GENE CO-EXPRESSION NETWORK DISCOVERY WITH CONTROLLED STATISTICAL AND BIOLOGICAL SIGNIFICANCE

Dongxiao Zhu [a,b] *and Alfred O Hero* [b]

[a]Bioinformatics Program,[b]Depatments of EECS, Biomedical Engineering and Statistics
University of Michigan,Ann Arbor, MI 48105

ABSTRACT

Many biological functions are executed as a module of co-expressed genes which can be conveniently viewed as a co-expression network. Genes are network vertices and significant pairwise co-expressions are network edges. Traditional network discovery methods controls either statistical significance or biological significance, but not both. We have designed and implemented a two-stage algorithm that controls both statistical significance (False Discovery Rate, FDR) and biological significance (Minimum Acceptable Strength, MAS) of the discovered network. Based on the estimation of pairwise gene profile correlation, the algorithm provides an initial network discovery that controls only FDR, which is then followed by a second network discovery which controls both FDR and MAS. We illustrate the algorithm for discovery of co-expression networks for yeast galactose metabolism with controlled FDR and MAS.

1. INTRODUCTION

Microarray gene expression data enable researchers to interrogate gene expression levels simultaneously on the genome scale. Detection of co-expressed genes from microarray data has attracted much attention since many co-expressed genes are found to have functional relationships, e.g. lying in the same signal transduction pathway. Many co-expression detection techniques such as relevance network and hierarchical clustering rely on the quantitative or qualitative assessment of similarities between the expression profiles of gene pairs, which is one of the fundamental objectives in functional genomics and system biology. Traditional methods either screen statistically significant or biologically significant co-expressed gene pairs. The former does not control error rate, and the latter screens many co-expressions with excessively low correlation.

In this paper, we present a two-stage algorithm that simultaneously controls statistical and biological significance of the discovered co-expression network. The algorithm implements Pearson correlation coefficients and Kendall correlation coefficients in order to capture both linear and nonlinear types of dependencies between all pairs of gene expression profiles. A two-stage error control procedure is then implemented through which a number of gene pairs are declared to be both statistically and biologically significant as measured by FDR and MAS of association. These gene pairs form the edges of the relevance network that represents the complicated web of gene co-expression among all pairs of genes.

We demonstrate the application of our two-stage algorithm by constructing relevance networks from yeast galactose metabolism data [1]. This data represents approximately 6200 gene expression levels on two-color cDNA microarrays over 20 physiological/genetic conditions (nine mutants and one wild type strains incubated in either GAL-inducing or non-inducing media).

The paper is organized into five parts: Introduction of Kendall and Pearson statistics for strength of association (Sec. 2); Formulation of the problem of network discovery as a composite hypothesis test with multiple comparisons (Sec 3); Introduction of two-stage procedure for testing these hypotheses (Sec 4); Validation of the two-stage algorithm and application to yeast data (Sec 5).

2. MEASURING THE STRENGTH OF ASSOCIATION

We use Γ to denote the true strength of association between a pair of gene expression profiles. Under a Gaussian linear hypothesis, the sample Pearson correlation coefficient $\hat{\rho}$ is an appropriate metric [2]. A robust distribution-free alternative is the sample Kendall rank correlation coefficient $\hat{\tau}$ [3]. The Pearson and Kendall correlation coefficients are special cases of the generalized correlation coefficient Γ. We define $\{g_p\}_{p=1}^{G}$ as the indices of G gene probes on the microarray; $\{X_{g_p}\}_{p=1}^{G}$ as normalized probe responses (random variables); and $\{\{x_{g_{p(n)}}\}_{p=1}^{G}\}_{n=1}^{N}$ as realizations of $\{X_{g_p}\}_{p=1}^{G}$ under N i.i.d. microarray experiments.

Kendall's τ statistic is a measure of correlation that captures both linear and nonlinear associations. The parameter τ is defined as $\tau = P_+ - P_-$, where, for any two independent pairs of observations $(x_{g_{i(n)}}, x_{g_{j(n)}})$, $(x_{g_{i(m)}}, x_{g_{j(m)}})$

from the population: $P_+ = P[(x_{g_{i(n)}} - x_{g_{i(m)}})(x_{g_{j(n)}} - x_{g_{j(m)}}) \geq 0]$ and $P_- = P[(x_{g_{i(n)}} - x_{g_{i(m)}})(x_{g_{j(n)}} - x_{g_{j(m)}}) < 0]$. An unbiased estimator of τ is given by the Kendall τ statistic:

$$\hat{\tau}_{i,j} = 2 \sum\sum_{1 \leq n \leq m \leq N} \frac{K_{nm}}{N(N-1)}, \qquad (1)$$

here K_{nm} is a indicator variable defined as $K_{nm} = \text{sgn}(x_{g_{i(n)}} - x_{g_{i(m)}})\text{sgn}(x_{g_{j(n)}} - x_{g_{j(m)}})$ for each set of pairs drawn from $\{X_{g_i}\}_{i=1}^{G}$ and $\{X_{g_j}\}_{j=1}^{G}$.

3. HYPOTHESIS TESTING SCHEME

For G genes on each microarray, we need to simultaneously test $\Lambda = \binom{G}{2}$ pairs of two-sided hypotheses:

$$H_0 : \Gamma_{g_i,g_j} \leq cormin \text{ versus } H_\alpha : \Gamma_{g_i,g_j} > cormin,$$
$$\text{for } g_i \neq g_j, \text{ and } g_i, g_j \in (1, 2, ...G) \qquad (2)$$

where $cormin$ is a minimum acceptable strength of correlation. The sample correlation coefficient $\hat{\Gamma}$ ($\hat{\rho}$ or $\hat{\tau}$) is used as a decision statistic to decide on pairwise dependency of two genes in the sample. For N realizations of any pair of gene probe responses, $\{x_{gi(n)}, x_{gj(n)}\}_{n=1}^{N}$, we first calculate $\hat{\tau}$ or $\hat{\rho}$. For large N, the Per Comparison Error Rate (PCER) p-values for ρ or τ are:

$$p_{\rho_{i,j}} = 2\left(1 - \Phi\left(\frac{\tanh^{-1}(\hat{\rho}_{i,j})}{(N-3)^{-1/2}}\right)\right) \qquad (3)$$

$$p_{\tau_{i,j}} = 2\left(1 - \Phi\left(\frac{K}{N(N-1)(2N+5)/18^{1/2}}\right)\right) \qquad (4)$$

where Φ is the cumulative density function of a standard Gaussian random variable, and $K = \sum\sum_{1 \leq n \leq m \leq N} K_{nm}$. The above expressions are based on asymptotic Gaussian approximations to $\hat{\rho}_{i,j}$ [2] and to $\hat{\tau}_{i,j}$ [3].

The PCER p-value refers to the probability of Type I error rate incurred in testing a single pair of hypothesis for a single pair of genes g_i, g_j. It is the probability that purely random effects would have caused g_i, g_j to be erroneously selected based on observing correlation between this pair of genes only. When considering the Λ multiple hypotheses for all possible pairs, as in previous studies, we adopt the FDR to control statistical significance of the selected gene pair correlations in our screening procedure [4].

4. TWO-STAGE SCREENING PROCEDURE

Select a level α of FDR and a level $cormin$ of MAS significance levels. We use a modified version of the two-stage screening procedure applied to gene screening [4]. This procedure consists of:

Stage I. Test the simple null hypothesis.

$$H_0 : \Gamma_{g_i,g_j} = 0 \text{ versus } H_\alpha : \Gamma_{g_i,g_j} \neq 0$$

at FDR level α. The step-down procedure of Benjamini and Hochberg [5] is used.

Stage II. Suppose Λ_1 pairs of genes pass the stage I procedure. In stage II, we first construct asymptotic PCER Confidence Intervals (PCER-CI's) :$I^\lambda(\alpha)$ for each Γ (ρ or τ) in subset $\mathcal{G}_1$, and convert into FDR Confidence Intervals (FDR-CI's) :$I^g(\Lambda_1\alpha/\Lambda)$ [6]. A gene pair in subset $\mathcal{G}_1$ is declared to be both statistically significant and biologically significant if its FDR-CI does not intersect the MAS interval $[-cormin, cormin]$ (see Fig 3).

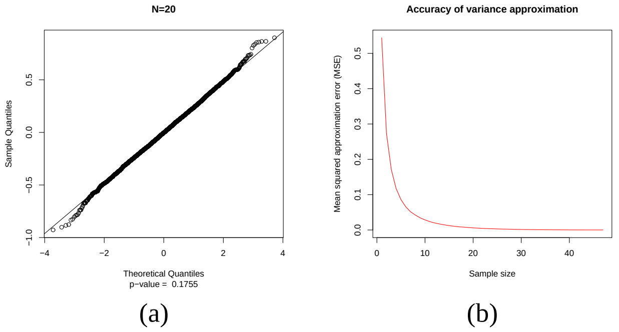


Fig. 1. *Verification of null sampling distribution (a) and variance approximation (b). (a) QQ plot of transformed sampling distribution of Pearson correlation coefficient versus normal distribution. (b) Variance approximation of transformed sampling distribution of Pearson correlation coefficient.*

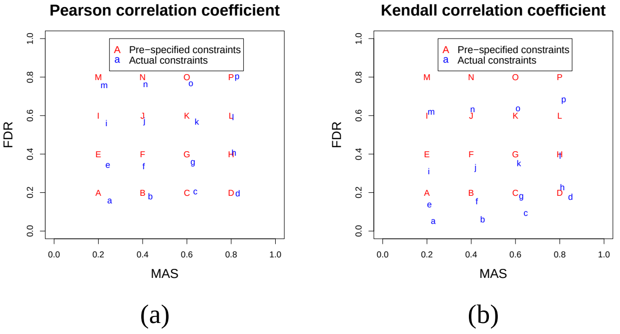
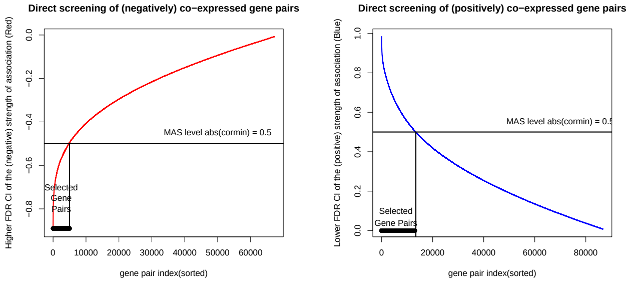


Fig. 2. *Verification of two-stage error control procedure based on Pearson correlation coefficient(a) and Kendall correlation coefficient(b). Sample size $N = 20$.*



(b) (b)

Fig. 3. *Segments of lower bounds (a) and upper bounds (b) specifying the* 5% *FDR-CI's on the positive Pearson correlation coefficients (a) and negative Pearson correlation coefficients (b) for the galactose metabolism study. Only those gene pairs whose FDR-CI's do not intersect* $[-cormin, cormin]$ *are selected by the second stage of screening. When the MAS strength of association criterion is* $cormin = 0.5$*, these gene pairs are obtained by thresholding the curves as indicated.*

5. VALIDATION OF TWO-STAGE ALGORITHM

5.1. Validating asymptotic null distribution

Here we verify that the proposed two-stage algorithm controls FDR at a specified MAS level using simulated data. Since the p-values are based on asymptotic distribution approximations (eq. 3 and eq. 4), we display in Fig. 1a that the $\hat{\rho}$ sampling distribution fits well to the Gaussian distribution for a small sample size of 20 using QQ plot. Moreover, since the construction of confidence intervals requires estimation of sampling distribution variance, the accuracy of the variance approximation is vital. This can be evaluated by the mean squared approximation error (MSE) at the sample size N:

$$MSE_{\rho}^{(N)} = \Lambda^{-1} \sum_{1 \le i < j \le G} (S^{(N)}_{\tanh^{-1}(\hat{\rho}_{i,j})} - (N-3)^{-1/2})^2, \tag{5}$$

$$MSE_{\tau}^{(N)} = \Lambda^{-1} \sum_{1 \le i < j \le G} (S^{(N)}_{\hat{\tau}_{i,j}} - \hat{\sigma}_{\tau_{i,j}})^2, \tag{6}$$

where $S^{(N)}_{\tanh^{-1}(\hat{\rho}_{i,j})}$ and $S^{(N)}_{\hat{\tau}_{i,j}}$ denote standard errors of $\tanh^{-1}(\hat{\rho}_{i,j})$ and $\hat{\tau}_{i,j}$ at the sample size N. The $\hat{\rho}$ variance approximations are seen to be in good agreement even for small sample sizes ($N > 10$) from Fig 1b.

5.2. Validating error control procedure

In order to validate our FDR and MAS error control procedure, we simulated pairwise gene expression data based on known population covariances. The actual FDR at a MAS level is calculated as a ratio of the number of screened gene pairs whose corresponding population correlation parameters $\Gamma_{i,j}$ are less than the MAS level specified, divided by the total number of screened gene pairs. The actual MAS is the minimum true discovery of population correlation $\Gamma_{i,j}$ among the screened pairs. We specified 16 pairs of (FDR,MAS) criteria (Four FDR levels: 0.2, 0.4, 0.6, 0.8; Four MAS levels: 0.2, 0.4, 0.6, 0.8), and each is plotted as a different upper case English alphabet (Red) in Fig 2. The 16 corresponding pairs of actual (FDR,MAS) criteria are also shown in Fig 2 using the same set of lower case English alphabets (Blue). It can be observed that generally the actual FDR's (lower case) fall below the specified constraint (upper case) and the actual MAS's (lower case) fall above the specified constraints (upper case). Any deviations of actual FDR's and MAS's from their specified levels are due to the conservative asymptotic approximation (eq.3 and eq.4). Observe that use of Kendall correlation (Fig 2b) leads to greater overestimation of error rates than the Pearson correlation (Fig 2a). Overestimation of error rates will translate into a reduction of power in discovering co-expressed pairs at the specified levels.

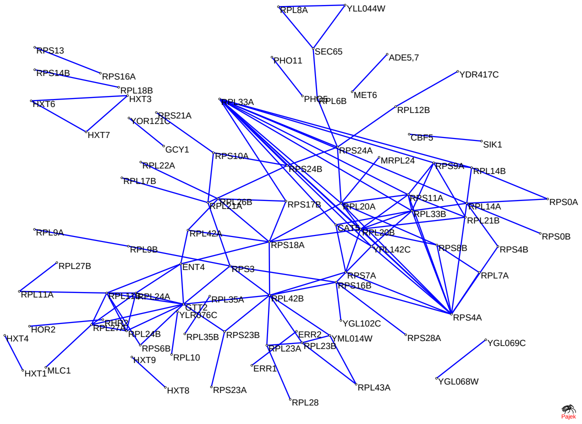


Fig. 4. *Network topology visualization. The network is discovered by constraining* $FDR \le 5\%$ *at a MAS level of 0.9. No significant negative correlation is discovered at this level. The graph is drawn using Pajek [10].*

6. CONSTRUCTING A RELEVANCE NETWORK WITH CONTROLLED FDR AND MAS

Relevance networks are implemented as a graph where n nodes (genes) are connected by p sets of edges (co-expressions) [7]. Our constructed networks are mixed networks with $p = 2$ in which edges are discovered using either Pearson or Kendall correlation coefficients constrained by the same set

Table 1. *Top ten "hub genes". The rank of each gene is the average rank over five networks. Each of five networks is constraint by a different pair of (FDR,MAS) criteria. Highest rank is the most connected and most stable gene under varying constraints of (FDR,MAS)*

Gene Name	Average Rank
RPL42B	4.2
RPS3	5.8
RPL14A	7.0
RPS16B	7.6
GTT2	8.4
RPS4A	9.8
RPL33A	11.8
RPL23B	15.8
RPS7A	16
RPL27A	17.4

of (FDR,MAS) criteria. For the yeast galactose metabolism dataset, a subset of 997 genes were identified by Ideker et al using generalized likelihood ratio test in [1]. Genes having a likelihood statistic $\lambda \leq 45$ were selected as differentially expressed, whose mRNA levels differed significantly from reference under one or more treatments.

Fig. 3a and Fig. 3b illustrate the direct implementation of the two-stage procedure to screen positively or negatively correlated gene pairs based on the Pearson correlation coefficient. See [4] for more details on how to interpret these plots. The direct screening procedure is constrained by FDR criterion $\alpha = 0.05$ and MAS criterion $cormin = 0.5$.

Fig. 4 presents the discovered network topology with a FDR level of 0.05 (5% discovered edges are expected to be false positive) at the MAS level of 0.9 ($cormin = 0.9$). The network is composed of 89 connected vertices and 132 edges. Similar to some other biological networks, the network marginal degrees appear power-law distributed. This was tested by verifying goodness of fit to the log-transformed power-law model, (goodness of fit criterion $R^2 = 0.95$) [9].

Genes that are of considerable interest to the biologist are the highly connected genes that dominates the network topology. These are called "hub genes", such as RPL33A and RPS4A in Fig 4. "Hub genes" are best connected genes that dominate a large part of the network topology. Most of the "hub genes" in each discovered network fall into two categories: "RPL" and "RPS". The former encodes "Ribosome Protein Large (60S) subunit", and the latter encodes "Ribosome Protein Small (40S) subunit" (Table 1). Both are structural components of the ribosome that is responsible for protein biosynthesis. Protein biosynthesis plays the central role in galactose metabolism because galactose is not a primary carbon source for yeast, when switching from primary carbon sources (glucose) to secondary carbon source (e.g. galactose), many different types of proteins including transporters, enzymes, and regulators have to be synthesized to be able to degrade the secondary carbon source [8]. Interestingly, the list of "hub genes" contains many hypothetical Open Reading Frames (ORFs)(data not shown), which are presumably indispensable for galactose metabolism [9].

7. CONCLUSION

We have introduced a method to construct gene co-expression networks with controlled FDR at different levels of MAS. By replacing correlation coefficient with partial correlation coefficient, the method can be naturally extended to the Gaussian Graphic Model framework.

8. REFERENCES

[1] Ideker,T., Thorsson,V. et al. (2000) Testing for differen tially-expressed genes by maximum-likelihood analysis of microarray data. *J Comput Biol*, 7(6): 805-17.

[2] Bickel,P.J. and Doksum,K.A. (2000) Mathematical statistics: basic ideas and selected topics. 2nd Edition. *Prentice Hall*, Upper Saddle River, NJ, USA.

[3] Hollander,A. and Wolfe,D. (1999) Nonparametric Statistical Methods. 2nd Edition *Wiley-Interscience*.

[4] Hero,A.O., Fleury,G. et al. (2004) Multicriteria gene screening for analysis of differential expression with DNA microarrays, *EURASIP Journal on Applied Signal Processing*, 2004(1):43-52.

[5] Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J Roy Stat Soc B Met*, 57 (1): 289-300.

[6] Benjamini,Y. and Yekutieli,D. (2004) False Discovery Rate adjusted multiple confidence intervals for selected parameters. Submitted to *Journal of American Statistical Association*.

[7] Butte,A.J., Tamayo,P. et al. (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc Natl Acad Sci U S A*, 97(22): 12182-6.

[8] Wieczorke,R., Krampe,S. et al. (1999) Concurrent knockout of at least 20 transporter genes is required to block uptake of hexoses in *Saccharomyces cerevisiae*. *FEBS Lett*, 464(3): 123-8.

[9] Barabàsi,A. (2004) Network biology: understanding the cell's functional organization. *Nat.Rev.Genet.*,5:101-113.

[10] Batagelj,A., Mrvar,A. (1998) Pajek - Program for Large Network Analysis. *Connections*, 21(2): 47-57.