A Null Model for Pearson Coexpression Networks

Andrea Gobbi, Giuseppe Jurman*
Fondazione Bruno Kessler, Trento, Italy

* jurman@fbk.eu

Abstract

Gene coexpression networks inferred by correlation from high-throughput profiling such as microarray data represent simple but effective structures for discovering and interpreting linear gene relationships. In recent years, several approaches have been proposed to tackle the problem of deciding when the resulting correlation values are statistically significant. This is most crucial when the number of samples is small, yielding a non-negligible chance that even high correlation values are due to random effects. Here we introduce a novel hard thresholding solution based on the assumption that a coexpression network inferred by randomly generated data is expected to be empty. The threshold is theoretically derived by means of an analytic approach and, as a deterministic independent null model, it depends only on the dimensions of the starting data matrix, with assumptions on the skewness of the data distribution compatible with the structure of gene expression levels data. We show, on synthetic and array datasets, that the proposed threshold is effective in eliminating all false positive links, with an offsetting cost in terms of false negative detected edges.

Introduction

Universally acknowledged by the scientific community as the basic task of systems biology, network inference is the prototypical procedure for moving from the classical reductionist approach to the novel paradigm of data-driven complex systems in the interpretation of biological processes [1].

The goal of all network inference (or network reconstruction) procedures is the detection of the topology (i.e., the wiring diagram) of a graph whose nodes belong to a given set of biological entities, starting from measurements of the entities themselves. In the last fifteen years, the reconstruction of the regulation mechanism of a gene network and of the interactions among proteins from high-throughput data such as expression microarray or, more recently, from Next Generation Sequencing data has become a major line of research for laboratories worldwide. The proposed approaches rely on techniques ranging from deterministic to stochastic, and their number is constantly growing. Nonetheless, network inference is still considered an open, unsolved problem [2]. In fact, in many practical cases, the performance of the reconstruction algorithms are poor, due to several factors limiting inference accuracy [3, 4] to the point of making it no better than a coin toss in some situations [5]. The major problem is the
under-determinacy of the task [6], due to the overwhelming number of interactions to be predicted starting from a (usually) small number of available measurements. In general, size and quality of available data are critical factors for all inference algorithms.

In what follows the impact of data size is discussed for one of the simplest inference techniques, i.e., the gene coexpression network, where interaction strength between two genes is a function of the correlation between the corresponding expression levels across the available tissue samples. Despite its simplicity, the number of studies in the literature based on coexpression networks is still a large fraction of all manuscripts in systems biology [7–13], including Next Generation Sequencing data [14]. The underlying biological hypothesis is that functionally related genes have similar expression patterns [15], and thus that coexpression is correlated with functional relationships, although this does not imply causality. This implies for instance that genes that are closer in a biological network are likely to have more similar expression [16, 17]: this hypothesis has been validated to some extent, for example by Jensen et al. in [18]. However, a caveat is mandatory here: since correlation is a univariate method, it is unable to capture the relations occurring among genes when the independence hypothesis does not hold, which is a common situation in the -omics datasets [19, 20]. To overcome this problem, characterized by small correlation values between functionally related genes, multivariate approaches are required [21], also in the network case [22].

Notwithstanding this limitation, as highlighted in [23], correlation can help unveil the underlying cellular processes, since coordinated coexpression of genes encode interacting proteins. The Pearson Correlation Coefficient (PCC for short) is used as the standard measure, although alternative correlation measures can be also employed: see for instance [24] for a comparative review of eight well known association metrics. Furthermore, coexpression analysis has been intensively used as an effective algorithm to explore the system-level functionality of genes, sometimes outperforming much more refined approaches [25, 26]. The observation that simpler approaches such as correlation can be superior even on synthetic data has been explained by some authors [27, 28] with the difficulties of a complex algorithm in detecting the subtleties of combinatorial regulation. Coexpression networks can also capture more important features than the conventional differential expression approach [29], and their use has been extended to other tasks, for instance the investigation of complex biological traits [30]. Finally, these networks can be crucial for understanding regulatory mechanisms [31], for the development of personalised medicine [32] or, more recently, in metagenomics [33].

However, as noted in [34], correlation between genes may sometimes be due to unobserved factors affecting expression levels. Moreover, deciding when a given correlation value between two nodes can be deemed statistically significant and thus worthwhile for assigning a link connecting them is a major issue affecting coexpression networks [35]. This translates mathematically into choosing (a function of) a suitable threshold, as in the case of mutual information and relevance networks [36]. As reported in [37], in the literature statistical methods for testing the correlations are underdeveloped, and thresholding is often overlooked even in important studies [38]. The two main approaches known in the literature to solve this problem can be classified as soft or hard thresholding. Soft thresholding is adopted in a well-known framework called Weighted Gene Coexpression Network Analysis (WGCNA) [39], recently used also for other network types [40, 41]. All genes are mutually connected, and the weight of a link is a positive power of the absolute value of PCC, where the exponent is chosen as the best fit of the resulting network according to a scale-free model [42, 43]. This approach, without discarding any correlations, promotes high correlation values and penalises low values. In the hard thresholding approach, instead, only correlation values larger than the threshold are taken into account, and an unweighted link is set for each of these values, so that a binary network is generated (see [44] for one of the earliest references). Clearly, an incorrectly chosen threshold.
A Null Model for Pearson Coexpression Networks

value can jeopardise the obtained results with false negative links (for too strict a threshold) or false positive links (for too loose a threshold). Many heuristics have been proposed for setting the threshold values, such as using the False Discovery Rate [45–47], using a cross-validation strategy [48], the p-value of the correlation test [15, 32], employing partial correlation [49, 50], using rank-based techniques [51–53], more complex randomisation techniques [54], or keeping only values exceeding a minimum acceptable strength (MAS) level specified by the threshold [55]. Alternative approaches have been recently studied, evaluating the correlation distribution, both experimentally [56] and theoretically [57]. However, these studies focus on the level of single interaction rather than considering the whole network.

Moreover, in many studies in the literature, the threshold is not chosen according to an algorithmic procedure, but referring to standard choices [58–61], or to heuristics not directly related to the correlation values, but rather with the resulting network topology [62–71]. In [72] a comparison of some coexpression thresholds is shown on a few microarray datasets. Furthermore, each threshold choice yields a compromise between detecting artefacts (false positives) and neglecting existing connections (false negatives) [56], which can be driven by the cost attached to the task studied. For instance, the optimized threshold selection procedure in [73] attains a very low false positive rate by using a permutation-based strategy [74].

Here we propose a new a priori model for the computation of a hard threshold based on the assumption that a random coexpression graph should not have any edges. This threshold follows from the work of Fisher [75] and Bevington [76], and, as a deterministic independent null model, it depends only on the dimensions of the starting data matrix, with assumptions on the skewness of the data distribution compatible with the structure of gene expression levels data [77, 78]. Hence, the procedure to obtain the threshold is not stochastic and thus deeply different from approaches based on permutation tests [74]. Further, this model is non-parametric, because its only input is the data themselves and no additional quantity needs to be tuned. By definition, this threshold is aimed at minimizing the possible false positive links, paying a price in terms of false negative detected edges. This characterising property makes this method especially useful when the intrinsic cost function associated with the studied task is biased towards penalising false positives. We conclude by demonstrating the procedure first on a synthetic dataset and then on an ovarian epithelial carcinoma dataset on a large cohort of 285 cases [79, 80].

A Motivating Example

We show hereafter an example of a common situation arising in the small sample size setting, where PCC can reach extremely high values possibly leading to the detection of false positives.

Consider the HepatoCellular Carcinoma (HCC) dataset, first introduced in [81] and later used in [82] and publicly available at the Gene Expression Omnibus (GEO) http://www.ncbi.nlm.nih.gov/geo with accession number GSE6857. The dataset collects 482 tissue samples from 241 patients affected by HCC. For each patient, a sample from cancerous hepatic tissue and a sample from surrounding non-cancerous hepatic tissue were extracted, hybridised on the Ohio State University CCC MicroRNA Microarray Version 2.0 platform consisting of 11,520 probes collecting expressions of 250 non-redundant human and 200 mouse miRNA. After a preprocessing phase including imputation of missing values [83] and discarding probes corresponding to non-human (mouse and controls) miRNA, the resulting dataset HCC includes 240 + 240 paired samples described by 210 human miRNA, with the cohort consisting of 210 male and 30 female patients. In particular, consider now the three microRNA identified as hsa.mir.010b.precNo1, hsa.mir.016a.chr13 and hsa.mir.016b.chr3. A search by sequence with the miRBase web-service [84] shows that the two probes hsa.mir.016a.
chr13, and hsa.mir.016b.chr3 share a high alignment score ($e$-value: $5 \cdot 10^{-4}$[85]), while no link due to coherent coexpression is known between each of these probes and hsa.mir.010b.precNo1. Consistent with this observation, the absolute PCC $|\rho|$ on the dataset H consisting of the 210 tumoral samples from male patients is

\[
|\rho_{H}(\text{hsa.mir.016a.chr13}, \text{hsa.mir.016b.chr3})| = 0.969
\]

\[
|\rho_{H}(\text{hsa.mir.010b.precNo1}, \text{hsa.mir.016b.chr3})| = 0.536.
\]

When we restrict the analysis to a suitable small subset of patients, different situations can occur. Consider in fact the following sub-cohort of 10 patients

\[ S = \{03 - 457, 02 - 354, 03 - 146, 03 - 467, 03 - 037, 03 - 033, 03 - 280, 03 - 205, 02 - 421, 02 - 432\}.
\]

On S, PCC reads as follows:

\[
|\rho_{S}(\text{hsa.mir.016a.chr13}, \text{hsa.mir.016b.chr3})| = 0.907
\]

\[
|\rho_{S}(\text{hsa.mir.010b.precNo1}, \text{hsa.mir.016b.chr3})| = 0.941,
\]

that is, the correlation between the two unrelated miRNA on S is very high and even higher than the correlation of the two aligned probes. The curves (averaged on 1000 runs) of PCC versus the sample size when the remaining samples are increasingly and randomly added (while S remains the same across the 1000 runs) are shown in Fig 1. When sample size increases, $|\rho(\text{hsa.mir.010b.precNo1}, \text{hsa.mir.016b.chr3})|$ quickly drops down to 0.536, while the correlation of the two aligned probes remains almost constant.

![Fig 1. HCC dataset: PCC versus sample size for the coexpression of hsa.mir.016a.chr13 with hsa.mir.016b.chr3 (blue line) and with hsa.mir.010b.precNo1 (red line). When the correlation is computed on the 10-sample set S the correlation between the two uncoexpressed probes is even higher than the correlation between the two probes sharing an almost identical alignment. When more samples are randomly added, the blue line slightly increases, while the red line quickly drops to the final value 0.536 on the whole dataset H. Both curves are averaged over 1000 randomisations of the added samples, keeping the sub-cohort S constant.](image-url)
Clearly, if the task were to build a PCC coexpression network \( N_S \) on \( S \), whatever the chosen hard threshold \( t \), the network \( N_S \) would include either a false positive or a false negative link (or even both, for \( 0.907 \leq t \leq 0.941 \)). Note that the probability of extracting 10-sample sub-cohorts \( S_0 \) satisfying

\[
|\rho_{S'_0}(\text{hsa.mir.010b.precNo1, hsa.mir.016b.chr3})| > 0.9
\]

\[
|\rho_{S'_0}(\text{hsa.mir.016a.chr13, hsa.mir.016b.chr3})| > 0.9
\]

is about 0.03%, but when the constraints are relaxed the probability of detecting artificially high correlation values \( |\rho_{S'_0}(\text{hsa.mir.010b.precNo1, hsa.mir.016b.chr3})| > 0.75 \) raises to about 6.2% for this dataset. Situations like those shown in this example regularly occur when building coexpression networks, supporting the need for an algorithmic solution to the problem of hard thresholding in the small sample size setup.

### Methods

#### Distribution of PCC

Given a real number \( 0 < p < 1 \), define the function

\[
F(n, p) = P(|\rho(x, y)| > p)
\]

where \( x \) and \( y \) are two independent normal vectors of length \( n \) and \( \rho \) is the PCC.

The results in [75, 86–92] and the symmetry of \( \rho \) yield that \( F(n, p) \) has the following close form:

\[
F(n, p) = P(|\rho(x, y)| > p) = \frac{2}{\sqrt{n}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)} \int_0^{\arccos p} \sin^{n-3}(\theta) d\theta,
\]

where \( \Gamma(x) \) is the Gamma function \( \Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt \). In Text A in S1 Text we also propose a novel proof of Eq (2) by means of an analytic argument stemming from the work of Bevington in [76, Ch. 7]. Moreover, Eq (2) is also a good approximation in the case of a non-normal distribution of \( x \) and \( y \) when data skewness can be bounded [93], because of the generalization shown in [89, 94–96]. Non-Gaussian asymmetric distributions can occasionally be detected in some array studies [78]; however, techniques for reducing the skewness are routinely applied during preprocessing [77], and thus the aforementioned results can be safely used in the microarray framework.

#### Coexpression network and threshold selection

The results derived in the previous section are used here to construct a null model for the correlation network, thus yielding a threshold for the inference of a coexpression network from the nodes’ data. As mentioned in the Introduction, these correlation networks are subject to the hypothesis of independence between genes, so they are not detecting higher-order relations for which a multivariate method is needed. In Text E in S1 Text we show a few synthetic and -omics examples of relations among genes that are not captured by coexpression networks.

In the general situation, we are measuring the signal (expression) of \( m \) probes (genes), in \( n \) different instances/conditions (samples), to infer the corresponding correlation networks on \( m \) nodes.
Formally, let \( X = \{x_i\}_{i=1}^m \) be a set such that \( x_i \in \mathbb{R}^n \forall i = 1, \ldots, m \). Then the coexpression graph \( G_p = \{V, E_p\} \) is the graph where

\[
V = \{v_1, \ldots, v_m\} \quad \text{and} \quad (v_i, v_j) \in E_p \iff |\rho(x_i, x_j)| > p.
\]

The first result characterizes the coexpression graphs in terms of null models:

**Proposition 1** If the vectors \( x_i \) are sampled from the uniform distribution, the graph \( G_p \) is an Erdős-Rényi model \([97]\) with \( m \) nodes and probability \( F(n, p) \) as in Eq (2).

The proof follows immediately from the definition of \( G_p \) and Eq (2).

**Definition** Using the results in the previous section, the secure threshold \( \tilde{p} \) is defined as follows, for an arbitrarily small \( \epsilon > 0 \):

\[
\tilde{p} = \min_{\rho \in (0, 1)} \left\{ F(n, p) \frac{m(m-1)}{2} \leq 1 - \epsilon \right\}, \tag{3}
\]

for \( m \) nodes measured on \( n \) samples. As a consequence of \([98]\), stating that the median of the binomial distribution is the integer part of the mean when \( F(n, p) < 1 - \log 2 \), the secure threshold \( \tilde{p} \) is the minimum value of \( p \) such that the corresponding random coexpression network is on average an empty graph \( E_m \), i.e., \( P(G_p = E_m) \geq \frac{1}{2} \).

The underlying hypothesis for Eq (3) is the assumption that in a random dataset no edge is expected, since no relation should occur between nodes. Due to its definition, the secure threshold \( \tilde{p} \) is biased towards avoiding the false positive links, paying a price in terms of false negatives. In Table 1 a collection of values of \( \tilde{p} \) is listed for different \( m \) and \( n \), while in Fig 2 the contour plot of the function \( \tilde{p}(n, m) \) is shown first on a large range of values and then zooming on the small sample size area.

As a first example, we compare the secure threshold \( \tilde{p} \) with the threshold derived by applying the Bonferroni correction procedure \([99, 100]\) to the \( p \)-value of PCC. Let now \( \Phi_F \) and \( \Phi_{t, n} \) be the cumulative distribution functions for the Fisher transformation and the Student’s \( t \)-distribution with \( n - 2 \) degrees of freedom, respectively. For a given PCC \( \rho \) computed between vectors in \( \mathbb{R}^n \), let

\[
FT_n(\rho) = 2\Phi_F(-|\text{arctanh}(\rho)|\sqrt{n-3}) \quad \text{and} \quad tT_n(\rho) = 2\Phi_{t,n-2}\left(-|\rho|\sqrt{\frac{n-2}{1-r^2}}\right) \tag{4}
\]

be the functions computing the \( p \)-value associated to \( \rho \) following either the Fisher approximation \( FT_n \) \([75, 89, 101]\) or the Student’s \( t \)-distribution \( tT_n \) \([102–106]\). Set now the global

<table>
<thead>
<tr>
<th>( m/n )</th>
<th>100</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>50000</th>
<th>100000</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>0.95629</td>
<td>0.98520</td>
<td>0.99070</td>
<td>0.99415</td>
<td>0.99800</td>
<td>0.99932</td>
</tr>
<tr>
<td>15</td>
<td>0.81681</td>
<td>0.89170</td>
<td>0.91323</td>
<td>0.93036</td>
<td>0.95800</td>
<td>0.97456</td>
</tr>
<tr>
<td>20</td>
<td>0.73825</td>
<td>0.82388</td>
<td>0.85077</td>
<td>0.87330</td>
<td>0.91286</td>
<td>0.93973</td>
</tr>
<tr>
<td>30</td>
<td>0.62814</td>
<td>0.71776</td>
<td>0.74817</td>
<td>0.77485</td>
<td>0.82534</td>
<td>0.86367</td>
</tr>
<tr>
<td>50</td>
<td>0.50225</td>
<td>0.58534</td>
<td>0.61513</td>
<td>0.64213</td>
<td>0.69607</td>
<td>0.74036</td>
</tr>
<tr>
<td>75</td>
<td>0.41647</td>
<td>0.49026</td>
<td>0.51740</td>
<td>0.54238</td>
<td>0.59353</td>
<td>0.63709</td>
</tr>
<tr>
<td>100</td>
<td>0.36343</td>
<td>0.42999</td>
<td>0.45477</td>
<td>0.47774</td>
<td>0.52537</td>
<td>0.56662</td>
</tr>
</tbody>
</table>

Table 1. A subset of values of the secure threshold \( \tilde{p} \) for different number of samples \( n \) and genes \( m \).

doi:10.1371/journal.pone.0128115.t001
Fig 2. Contour plot of the function $p(n, m)$ in the Samples × Genes space on (a) a wide $(n, m)$ range and (b) zoomed on the small sample size area.

doi:10.1371/journal.pone.0128115.g002
familywise error rate to 5% significance level (FWER < 0.05): for a network with \( m \) nodes, by Bonferroni correction, this yields a \( p \)-value at most \( p_m = \frac{0.1}{m(n-1)} \) for each link in the graph. By using the inverse functions \( B_{5,F} = FT_{n}^{-1}(p_m) \) and \( B_{5,t} = tT_{n}^{-1}(p_m) \) we obtain the corresponding thresholds (Fisher and Student’s t, respectively) attaining the desired significance level. In Table 2 the values of \( B_{5,F} \) and \( B_{5,t} \) are reported for a selection of pairs \((n, m)\), together with the corresponding secure threshold \( \tilde{p} \), while in Fig 3 the surface plot of \( \tilde{p} \) is compared to \( B_{5,t} \). 

In all cases, the relation \( \tilde{p} < B_{5,F} < B_{5,t} \) holds, and when \( n < 10 \) both \( B_{5,F} \) and \( B_{5,t} \) cannot be defined, since the 5% significance level is not reached by any threshold. Both these results do not come unexpected, since the Bonferroni correction is known to be a very conservative procedure, especially when the number of tests is large (see for instance [107–109]) because it ignores existing relations between measurements. This results in a large False Negative Rate. Thus, in many situations in general and in -omics studies in particular, the Bonferroni correction should be substituted by more refined multiple testing approaches [110].
induced by the influence of subpopulations on network generation, and comparing to permuted data. Note that for a given pair \((n, m)\), the threshold \(\hat{p}\) is constant, while the threshold \(O\) depends on the particular dataset. In almost all cases, the threshold \(\hat{p}\) is the strictest.

As shown in the previous section, for a not very skewed distribution, the good approximation provided by the exact formula for \(F(n, p)\) given in Eq (2) guarantees the effectiveness of the secure threshold \(\hat{p}\) in detecting actual links between nodes. Nonetheless, whenever a stricter threshold is needed, it is still possible to follow the construction proposed, with the following refinement: the edge-creation process in the Erdős-Rényi model follows a binomial distribution, where \(n\) is the number of trials and \(p\) the probability associated with the success of a trial. The mean \(np\) of this distribution is one of the contributing terms in the definition Eq (3) of the secure threshold. To further restrict the number of falsely detected links, the variance term \((np(1-p))\) for the binomial distribution) can be added to the original formula through Chebyshev’s
inequality:
\[ P(\left| X - \mu \right| \geq k\sigma) \leq \frac{1}{k^2}, \]

where \( \mu \) and \( \sigma \) are the mean and the standard deviation of \( X \). Thus, the definition of secure threshold can be sharpened to \( \tilde{p}_k \) as follows, for an arbitrarily small \( \epsilon > 0 \):
\[
\tilde{p}_k = \min_{p \in (0,1)} \left\{ F(n,p) \frac{m(m-1)}{2} + k \sqrt{1 - F(n,p)} F(n,p) \frac{m(m-1)}{2} \leq 1 - \epsilon \right\}.
\]

For instance, the binomial distribution, for large values of \( n \), can be approximated as a normal distribution for which 95.45% of the values lie in the interval \((\mu - 2\sigma, \mu + 2\sigma)\). Moreover,

Table 3. Comparison of the secure threshold \( p \) with the clustering coefficient-based threshold \( C^* \) [67] and the statistical thresholds based on the adjusted \( p \)-values B0.01, B0.05 or 0.1 on a collection of synthetic and array datasets.

<table>
<thead>
<tr>
<th>Dataset type</th>
<th>( n )</th>
<th>( m )</th>
<th>( C^* )</th>
<th>B0.01</th>
<th>B0.05</th>
<th>B0.1</th>
<th>( \rho )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulated</td>
<td>50</td>
<td>1000</td>
<td>0.57</td>
<td>0.58</td>
<td>0.54</td>
<td>0.52</td>
<td>0.6152</td>
</tr>
<tr>
<td>Simulated</td>
<td>25</td>
<td>1000</td>
<td>0.69</td>
<td>0.76</td>
<td>0.72</td>
<td>0.70</td>
<td>0.7956</td>
</tr>
<tr>
<td>H-U133P</td>
<td>23</td>
<td>897</td>
<td>0.72</td>
<td>0.78</td>
<td>0.74</td>
<td>0.72</td>
<td>0.8125</td>
</tr>
<tr>
<td>H-U133P</td>
<td>10</td>
<td>897</td>
<td>0.78</td>
<td>0.96</td>
<td>0.94</td>
<td>0.93</td>
<td>0.9723</td>
</tr>
<tr>
<td>H-U133P</td>
<td>10</td>
<td>675</td>
<td>0.77</td>
<td>0.96</td>
<td>0.93</td>
<td>0.92</td>
<td>0.9681</td>
</tr>
<tr>
<td>H-U133P</td>
<td>9</td>
<td>897</td>
<td>0.79</td>
<td>0.97</td>
<td>0.96</td>
<td>0.95</td>
<td>0.9821</td>
</tr>
<tr>
<td>H-U133P</td>
<td>8</td>
<td>897</td>
<td>0.81</td>
<td>0.98</td>
<td>0.97</td>
<td>0.96</td>
<td>0.9899</td>
</tr>
<tr>
<td>H-U133P</td>
<td>7</td>
<td>897</td>
<td>0.81</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
<td>0.9955</td>
</tr>
<tr>
<td>H-U133P</td>
<td>6</td>
<td>897</td>
<td>0.86</td>
<td>&gt;0.99</td>
<td>&gt;0.99</td>
<td>&gt;0.99</td>
<td>&gt;0.999</td>
</tr>
<tr>
<td>H-U133P</td>
<td>5</td>
<td>897</td>
<td>0.92</td>
<td>&gt;0.99</td>
<td>&gt;0.99</td>
<td>&gt;0.99</td>
<td>&gt;0.999</td>
</tr>
<tr>
<td>H-U133P</td>
<td>4</td>
<td>897</td>
<td>0.99</td>
<td>&gt;0.99</td>
<td>&gt;0.99</td>
<td>&gt;0.99</td>
<td>&gt;0.999</td>
</tr>
<tr>
<td>H-U133A</td>
<td>4</td>
<td>675</td>
<td>0.99</td>
<td>&gt;0.99</td>
<td>&gt;0.99</td>
<td>&gt;0.99</td>
<td>&gt;0.999</td>
</tr>
<tr>
<td>H-I6</td>
<td>4</td>
<td>675</td>
<td>0.99</td>
<td>&gt;0.99</td>
<td>&gt;0.99</td>
<td>&gt;0.99</td>
<td>&gt;0.999</td>
</tr>
<tr>
<td>M-U74</td>
<td>4</td>
<td>401</td>
<td>0.97</td>
<td>&gt;0.99</td>
<td>&gt;0.99</td>
<td>&gt;0.99</td>
<td>&gt;0.999</td>
</tr>
</tbody>
</table>

Table 4. Comparison of the secure threshold \( p \) with the optimal threshold \( O \) [73] on a collection of array datasets on brain tissues. Note that for a given pair \((n, m)\), the threshold \( p \) is constant, while the threshold \( O \) depends on the particular dataset.

<table>
<thead>
<tr>
<th>Dataset type</th>
<th>( n )</th>
<th>( m )</th>
<th>( O )</th>
<th>( \rho )</th>
</tr>
</thead>
<tbody>
<tr>
<td>H-U133P</td>
<td>30</td>
<td>22277</td>
<td>0.82</td>
<td>0.8457</td>
</tr>
<tr>
<td>H-U133P</td>
<td>28</td>
<td>26199</td>
<td>0.85</td>
<td>0.8649</td>
</tr>
<tr>
<td>H-U133P</td>
<td>58</td>
<td>29211</td>
<td>0.8</td>
<td>0.6884</td>
</tr>
<tr>
<td>H-U133P</td>
<td>56</td>
<td>29211</td>
<td>0.8</td>
<td>0.6974</td>
</tr>
<tr>
<td>H-U95av2</td>
<td>50</td>
<td>12453</td>
<td>0.8</td>
<td>0.7026</td>
</tr>
<tr>
<td>Agilent</td>
<td>39</td>
<td>12235</td>
<td>0.82</td>
<td>0.7659</td>
</tr>
<tr>
<td>H-U133A</td>
<td>44</td>
<td>22383</td>
<td>0.85</td>
<td>0.7520</td>
</tr>
<tr>
<td>H-U95av2</td>
<td>47</td>
<td>12453</td>
<td>0.79</td>
<td>0.7186</td>
</tr>
<tr>
<td>H-U95av2</td>
<td>46</td>
<td>12453</td>
<td>0.84</td>
<td>0.7241</td>
</tr>
<tr>
<td>H-U95av2</td>
<td>50</td>
<td>12453</td>
<td>0.82</td>
<td>0.7026</td>
</tr>
<tr>
<td>H-U95av2</td>
<td>59</td>
<td>12453</td>
<td>0.79</td>
<td>0.6601</td>
</tr>
<tr>
<td>MOE 430_2</td>
<td>24</td>
<td>25859</td>
<td>0.83</td>
<td>0.8968</td>
</tr>
<tr>
<td>MOE 430_2</td>
<td>24</td>
<td>25859</td>
<td>0.78</td>
<td>0.8968</td>
</tr>
<tr>
<td>MOE 430_2</td>
<td>24</td>
<td>25859</td>
<td>0.87</td>
<td>0.8968</td>
</tr>
</tbody>
</table>

\[ \text{doi:10.1371/journal.pone.0128115.t003} \]

\[ \text{doi:10.1371/journal.pone.0128115.t004} \]
the Chebyshev’s inequality implies that at least 96% of the values lie in the interval \((\mu - 5\sigma, \mu + 5\sigma)\). Finally, in Text D of S1 Text we show the analogue of Table 1 for \(\tilde{p}_2\) and \(\tilde{p}_\tau\), respectively.

**Results and Discussion**

To conclude, we show the application of the secure threshold \(\tilde{p}\) in two datasets—synthetic and array—demonstrating its behaviour as a function of data subsampling.

**Synthetic dataset**

In order to construct a simulated microarray dataset \(\mathcal{G}\), we first created a correlation matrix \(M_\mathcal{G}\) on 20 genes \(G_1, \ldots, G_{20}\), together with a dataset \(\mathcal{G}\) of the corresponding expression \(G_{1000}^{i}\) across 1000 synthetic samples, so that \(M_\mathcal{G}(i, j) = |\rho(G_{1000}^{i}, G_{1000}^{j})|\) is the absolute PCC between the expression of the genes \(G_i\) and \(G_j\) from \(\mathcal{G}\). In particular, \(M_\mathcal{G}\) has two 10 \(\times\) 10 blocks highly correlated on the main diagonal, and two 10 \(\times\) 10 poorly correlated blocks on the minor diagonal, as shown in Fig 4(a).

These blocks are derived from the following generating rule, given uncorrelated starting element \(G_{1000}^1\) and \(G_{1000}^{11}\):

\[
|\rho(G_{1000}^k, G_{1000}^j)| \approx \begin{cases} 
1 - 0.03j & \text{for } k = 1, 2 \leq j \leq 10 \\
0.7 - 0.015j & \text{for } k = 11, 12 \leq j \leq 20 
\end{cases}
\]

Outside the two main blocks, all correlation values range between 0.002 and 0.074. In order to use \(M_\mathcal{G}\) as the ground truth in what follows, all values outside the two blocks on the diagonal are thresholded to zero \((M_\mathcal{G})_{10 + i, 10 + j} = (M_\mathcal{G})_{10 + i, 10 + j} = 0\) for \(1 \leq i, j \leq 10\), while entries in the two main blocks are considered as real numbers when evaluating HIM distance and binarized to one when evaluating True/False Positive/Negative links.

In Fig 4(b) we also show the heatmap of the gene expression dataset \(\mathcal{G}\). Then a subset of \(n_s\) samples is selected from the starting 1000, and the corresponding coexpression networks is built, for the 100 hard threshold values 0.01\(j\), for \(1 \leq j \leq 100\). The secure threshold for these cases are respectively 0.799, 0.596 and 0.389. This procedure is repeated 500 times for each value \(n_s = 10, 20, 50\). The same experiment is then repeated adding a 20% and a 40% level of Gaussian noise to the original data: for a given signal \(s\), we build \(s + \varepsilon\) with \(\varepsilon \sim N(0, \alpha \cdot (\max(s) - \min(s)))\) for \(\alpha = 0.2, 0.4\) respectively.

Using \(M_\mathcal{G}\) as the ground truth, for each hard threshold 0.01\(j\) we evaluate the ratio of False Positive links (i.e., the quotient between the number of False Positive links and 190, the number of all possible links in a complete undirected network on 20 nodes), the ratio of False Negative links and the Hamming-Ipsen-Mikhailov (HIM) distance from the gold standard. The HIM distance \([111, 112]\) is a metric between networks having the same nodes, ranging from 0 (distance between two identical networks) to 1 (comparison between the complete and the empty graph). The graphs summarizing the experiments are displayed in Fig 5.

In all cases, the secure threshold \(\tilde{p}\) corresponds to the strictest value yielding a coexpression network with no false positive links included, which is its characterizing property. Moreover, in almost all displayed situations, thresholding at \(\tilde{p}\) still guarantees an acceptable HIM distance from the ground truth, and a false negative ratio that is always less than 0.4.

**Ovarian cancer**

The aforementioned results obtained in the synthetic case are then tested here in a large array study on 285 patients of ovarian cancer at different stages [80], recently used in a comparative study on conservation of coexpressed modules across different pathologies [79]. In detail, the
authors profiled the gene expression of 285 predominately high-grade and advanced stage serous cancers of the ovary, fallopian tube, and peritoneum; the samples were hybridized on the Affymetrix Human Genome HG-U133 Plus 2.0 Array, including 54,621 probes. The goal of the original study was to identify novel molecular subtypes of ovarian cancer by gene expression profiling with linkage to clinical and pathologic features. As a major result, the authors presented two ranked gene lists supporting their claim that molecular subtypes show distinct survival characteristics. The two gene lists characterize the Progression Free Survival (PSF) and the Overall Survival (OS) patients, respectively. In each list genes are ranked according to a score weighting their association to target phenotype (PSF or OS): association is stronger for larger absolute values of the score, while the score’s sign is negative for associations with good outcome and positive for associations with poor outcome.

Following the procedure of the previous, synthetic example, first we individuate the sample subset corresponding to the homogeneous cohort of 161 patients with grade three cancer and a set \( T \) of 20 genes, with 11 genes strongly associated with good PSF or good OS (EDG7,
Fig 5. Synthetic dataset $G$. Coexpression inference of the $M_G$ network from random subsampling of the $G$ dataset, without noise (a,b,c), with 20% Gaussian noise (d,e,f) and with 40% Gaussian noise (g,h,i), on 10 (a,d,g), 20 (b,e,h) and 50 (c,f,i) samples. Solid lines indicate the mean over 500 replicated instances of HIM distance (black), ratio of False Positive (blue) and ratio of False Negative (red); dotted lines of the same color indicate confidence bars ($+/-\sigma$), while grey vertical dashed lines correspond to the secure threshold $p$.

doi:10.1371/journal.pone.0128115.g005
LOC649242, SCGB1D2, CYP4B1, NQO1, MYCL1, PRSS21, MGC13057, PPP1R1B, KIAA1324, LOC646769) and 9 genes strongly associated to poor PFS or poor OS (THBS2, SFRP2, DPSG3, COL11A1, COL10A1, COL8A1, FAP, FABP4, POSTN), thus generating a dataset \( O_T \) of dimension 161 samples and 20 features. The corresponding absolute PCC matrix \( O_T \) is then used as the ground truth for the subsampling experiments, thresholding to zero all values smaller than 0.1: the levelplot of \( O_T \) and the heatmap of \( O_T \) are displayed in Fig 6.

In these experiments, a random subdataset of \( n_s \) samples is extracted from \( O_T \), and the corresponding absolute PCC coexpression network on the nodes \( T \) is built, for increasing threshold values. In Fig 7 we report the HIM and the ratio of False Positive and False Negative links for 500 runs of the experiments, separately for \( n_s = 5, 10, 20 \) and 50. Again, the secure threshold \( \bar{p} \) corresponds to the smallest PCC value warranting that no false positive links are included. Finally, on average, for threshold values greater than \( \bar{p} \), the derivative of HIM distance is larger than before \( \bar{p} \), while the false negative rate remains under 0.8.

**Conclusions**

We have proposed a simple a priori, theoretical and non-parametric method for the selection of a hard threshold for the construction of correlation networks. This model is based on the requirements of filtering random data due to noise and reducing the number of false positives,
Fig 7. Ovarian cancer dataset. Coexpression inference of the coexpression network from subsampling of C7, on 5 (a), 10 (b), 20 (c) and 50 (d) samples. Solid lines indicate the mean over 500 replicated instances of HIM distance (black), ratio of False Positive (blue) and ratio of False Negative (red); dotted lines of the same color indicate confidence bars (+/-σ), while grey vertical dashed lines correspond to the secure threshold p.

doi:10.1371/journal.pone.0128115.g007
and it is implemented by means of analytic properties of PCC. This new approach can be especially useful where there is a small sample size and when the task requires minimising the number of false positive links, probably the most common situation in profiling studies in functional genomics. Finally, when the number of samples increases, coupling this method with soft thresholding approaches, can help recovering false negative links neglected by overly strict thresholds.

Supporting Information

S1 Text. Supplementary Text and Figures.
(PDF)

Author Contributions

Conceived and designed the experiments: AG GJ. Performed the experiments: AG GJ. Analyzed the data: AG GJ. Wrote the paper: AG GJ.

References


