a bit more sensitive to the thresholds, mostly in the range of (0.25, 0.45). In (8), the second module could also be clearly separated from the other modules since the difference between within and between module $mCCor$ is larger. Therefore, $mCCor$ can better distinguish modules when the module sizes are not balanced.

$$W_{est}^{mCCor} = \begin{bmatrix} 0.79 & 0.44 & 0.04 & -0.75 & -0.12 \\ & 0.69 & 0.06 & -0.57 & 0.12 \\ & & 0.81 & -0.40 & -0.72 \\ & & & 0.87 & -0.53 \\ & & & & 0.82 \end{bmatrix} \quad (8)$$

4. Simulation Study

In this section, we use simulated data to compare $CCor$ with several other similarity measures, namely Pearson correlation ($PCor$), partial correlation ($PartialCor$) and topological overlap measure (TOM , Zhang and Horvath (2005); Langfelder and Horvath (2008)). Partial correlation measures the association between two random variables by controlling the other random variables considered. It is widely used in constructing networks under the Gaussian assumption, which is also termed as Gaussian graphic models (Giudici and Green, 1999; Dobra et al., 2004; Jones et al., 2005). Under this model, all nodes are assumed to follow a multivariate normal distribution and the partial correlation matrix, also called precision matrix, is the inverse of correlation matrix of all nodes. Under the Gaussian graphical model, the two nodes are linked only when their partial correlation is not zero. Therefore, the estimated partial correlations could be used as a similarity measure between genes in the gene expression network. In this section, $PartialCor$ is calculated using the *GeneNet* R package (Schäfer and Strimmer, 2005). Information of TOM can be found in

Introduction and the values of *TOM* are calculated using the *WGCCA* R package (Langfelder and Horvath, 2008).

First, we perform a stability analysis to study how the *CCor* of a given gene pair changes when using different sets of variables. This is of interest as the set of genes may vary depending on the platforms for microarray analysis. We first generate a data set with p genes and n samples and then we fix a pair of genes and randomly delete 50 genes from the rest of genes in the data set and recalculate the *CCor*. We repeat this procedure 100 times and calculate the mean and standard deviation of the estimated *CCors* and compare them to the *CCor* using all the genes. We consider two different settings in our simulations: 1) two-module case; 2) five-module case and both with correlation structures as shown in (2). For each simulation setting, the representing *CCors* and corresponding concentrating values of all the within and between-module pairs are calculated as that in the example shown in Section 3. Results are shown in the Supplementary Materials. It can be seen that in all the settings, the estimation of *CCor* is stable and close to the one calculated using all the genes with little variations across 100 replicates. We note that the concentration inequality (5) always holds irrespective of the particular subset of genes used. However, the concentrating values do depend on the correlation structure of the subset genes. The results show that there is little difference in the concentrating values based on all the genes and those based on a subset of genes.

After stability analysis, we use simulated data to compare the performance of the four similarity measures mentioned above. For each setting, 100 independent datasets are generated. For each dataset we label within-module pair as 1 and between-module pair as 0. Then we use different similarity measures to classify each pair into either a within-module pair or between-module pair and calculate the mean *AUC* across the 100 replicates for each similarity measure. We generate data with different sample sizes and noise levels. We consider a 5-module case and a 10-module case. Results are shown in Tables 1 and 2. In both cases, *CCor* consistently outperforms the other similarity measures especially when the sample size is much less than the total number of genes. When the sample size increases, the performance of *CCor*, *PCor* and *TOM* becomes similar, with the *PartialCor* has the worst performance across all cases. In our simulations, the correlation matrix has a blockwise structure and thus the precision matrix also has such structure, that is, the partial correlations in the diagonal blocks (within-module) are larger than those in the off-diagonal blocks (between-modules).

5. Real Data Examples

In this section, we evaluate similarity measures using the approach discussed in Song et al. (2012). We first performed clustering using different similarity measures and then use the resulting clusters (modules) to evaluate similarity measures in three data sets. In real data applications, the underlying module structure is not known. The key to evaluate a detected module is the biological interpretability. Pathway enrichment analysis is a common way to evaluate the quality and interpretability of a detected module since in gene co-expression networks, modules are comprised of highly interconnected genes and are often highly enriched in genes with shared functional annotations such as pathways. An enrichment test

can be used to test the overlap between a detected module and known pathways, and a smaller p value suggests more statistical evidence of overlap between the inferred modules and known pathways. Therefore, for each inferred module, we record the p values of the five most significant enriched pathways and pool these p values of all detected modules together. We use the mean and variance of these p values to evaluate the performance of a measure in real applications. Adopting the general strategy in Song et al. (2012), we assign modules by hierarchical clustering with $1 - \text{similarity}$ as the distance measure and using a dynamic tree cutting method (Langfelder et al., 2008) to cut the resulting hierarchical tree. Note that for *PCor* and *PartialCor*, we used $1 - |PCor|$ and $1 - |PartialCor|$ as the distance measure for clustering. We chose Gene Ontology (GO) pathways for enrichment analysis and for GO enrichment analysis of resulting modules, we used *GOSTats* R package (Falcon and Gentleman, 2007).

The first data set is from Kumari et al. (2012), which measured the gene expression profiles of 108 Arabidopsis samples under the same condition to study stress response (denoted as salt, series GSE7636, 7639, 7641, 7642, 8787, 5623 in GEO). The second data set, reported by Booker et al. (2012), contains expression profiles of 27 Arabidopsis samples (denoted as GSE34667). The third data set is from Robinson et al. (2012), which contains expression profiles of 76 pediatric medulloblastoma samples (denote as medulloblastoma, series GSE37418). For all studies, we downloaded the raw data and performed normalization using the *a y* R package (Gautier et al., 2004). For each data set, we selected genes with the largest variance. More specifically, we included 2000 genes in the salt data set, 1000 genes in the GSE34667 data set and 1500 for the medulloblastoma data set.

For each similarity measure, we recorded the five best GO enrichment p-values from all resulting modules and pooled them together. The average values of the pooled p values of all similarity measures are compared and the results are shown in Figure 4. It can be seen that in these real data sets, *CCor* outperforms *PCor* and *PartialCor* and has similar performance with *TOM* in terms of gene enrichment analysis. These results suggest that by borrowing information across all the other genes, *CCor* could better capture the similarity between two genes and identify more biologically meaningful modules. Among all the methods compared, *mCCor* performs the best, which is in accordance with the discussion in Section 2. In gene co-expression analysis, the sizes of modules are usually unbalanced and the number of modules may also be large. Therefore, *mCCor* might be a better alternative of *CCor* when the number of genes included in analysis is large. In these three applications, the threshold of *mCCor* was chosen as the 60% quantile of Pearson correlation coefficients, which resulted in 0.4 for the salt data, 0.45 for the GSE34667 and 0.25 for the medulloblastoma data. The weight coefficient was chosen as 5 for these data sets. As shown in the Supplementary Materials, *mCCor* is stable across different choices of threshold and weight. However, it is worth further investigation on the ‘optimal’ choices of threshold and weight.

6. Discussion

This article presents a new similarity measure *CCor* to quantify the relationship between a pair of genes. We derive a concentration inequality of *CCor* with the assumption that the

expression levels are Gaussian distributed and modules are defined by covariance matrices, i.e. blocks with higher correlations. Simulation studies show the good approximation of this concentration inequality. Compared to Pearson correlation, we have found that *CCor* could amplify the difference between gene pairs from the same module and those from different modules, and therefore, better infer gene modules. In comparison with other methods, *CCor* yields the best performance in detecting gene pairs in the same module in simulation. And in real data applications where module sizes are usually unbalanced, *mCCor* shows its power and performs the best across three datasets in terms of finding more biologically meaningful results.

We note that there has been some work on the analysis of correlation coefficients (Elston, 1975; Parmigiani et al., 2004; Shankavaram et al., 2007). In Elston (1975), the author performed an analysis on the distribution of a pair of correlation coefficients. He derived an asymptotic expression for the covariance between any two correlation coefficients when the covariance matrix has a specific structure. In Parmigiani et al. (2004), the main purpose was to measure the reproducibility of co-expression patterns from different datasets. The co-expression pattern of a dataset was represented by the Pearson correlation coefficients of genes. Then, the correlation coefficient of two vectors of correlation coefficients was used to measure the closeness of two co-expression patterns of two datasets. The corresponding elements in the two vectors are the correlation coefficients between the same pair of genes but calculated from different datasets. Shankavaram et al. (2007) focused on incorporating multiple microarrays to build a ‘consensus’ expression profile, which was later used to study the correlation between transcripts and protein expression. Briefly speaking, for each probe the authors calculated the correlation coefficients between each pair of datasets and chose the average of the two datasets with the largest correlation as the ‘consensus’ transcript. Compared with these studies, our work is unique as *CCor* is the correlation coefficient of two vectors of Pearson correlations (v_i and v_j in (1)) and the paired elements of the two vectors measure the correlations of different gene pairs from the same dataset. Furthermore, the idea of *CCor* could be extended to various forms, that is, the sample Pearson correlation in *CCor* can be substituted by any other similarity measure, for example, correlation of mutual information.

There are several limitations of *CCor*: first, its performance could be affected when the module sizes are unbalanced. In this situation, it may be hard to distinguish relatively small modules from larger modules. Our modified version of *CCor*, called *mCCor*, could address this issue to some extent. However, the choice of weight or threshold is not trivial and requires further investigation. Second, the computation burden of *CCor* is high. The time complexity for calculating the *CCor* of n genes is n^3 . Third, the theoretical advantage of *CCor* is based on the blockwise structure in the covariance matrix, by which the *CCor* could make full use of information of all other genes to increase the power of distinguishing different modules.

There are also challenges to compare the performance of different similarity measures as it is non-trivial to validate a co-expression measure. Previous literature has used pathway analysis, enrichment analysis and transcription factor analysis, which are based on the correlation between co-regulation and co-expression (Song et al., 2012; Zhang and Horvath,

2005; Wang et al., 2014; Kumari et al., 2012; Bar-Joseph et al., 2003). However, the relationship between co-regulation and co-expression is complicated and sometimes it might not be fair to use this relationship to measure the performance of a co-expression measure. With more ground truth in genetics uncovered, we could one day use a more rigorous and comprehensive way to compare the performance of different similarity measures between genes to uncover network structures from expression profiles.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would like to thank Jina Li, Jinzhe Li, Yu Lu, Qiongshi Lu, Tianqi Liu and Tao Wang for their insightful suggestions. Research was funded in part by a fellowship from the Yale World Scholars Program through the China Scholars Council, and the National Institutes of Health (NIH) grants R01 GM59507, P01 CA154295, and P30 CA016359 to the authors.

References

- Allen JD, Xie Y, Chen M, Girard L, Xiao G. Comparing statistical methods for constructing large scale gene networks. *PloS one*. 2012; 7:e29348. [PubMed: 22272232]
- Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, et al. Computational discovery of gene modules and regulatory networks. *Nature biotechnology*. 2003; 21:1337–1342.
- Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human b cells. *Nature genetics*. 2005; 37:382–390. [PubMed: 15778709]
- Booker F, Burkey K, Morgan P, Fiscus E, Jones A. Minimal influence of g-protein null mutations on ozone-induced changes in gene expression, foliar injury, gas exchange and peroxidase activity in *arabidopsis thaliana* l. *Plant, cell & environment*. 2012; 35:668–681.
- Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS. Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences*. 2000; 97:12182–12186.
- Cadeiras M, von Bayern M, Sinha A, Shahzad K, Latif F, Lim WK, Grenett H, Tabak E, Klingler T, Califano A, et al. Drawing networks of rejection—a systems biological approach to the identification of candidate genes in heart transplantation. *Journal of cellular and molecular medicine*. 2011; 15:949–956. [PubMed: 20497491]
- Daub CO, Steuer R, Selbig J, Kloska S. Estimating mutual information using b-spline functions—an improved similarity measure for analysing gene expression data. *BMC bioinformatics*. 2004; 5:118. [PubMed: 15339346]
- Dobra A, Hans C, Jones B, Nevins JR, Yao G, West M. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*. 2004; 90:196–212.
- Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*. 1998; 95:14863–14868.
- Elston R. On the correlation between correlations. *Biometrika*. 1975; 62:133–140.
- Falcon S, Gentleman R. Using gostats to test gene lists for go term association. *Bioinformatics*. 2007; 23:257–258. [PubMed: 17098774]
- Gautier L, Cope L, Bolstad BM, Irizarry RA. affyanalysis of affymetrix genechip data at the probe level. *Bioinformatics*. 2004; 20:307–315. [PubMed: 14960456]
- Giudici P, Green PJ. Decomposable graphical gaussian model determination. *Biometrika*. 1999; 86:785–801.
- Hoheisel JD. Microarray technology: beyond transcript profiling and genotype analysis. *Nature reviews genetics*. 2006; 7:200–210.

- Horvath S, Zhang B, Carlson M, Lu K, Zhu S, Felciano R, Laurance M, Zhao W, Qi S, Chen Z, et al. Analysis of oncogenic signaling networks in glioblastoma identifies aspm as a molecular target. *Proceedings of the National Academy of Sciences*. 2006; 103:17402–17407.
- Jones B, Carvalho C, Dobra A, Hans C, Carter C, West M. Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*. 2005:388–400.
- Kumari S, Nie J, Chen H-S, Ma H, Stewart R, Li X, Lu M-Z, Taylor WM, Wei H. Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. *PloS one*. 2012; 7:e50411. [PubMed: 23226279]
- Langfelder P, Horvath S. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*. 2008; 9:559. [PubMed: 19114008]
- Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r. *Bioinformatics*. 2008; 24:719–720. [PubMed: 18024473]
- Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, Califano A. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*. 2006; 7:S7. [PubMed: 16723010]
- Meyer PE, Lafitte F, Bontempi G. minet: Ar/bioconductor package for inferring large transcriptional networks using mutual information. *BMC bioinformatics*. 2008; 9:461. [PubMed: 18959772]
- Parmigiani G, Garrett-Mayer ES, Anbazhagan R, Gabrielson E. A cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clinical cancer research*. 2004; 10:2922–2927. [PubMed: 15131026]
- Priness I, Maimon O, Ben-Gal I. Evaluation of gene-expression clustering via mutual information distance measure. *BMC bioinformatics*. 2007; 8:111. [PubMed: 17397530]
- Robinson G, Parker M, Kranenburg TA, Lu C, Chen X, Ding L, Phoenix TN, Hedlund E, Wei L, Zhu X, et al. Novel mutations target distinct subgroups of medulloblastoma. *Nature*. 2012; 488:43–48. [PubMed: 22722829]
- Schäfer J, Strimmer K. An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics*. 2005; 21:754–764. [PubMed: 15479708]
- Schulze A, Downward J. Navigating gene expression using microarrays: a technology review. *Nature cell biology*. 2001; 3:E190–E195. [PubMed: 11483980]
- Shankavaram UT, Reinhold WC, Nishizuka S, Major S, Morita D, Chary KK, Reimers MA, Scherf U, Kahn A, Dolginov D, et al. Transcript and protein expression profiles of the nci-60 cancer cell panel: an integrative microarray study. *Molecular cancer therapeutics*. 2007; 6:820–832. [PubMed: 17339364]
- Song L, Langfelder P, Horvath S. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC bioinformatics*. 2012; 13:328. [PubMed: 23217028]
- Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. *science*. 2003; 302:249–255. [PubMed: 12934013]
- Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. 2010. arXiv preprint arXiv:1011.3027
- Wang YR, Waterman MS, Huang H. Gene coexpression measures in large heterogeneous samples using count statistics. *Proceedings of the National Academy of Sciences*. 2014; 111:16371–16376.
- Wang Z, Gerstein M, Snyder M. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*. 2009; 10:57–63.
- Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*. 2005; 4
- Zhou X, Kao M-CJ, Wong WH. Transitive functional annotation by shortest-path analysis of gene expression data. *Proceedings of the National Academy of Sciences*. 2002; 99:12783–12788.

Figure 1. A simple illustrative example

upper left: observations of two genes from same module; upper right: observations of two genes from different modules; lower left: x axis: Pearson correlation coefficients of X_1 and $X \setminus (X_1, X_2)$ (all the genes except genes 1 and 2), y axis: X_2 and $X \setminus (X_1, X_2)$; lower right: x axis: Pearson correlation coefficients of X_1 and $X \setminus (X_1, X_{p1+1})$ (all the genes except genes 1 and $p_1 + 1$), y axis: X_{p1+1} and $X \setminus (X_1, X_{p1+1})$

Figure 2. An example with unbalanced module sizes

Notations are similar to Fig. 1. upper left: x axis: $PCor$ of X_1 and $X \setminus (X_1, X_2)$, y axis: X_2 and $X \setminus (X_1, X_2)$ ($CCor_{1,2}$, similarity within first module); upper right: x axis: $PCor$ of X_1 and $X \setminus (X_1, X_{p1+1})$, y axis: X_{p1+1} and $X \setminus (X_1, X_{p1+1})$ ($CCor_{1,p1+1}$, similarity between first and second module); lower left: x axis: $PCor$ of $X_{p1+p2+1}$ and $X \setminus (X_{p1+p2+1}, X_{p1+p2+2})$, y axis: $X_{p1+p2+2}$ and $X \setminus (X_{p1+p2+1}, X_{p1+p2+2})$ ($CCor_{p1+p2+1,p1+p2+2}$, similarity within third module); lower right: x axis: $PCor$ of X_1 and $X \setminus (X_1, X_{p1+p2+1})$, y axis: $X_{p1+p2+1}$ and $X \setminus (X_1, X_{p1+p2+1})$ ($CCor_{1,p1+p2+1}$, similarity between first and third module);

Figure 3. The theoretical and empirical values of $CCor$

$p_1 = p_2 = 100$ and $n = 100$. Within-module and between-module similarities correspond to $CCor_{1,2}$ and $CCor_{1,p_1+1}$. Points in plots represent the empirical estimates by taking mean of $CCor$ calculated from 100 simulated datasets with error bars representing the standard deviation of the 100 estimates.

Figure 4. GO enrichment analysis comparing *CCor* and *mCCor* with other three measures in 3 data sets

5 best GO enrichment p values from all modules identified using each similarity measures are log transformed. Bar plots show the mean of these pooled p values and error bars stand for 95% confidence intervals. P values shown on top of the panels are obtained by applying Fisher's method on pairwise comparison p values of *mCCor* vs other three methods.

Table 1
Comparison of different methods based on AUC when sample size and noise level vary for the 5-module case. Noise level refers to the standard deviation of the error term

Setting	p	noise sd	n	$CCor$	$PCor$	$PartialCor$	TOM
$\begin{bmatrix} .5 & .3 & .3 & .2 & .3 \\ & .6 & .3 & .2 & .3 \\ & & .55 & .3 & .2 \\ & & & .55 & .25 \\ & & & & .5 \end{bmatrix}$	50	1	10	.64 (.05)	.61(.04)	.60(.02)	.57(.03)
			15	.69 (.04)	.64(.03)	.61(.02)	.62(.04)
			30	.80 (.04)	.70(.03)	.63(.01)	.70(.04)
			60	.91 (.02)	.78(.03)	.62(.01)	.80(.02)
			150	.99 (.01)	.89(.02)	.60(.00)	.90(.02)
		2	300	1.0 (.00)	.96(.01)	.60(.00)	.96(.01)
			10	.55 (.02)	.54(.02)	.54(.01)	.51(.01)
			15	.57 (.02)	.55(.02)	.55(.01)	.52(.01)
			30	.61 (.02)	.58(.01)	.57(.01)	.54(.01)
			60	.68 (.03)	.61(.02)	.58(.01)	.58(.02)
			150	.83 (.03)	.68(.01)	.59(.01)	.67(.02)
			300	.94 (.01)	.75(.01)	.59(.01)	.75(.02)

Table 2
Comparison of different methods using AUC when sample size and noise level vary, a 10-module case. Noise level refers to the standard deviation of error term

Setting		p	noise sd	n	$CCor$	$PCor$	$PartialCor$	TOM
[.65 .82 .7 .67 .75 .59 .75 .97 .58 -.02 .51 60]	.23	.15	1	10	.78(.04)	.75(.04)	.71(.02)	.67(.04)
				25	.91(.02)	.86(.02)	.73(.01)	.83(.03)
				50	.97(.01)	.93(.01)	.72(.00)	.92(.01)
				100	.99(.00)	.97(.00)	.70(.00)	.97(.01)
				200	1.0(.00)	.99(.00)	.64(.00)	.99(.00)
				300	1.0(.00)	.99(.00)	.61(.00)	1.0(.00)
	.27	.15	2	10	.62(.03)	.60(.03)	.59(.02)	.53(.02)
				15	.71(.02)	.67(.02)	.63(.01)	.55(.02)
				30	.82(.02)	.73(.02)	.66(.01)	.61(.02)
				60	.91(.01)	.81(.01)	.66(.01)	.69(.02)
				150	.97(.01)	.89(.01)	.65(.00)	.82(.01)
				300	.98(.00)	.92(.01)	.65(.00)	.91(.01)