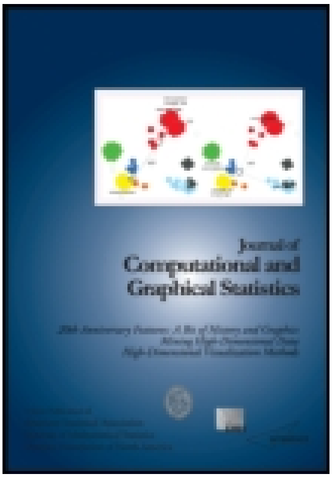

Journal of Computational and Graphical Statistics

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ucgs20

A Nonparametric Approach to Detect Nonlinear Correlation in Gene Expression

Y. Ann Chen[a], Jonas S. Almeida[a], Adam J. Richards[a], Peter Müller[a], Raymond J. Carroll[a] & Baerbel Rohrer[a]

[a] Y. Ann Chen is Assistant Member, Department of Biostatistics, Moffitt Cancer Center, Tampa, FL 33612 . Jonas S. Almeida is Professor, Department of Bioinformatics and Computational Biology, The University of Texas, M. D. Anderson Cancer Center, Houston, TX 77030. Adam J. Richards is Doctoral Candidate, Department of Biostatistics, Bioinformatics, and Epidemiology, Medical University of South Carolina, Charleston, SC 29425. Peter Müller is Professor, Department of Biostatistics, The University of Texas, M. D. Anderson Cancer Center, Houston, TX 77030. Raymond J. Carroll is Distinguished Professor, Department of Statistics, Texas A&M University, College Station, TX 77843. Baerbel Rohrer is Associate Professor, Departments of Ophthalmology and Neurosciences, Medical University of South Carolina, Charleston, SC 29425 .

Published online: 01 Jan 2012.

To cite this article: Y. Ann Chen, Jonas S. Almeida, Adam J. Richards, Peter Müller, Raymond J. Carroll & Baerbel Rohrer (2010) A Nonparametric Approach to Detect Nonlinear Correlation in Gene Expression, Journal of Computational and Graphical Statistics, 19:3, 552-568, DOI: 10.1198/jcgs.2010.08160

To link to this article: http://dx.doi.org/10.1198/jcgs.2010.08160

Supplementary materials for this article are available online.
Please click the JCGS link at http://pubs.amstat.org.

A Nonparametric Approach to Detect Nonlinear Correlation in Gene Expression

Y. Ann CHEN, Jonas S. ALMEIDA, Adam J. RICHARDS, Peter MÜLLER, Raymond J. CARROLL, and Baerbel ROHRER

We propose a distribution-free approach to detect nonlinear relationships by reporting local correlation. The effect of our proposed method is analogous to piecewise linear approximation although the method does not utilize any linear dependency. The proposed metric, maximum local correlation, was applied to both simulated cases and expression microarray data comparing the *rd* mouse with age-matched control animals. The *rd* mouse is an animal model (with a mutation for the gene *Pde6b*) for photoreceptor degeneration. Using simulated data, we show that maximum local correlation detects nonlinear association, which could not be detected using other correlation measures. In the microarray study, our proposed method detects nonlinear association between the expression levels of different genes, which could not be detected using the conventional linear methods. The simulation dataset, microarray expression data, and the Nonparametric Nonlinear Correlation (NNC) software library, implemented in Matlab, are included as part of the online supplemental materials.

Key Words: Local correlation; Microarray gene expression; Nonlinear correlation; Nonparametric.

1. INTRODUCTION

Although nonlinear relationships between the expression levels of genes or gene products are expected (Kitano 2002a, 2002b) and observed (Rohrer et al. 2004) in biological datasets, there is no commonly used statistic quantifying nonlinear correlation that can find a similarly generic use as Pearson's correlation coefficient for quantifying linear cor-

Y. Ann Chen is Assistant Member, Department of Biostatistics, Moffitt Cancer Center, Tampa, FL 33612 (E-mail: *ann.chen@moffitt.org*). Jonas S. Almeida is Professor, Department of Bioinformatics and Computational Biology, The University of Texas, M. D. Anderson Cancer Center, Houston, TX 77030. Adam J. Richards is Doctoral Candidate, Department of Biostatistics, Bioinformatics, and Epidemiology, Medical University of South Carolina, Charleston, SC 29425. Peter Müller is Professor, Department of Biostatistics, The University of Texas, M. D. Anderson Cancer Center, Houston, TX 77030. Raymond J. Carroll is Distinguished Professor, Department of Statistics, Texas A&M University, College Station, TX 77843. Baerbel Rohrer is Associate Professor, Departments of Ophthalmology and Neurosciences, Medical University of South Carolina, Charleston, SC 29425 (E-mail: *rohrer@musc.edu*).

Journal of Computational and Graphical Statistics, Volume 19, Number 3, Pages 552–568
DOI: 10.1198/jcgs.2010.08160

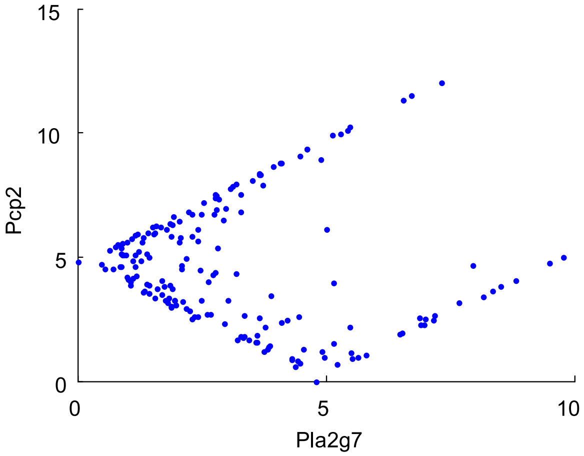

Figure 1. An example of nonlinear correlation observed in the microarray study (in Section 4). Nonlinear correlation can be detected by maximal local correlation ($M = 0.93$, $p = 0.007$), but not by Pearson correlation ($C = -0.08$, $p = 0.88$) between genes *Pla2g7* and *Pcp2* (i.e., between two columns of the distance matrix). *Pla2g7* and *Pcp2* are negatively correlated when their transformed levels are both less than 5. These two genes are, otherwise, positively correlated. More experimental and data processing details are provided in Section 4. A color version of this figure is available in the electronic version of this article.

relation. Many analyses of complex biological phenomena still approach the biological question of correlation using global linear association only. Mutual information has been used to detect nonlinear correlation in a microarray expression dataset, implicitly (Kasturi and Acharya 2005) and explicitly (Steuer et al. 2002; Daub et al. 2004). However, different estimation methods of the mutual information lead to different conclusions. We propose a method to quantify a global nonlinear relationship by reporting local correlation. Local here refers to the correlation that exists either at a certain scale or under certain condition(s), and is therefore also referred to as transient correlation. An example is illustrated in Figure 1 that includes partial signals that are positively correlated and partial signals that are negatively correlated. However, neither linear dependency is assumed nor used in our method. That is, the association need not be linear to be detected at either the local or global scale. Our proposed method quantifies correlation by examination of bivariate distances between data points and was inspired by the use of the correlation integral in a fluctuating dynamic system (Grassberger and Procaccia 1983).

The correlation integral examines the cumulative distribution of distances between data points of a time series. With proper modification, we show that the correlation integral can be used to estimate global patterns and global association. We develop statistics using the correlation integral to estimate local correlation. We only focus on bivariate correlation.

The need for quantification of nonlinear relationships is particularly acute for expression microarray data, where a massive number of variables and a wide range of biological processes involved in a typical experiment represent a particularly convoluted version of the proverbial "search for a needle in a haystack." Aside from the nonlinearity, several additional challenges (Quackenbush 2001) are commonly found in high-throughput biological datasets: (1) different datasets have different scales, (2) the scale that is potentially biologically relevant is often unknown at the exploratory stage of the research, (3) there exist high noise levels (Marshall 2004), and (4) the sampling distribution is generally unknown; very seldom if ever is it normally distributed. Multimodality is not uncommon. We propose a generic method to quantify nonlinear correlation by reporting local correlation, with the option of removal of the scaling effects between different measurements, which will (1) detect association at multiple scales, (2) be insensitive to noise, and (3) not rely on distribution assumptions, that is, a nonparametric method.

In Section 2, the correlation integral is introduced and we define the proposed measures of local correlation and of correlation change. Local correlation measures are used to describe the relationships across experimental units, genes in the case of our motivating application. On the other hand, correlation change allows us to identify experimental units whose relationships differ across two conditions, in our case gene expression in a treatment group versus a control group. In Section 3 we validate our method on simulated cases. Finally, in Section 4, we use the proposed measures to quantify nonlinear association change in microarray expression between a treatment group and a control group in an animal experiment. The treatment group are mice exhibiting photoreceptor degeneration (*rd*) and the control group are wild type mice (*wt*). The *rd* mouse is also referred to as the *rd1* mouse in the literature. The generality of the proposed method makes it appropriate to many other types of data, such as those generated by proteomics or metabolomics.

2. LOCAL CORRELATION

We first describe the proposed summary of bivariate local correlation in words. Formal definition of the proposed method and the notations will follow in the next few paragraphs. First, each variable is transformed such that the marginal distribution is uniform. This is achieved by transforming to ranks (in ascending order) followed by a linear transformation. Let N denote the sample size. The linear transformation is achieved by subtracting the minimum rank (i.e., 1, in the absence of ties) from the ranks and then divided by the difference between maximum rank and minimum rank (i.e., $N-1$, in the absence of ties). The rank transformation is the default setting in our implementation. Alternatively, this preprocessing step can be omitted if the raw data are already on comparable and nonarbitrary scales. Next, we evaluate the neighbor density, which records the rate of change of the number of observations within a neighborhood of a given radius. We then compare the observed neighbor density with the neighbor density under the null hypothesis of no linear or nonlinear association. The difference defines the proposed measure of local correlation. Finally, we define maximum local correlation to describe overall bivariate nonlinear correlation.

The definition of the proposed method is based on the concept of correlation integrals. Consider a time series, z_i, $i = 1, \ldots, N$. The correlation integral $I(r)$ is defined as (Grassberger and Procaccia 1983)

$$I(r) = \lim_{N\to\infty}\left\{\frac{1}{N^2}\sum_{i,j=1}^{N} I(|z_i - z_j| < r)\right\}.$$

The correlation integral quantifies the average cumulative number of neighbors within radius r. The definition remains meaningful also when the data are not a time series.

To develop a measure of association between vectors, x and y, we modify the definition of $I(r)$ as follows. Let $z_i = (x_i, y_i)$, $i = 1, \ldots, N$, denote the observations in the dataset, and let $|z_i - z_j|$ denote Euclidean distance. We define $\hat{I}(r) = \frac{1}{N^2}\sum_{i,j=1}^{N} I(|z_i - z_j| < r)$. The observed distances are further linearly transformed between 0 and 1 before quantifying $\hat{I}$. Note that $\hat{I}$ has the property of a cumulative distribution function (cdf). It is nondecreasing from 0 to 1 and continuous from the right. The function $\hat{I}(r)$ describes the global pattern of neighboring distances.

Our primary interest is the definition of a metric to quantify nonlinear association by reporting local patterns. Therefore the neighbor density D is devised as the derivative of $\hat{I}$:

$$\widehat{D}(r) = \Delta\hat{I}(r)/\Delta r,$$

where $\Delta\hat{I}(r)$ denotes change in $\hat{I}(r)$. The observed neighbor density is evaluated at the discrete radius r, where $r = 0, 1/m, 2/m, \ldots, 1$, and m is an arbitrary grid size that determines $\Delta r = 1/m$. An automatic smoother using cross-validation to choose an optimal window size (Vilela et al. 2007) is applied to smooth $\widehat{D}(r)$. Any traditional smoothing algorithms with proper choice of smoothing window size could alternatively be used here. In our study, the default size m is set as N, the number of observations. The statistic $\widehat{D}$ is a discrete approximation of $d\hat{I}(r)/dr$, which has the formal properties of a probability density function (pdf). Therefore, with a slight abuse of terminology we refer to $\widehat{D}(r)$ as a distribution. An example of a correlation integral and a neighbor density is illustrated in Figure 2.

Local correlation (ℓ). Intuitively, the distances between data points between two correlated variables would differ from that between two uncorrelated variables. Let $\widehat{D}_0(r)$ denote the estimate of a null distribution, which is composed of two vectors without any association. We define local correlation (ℓ) as the deviation of $\widehat{D}$ from that of the null

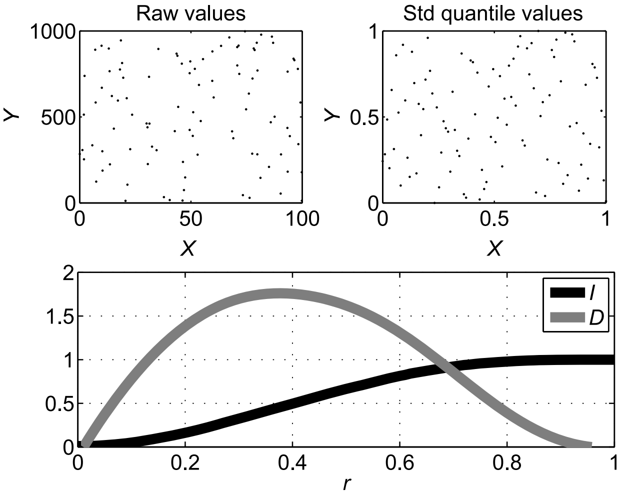

Figure 2. An example of the correlation integral ($\hat{I}$) and neighbor density ($\widehat{D}$) for a null distribution, that is, in the absence of correlation between x and y.

distribution at a given neighboring distance r:

$$\ell(r) = \widehat{D}(r) - \widehat{D}_0(r). \tag{2.1}$$

Our approach does not assume any parametric distribution. The flexibility of this method makes it easy to change the null distribution to any distribution of interest. The null hypothesis H_0 of no difference between the observed neighbor density $\widehat{D}$ and $\widehat{D}_0$ is $H_0 : \ell(r) = 0$ for $r = 0, 1/m, \ldots, 1$. Let $z^\star(b)$, $b = 1, 2, \ldots, B$, denote permutation replicates (see Appendix A for details) and let $\ell^\star(r, b)$ be corresponding permutation evaluations of $\ell(r)$. A two-sided p-value can be evaluated by (Efron and Tibshirani 1993)

$$p(\ell, r) = \#\bigl\{|\ell^\star(r, b)| > |\ell(r)|\bigr\}/B, \tag{2.2}$$

where $r = 0, 1/m, \ldots, 1$. Multiplicity correction could be carried out to adjust for the m tests performed. This could be done through the control of either false discovery rate (FDR) or experiment-wide error. In the following sections, we make use of these two alternative approaches for different aspects of the inferences.

Maximal local correlation (M). Local correlation is further used for the definition of a summary statistic, maximum local correlation M, to describe overall nonlinear association. Using an idea similar to the Kolmogorov–Smirnov statistic, maximum local correlation is defined, as its name implies, by

$$M = \max_r \bigl\{|\ell(r)|\bigr\}.$$

The interpretation of $\ell(r)$ as the difference of two pdf's implies that M can be interpreted as distance under the supremum norm of D and D_0. In other words, we define the statistic M as the maximum deviation between two underlying neighbor densities. Statistical significance of M is assessed in a permutation procedure similar to (2.2). See Appendix B for details.

Correlation change (δ_M). Recall that the motivating application for the development of the local correlation measures ℓ and M is an application for microarray data analysis. We will report details of the experiment and the data later, in Section 4. The setup is such that local correlation between responses for gene i and responses for gene j can be interpreted as measuring the relationship between genes i and j. Like many microarray experiments, the data include measurements under two biologic conditions, *wt* and *rd*. The ultimate inference goal is to understand how the two different biologic conditions affect the gene functions. We formalize this inference goal by considering the change in local correlation between the two conditions.

Building upon maximal local correlation M, we propose a statistic δ_M to measure association change. The motivation is that nonlinear behavior in biology can reflect the fact that the molecular machinery that is underlying biological processes is reconfigured to changes occurring in physiological or diseased states. In the microarray data example, for each pair of genes (ij) we want to test whether there is a change of nonlinear correlation between *rd* and *wt* mice. We therefore define a statistic δ_M to identify changes in maximal local correlation. We will later use it to identify critical genes in the disease development process.

The statistic δ_M is defined as

$$\delta_M = M^{rd} - M^{wt}. \tag{2.3}$$

The results using our proposed methods are compared to those obtained using existing methods, including Pearson's linear correlation (C), Spearman's rank correlation coefficient, and mutual information (MI) as a nonlinear approach. Mutual information is evaluated using the R function `mutual_information2()` (Daub et al. 2004). The default setting was used for the R function. See Appendix C for a definition of MI. MI is a summary of (nonlinear) association between x and y. We use it for a comparison of various performance summaries in the examples. No details are needed for the upcoming discussion. See Daub et al. (2004) for more details.

The statistical significance for C and MI is estimated using the same permutation procedures as for M to ensure comparability. See (A.1) and (A.2) in Appendix A. Similarly to (2.3) we define

$$\delta_S = S^{rd} - S^{wt}, \quad \text{where } S = M, C, \text{MI}. \tag{2.4}$$

Permutation tests for δ_S are defined in (B.1) in Appendix B.

3. SIMULATION STUDY

Twelve cases representing linear or different nonlinear relationships were considered (Figure 3). Case 1 is composed of independent vectors x and y. The observations of x and y in Case 2 have a perfect linear relationship. Case 3 is composed of seven nondecreasing broken straight lines while Case 4 has eight broken straight lines. Case 5 is a continuous monotonically increasing curve and Case 6 has three broken monotonically increasing curves; Case 7 is composed of three segments of sine waves. Case 8 is a mixture distribution of Cases 1 and 4 while Case 9 is a mixture of Cases 1 and 7. The three clusters in Case 10 are randomly sampled 100 data points from dataset 1 from the work of Jonnalagadda and Srinivasan (2004). Case 11 is a mixture distribution of Cases 1 and 10, that is, a mixture of three clusters along with some background noise. Two-fifths of the data points in Case 11 are randomly sampled from Case 1 while $3/5$ of the data are randomly sampled from Case 10. Case 12 includes five clusters with $3/5$ of the data points randomly sampled from Case 10 and the remaining $2/5$ of additional clusters (Figure 3). The dataset (Jonnalagadda and Srinivasan 2004) and the Matlab scripts to generate the simulated cases are included as supplemental materials. We evaluate ℓ and M for each case. For each case with stochastic components, that is, Cases 1 and 8 through 12, we simulated 100 realizations to estimate positive rates of testing for significant maximum local correlation M under repeat simulation. For Case 1 the (false) positive rate is a Type I error. For Cases 8 through 12 the (true) positive rate is interpreted as statistical power. The results are summarized in Table 1. Maximum local correlation M was computed for the raw data values, without rank transformation, to keep results comparable with other correlation methods (Pearson's correlation, Spearman's correlation, and mutual information, which will be described later).

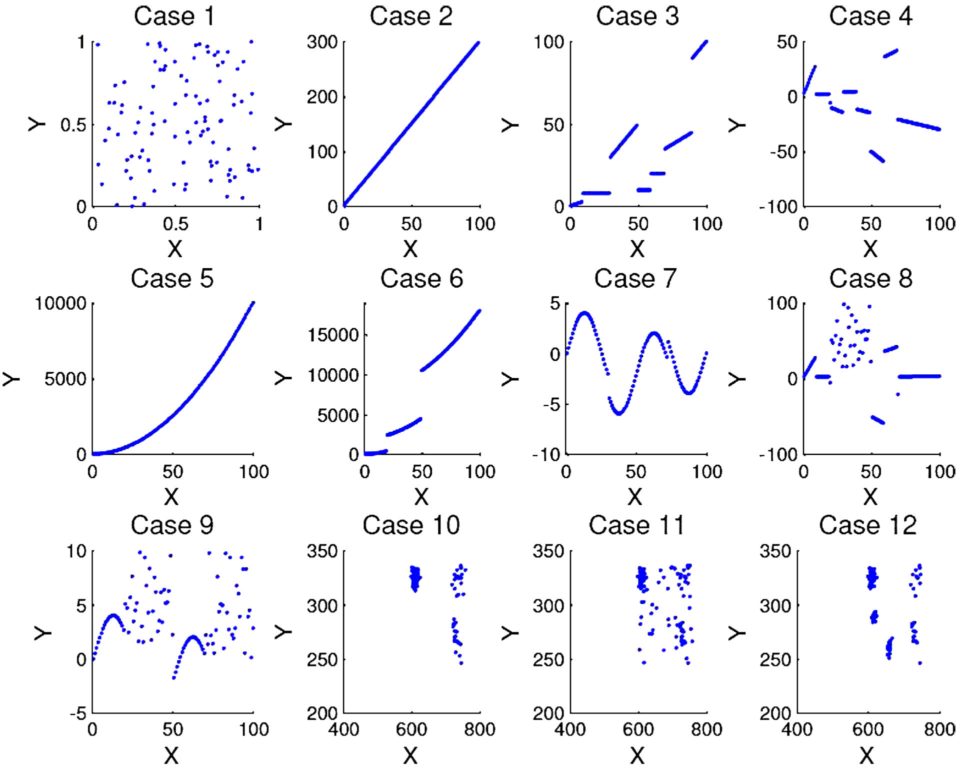

Figure 3. Twelve simulated cases. (1) Random points; (2) a straight line; (3) broken nondecreasing straight lines; (4) broken straight lines; (5) a continuous monotonically increasing curve; (6) broken monotonically increasing curves; (7) three partial sine waves; (8) a mixture distribution of Cases 1 and 4; (9) a mixture distribution of Cases 1 and 7; (10) three clusters; (11) a mixture distribution of Cases 1 and 10; (12) five clusters. A color version of this figure is available in the electronic version of this article.

Local correlation (ℓ). Figure 4 shows the local correlations $\ell(r)$ from one simulation for each of the 12 cases. Significant local correlations were found in all simulated cases, except (correctly) for Case 1. Here, significance is determined after a Bonferroni correction ($p(\ell, r) \leq 5 \cdot 10^{-4}$).

Positive local correlation means that the observed neighbor density is greater than the neighbor density under the null hypothesis when evaluated at radius r. Negative local correlation means that the observed neighbor density is lower than the neighbor density under the null hypothesis. Significant local correlations are detected at almost all scales (of radius r) in Case 2, which is a boundary condition with perfectly linear relationship. The profile of ℓ peaks near 0 and decreases along r. Observed multimodal $\ell(r)$ for all cases except Case 1 showed significant correlation at multiple scales. Statistically significant local correlation was observed for Case 11 although it contains both the signals of Case 10 and the noise from Case 1.

Maximal local correlation (M). Maximal local correlation is a summary statistic to test global nonlinearity while ℓ reports local correlation as a function of the neighborhood width of interest r. All cases with only deterministic components, that is, Cases 2 to 7, had statistically significant maximum local correlation M at significance level $\alpha = 0.05$

Table 1. Comparison of local and global correlations. Listed are positive rates for Cases 1, 8 to 12, and correlation measurement (with estimated p-value) for Cases 2 to 7. For Case 1 the (false) positive rate is Type I error. For Cases 8 through 12 the (correct) positive rate is statistical power.

Case	M	MI	C	Sp
		Positive rates		
1	3%	1%	3%	3%
8	100%	100%	96%	3%
9	100%	100%	22%	11%
10	100%	100%	100%	99%
11	100%	99%	100%	98%
12	100%	100%	98%	100%
		Correlation measurement (p-values)		
2	1.60 ($< \epsilon$)	3.32 ($< \epsilon$)	1.00 ($< \epsilon$)	1.00 ($< \epsilon$)
3	1.55 ($< \epsilon$)	2.01 ($< \epsilon$)	0.76 ($< \epsilon$)	0.82 ($< \epsilon$)
4	1.47 ($< \epsilon$)	2.14 ($< \epsilon$)	−0.40 ($< \epsilon$)	−0.59 ($< \epsilon$)
5	0.08 ($< \epsilon$)	2.39 ($< \epsilon$)	0.97 ($< \epsilon$)	1.00 ($< \epsilon$)
6	0.19 (0.0005)	2.60 ($< \epsilon$)	0.97 ($< \epsilon$)	1.00 ($< \epsilon$)
7	1.60 ($< \epsilon$)	1.67 ($< \epsilon$)	−0.37 ($< \epsilon$)	−0.40 ($< \epsilon$)

NOTE: $\epsilon = 2.50 \cdot 10^{-4}$. For the cases with stochastic components, that is, Cases 1 and 8 to 12, the estimated positive rate is listed based on 100 simulations for each case. M: maximum local correlation; MI: mutual information; C: Pearson correlation; Sp: Spearman correlation.

(Table 1). The stochastic component in the remaining Cases 1 and 8 through 12 allows us to evaluate positive rates across repeat experimentation. Except for Case 1, we (correctly) detect significant maximal local correlation for all 100 repeat simulations performed for each of these cases, that is, 100% true positive rates for Cases 8 through 12. Positive rate for these cases is interpreted as power. The (false) positive rate for maximal local correlation for Case 1 is 3% (Table 1). It is interpreted as Type I error.

Comparison with other correlation coefficients. Pearson's (C), Spearman's (Sp), maximum local correlation (M), and mutual information (MI) for the 12 simulated cases are compared in Table 1. The results of Pearson's and Spearman's correlations showed that the majority of cases have significant global linear correlation, that is, downward or upward trends, except for Case 1. When the global upward or downward trend is weak, that is, Cases 8 and 9, the statistical power for Pearson's and Spearman's correlations is lower than that of M and MI. The performances of M and MI are comparable.

4. APPLICATION TO MICROARRAY EXPRESSION DATA

We analyze microarray expression data that were collected to identify critical genes in photoreceptor degeneration (Rohrer et al. 2004). The samples were collected from age-matched wild type controls (C57BL/6; abbreviated as *wt*) and *rd* mice with rod degeneration at five postnatal time points (6, 10, 14, 17, and 21 days of age). Retinas from four animals per genotype per time point were pooled, and biological duplicates were obtained. Each probe is treated as an independent unit (although some genes have more than one probe on the array). Gene expression data were filtered based on reproducibility for the original analysis, leaving 181 genes for the analysis.

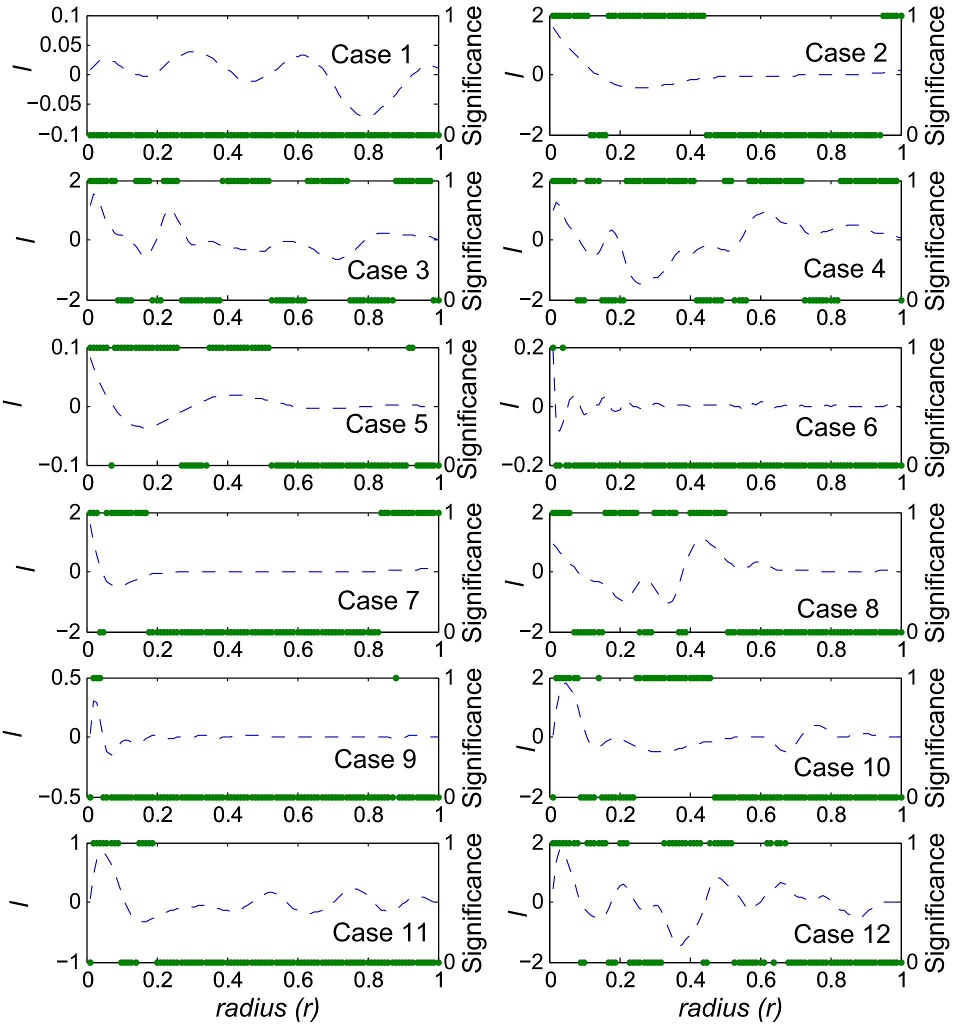

Figure 4. Local correlation between x and y for one simulated dataset from each of the 12 simulated cases. The dashed curves are the estimated local correlations ($\ell(r)$) (using the y-axis on the left). The solid dots indicate the statistical significance of local correlations (using the y-axis on the right). It is labeled as 1 if local correlation is significant after Bonferroni correction, 0 otherwise. Local correlations (and the associated significance) are plotted against the radius (r). A color version of this figure is available in the electronic version of this article.

Correlation change (δ_M, δ_C, and δ_{MI}) was evaluated based on the Euclidean distances of expression profiles between *wt* and *rd* mice. We proceeded as follows. Average expression values of the duplicates were calculated for *wt* and *rd* mice, respectively. Next, distance matrices, d_{ij}^{wt} and d_{ij}^{rd}, were generated for the *wt* and *rd* mouse, respectively, with the (i, j) entry denoting the Euclidean distance between the ith and the jth genes using the data from the time course (Figure 5). For each pair of the genes (i, j) we evaluated M, C, and MI between the ith and jth columns of the distance matrix, resulting in a total of 16,290 pairwise correlations pairs, one for *wt* and one for *rd*. The correlation is measuring the relationship between how gene i interacts with all the other genes versus how gene j interacts with all the other genes. Finally, correlation change between *wt* and *rd* mice, δ_S

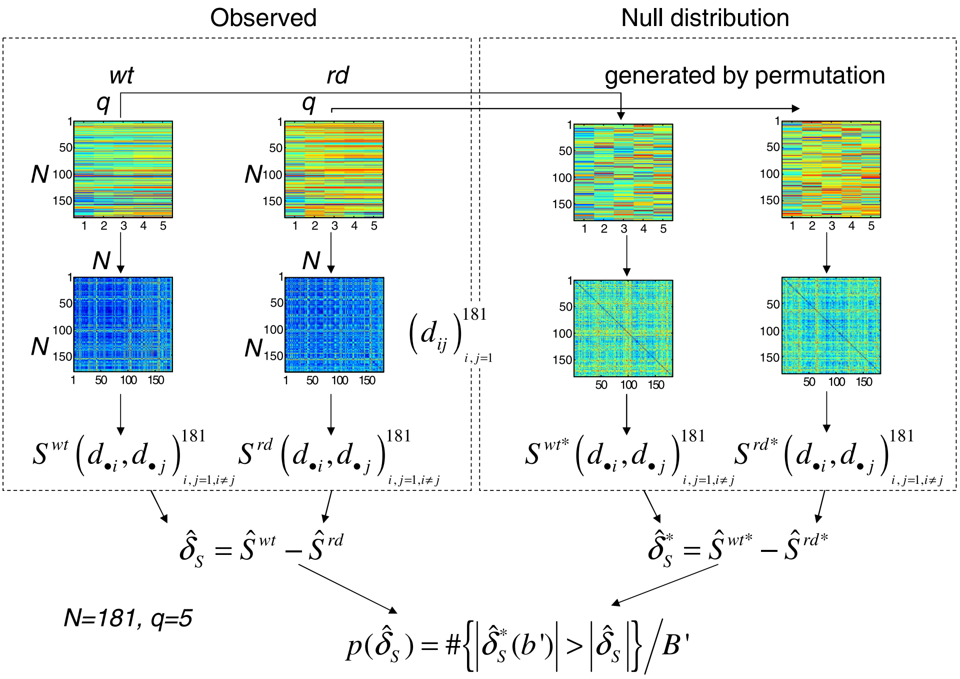
Observed
Null distribution
wt
rd
generated by permutation
q
q
N
N
N
N
1
50
100
150
1 2 3 4 5
1 50 100 150
$(d_{ij})_{i,j=1}^{181}$
$S^{wt}(d_{\bullet i}, d_{\bullet j})_{i,j=1,i\neq j}^{181}$
$S^{rd}(d_{\bullet i}, d_{\bullet j})_{i,j=1,i\neq j}^{181}$
$S^{wt*}(d_{\bullet i}, d_{\bullet j})_{i,j=1,i\neq j}^{181}$
$S^{rd*}(d_{\bullet i}, d_{\bullet j})_{i,j=1,i\neq j}^{181}$
$\hat{\delta}_S = \hat{S}^{wt} - \hat{S}^{rd}$
$\hat{\delta}_S^* = \hat{S}^{wt*} - \hat{S}^{rd*}$
N=181, q=5
$p(\hat{\delta}_S) = \#\{|\hat{\delta}_S^*(b')| > |\hat{\delta}_S|\}/B'$

Figure 5. Data setup and permutation test to test association change for microarray expression data. Association change δ_S ($S = M, C$, and MI) between distances (d_{ij}) of expression profiles between the control (*wt*) and treatment (*rd*) group. In our study, $N = 181$, and $q = 5$. A color version of this figure is available in the electronic version of this article.

($S = M, C,$ MI), was evaluated using (2.4). The values of association change based on Spearman's rank correlation coefficient and Pearson's correlation coefficient are almost identical. Also, these two methods had very similar performance on 12 simulated cases. Therefore, $\delta_{Spearman}$ is not included in the results for the expression data.

Nonlinear correlation in expression data. Plotting Pearson's correlation C against M for each of the pairwise correlations results in a W-shaped relationship (Figure 6). That is, when Pearson correlation C is low, maximal local correlation M could be high, the portion reflected by the middle of the W. This phenomenon, however, is not vice versa. In addition, two sides of the W indicate that when C is high, the value of M is also high as well. In other words, this W-shaped relationship indicates that maximum local correlation can capture local correlation observed in the biological dataset, which cannot be detected using Pearson's correlation C. Figure 1 shows an example of correlation between two genes that can only be detected by nonlinear correlation measures ($M = 0.93$, $p = 0.007$; MI $= 1.15$, $p = 0.01$) but not by linear correlation ($C = -0.08$, $p = 0.88$; $S = -0.21$, $p = 0.67$). This is because signals are partially positively and partially negatively correlated (e.g., Figure 1). The agreement between M and MI is high across all pairs of genes ($C = 0.81$, $p \approx 0$).

4.1 IDENTIFICATION OF CRITICAL GENES BY CORRELATION CHANGE

Association changes (δ_M, δ_C, and δ_{MI}) between *wt* and *rd* animals were estimated for each pair of genes. Their p-values were estimated using (B.1) in Appendix B and corre-

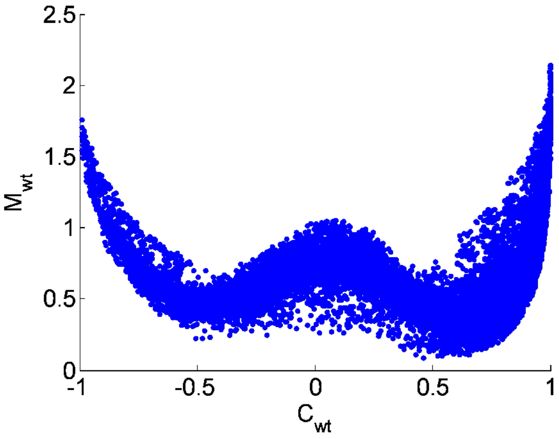

Figure 6. The "W" shape relationship between Pearson correlation (C) and maximum local correlation (M). This indicates that M captures additional information. Some local correlation can only be detected by M, but not by C. A color version of this figure is available in the electronic version of this article.

sponding q-values were estimated (Storey and Tibshirani 2003). A total of 75 genes with at least one significant association change was detected using δ_C when controlling FDR at 0.05. In a similar manner, 137 genes with at least one significant association change were detected using δ_M, and 67 genes were detected using δ_{MI} when controlling FDR for each statistic at 0.05. The ranked importance of the candidate genes involved in rod degeneration is based on the frequency of association change with statistical significance. Some association changes are detected by both δ_M and δ_C statistics, but not all (Figure 7(a)). The linear correlation between δ_M and δ_C is weak but significant ($C = 0.03$, $p \approx 4.4 \cdot 10^{-5}$). Out of the selected candidate genes with significant association changes between control and disease states using δ_M and δ_C, 62 of the probes were selected by both statistics, 75 of them could only be detected using δ_M, whereas 13 of them were only detected using δ_C. The overall agreement between δ_M and δ_{MI}, in contrast, is high ($C = 0.74$, $p \approx 0$), especially for large values (Figure 7(b)). This is consistent with the earlier observation that M and MI have relatively high concordance. When controlling FDR at 0.05 for each statistic, δ_M can detect more association changes (170 changes) between *rd* and *wt* mouse with statistical significance than those detected by δ_{MI} (58 changes; Figure 7(b)). Figure 8 shows an example of an association change between genes, and this association change can only be detected using δ_M ($\delta_M = 1.22$, $p \approx 0$), but not by δ_C ($\delta_C = 0.47$, $p = 0.32$) nor by δ_{MI} ($\delta_{\text{MI}} = 1.32$, $p = 0.001$). In the *wt* mouse, the correlation between the two genes is linear, but the association becomes a nonlinear (partially positively and partially negatively) correlation in the *rd* mouse.

4.2 Validation of the Selected Candidate Genes

Given the current status of the literature on photoreceptor degeneration, much of the physiology still remains unclear. Knowing this, the results for δ statistics were evaluated through the following means. First, based on the fact that photoreceptor cell death in the *rd* mouse retina is caused by a mutation in *Pde6b*, leading to the complete loss of *Pde6b*

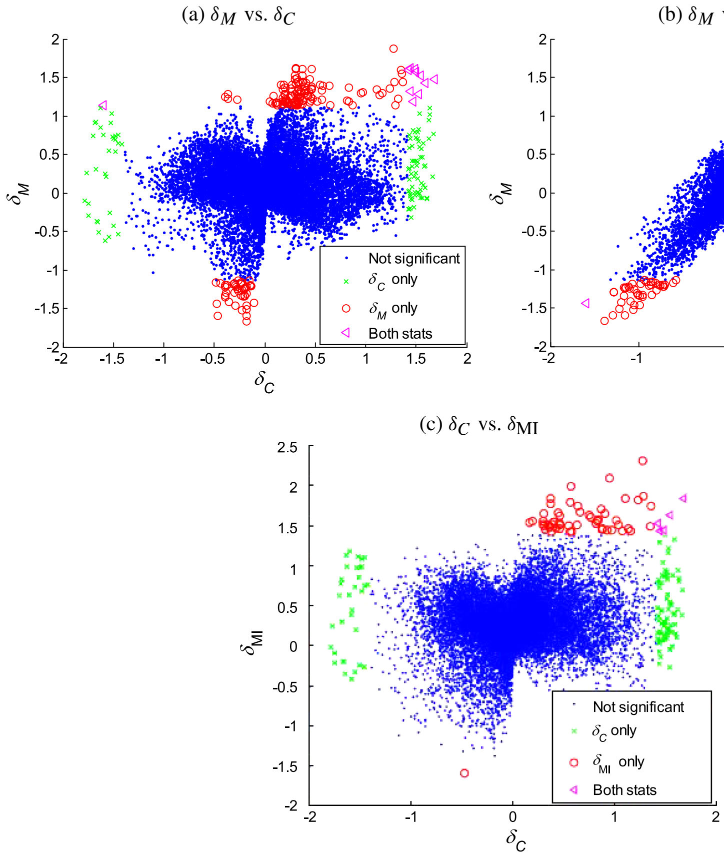

Figure 7. Scatterplot between different association changes measured between *wt* and *rd* mouse. (a) δ_M and δ_C can identify different association changes; (b) δ_M and δ_{MI} have high concordance, and identify more significant correlation changes than δ_C; (c) the relationship between δ_C and δ_{MI} is similar to that between δ_C and δ_M in (a). A color version of this figure is available in the electronic version of this article.

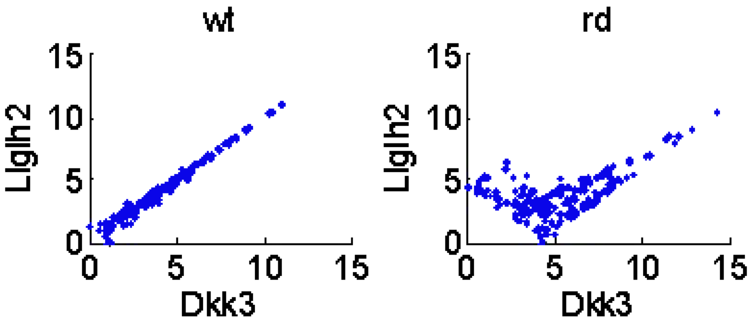

Figure 8. An example of the local association change between two states (*rd* and *wt*). This association change can only be detected with statistical significance using δ_M but not δ_C or δ_{MI}. A color version of this figure is available in the electronic version of this article.

Table 2. Top ranked candidate genes in rod degeneration for *rd* mouse.

	δ_M		δ_C		δ_{MI}	
Rank	Gene	f	Gene	f	Gene	f
1	*Pde6b*	16	*Pde6b*	54	*Pde6b*	14
2	*Dp1l1*	15	*Cnga1*	22	*Stx3*	4
3	*Fos*	8	*Pde6b*	19	*Prom1*	4
4	*Ndrg4*	7	*Gnb1*	10	*Akp2*	4

NOTE: f: observed frequency of association change with other genes between *wt* and *rd* mouse with statistical significance.

mRNA expression, it is expected that a successful method will identify *Pde6b* as a high-ranked gene; that is, *Pde6b* serves as a positive control. Indeed, *Pde6b* is ranked as the number 1 gene using any of the three statistics (Table 2). Second, the ultimate goal is to identify marker genes that are able to correctly recognize physiological processes (i.e., photoreceptor degeneration) and to identify key regulatory genes involved in this process. Thus, criteria for genes that accurately reflect the biological process might be (a) sharing similar expression profiles with genes involved in photoreceptor development, and (b) the involvement of their gene products in processes such as cellular stress response, cellular homeostasis, or others. These key regulatory genes might encode for transcription factors, growth factors or their receptors, anti-apoptotic genes, etc. To identify genes induced during the period of photoreceptor development, *Pde6b* was used as the bait gene to identify the 180 neighboring genes whose expression profiles cluster with *Pde6b* during development (*http://cepko.med.harvard.edu/default.asp*) (Blackshaw et al. 2001). Twenty-two of these neighboring genes were in the list of genes analyzed in this study. We then compared these genes against the candidates selected by each δ statistic. Among those, δ_C identified 9 of them, δ_{MI} identified 11, whereas δ_M identified 21 of the 22 possible genes, suggesting that δ_M was able to identify more photoreceptor-specific genes. The known biological roles of the top ranked genes for each method are summarized in Table 3. Interestingly, δ_M identified three transcription factors known to be involved in neuronal degeneration and cellular stress responses, whereas δ_C and δ_{MI} identified only one each, again suggesting that δ_M could identify more regulatory genes. To further evaluate the performance of the developed statistic, a gene regulatory network was constructed for the 181 analyzed genes using transcription factor binding sites to establish node connectivity (see Appendix D for details). Briefly, the top ranked genes generated by each of the statistics (Table 2) were used as "seed genes" to find a subnetwork of pairwise shortest paths in order to summarize the genes' relationships. Each set of top ranked genes generated by each statistic (Table 2) was used to establish each of the subnetworks, subnetwork-M, subnetwork-MI, and subnetwork-C, depicted in Figure 9. Let A_M, A_{MI}, and A_C denote the sets of "discovered genes," discovered independently of their expression levels, in the three subnetworks. The underlying assumption is that the discovered genes in A_S $(S = M, \mathrm{MI}, C)$ may also play important roles as the "seed genes" since they are likely regulated by the same transcription factors. We asked whether these genes in A are also recognized by the δ statistics, and how do they compare to each other. Almost all of the genes in A_M, A_{MI}, and A_C are also identified as

Table 3. Known biological function of the top 25* candidate genes identified by δ_M, δ_C, and δ_{MI} for six criteria (transcription factor, mutations in genes that are known to cause retinal degeneration, gene products involved in regulating apoptosis, gene products that are known photoreceptor structural genes or are involved in the photoreceptor signal transduction cascade, stress-induced genes, and genes involved in calcium or ion binding as the genetic mutation in the *rd* mouse causes calcium overload in the photoreceptors).

	Function	
δ_M	δ_{MI}	δ_C
	Transcription factors:	
Csda, *Fos*, *Gas6*	*Gas6*	*DKK3*
	Disease genes:	
Pde6b, *Prom1*, *Rho*	*Pde6a*, *Pde6b*, *Cnga1*, *Prom1*	*Pde6b*, *Gnat2* *Cnga1*
	Regulator of apoptosis:	
Aldh1a1, *Tnfsf12*	*Tnfsf12*	*Aldh1a1*, *Eef1a2*
	Photoreceptor structural genes:	
Pde6b (2), *Cacnb2*, *Rho*, *Prom1*	*Gnb1*, *Pde6b*, *Rom1*, *Cnga1*, *Pde6a*, *Cacnb2*, *Gnb1*	*Pde6b* (2), *Cnga1* *Gnb1*, *Gnat2*
	Stress genes:	
Usp2, *Pcp4*, *Clu*, *A2m*	*Clu*	*Clu*, *Mt1*, *Mt2*
	Ca2+ or ion binding:	
Pcp4, *Slc4a7*, *Calb2*, *Vsnl1*, *B2m*	*Vsnl1*, *Sparcl1*	*Spock2*, *Pcp4*, *Vsnl1* *Calb2*

NOTE: Due to tied ranks, 26 genes are used as cut off for δ_M, 25 for δ_{MI}, and 27 for δ_C.

candidate genes by δ_M (Figure 9), with a few exceptions: *Thrsp, Nefl*, and *Mt1*. Among those, *Thrsp* and *Nefl* were not identified by any of the statistics. Thus, the resulting "discovered" genes are in high agreement with the listed genes selected by no matter which set of "seeds" we started with. This is not the case for δ_{MI} and δ_C. There is at least one or more genes that cannot be identified by δ_{MI} and δ_C in each of the three subnetworks.

Is the subnetwork-M (Figure 9) plausible for further experimental assessment? *c-Fos*, a transcription factor, is known to be induced in the *rd* rods prior to degeneration (Rich, Zhan, and Blanks 1997), but was not essential for rod degeneration in this particular mouse model (Hafezi et al. 1998). *Itm2C*, a relatively unknown gene, interacts with the beta-site APP-cleaving enzyme 1, a protease important in the pathogenesis of Alzheimer's disease (Wickham et al. 2005). It is plausible that over-expression of this protein might be involved in rod degeneration. *Dp1l1* is a protein presumably associated with membrane trafficking (Sato et al. 2005). Loss of *Dp1l1* expression may be a reflection of cell loss, or alternatively suggest that disrupted cellular homeostasis in the *rd* rods leads to a secondary defect in membrane trafficking. Follow-up experiments are required to test the potential roles of these genes in the regulation of rod degeneration.

4.3 DISCUSSION

In Section 3, we demonstrated that our proposed metrics, local correlation (ℓ; Figure 4) and maximum local correlation M (Table 1), can quantify generic nonlinear associations

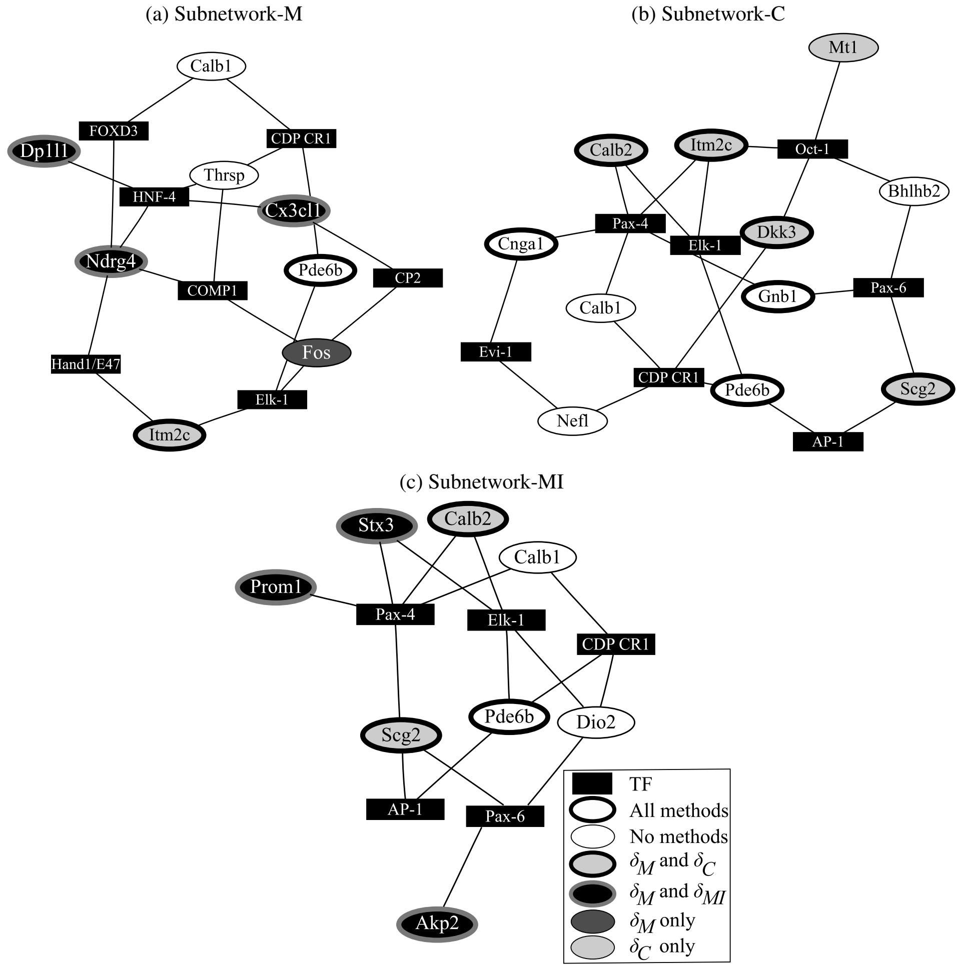

Figure 9. Three subnetworks, subnetwork-M, subnetwork-C, and subnetwork-MI generated using transcription factor binding information. These subnetworks were generated automatically using the top-ranked genes selected by (a) δ_M, (b) δ_C, and (c) δ_{MI}, respectively, as "seeds." Independent of correlation of expression, transcription factor (TF) binding information was then used to "discover" other important genes connected to the "seeds." The legend of the genes indicates whether these seed or discovered genes are also identified by statistics as important candidates. "All" means all three methods, and "none" means none of the statistics.

in all simulated cases that have been designed with some degree of association. This is in contrast to Pearson's or Spearman's correlations, which were not able to detect significant association when the data were not linear in nature.

In Section 4, we applied the proposed method to identify genes related to rod degeneration. The data were microarray expression data of healthy and mutant animals. All three δ statistics, δ_C, δ_M, and δ_{MI}, correctly identified *Pde6b*, the defective gene in the *rd* mouse, as the most central gene in the degeneration process, validating our method. Furthermore, δ_M selected more genes involved in photoreceptor development (the presumed inverse of photoreceptor degeneration), and more plausible key regulatory genes involved in the process of cell death when compared to δ_C, which selected genes whose expression appears changed as a response to death (Table 3). The discordant results between nonlinear and linear methods (Figure 7(a)) as well as the concordant findings between δ_M and δ_{MI} (Figure 7(b)) further emphasize the importance of the development of statistics to measure nonlinear association instead of global linear patterns. In addition, δ_M has higher statistical power, and can detect more local correlation change than δ_{MI} when controlling FDR at the same level (Figure 7(b)). Our method is not distributional based, and is instead an omnibus approach for detecting nonlinear correlations.

The challenges of nonlinearity and small sample size for the massive amount of data generated by modern high-throughput methods set the stage for the study described here. We have extended the use of correlation integrals to detect nonlinear correlation between any two variables reporting transient association. The proposed method is shown to have higher statistical power when compared with the reference use of mutual information and Pearson's correlation. Being distribution-free makes this tool applicable to a wide variety of problems. Although we only applied this approach to expression data in this study, the method can be applied to many other data, such as those generated by proteomics and metabolomics. In conclusion, the development of novel correlation methods that cope with the characteristic transient nonlinearity of biological dependencies, as assessed by the pairwise comparison study reported here, holds great promise.

The results of our study suggest a number of other avenues for future research. Our current work only focused on bivariate correlation. A natural extension would be the development of multivariate local correlations. Another extension would be to identify local group membership using the significant local correlation (ℓ) at particular scales of interest. Improvement of computational efficiency could also be a topic itself in the future since the computational complexity is $O(N^2)$ as it stands currently. More details on the computational complexity can be found in Appendix E. The relationship between linear and nonlinear association changes (Figure 7) indicates that using both correlations yield more information. This also leaves open the question of the relationship between these statistics or the development of a new metric.

SUPPLEMENTAL MATERIALS

Appendix, NNC toolbox, data, and scripts: The supplemental materials are available in a single zip file, which contains a readme file (README.txt) with a detailed explanation of its contents. (supplements.zip)

ACKNOWLEDGMENTS

Funding was provided by NIH grant EY-13520, a Core grant EY-14793, and an unrestricted grant to MUSC from Research to Prevent Blindness, Inc., New York. YAC was also supported by a National Cancer Institute training grant (CA90301). JSA acknowledges funding from the National Heart, Lung and Blood Institute (contract number N01-HV-28181). Doctoral studies of AJR are supported by a NLM training grant. Carroll's research was supported by a grant from the National Cancer Institute (CA57030). BR is a Research to Prevent Blindness Olga Keith Weiss Scholar. The authors appreciate the comments from Dr. Francisco Pinto. We thank Katie

Hulse for preparation of the microarray dataset. Finally, we also thank the anonymous reviewers for their suggestions.

[Received November 2008. Revised December 2009.]

REFERENCES

Blackshaw, S., Fraioli, R., Furukawa, T., and Cepko, C. (2001), "Comprehensive Analysis of Photoreceptor Gene Expression and the Identification of Candidate Retinal Disease Genes," *Cell*, 107, 579–589. [564]

Daub, C., Steuer, R., Selbig, J., and Kloska, S. (2004), "Estimating Mutual Information Using B-Spline Functions—An Improved Similarity Measure for Analysing Gene Expression Data," *BMC Bioinformatics*, 5, 118. [553,557]

Efron, B., and Tibshirani, R. J. (1993), "Permutation Tests," in *An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability*, Vol. 57, Boca Raton, FL: Chapman & Hall/CRC. [556]

Grassberger, P., and Procaccia, I. (1983), "Characterization of Strange Attractors," *Physical Review Letters*, 50, 346–349. [553,554]

Hafezi, F., Abegg, M., Grimm, C., Wenzel, A., Munz, K., Sturmer, J., Farber, D., and Reme, C. (1998), "Retinal Degeneration in the *rd* Mouse in the Absence of *c-fos*," *Investigative Ophthalmology and Visual Science*, 39, 2239–2244. [565]

Jonnalagadda, S., and Srinivasan, R. (2004), "An Information Theory Approach for Validating Clusters in Microarray Data," in *Proceedings of the 12th International Conference on Intelligent Systems for Molecular Biology*, Glasgow, Scotland. [557]

Kasturi, J., and Acharya, R. (2005), "Clustering of Diverse Genomic Data Using Information Fusion," *Bioinformatics*, 21, 423–429. [553]

Kitano, H. (2002a), "Computational Systems Biology," *Nature*, 420, 206–210. [552]

——— (2002b), "Systems Biology: A Brief Overview," *Science*, 295, 1662–1664. [552]

Marshall, E. (2004), "Getting the Noise Out of Gene Arrays," *Science*, 306, 630–631. [554]

Quackenbush, J. (2001), "Computational Analysis of Microarray Data," *Nature Review Genetics*, 2, 418–427. [553]

Rich, K. A., Zhan, Y., and Blanks, J. C. (1997), "Aberrant Expression of c-Fos Accompanies Photoreceptor Cell Death in the *rd* Mouse," *Journal of Neurobiology*, 32, 593–612. [565]

Rohrer, B., Pinto, F. R., Hulse, K. E., Lohr, H. R., Zhang, L., and Almeida, J. S. (2004), "Multidestructive Pathways Triggered in Photoreceptor Cell Death of the RD Mouse as Determined Through Gene Expression Profiling," *Journal of Biological Chemistry*, 279, 41903–41910. [552,559]

Sato, H., Tomita, H., Nakazawa, T., Wakana, S., and Tamai, M. (2005), "Deleted in Polyposis 1-Like 1 Gene (*Dp1l1*): A Novel Gene Richly Expressed in Retinal Ganglion Cells," *Investigative Ophthalmology and Visual Science*, 46, 791–796. [565]

Steuer, R., Kurths, J., Daub, C. O., Weise, J., and Selbig, J. (2002), "The Mutual Information: Detecting and Evaluating Dependencies Between Variables," *Bioinformatics*, 18, S231–S240. [553]

Storey, J. D., and Tibshirani, R. (2003), "Statistical Significance for Genomewide Studies," *PNAS*, 100, 9440–9445. [562]

Vilela, M., Borges, C. C. H., Vinga, S., Vasconcelos, A. T. R., Santos, H., Voit, E. O., and Almeida, J. S. (2007), "Automated Smoother for the Numerical Decoupling of Dynamics Models," *BMC Bioinformatics*, 8, 305. [555]

Wickham, L., Benjannet, S., Marcinkiewicz, E., Chretien, M., and Seidah, N. G. (2005), "Beta-Amyloid Protein Converting Enzyme 1 and Brain-Specific Type II Membrane Protein BRI3: Binding Partners Processed by Furin," *Journal of Neurochemistry*, 92, 93–102. [565]