Gene expression

Combining partial correlation and an information theory approach
to the reversed engineering of gene co-expression networks

Antonio Reverter* and Eva K. F. Chan

CSIRO Livestock Industries, Queensland Bioscience Precinct, 306 Carmody Road, Brisbane, Queensland 4067, Australia

Received on March 12, 2008; revised on August 1, 2008; accepted on September 9, 2008

Advance Access publication September 10, 2008

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: We present PCIT, an algorithm for the reconstruction of gene co-expression networks (GCN) that combines the concept of partial correlation coefficient with information theory to identify significant gene to gene associations defining edges in the reconstruction of GCN. The properties of PCIT are examined in the context of the topology of the reconstructed network including connectivity structure, clustering coefficient and sensitivity.

Results: We apply PCIT to a series of simulated datasets with varying levels of complexity in terms of number of genes and experimental conditions, as well as to three real datasets. Results show that, as opposed to the constant cutoff approach commonly used in the literature, the PCIT algorithm can identify and allow for more moderate, yet not less significant, estimates of correlation (r) to still establish a connection in the GCN. We show that PCIT is more sensitive than established methods and capable of detecting functionally validated gene–gene interactions coming from absolute r values as low as 0.3. These bona fide associations, which often relate to genes with low variation in expression patterns, are beyond the detection limits of conventional fixed-threshold methods, and would be overlooked by studies relying on those methods.

Availability: FORTRAN 90 source code to perform the PCIT algorithm is available as Supplementary File 1.

Contact: tony.reverter-gomez@csiro.au

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Gene expression profiling has led to the identification of genes that perform in a coordinated manner allowing researchers to reasonably predict the role of genes for which no biological function was previously attributed. It has been long demonstrated that co-expression patterns can provide insights into cellular processes (Eisen et al., 1998).

Gene co-expression networks (GCN) can be determined from expression experiments where a number of different experimental conditions have been profiled. These networks rely on the guilt-by-association heuristic, widely invoked in genomics (Wolfe et al., 2005). However, for maximum utility the input data should be derived from clearly defined systems and optimally designed experiments.

Microarray gene expression data are notorious for being noisy due to systematic biases. Accounting for systematic effects through normalization, and methods to adjust for such biases, has been a subject of great importance (Smyth et al., 2003), and alternative approaches are continuously being proposed (Baird et al., 2004; Benito et al., 2004; Dabney and Storey, 2007; Fan et al., 2004). Even when these effects have been accounted for, it might be of interest to standardize each vector to unit variance (dividing by its SD) to avoid non-interpretable correlations resulting from either one or both vectors having a negligible variation.

For a bivariate normal data, the distribution of the Pearson correlation coefficient (denoted by r) and density function under the null hypothesis of independence, is given by (Bevington and Robinson, 2003):

\[ P(r) = \frac{1}{\sqrt{\pi}} \frac{\Gamma((v+1)/2)}{\Gamma(v/2)} (1-r^2)^{(v-2)/2} \] (1)

where \( \Gamma(\cdot) \) is the gamma function and \( v = n - 2 \) is the degrees of freedom.

Furthermore, the standard error of an estimate of r is given by \( \sqrt{v/r^2/(v-2)} \), and \( r\sqrt{v-2} \sqrt{1-r^2} \) follows a t-distribution with \( n-2 \) degrees of freedom, where n is the number of paired elements in r, that is the number of experimental conditions being surveyed in the microarray experiment.

From the above formulae, it follows that extreme estimates of r (either positive or negative) resulting from a gene pair (network nodes) as well as those r computed across a large number of conditions are more likely to be statistically significant and, in turn, establish an edge between the two genes in a GCN.

For any given number of genes, the impact of the number of experimental conditions on the resulting distribution of r is further illustrated in Figure 1 for the case of 500 genes. In this case, virtually no correlations can be expected beyond magnitudes of 0.5 and 0.8, when dealing with 50 and 15 experimental conditions, respectively. Figure 1 also highlights the variation in the shape of the density that can be expected ranging from fat tails when a limited number of conditions are surveyed to grossly leptokurtic (i.e. most of the mass concentrated around zero) with large number of conditions. These results have important implications for choosing appropriate thresholds for r in the reconstruction of a network, and questions

*To whom correspondence should be addressed.
the suitability of using a single distribution (i.e. typically a Student-t or a standard normal) and a constant cutoff across all estimates of r and experimental scenarios.

While alternate methods of data normalization prior to computing co-expression measurements have been reported (Reverter et al., 2005a), choices for appropriate significance threshold above which gene–gene interactions are considered relevant remain a major challenge with most studies relying on constant, and possibly arbitrary, thresholds that are then applied across the entire set of r estimates, and without giving detailed guidance on how to interpret them.

In order to address this limitation weighted GCN have been proposed (Zhang and Horvath, 2005). However, potential drawbacks of such approaches include the need to estimate the parameters of the adjacency function, typically a sigmoid mapping of a co-expression measure to a connection weight, which in itself imposes the mandatory criterion of a scale-free topology.

More data-driven methodologies have also been proposed. For example, in the study of Elo et al. (2007), the authors introduced a systematic procedure for estimating a threshold to build GCN directly from their topological properties. The authors formulated a discrete optimization problem, in which a specified set of distinct threshold values were evaluated based on the difference between the clustering coefficients at consecutive thresholds. The clustering coefficient is defined as the proportion of existing connections from the total possible (Watts and Strogatz, 1998). However, their optimization algorithm results in the identification of a single cutoff for r which is then applied across the entire network.

Using a single cutoff for r suffers from several limitations that hinder a clear interpretation of the results because it amounts to assuming homogeneous within-gene variance. However, evidence is emerging against such assumption and in favor of heteroskedastic models due to their ability to reduce false discovery rate, while maintaining a high sensitivity (Lo and Gottardo, 2007).

The objective of this study is to evaluate the numerical and computational properties of a novel algorithm, which we refer to as PCIT, for the reconstruction of GCN. PCIT combines the concept partial correlation coefficient with information theory to identify significant gene to gene associations defining edges in the reconstruction of GCN. The properties of the algorithm are examined in the context of the topology of the reconstructed network including connectivity, clustering coefficient and sensitivity. We first examine the behavior of the PCIT algorithm using a series of simulated datasets with varying numbers of genes and experimental conditions, we then demonstrate its applicability and sensitivity using three publicly available gene expression datasets.

2 METHODS
2.1 The PCIT algorithm

We examine the use of partial correlation coefficients combined with an information theory approach to identify meaningful gene to gene associations. These two strategies have been applied independently in the reconstruction of gene networks (Basso et al., 2005; de la Fuente et al., 2004; Magwene and Kim, 2004); here, we utilize these two approaches simultaneously.

The proposed PCIT algorithm contains two distinct steps as follows:

**Step 1—Partial correlations:** For every trio of genes in x, y and z, the three first-order partial correlation coefficients are computed by:

\[
 r_{xy,z} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}},
\]

and similarly for \( r_{x,z} \) and \( r_{y,z} \).

The partial correlation coefficient between x and y given z (here denoted by \( r_{xy,z} \)) indicates the strength of the linear relationship between x and y that is independent of (uncorrelated with) z. Calculating the ordinary (or unconditional or zero-order) correlation coefficient and comparing it with the partial correlation, we might see that the association between the two variables has been sharply reduced after eliminating the effect of the third variable. Baba et al. (2004) provide a detailed account of the theoretical properties of the partial correlation coefficient as a measure of conditional independence. Strictly speaking, if the data (gene expression vectors in our context) are not multivariate normal, then zero partial correlations do not necessarily imply independence, but rather conditional uncorrelatedness. Furthermore, conditional independence is a key notion in network theory and graph modeling, where two vertices are connected by an edge if and only if the corresponding variables are not conditionally independent (see, for instance, Edwards, 1995, Section 1.3; Whittaker, 1990, Section 3.2). Within the context of gene networks, Zampieri et al. (2008) has recently explored the importance of conditional correlation and its ability to recover causal (regulatory) interactions.

**Step 2—Information theory:** We invoke the Data Processing Inequality (DPI) theorem of Information Theory which states that ‘no clever manipulation of the data can improve the inference that can be made from the data’ (Cover and Thomas, 2006). For every trio of genes, and in order to obtain the tolerance level (ε) to be used as the local threshold for capturing significant associations, the average ratio of partial to direct correlation is computed as follows:

\[
 ε = \frac{1}{3} \left( \frac{r_{xy} + r_{xz} + r_{yz}}{r_{xy} + r_{xz} + r_{yz}} \right).
\]

In the context of our network reconstruction, a connection between genes x and y is discarded if:

\[
 |r_{xy}| \leq |r_{x,z}| \text{ and } |r_{xy}| \leq |r_{y,z}|
\]

Otherwise, the association is defined as significant, and a connection between the pair of genes is established in the reconstruction of the GCN.

To ascertain the significance of the association between genes x and y, the above mentioned Steps 1 and 2 are repeated for each of the remaining n−2 genes (denoted here by z).
2.2 Simulated datasets
In order to evaluate the numerical efficiency and properties of the PCIT algorithm, we simulated a total of 300 networks to examine the effects of various numbers of conditions and genes. Five levels of experimental conditions (i.e., the dimension of each gene expression vector) were explored: 5, 10, 15, 25 and 50, and each with six numbers of genes: 100, 250, 500, 750, 1000 and 5000.

In each scenario, the vectors of gene expression values, one for each gene and with dimension equal to the number of conditions, were simulated from an independent standard normal distribution. Finally, and in order to account for sampling variance, we generated a total of 10 replications for each of the 30 combinations of experimental conditions and number of genes.

2.3 Human inflammation dataset
For our first real data example, we consider the Affymetrix GeneChip data from the study of Calvano et al. (2005) examining gene expression profiles in whole blood leukocytes immediately before and at 2 h, 4 h, 6 h, 9 h and 24 h after intravenous administration (i.e., a total of six time points or conditions) of bacterial endotoxin to four healthy human subjects.

Data acquisition criteria were as follows: probe sets identified as absent on all gene chip arrays were removed; expression signals below 128 were deemed as inaccurate and not included; and only probe sets providing signals at the six time points and for the two class treatments (infected and placebo) were included in the analyses. These criteria resulted in 510,672 expression signals from 12,192 probe sets corresponding to 3388 unique genes. Following the reanalysis of these data (Reverter et al., 2006a), intensity signals were jointly normalized by fitting a large-scale mixed model that included the fixed effects of chip and detection flag, and the random effects of probe, probe-by-individual interaction and residual.

2.4 Bovine muscle dataset
Our second real data consists of a collection of our own microarray gene expression experiments using RNA samples from bovine skeletal muscle which was previously described in Reverter et al. (2005b). In brief, these experiments span 147 hybridizations across 47 biological treatments (or conditions) relevant to genetic improvement of cattle.

The experiments were performed using an in-house bovine cDNA microarray platform comprising 9600 elements (clones) printed in duplicates onto glass slides (Lehnert et al., 2006). For the present study, we excluded readings with foreground signal less than background signal, and clones not observed in all 47 conditions. These criteria resulted in a total of 4059,807 fluorescent signals from 7898 clones of which 1947 contained accurate functional annotation for 822 unique genes.

2.5 B cell dataset
Our third real data was first used for the implementation of ARACNe (Basso et al., 2005). This data represents a collection of 336 gene expression profiles (conditions) corresponding to up to 60 different perturbations of the human B cell phenotypes, including normal and diseased B cells as well as experimentally manipulated and unadulterated B cell cell-lines. The data were downloaded from Gene Expression Omnibus database with accession number GSE2350, totaling 12,600 probes.

A modified subset of 7750 unique genes (or probesets) were used in the current study following: (1) the removal of 3191 probes with expression intensity <50 in >90% of the 336 samples, and (2) collapsing of the remaining 9409 probes to 7750 unique genes. Probes not mapped to a known gene were considered as unique entities.

2.6 Sensitivity analyses
We define sensitivity, not straight from statistical confidence (e.g., the ability to detect a prespecified change in a correlation coefficient), but rather by the minimum detectable change (i.e., equivalent to minimum detectable concentration in signal detection theory; Brown et al., 1996; O’Malley and Deely, 2003). We then ascertain when the probability of erroneously detecting a correlation (Type I error) equals that of not detecting a genuine one (Type II error). Our approach to inferring GCN sensitivity is based on locating the minimum change in correlation at which the probability of a gene pair showing this correlation equals the probability of the same gene pair being identified as significantly interacting. This equilibrium-based approach of assessing sensitivity has been successfully applied in the past in the context of gene expression data (Reverter et al., 2005b).

Sensitivity analyses in GCN are often difficult to perform due to limited validated network datasets. In order to ascertain the biological relevance of the networks that were reconstructed using PCIT and ARACNE on the real datasets, we downloaded the set of 55,606 functionally validated interactions among 7196 genes described by Franke et al. (2006).

This interactions dataset includes 2788 confirmed, direct, physical protein–protein interactions derived from the Biomolecular Interaction Network Database (BIND; http://bind.ca), 18,176 confirmed human protein interactions from the Human Protein Reference Database (HPRD; http://www.hprd.org/), 22,012 direct functional interactions from the Kyoto Encyclopedia of Genes and Genomes (KEGG; http://www.genome.jp/kegg) and 16,295 interactions derived from Reactome (http://www.reactome.org).

While this set is by no means complete, we use it as the collection of true positives upon which sensitivity and specificity analyses can be performed and placing emphasis on significant interactions that were captured by PCIT even though small to moderate correlation coefficients were obtained among the genes involved in the interaction.

3 RESULTS
3.1 Simulated datasets
The simulated datasets were generated assuming independent vectors and hence no true associations exist. Within these settings and as expected, the PCIT recovered a random network (Fig. 2B) with the majority of genes having an average number of connections (~90 in the example shown) and few genes having very many or very few connections.

Also within these settings, the task of identifying significant correlations equates to identifying outliers. Without imposing predefined global and constant nominal thresholds (e.g., top 1% or 5%), the PCIT algorithm identifies significant associations ranging from 13.23% ± 0.13% for the scenario with 5 conditions and 100 genes, to 0.94% ± 0.01% for 50 conditions and 5000 genes (Fig. 2). As opposed to the constant cutoff approach commonly used in the literature, the PCIT algorithm can identify and allow for more moderate, yet not less significant, values of r to still establish a connection in the reconstruction of the GCN.

Figure 2C illustrates the observed distributions of the resulting tolerance levels (ε) for all trios of genes examined at 5 and 50 conditions. The two distributions overlap. However, on average, higher values of ε were observed with increased number of conditions: averaging 0.85 versus 0.65 for 50 versus 5 conditions. We further note that these two distributions were unaffected by varying number of genes.

We conclude from the decrease in clustering coefficients along with increasing number of experimental conditions and genes (Fig. 2A) that PCIT provides a consistent estimate of the number of significant connections because its bias tends to zero as sample size increases. However, caution should be taken when reconstructing networks from few experimental conditions and limited numbers of genes as this combination is likely to inflate spurious connections.
A. Reverter and E. K. F. Chan

Fig. 2. Performance of PCIT on simulated datasets: (A) pattern of clustering coefficient, measured as the percentage of significant correlation coefficients, with increasing numbers of genes in the gene coexpression network. (B) Connectivity distribution of a random network recovered by the PCIT algorithm from a simulated dataset of 5000 genes and 5 conditions. (C) Frequency distributions of PCIT local thresholds ($\varepsilon$) for 5 and 50 conditions.

Fig. 3. Sensitivity analysis of PCIT on simulated datasets: the theoretical probability of finding an absolute correlation coefficient as extreme is given by the black curve. This is compared with the probability of identifying such a correlation as significant in a network with either 100 genes (green curve) or 5000 genes (red curve) the expression of each surveyed at either 5 conditions (A) or 50 conditions (B). The point where the two curves intersect defines the sensitivity of the experiment.

The raise of spurious connections is further illustrated in Figure 3, where the results from the sensitivity analyses at 100 or 5000 genes and with 5 or 50 conditions are presented. As expected, better sensitivity is observed with an increased number of experimental conditions. With only five conditions (Fig. 3A), a change in correlation can be reliably detected if $|r| > 0.55$ and $|r| > 0.75$ for 100 and 5000 genes, respectively. With 50 conditions (Fig. 3B), these values drop to 0.15 and 0.25. Interestingly, sensitivity is worse with more genes; this is attributed to the inflation of false positives resulting from the exponential increase of all possible pairwise correlations being tested.

3.2 Application to real datasets

Figure 4 illustrates the density distribution function of $r$ resulting from the human inflammation, the bovine muscle and the human B cells datasets. Black-shaded areas correspond to the significant $r$ as identified by PCIT.

For the bovine muscle and the B cell datasets, these distributions were uni-modal and similar to that obtained from random data (Fig. 1), suggesting an absence of correlation (interaction) between the majority of gene pairs. For the inflammation dataset, a bimodal distribution was observed with modes corresponding to extreme values of $r$. Systemic inflammation represents a major perturbation to homeostatic conditions and results in most genes becoming either strongly correlated or anti-correlated. These findings are in complete agreement with previous studies dealing with bacterial endotoxins (Remondini et al., 2005).

The results from the sensitivity analysis are illustrated by Figure 5, where the receiver operating characteristic (ROC) curves obtained after applying the PCIT algorithm are presented. While the three datasets yielded a discrimination power (as measured by the area under the ROC curve) $>50\%$ (being 66.7%, 57.1% and 59.8% for the bovine muscle, the human inflammation and the human B cell datasets, respectively), the seemingly better performance of the bovine muscle dataset was attributed to the smaller number of genes (822) combined with the relatively large number of conditions (47). Higher areas under the ROC curves could be anticipated using an ideal complete set of true positive interactions. We also noted that saturation of precision (proportion of true positives to total positives) was rapidly obtained with early declines of specificity.

To highlight how PCIT is able to recover significant interactions of relatively small correlation coefficients, we note that in the bovine muscle dataset, a total of 265 true positive (i.e. previously experimentally validated) interactions among 110 genes were identified from pairwise correlations $< |0.50|$. Similarly, at the same threshold, 707 true positive interactions among 816 genes were identified in the B cell data. To the best of our knowledge, such a small threshold is rarely used in the literature.

Importantly, small in magnitude but significant estimates of $r$ were more likely to affect lowly variable and poorly connected genes. This observation is illustrated by the results presented in Table 1, where the relationship between the variation of a gene’s expression level and the average absolute $r$ of its significant connections is given for the three datasets, and separately for genes with fewer and more than average number of connections. The higher number of genes with fewer than average number of connections (61%, 28% and 30% more for the bovine muscle, inflammation and B cell datasets, respectively) reflects the non-randomness, power-law distribution of the resulting GCN.

3.3 PCIT versus ARACNE

The conception of PCIT was partly motivated by a previously reported approach, ARACNE, that uses mutual information (MI) and DPI to identify statistically significant direct gene–gene correlations (Margolin et al., 2008). A major difference between PCIT and
Reconstruction of GCN

Fig. 4. Empirical density distribution functions for all (gray) and significant (black) correlation coefficients defined by the PCIT algorithm and for the human inflammation (A), the bovine muscle (B) and the human B cell (C) datasets.

ARACNE is in their use of the tolerance level, ε, for selecting significant correlations. ARACNE applies a global ε threshold, while PCIT adopts local ε thresholds. Three major arguments in favor of local ε thresholds are: (1) the distribution of ε varies depending on the number of conditions examined (Fig. 2) as well as the underlying distribution of the correlations per gene pair; (2) the strengths of interactions (correlation) are expected to differ between gene pairs and so tolerances are also expected to vary (Table 1); and (3) local ε thresholds are driven purely by the data and so eliminate human intervention. Another difference between the two approaches is that, prior to filtering on ε, ARACNE has the option of applying a global probability threshold on MI.

We compared results from these two algorithms using the three aforementioned real datasets (Table 2). To ensure the comparisons are sensible, ARACNE was implemented without thresholding on the MI values and we set the global ε threshold to 0.25. This value was empirically chosen from analyses with PCIT: examination of the resulting local ε distributions (e.g. Fig. 2) suggested the average percent tolerance to be ~25% (i.e. local ε = 0.75).

In general, we found that ARACNE tended to return fewer significant interactions; as the original authors pointed out, this is likely due to the stringency of the chosen global threshold that may ‘hinder the detection of genes with lower mutual information’ (Basso et al., 2005). Despite this, we noted that the increase in significant interactions was proportionate across all genes; i.e. there is strong correlation between the numbers of interactions per gene identified by the two methods (Table 2, bottom rows). In general and based on our set of true positive interactions, the extent of

Fig. 5. ROC curves resulting from applying the PCIT algorithm to the human inflammation (green) and bovine muscle (red) and human B cell (blue) datasets.

Table 1. Comparison of the extent of significant interactions between lowly and highly connected genes for each of the three real datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Mean&lt;sup&gt;a&lt;/sup&gt;</th>
<th>≤Mean&lt;sup&gt;b&lt;/sup&gt;</th>
<th>&gt;Mean&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bovine</td>
<td>62</td>
<td>507</td>
<td>315</td>
</tr>
<tr>
<td>Inflammation</td>
<td>222</td>
<td>1902</td>
<td>1486</td>
</tr>
<tr>
<td>B Cell</td>
<td>671</td>
<td>4382</td>
<td>3368</td>
</tr>
</tbody>
</table>

<sup>a</sup>Mean, Average number of connections; <sup>b</sup>N, Number of genes; <sup>c</sup>Corr, Correlation between SD of a gene’s expression level and its average absolute correlation of its significant connections.
true and false connections was similar between the two methods, except for the human inflammation data. Analysis of this data with ARACNE appears to have led to higher false positive rate and comparatively lower specificity. We attribute this to the combination of the abnormal $r$ distribution corresponding to this dataset with the hard-thresholding inherent in ARACNE. Due to the higher density of large $|r|$, a constant and global threshold on $\varepsilon$ would comparatively identify more interactions as significant in ARACNE than PCIT.

Authors of ARACNE suggest exploring different choices of $\varepsilon$ for different datasets, and as such a different $\varepsilon$ threshold may have been chosen for the Inflammation data than the other two datasets. While a different choice of $\varepsilon$ may have resulted in more comparable FDR and specificity for this dataset, the final selection of $\varepsilon$ is often less than intuitive. Conversely, PCIT is completely data driven and does not rely on human intervention and so is more objective.

### 4 DISCUSSION

Most research involving the development of GCN based on microarray data ignores alternate methods of identifying the optimal cutoff in the threshold for the correlation coefficient between the expression of two genes because they are based on the reasonable, yet non-testable, assumption that a sufficiently large stringent threshold across the entire network is more likely to reveal functionally relevant (true positives) interactions. Yeung et al. (2004) concluded that the proportion of co-regulated genes is directly related with the number of hybridization experiments and not so much with the diversity of experiments. This could be attributed to the fact that noise can be better described with more technical replicates (i.e. power increases with more data), while diverse microarray experiments do not help in characterizing such noise. To standardize data, the authors used a $z$-score as their primary evaluation criterion.

Asymptotically, and by means of the Central Limit Theorem (for details see Mood et al., 1974), such standardization is expected to yield a ‘normalized’ set of data that will follow a standard normal distribution (i.e. with zero mean and unit variance). As a consequence, such standardization amounts to giving equal prior importance to all genes (Frank and Friedman, 1993), which contradicts the emerging evidence of heterogeneous within-gene variance (Lo and Gottardo, 2007, and references there in).

In the present study, we have introduced PCIT, an algorithm to estimate significant interactions on a by-gene basis and based on the concept of partial correlation and an information theoretic approach. This algorithm falls in the known category of ‘weighted networks’. The inherent strength of weighted networks is their ability to judge every realization of a similarity measure (e.g. a correlation coefficient in this instance) using an information-driven criterion, rather than applying the same hard-coded threshold across the entire space. In PCIT, the weighing, based on conditional information, is given by a local threshold computed from the average ratio of conditional to direct correlation. PCIT is a substantial improvement to existing methods with the following appealing features: (1) the tolerance-level epsilon is estimated separately for each triplet of genes; (2) it is fully data-driven; and hence (3) it obviates the need for one to define a critical threshold for determining conditional independence.

The choice to use the average ratio of conditional to direct correlation, as opposed to other location parameters that are functions of these ratios is justified in Supplementary File 2. In brief, the average was selected because: (1) its bell-shaped distribution lies well within the [0, 1] interval and is only affected by the number of experimental conditions being explored and not the number of genes (Fig. 2C); (2) it is less affected by extreme values resulting from near-zero denominators; and (3) it represents a sound compromise between the more conservative minimum and the more restrictive median and maximum.

We have tested the performance of the PCIT algorithm across a variety of simulated scenarios generated from independent normally distributed vectors of gene expression. The reasoning behind such a strategy is that, within these settings the expectation is to recover a random network (and PCIT does exactly this; Fig. 2B) with the majority of genes having an average number of connections and few genes having very many or very few connections. Also within these settings, the task of identifying significant correlations equates to identifying outliers.

Commonly used approaches (i.e. hypothesis driven) will call outliers those realizations outside a predefined (e.g. tabled value) number of SDs beyond the mean. Without imposing predefined global and constant nominal thresholds (e.g. 1% or 5% significance) the PCIT algorithm (information data driven) identifies significant associations ranging from ~10% for the scenario with 5 conditions and 100 genes, to ~1% for 50 conditions and 5000 genes (Fig. 2A).

We have further applied PCIT to three real datasets representing a large dynamic range in the number of genes (822, 3388 and 7750) as well as in the lengths of the gene expression vectors corresponding to experimental conditions being surveyed (and not necessarily the number of arrays) and used for calculating the correlations (6, 47 and 336).

The performance of PCIT on these real datasets was evaluated by a series of error rates (Table 2) computed from a collection of functionally validated interactions. We emphasize here that, unknown hence unreported interactions that were nonetheless identified by PCIT (or any other algorithm), were consequently and necessarily considered as true negatives.

### Table 2. PCIT versus ARACNE comparison across the bovine muscle, human inflammation and human B cell datasets

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Bovine (822 genes)</th>
<th>Inflammation (3378 Genes)</th>
<th>B Cell (7750 Genes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCIT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>7.6</td>
<td>6.5</td>
<td>16.3</td>
</tr>
<tr>
<td>FDR</td>
<td>8.1</td>
<td>6.6</td>
<td>8.6</td>
</tr>
<tr>
<td>SENS</td>
<td>37.3</td>
<td>16.6</td>
<td>20.0</td>
</tr>
<tr>
<td>SPEC</td>
<td>91.9</td>
<td>93.4</td>
<td>91.4</td>
</tr>
<tr>
<td>PREC</td>
<td>8.7</td>
<td>1.0</td>
<td>0.8</td>
</tr>
<tr>
<td>ARACNE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>6.1</td>
<td>26.9</td>
<td>2.3</td>
</tr>
<tr>
<td>FDR</td>
<td>5.9</td>
<td>27.1</td>
<td>2.6</td>
</tr>
<tr>
<td>SENS</td>
<td>16.2</td>
<td>33.0</td>
<td>12.4</td>
</tr>
<tr>
<td>SPEC</td>
<td>94.0</td>
<td>72.9</td>
<td>97.4</td>
</tr>
<tr>
<td>PREC</td>
<td>5.4</td>
<td>0.5</td>
<td>1.9</td>
</tr>
<tr>
<td>Connection correlation</td>
<td>0.23</td>
<td>0.78</td>
<td>0.79</td>
</tr>
<tr>
<td>Proportion overlap</td>
<td>0.51</td>
<td>0.84</td>
<td>0.96</td>
</tr>
</tbody>
</table>

*CC, clustering coefficient; FDR, false discovery rate = FP/(FP+TN); SENS, sensitivity = TP/(TP+FN); SPEC, specificity = TN/(TN+FP); PREC, precision = TP/(TP+FP); Connection correlation, correlation between the number of connections for each gene identified by PCIT and ARACNE; Proportion overlap, proportion of connections that are identified by both methods for each gene.*
The PCIT algorithm operates such that, for any given scenario (in terms of number of experimental conditions and genes), the probability that an estimate of \( r \) will be deemed significant, and hence results in the establishment of an edge during the network reconstruction, is proportional to its value in relation to the values of its neighbors. Small (in magnitude) but significant estimates of \( r \) correspond to lowly variable poorly connected genes. In our opinion, this data-driven phenomenon is an unsupervised numerically elegant feature of PCIT. These small but significant estimates of \( r \) (representing gene pairs that are tightly correlated over a small expression range) will not be detected by existing fixed-threshold methods. However, as their expression changes are subtle across various treatments, it is likely they represent genes whose expression is of fundamental importance. Consequently, it is imperative that these particular interactions are not overlooked during network assembly.

ACKNOWLEDGEMENTS

The authors are grateful to John Henshall and Nick Hudson for insightful suggestions in the preparation of this article. The authors thank three anonymous reviewers who by suggestions and advises contributed to the significant improvement of this article.

Funding: Cooperative Research Centre for Beef Genetic Technologies (http://www.beef.crc.org.au/).

Conflict of Interest: none declared.

REFERENCES