Statistical Applications in Genetics and Molecular Biology

Volume 4, *Issue* 1 2005 *Article* 17

A General Framework for Weighted Gene Co-Expression Network Analysis

Bin Zhang* Steve Horvath†

*Departments of Human Genetics and Biostatistics, University of California at Los Angeles, binzhang.ucla@gmail.com

†Departments of Human Genetics and Biostatistics, University of California at Los Angeles, shorvath@mednet.ucla.edu

A General Framework for Weighted Gene Co-Expression Network Analysis*

Bin Zhang and Steve Horvath

Abstract

Gene co-expression networks are increasingly used to explore the system-level functionality of genes. The network construction is conceptually straightforward: nodes represent genes and nodes are connected if the corresponding genes are significantly co-expressed across appropriately chosen tissue samples. In reality, it is tricky to define the connections between the nodes in such networks. An important question is whether it is biologically meaningful to encode gene co-expression using binary information (connected=1, unconnected=0). We describe a general framework for 'soft' thresholding that assigns a connection weight to each gene pair. This leads us to define the notion of a weighted gene co-expression network. For soft thresholding we propose several adjacency functions that convert the co-expression measure to a connection weight. For determining the parameters of the adjacency function, we propose a biologically motivated criterion (referred to as the scale-free topology criterion).

We generalize the following important network concepts to the case of weighted networks. First, we introduce several node connectivity measures and provide empirical evidence that they can be important for predicting the biological significance of a gene. Second, we provide theoretical and empirical evidence that the 'weighted' topological overlap measure (used to define gene modules) leads to more cohesive modules than its 'unweighted' counterpart. Third, we generalize the clustering coefficient to weighted networks. Unlike the unweighted clustering coefficient, the weighted clustering coefficient is not inversely related to the connectivity. We provide a model that shows how an inverse relationship between clustering coefficient and connectivity arises from hard thresholding.

We apply our methods to simulated data, a cancer microarray data set, and a yeast microarray data set.

KEYWORDS: Scale-free Topology, Network Analysis, Clustering Coefficient, Hierarchical Organization, Module, Topological Overlap, Microarrays

---

*Please send correspondence to shorvath@mednet.ucla.edu. We would like to acknowledge the grant support from NINDS/NIMH 1U24NS043562-01 (PI Stanley Nelson). We are grateful for discussions with our UCLA collaborators Marc Carlson, Jun Dong, Tom Drake, Dan Geschwind, Jake Lusis, Ai Li, Paul Mischel, Stanley Nelson, and Andy Yip.

1 Introduction

Networks provide a straightforward representation of interactions between the nodes. Intuitive network concepts (e.g. connectivity and module) have been found useful for analyzing complex interactions.

Network based methods have been found useful in many domains, e.g. gene co-expression networks (Stuart *et al.*, 2003; Carter *et al.*, 2004; Butte and Kohane, 2000), protein-protein interaction networks (Jeong *et al.*, 2001; Rzhetsky and Gomez, 2001; Yook *et al.*, 2004), cell-cell interaction networks (Hartwell *et al.*, 1999), the world wide web and social interaction networks (Barabasi and Bonabeau, 2003; Csanyi and Szendroi, 2004).

In many real networks, the probability that a node is connected with k other node (the degree distribution $p(k)$ of a network) decays as a power law $p(k) \sim k^{-\gamma}$, which is the defining property of scale-free networks (Barabasi and Albert, 1999; Barabasi and Bonabeau, 2003; Jeong *et al.*, 2000). More details can be found in section 3. Scale-free networks are extremely heterogeneous, their topology being dominated by a few highly connected nodes (hubs) which link the rest of the less connected nodes to the system. For example, analysis of the yeast protein-protein interaction network revealed that highly connected nodes are more likely to be essential for survival (Jeong *et al.*, 2001; Han *et al.*, 2004; Carter *et al.*, 2004)

The emergence of power-law distribution (scale free topology) is intimately linked to the growth of the network in which new nodes are preferentially attached to already established nodes, a property that is also thought to characterize the evolution of biological systems (Barabasi and Albert, 1999; Albert and Barabasi, 2000), e.g. there is evidence that the scale-free topology of protein interaction networks originates from gene duplication (Barabasi and Oltvai, 2004; Rzhetsky and Gomez, 2001). Scale free networks display a surprising degree of tolerance against errors. For example, relatively simple organisms grow, persist and reproduce despite drastic pharmaceutical or environmental interventions, an error tolerance attributed to the robustness of the underlying metabolic network. The ability of nodes to communicate is unaffected even by very high failure rates in scale free networks(Albert *et al.*, 2000). But error tolerance comes at a high price in that these networks are extremely vulnerable to attacks, i.e. to the selection and removal of a few nodes that play a vital role in maintaining the network's connectivity (Albert *et al.*, 2000).

In this article, we focus on gene co-expression networks based on the tran-

scriptional response of cells to changing conditions. Since the coordinated co-expression of genes encode interacting proteins, studying co-expression patterns can provide insight into the underlying cellular processes (Eisen *et al.*, 1998). It is standard to use the (Pearson) correlation coefficient as a co-expression measure, e.g., the absolute value of Pearson correlation is often used in a gene expression cluster analysis.

Recently, several groups have suggested to threshold this Pearson correlation coefficient in order to arrive at gene co-expression networks, which are sometimes referred to as 'relevance' networks (Butte and Kohane, 2000; Carter *et al.*, 2004; Davidson *et al.*, 2001). In these networks, a node corresponds to the gene expression profile of a given gene. Thus nodes are connected if they have a significant pairwise expression profile association across the environmental perturbations (cell- or tissue- samples).

There are several questions associated with thresholding a correlation to arrive at a network. On the simplest level, how to pick a threshold? In section 4, we review several strategies for picking a 'hard' threshold (a number) based on the notion of statistical significance. Drawbacks of 'hard' thresholding include loss of information and sensitivity to the choice of the threshold (Carter *et al.*, 2004). On a more fundamental level, the question is whether it is biologically meaningful to encode gene co-expression using binary information (connected=1, unconnected=0).

We propose a general framework for 'soft' thresholding that weighs each connection by a number in [0,1]. Using simulated and empirical data, we provide evidence that weighted networks can yield more robust results than unweighted networks.

Below, we describe a general framework for constructing gene co-expression networks. We introduce three adjacency functions for converting a co-expression similarity measure into a connection strength. Then we propose a biologically motivated criterion for estimating the parameters of an adjacency function. Then, we generalize important network concepts (connectivity, clustering coefficient, topological overlap) to weighted networks. We use simulated and empirical data to provide evidence that the proposed methods are useful.

2 Steps of the Network Analysis

In gene co-expression networks, each gene corresponds to a node. A flowchart for constructing a gene co-expression networks is presented in Figure 1. We

assume that the gene expression data have been suitably quantified and normalized. For computational reasons, the network analysis is often restricted to a subset of genes (e.g. the 4000 most varying genes).

As we will detail below, each co-expression network corresponds to an adjacency matrix. The adjacency matrix encodes the connection strength between each pair of nodes. In unweighted networks, the adjacency matrix indicates whether or not a pair of nodes is connected, i.e. its entries are 1 or 0.

To begin with, one needs to define a measure of similarity between the gene expression profiles. This similarity measures the level of concordance between gene expression profiles across the experiments. The $n \times n$ similarity matrix $S = [s_{ij}]$ is transformed into an $n \times n$ adjacency matrix $A = [a_{ij}]$, which encodes the connection strengths between pairs of nodes. Since, the networks considered here are undirected, A is a symmetric matrix with non-negative entries. By convention, the diagonal elements of A are set to 0, i.e. $a_{ii} = 0$. Without loss of generality, we assume $a_{ij} \in [0, 1]$ for weighted networks. The adjacency matrix is the foundation of all subsequent steps. In particular, it is used to define node connectivity (as the row sum).

To define the adjacency matrix, one makes use of an adjacency function, which transforms the co-expression similarities into connection strengths. The adjacency function depends on certain parameters, which can be determined using different statistical or biological criteria. The resulting adjacency matrix is used to define a measure of node dissimilarity (distance). The node dissimilarity measure is used as input of a clustering method to define network modules (clusters of nodes). Once the modules have been defined, one can define additional network concepts, e.g. the intramodular connectivity.

Finally, the modules and their highly connected (hub-) genes are often related to external gene information. For example, we show in section 5 that the hub genes of a certain module are highly predictive of cancer survival (Mischel *et al.*, 2005). In section 6, we relate intramodular connectivity to a binary variable which encodes whether or not a gene is essential for yeast survival. Further, certain modules can be used to stratify the samples. We provide more details on these steps in the following.

2.1 The Definition of a Gene Co-expression Similarity

First, one needs to define a measure of similarity between the gene expression profiles. This similarity measures the level of concordance between gene

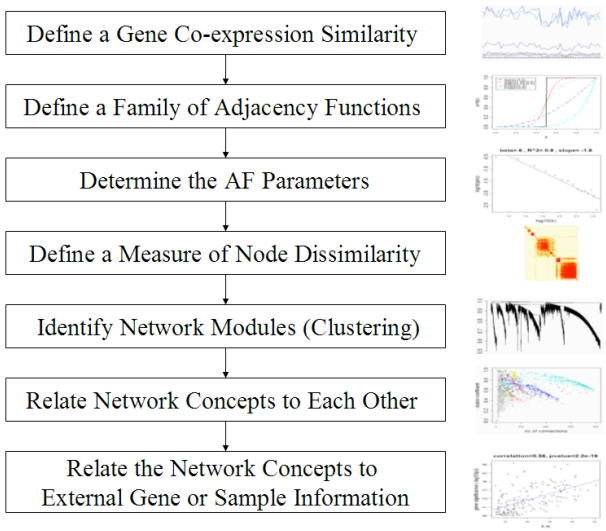


Figure 1: *Flowchart and illustration of gene co-expression network analysis. A typical figure has been placed to the right of each step. The meaning of the figures will be explained in the following sections.*

expression profiles across the experiments. Specifically, for each pair of genes i and j denote this similarity measure by s_{ij}. In the examples below, we will use the absolute value of the Pearson correlation $s_{ij} = |cor(i, j)|$. To protect against outliers, one could use a jacknifed correlation coefficient. To preserve the sign of the correlation $cor(i, j)$, one could use $s_{ij} = \frac{1+cor(i,j)}{2}$. The only mathematical restriction on the similarity measure is that its values lie between 0 and 1. We denote the similarity matrix by $S = [s_{ij}]$.

2.2 The Definition of a Family of Adjacency Functions

To transform the similarity matrix into an adjacency matrix, one needs to define an *adjacency function*. This choice determines whether the resulting network will be weighted (soft-thresholding) or unweighted (hard thresholding).

The adjacency function is a monotonically increasing function that maps the interval $[0, 1]$ into $[0, 1]$. The most widely used adjacency function is the signum function which implements 'hard' thresholding involving the thresh-

old parameter τ. Specifically,

$$a_{ij} = signum(s_{ij}, \tau) \equiv \begin{cases} 1 & if\ s_{ij} \geq \tau \\ 0 & if\ s_{ij} < \tau \end{cases} \quad (1)$$

Below, we discuss several approaches for choosing the threshold parameter τ.

Hard thresholding using the signum function leads to intuitive network concepts (e.g. the node connectivity equals the number of direct neighbors) but it may lead to a loss of information: if τ has been set to 0.8, there will be no connection between two nodes if their similarity equals 0.79.

To avoid the disadvantages of hard thresholding, we propose two types of 'soft' adjacency functions: the sigmoid function

$$a_{ij} = sigmoid(s_{ij}, \alpha, \tau_0) \equiv \frac{1}{1 + e^{-\alpha(s_{ij} - \tau_0)}} \quad (2)$$

with parameters α and τ_0, and the power adjacency function

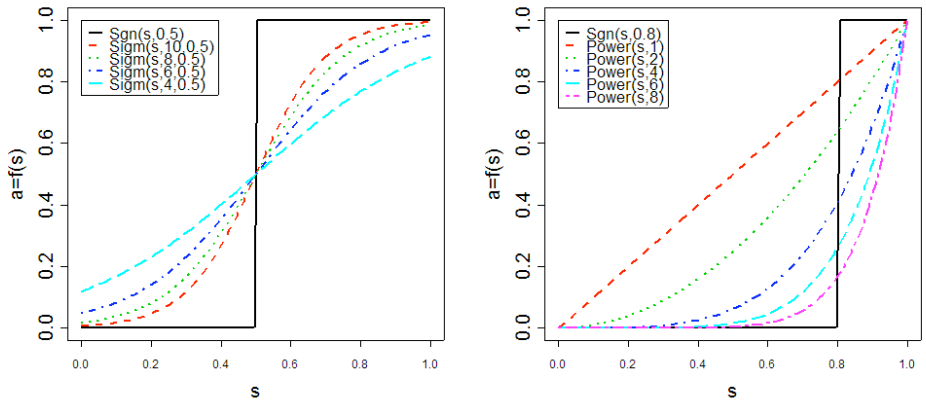
$$a_{ij} = power(s_{ij}, \beta) \equiv |s_{ij}|^{\beta} \quad (3)$$

with the single parameter β. The power adjacency function has the potentially attractive *factorization property* (see section 7.2). If s_{ij} factors into the contributions of nodes i and j (i.e. $s_{ij} = s_i s_j$) then a_{ij} factors as well $a_{ij} = a_i a_j$ where $a_i = (s_i)^{\beta}$.

The parameters of the adjacency functions can be chosen such that they approximate each other, see Figure 2.

As illustrated in appendix A, we find that the power- and the sigmoid adjacency functions lead to very similar results if the parameters are chosen with the scale-free topology criterion proposed in section 4.2.

One potential drawback of soft thresholding is that it is not clear how to define the directly linked neighbors of a node. A soft adjacency matrix only allows one to rank all the nodes of the network according to how strong their connection strength is with respect to the node under consideration. If a list of neighbors is requested, one needs to threshold the connection strengths, i.e. the values in the adjacency matrix. When dealing with an unweighted network, this is equivalent to the standard approach of hard thresholding the co-expression similarities since the adjacency function is monotonically increasing by definition.



(a) (b)

Figure 2: *Adjacency functions for different parameter values. a) Sidmoid and signum adjacency functions. b) Power and signum adjacency functions. The value of the adjacency function (y-axis) is plotted as a function of the similarity (co-expression measure). Note that the adjacency function maps the interval [0,1] into [0,1].*

2.3 Determining the Parameters of the Adjacency Function

The adjacency function depends on certain parameters, e.g. the signum function depends on the threshold parameter τ and the power function on the parameter β. How to determine these parameters is the subject of section 4. The choice of the parameters determines the sensitivity and specificity of the pairwise connection strengths. For example, increasing the value of τ leads to fewer node connections, which may reduce the noise in the network. If τ is chosen too high the resulting network may be too sparse for detecting the presence of gene modules (clusters of nodes).

2.4 Defining a Measure of Node Dissimilarity

An important aim of co-expression network analysis is to detect subsets of nodes (modules) that are tightly connected to each other. It is important to point out that authors differ on how they define gene modules. Here we

will consider module identification methods that are based on using a node dissimilarity measure in conjunction with a clustering method. Several such dissimilarity measures have been proposed. In this paper, we will use the topological overlap dissimilarity measure (Ravasz *et al.*, 2002) since it was found to result in biologically meaningful modules (see also our 2 real data applications). In appendix A, we provide a limited comparison of different node dissimilarity measures. A comprehensive comparison is beyond the scope of this article.

The **topological overlap of two nodes** reflects their relative interconnectedness. The topological overlap matrix (TOM) $\Omega = [\omega_{ij}]$ provides a similarity measure (opposite of dissimilarity), which has been found useful in biological networks (Ravasz *et al.*, 2002; Ye and Godzik, 2004). For unweighted networks (i.e. $a_{ij} = 1$ or $= 0$), Ravasz and colleagues report the following topological overlap matrix in the methods supplement of their paper (there is a typo in the main paper):

$$\omega_{ij} = \frac{l_{ij} + a_{ij}}{\min\{k_i, k_j\} + 1 - a_{ij}} \tag{4}$$

where $l_{ij} = \sum_u a_{iu} a_{uj}$, and $k_i = \sum_u a_{iu}$ is the node connectivity, see equation (6). In the case of hard thresholding, l_{ij} equals the number of nodes to which both i and j are connected. Note that $\omega_{ij} = 1$ if the node with fewer connections satisfies two conditions: (a) all of its neighbors are also neighbors of the other node and (b) it is connected to the other node. In contrast, $\omega_{ij} = 0$ if i and j are un-connected and the two nodes do not share any neighbors.

Formula (4) does not require that the adjacency matrix $A = [a_{ij}]$ contain binary entries (1 or 0). We propose to generalize the topological overlap matrix to weighted networks by simply using the real numbers $0 \leq a_{ij} \leq 1$ in formula (4).

Since $l_{ij} \leq \min(\sum_{u \neq j} a_{iu}, \sum_{u \neq i} a_{uj})$, it follows that $l_{ij} \leq \min(k_i, k_j) - a_{ij}$. Therefore, $0 \leq a_{ij} \leq 1$ implies $0 \leq \omega_{ij} \leq 1$.

The topological overlap matrix $\Omega = [\omega_{ij}]$ is a similarity measure (Kaufman and Rousseeuw, 1990) since it is non-negative and symmetric. To turn it into a dissimilarity measure, it is subtracted from one, i.e, the topological overlap based dissimilarity measure is defined by

$$d^{\omega}_{ij} = 1 - \omega_{ij}. \tag{5}$$

When it comes to clustering gene expression profiles (module definition), the TOM-based dissimilarity d_{ij}^{ω} leads to more distinct gene modules than the current standard measure (1 minus the absolute value of correlation coefficient) as illustrated in appendix A (Figure 15).

2.5 Identifying Gene Modules

Authors differ on how they define modules. Intuitively speaking, we assume that modules are groups of genes whose expression profiles are highly correlated across the samples. Our definition is slightly different from that of Bergmann et al (Bergman *et al.*, 2004). We adopt the definition of Ravasz et al (Ravasz *et al.*, 2002): modules are groups of nodes with high topological overlap.

To group genes with coherent expression profiles into modules, we use average linkage hierarchical clustering coupled with the TOM-based dissimilarity d_{ij}^{ω}. As an aside, we mention that we have also used the TOM-based dissimilarity in conjunction with partitioning around medoid clustering. A discussion of alternative cluster procedures and node dissimilarity measures is beyond the scope of this paper. In this article, gene modules correspond to branches of the hierarchical clustering tree (dendrogram). The simplest (not necessarily best) method is to choose a height cutoff to cut branches off the tree. The resulting branches correspond to gene modules, i.e. sets of highly co-expressed genes.

The choice of the height cut-off can be guided by the topological overlap matrix (TOM) plot (Ravasz *et al.*, 2002), which will be briefly reviewed in the following.

Topological Overlap Matrix Plots: A TOM plot provides a 'reduced' view of the network that allows one to visualize and identify network modules. The TOM plot is a color-coded depiction of the values of the TOM-based dissimilarity $[d_{ij}^{\omega}]$ for which the rows and columns are sorted by the hierarchical clustering tree that used the TOM-based dissimilarity as input. As an example, consider Figure 3b) where red/yellow indicate low/high values of d_{ij}^{ω}. Both rows and columns of $[d_{ij}^{\omega}]$ have been sorted using the hierarchical clustering tree. Since $[d_{ij}^{\omega}]$ is symmetric, the TOM plot is also symmetric around the diagonal. Since modules are sets of nodes with high topological overlap, modules correspond to red squares along the diagonal. As in all hierarchical clustering analyses, it is a judgement call where to cut the tree branches. Here the modules are found by inspection: a height cutoff value

is chosen in the dendrogram such that some of the resulting branches correspond to dark squares (modules) along the diagonal of the TOM plot. The genes of the resulting modules correspond to the color-coded bands along the rows and columns of the TOM plot, see Figure 3b).

2.6 Relating Network Concepts to Each Other

Once the network has been constructed (i.e. the adjacency matrix has been defined), several biologically important network concepts can be defined. In section 3, we generalize node connectivity and scale-free topology to weighted networks. In section 7, we generalize the clustering coefficient to weighted networks. The relationship between clustering coefficient and connectivity has important implications on the overall organization of the network, see Section 7.3. We have found that one can relate modules to each other by correlating the corresponding module eigengenes (Horvath *et al.*, 2005). If two modules are highly correlated, one may want to merge them. These types of analyses may allow one to define network 'diagnostics' that may aid in the network construction.

2.7 Relating Network Concepts to External Gene or Sample Information

A main purpose of many network analyses is to relate a connectivity measure to external gene information. For example, in the yeast network application, we show that intramodular connectivity in the turquoise module is highly correlated with gene essentiality, which was determined by gene knock-out experiments. In our cancer microarray application, we show that for brown module genes intramodular connectivity is highly correlated with prognostic significance for cancer survival. This facilitates novel strategies for screening for therapeutic targets (Brummelkamp and Bernards, 2003; Sonoda, 2003; Carter *et al.*, 2004; Mischel *et al.*, 2005). Standard statistical methods (e.g. regression models or multi-group comparison tests) can be used for these types of analyses.

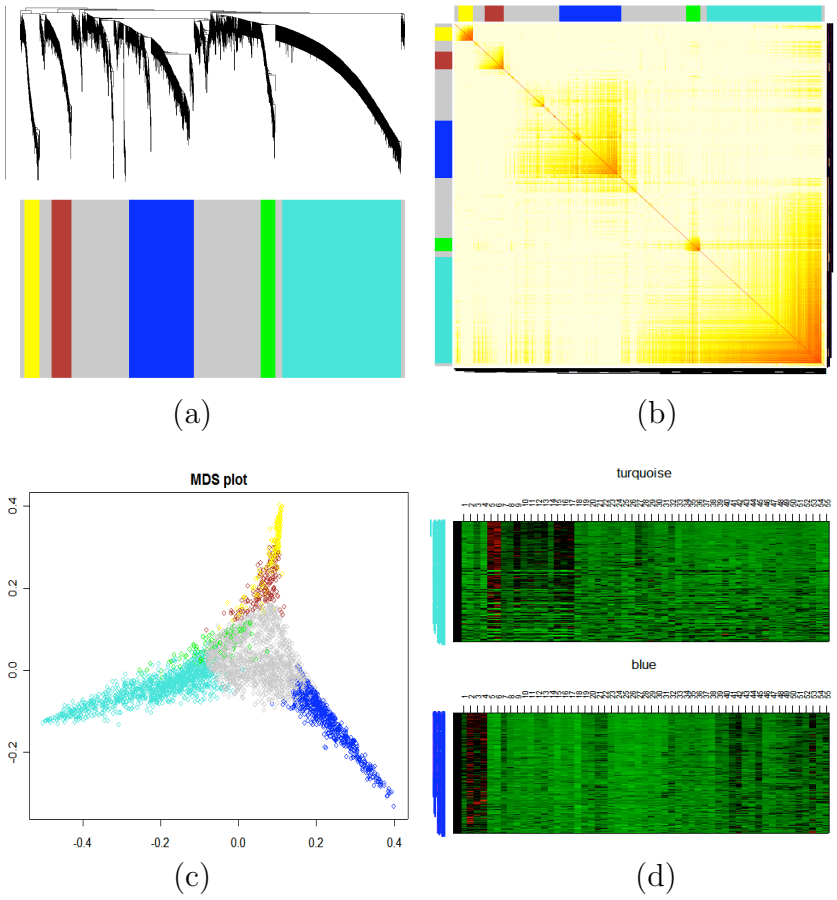


Figure 3: *Modules in the cancer co-expression network. (a) Hierarchical clustering tree using the topological overlap dissimilarity* $[1-\omega_{ij}]$. *Tree branches have been colored by module membership. b) Topological overlap matrix plot. Genes in the rows and columns are sorted by the clustering tree in a). Clusters correspond to squares along the diagonal. (c) The proposed classical multidimensional scaling plots of the TOM-based dissimilarity. Genes are colored according to the module membership defined in plot (a). (d) The gene expression profiles for the turquoise and blue module genes. Since the genes of a module (rows) are highly correlated across microarray samples (columns), one observes vertical bands. The expression profiles are standardized across arrays. Red corresponds to high- and green to low expression values.*

3 Connectivity and Scale-free Topology

Connectivity in Weighted Networks

For simple (unweighted) networks the connectivity k_i of node i equals the number of its direct connections to other nodes. In terms of the adjacency matrix, this can be expressed as follows

$$k_i = \sum_{j=1}^{n} a_{ij} \tag{6}$$

It is natural to use formula (6) to define the connectivity of node i in a general, weighted network. In this case, k_i is a non-negative real number.

A TOM-based Connectivity Measure

A key concept of network analysis is node connectivity (centrality). A central node (referred to as hub) is one with many connections to other nodes. The standard connectivity measure is given by equation (6), but many alternatives are possible, e.g. we propose the following TOM-based measure of connectivity ω_i

$$\omega_i = \sum_{j=1}^{n} \omega_{ij} \tag{7}$$

where ω_{ij} is the topological overlap between two nodes i and j. Thus, a node has high TOM-based connectivity ω_i if it has high overlap with many other nodes. In our cancer network application (section 5), we provide empirical evidence that ω_i can be superior to the standard connectivity k_i.

Intramodular Connectivity

A network connectivity measure can be defined with respect to the whole network (whole-network connectivity) or with respect to the genes of a particular module (intramodular connectivity). The distinction between whole-network connectivity and intramodular connectivity can be made for the standard connectivity measure of equation (6) and the topological overlap based connectivity of equation (7). In obvious notation, we denote the standard and TOM-based intramodular connectivity by $k.in$ and $\omega.in$, respectively.

We find the distinction between intramodular and whole-network node properties very important. In our applications, we provide evidence that the intramodular connectivity measures are biologically more meaningful than their whole-network analogs. But on a more theoretical level, it is not clear to us whether it is meaningful to compare whole-network connectivities across modules: a gene that is highly connected within a small but important module may have far fewer whole-network connections than a moderately connected gene in a large but unimportant module.

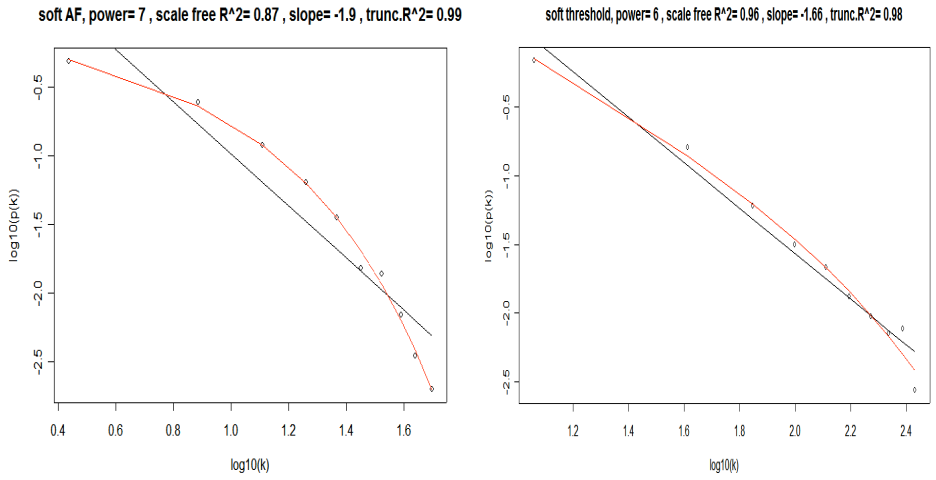
Generalized Scale-free Topology

As mentioned above, it has been found that the frequency distribution $p(k)$ of the connectivity follows a power law: $p(k) \sim k^{-\gamma}$. This definition naturally generalizes to weighted networks where k takes on non-negative real numbers.

To visually inspect whether approximate scale-free topology is satisfied, one plots $log_{10}(p(k))$ versus $log_{10}(k)$. A straight line is indicative of scale-free topology, see Figure 4).

To measure how well a network satisfies a scale-free topology, we propose to use the square of the correlation between $log(p(k))$and $log(k)$, i.e. the model fitting index R^2 of the linear model that regresses $log(p(k))$ on $log(k)$. If R^2 of the model approaches 1, then there is a straight line relationship between $log(p(k))$ and $log(k)$. This R^2 measure will play an essential role in the scale-free topology criterion described below.

Many co-expression networks satisfy the scale-free property only approximately. Figure 4a) shows that for our yeast network application, the connectivity distribution $p(k)$ is better modelled using an exponentially truncated power law $p(k) \sim k^{-\gamma} exp(-\alpha k)$, see also (Csanyi and Szendroi, 2004). In practice, we find that the 2 parameters α and γ provide too much flexibility in curve fitting: as illustrated by column 5 in Table 1, the truncated exponential model fitting index R^2 tends to be high irrespective of the adjacency function parameter. For this reason, we focus on the scale free topology fitting index in our scale free topology criterion. Exploring the use of the truncated exponential fitting index is beyond the scope of this article.



(a) (b)

Figure 4: *a) Scale free topology plot of the weighted yeast co-expression network that was constructed with the power adjacency function* $power(s, \beta = 7)$. *This scatter plot between* $log_{10}(p(k))$ *and* $log_{10}(k)$ *shows that the network satisfies a scale free topology approximately (black linear regression line,* $R^2 = 0.87$*). But a better fit is provided by an exponentially truncated power law (red line,* $R^2 = 0.99$*). b) Analogous plot for the weighted cancer network based on* $power(s, \beta = 6)$.

4 Choosing the Parameters of the Adjacency Function

4.1 Choosing the Threshold τ of the Signum Adjacency Function

Currently, the signum function is most frequently used to convert pairwise correlations into an adjacency matrix. The signum adjacency function leads to an unweighted network.

The estimation of the adjacency function parameter τ is tricky. Several authors have proposed to threshold the *significance level* of the correlation instead of the correlation coefficient itself. The significance level of a correlation coefficient can be estimated by using the Fisher transformation (Davidson

et al., 2001) or by using a permutation test procedure (Butte and Kohane, 2000; Carter *et al.*, 2004). Thus thresholding a correlation coefficient is replaced by thresholding the corresponding p-value. The significance level of a given correlation coefficient is a monotonic function of the underlying sample size (the number of microarrays). The network size (total number of connections) decreases as a function of the correlation threshold and its corresponding significance level. Thus when thresholding the significance level, the network size is highly dependent on the number of samples (microarrays). Another approach (Bergman *et al.*, 2004) to choosing a threshold is based on setting the network size equal a constant. Since the network size is a monotonically decreasing function of the correlation threshold, this is easily implemented.

Using statistical significance for determining the parameters of an adjacency function only works for hard thresholding. For soft thresholding, we propose the following criterion for determining the parameters of an adjacency function.

4.2 The Scale-free Topology Criterion

Instead of focusing on the significance of the correlation or the network size, we propose to pick the threshold by making use of the fact that despite significant variation in their individual constituents and pathways, metabolic networks have been found to display approximate scale free topology (Jeong *et al.*, 2000; Bergman *et al.*, 2004). This may indicate that metabolic organization is not only identical for all living organisms, but also complies with the design principles of robust and error-tolerant scale-free networks, and may represent a common blueprint for the large-scale organization of interactions among all cellular constituents (Jeong *et al.*, 2000; Bergman *et al.*, 2004; Ravasz *et al.*, 2002)

Most biologists would be very suspicious of a gene co-expression network that does not satisfy scale-free topology at least approximately. Therefore, adjacency function parameter values that give rise to networks that do not satisfy approximate scale-free topology should not be considered.

Earlier, we have described that the linear regression model fitting index R^2 can be used to quantify how well a network satisfies a scale-free topology. There is a natural trade-off between maximizing scale-free topology model fit (R^2) and maintaining a high mean number of connections: parameter values that lead to an R^2 value close to 1 may lead to networks with very few

connections. This trade-off is visualized in Figures 5 and 7 for our cancer- and yeast network application, respectively. Actually, we consider a signed version of the scale free topology fitting index. Since it is biologically implausible that a networks contains more hub genes than non-hub genes, we multiply R^2 with -1 if the slope of the regression line between $log_{10}(p(k))$ and $log_{10}(k)$ is positive.

These considerations motivate us to propose the following **scale-free topology criterion** for choosing the parameters of an adjacency function: *Only consider those parameter values that lead to a network satisfying scale-free topology at least approximately, e.g. signed* $R^2 > 0.80$. In addition, we recommend that the user take the following additional considerations into account when choosing the adjacency function parameter. First, the mean connectivity should be high so that the network contains enough information (e.g. for module detection). Second, the slope $-\hat{\gamma}$ of the regression line between $log_{10}(p(k))$ and $log_{10}(k)$ should be around -1.

When considering the signum and power adjacency functions, we find the relationship between R^2 and the adjacency function parameter (τ or β) is characterized by a saturation curve type of relationship (see Figures 5 and 7). *In our applications, we use the first parameter value where saturation is reached as long as it is above* 0.8. In our microarray data applications described in sections 5 and 6, we show that using an R^2 value of 0.95 and 0.85, respectively, leads to a high biological signal.

Example Table 1 reports the results for varying the signum parameter τ for the cancer microarray data (section 5). Here we focus on hard thresholding since it allows us to attach a significance level (p-values) to each threshold (using the Fisher transformation of the correlation coefficient). Values of $\tau \geq 0.3$ correspond to significant correlations. In our opinion, it would be difficult to use the p-values to argue that a parameter value of $\tau = 0.70$ ($p = 1.9 \times 10^{-9}$) is superior to $\tau = 0.50$ ($p = 8.7 \times 10^{-5}$) since both are highly significant. However, the scale-free topology fitting index R^2 clearly favors $\tau = 0.70$ ($R^2 = 0.97$) over $\tau = 0.50$ ($R^2 = 0.79$). In this application, we would choose $\tau = 0.70$ since this is where the R^2 curve starts to reach its 'saturation' point, see Figure 5. It leads to a good scale-free topology fit and also a high number of connections. There is indirect biological evidence that $\tau = 0.70$ leads to a network with a stronger biological signal than the network corresponding to $\tau = 0.50$. As can be seen from the last column in the Table, $\tau = 0.70$ and $\tau = 0.50$ lead to networks for which the intramodular connectivity $\omega.in$ has a Spearman correlation with prognostic

τ	p-value	signed scalefree R^2	slope	signed truncated exp. R^2	mean(k)	median(k)	max(k)	Biological Signal
0.20	1.4e-01	-0.68	1.79	-0.95	3530	3580	5520	0.08
0.30	2.5e-02	0.12	0.07	-0.95	1960	1890	4200	0.18
0.40	2.3e-03	0.60	-0.84	0.94	947	787	2940	0.40
0.50	8.7e-05	0.79	-1.21	0.90	395	232	1860	0.46
0.55	1.1e-05	0.85	-1.23	0.90	243	110	1410	0.51
0.60	1.0e-06	0.84	-1.24	0.87	145	43	1080	0.58
0.65	5.9e-08	0.95	-1.11	0.95	85.9	14	795	0.62
0.70	1.9e-09	0.97	-1.05	0.97	50.2	4	616	0.64
0.75	2.9e-11	0.98	-1.01	0.98	28.9	1	480	0.65
0.80	1.4e-13	0.95	-1.03	0.94	15.7	0	383	0.64
0.85	2.2e-16	0.97	-1.03	0.97	7.2	0	262	0.59
0.90	<1e-22	0.98	-1.09	0.98	2.2	0	154	0.55
0.95	<1e-22	0.93	-1.26	0.97	0.2	0	47	0.15

Table 1: *Cancer network characteristics for different hard thresholds τ. The asymptotic p-value for τ were calculated using the Fisher transform of the correlation coefficient. The sign of the scale-free model fitting index R^2 is determined by minus the sign of the slope. The last column contains measures the biological signal of interest in this analysis: the Spearman correlation between intramodular gene connectivity $\omega.in$ and prognostic gene significance, see Section 5.*

gene significance of $r = 0.64$ and $r = 0.46$, respectively. More details about this example are presented in the following section.

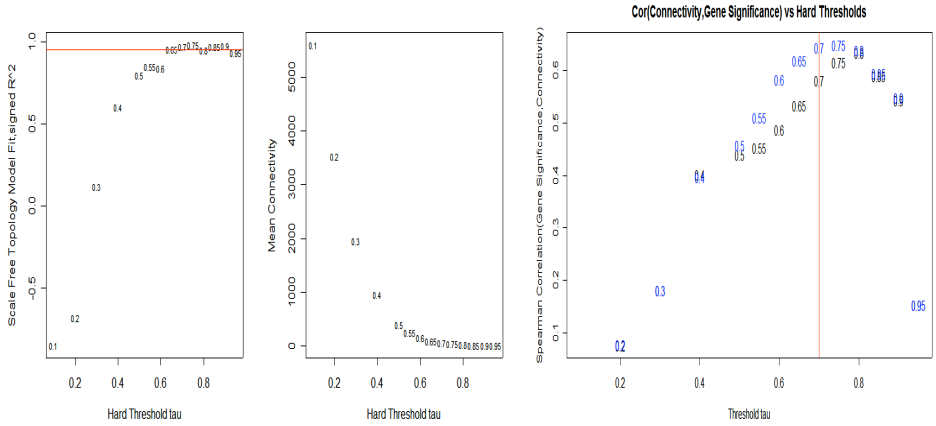
We are hesitant of formulating the scale free topology criterion as an optimization problem because noise affects the relationship between R^2 and the AF parameters (see Figures 5 and 7). These Figures and Table 1 show that unlike the mean number of connections, R^2 is not a *strictly* monotonic function of τ. Although the biological findings are fairly robust with respect to the R^2 cut-off, Figures 5 and 7 and Table 1 provide empirical evidence that the scale free topology criterion results in adjacency function parameter estimates that result in networks with a high biological signal.

5 Application I: Cancer Microarray Data

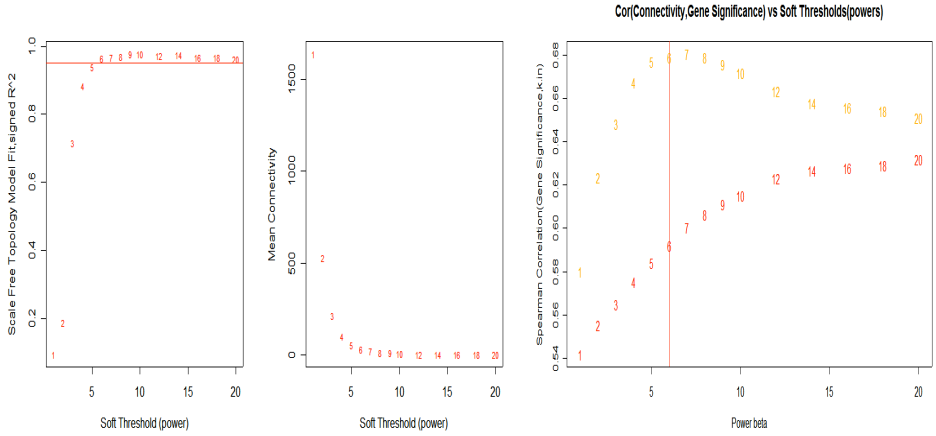
The proposed framework was used to analyze microarray data of 55 brain cancer microarray samples. The biological findings are described in (Mischel *et al.*, 2005). Here we focus on the statistical aspects of this analysis. To eliminate noise and for computational convenience, the analysis was restricted to the 8000 most varying genes (highest variance). The absolute value of the Pearson correlation was used as co-expression similarity measure. The power and signum adjacency function were used to construct weighted and unweighted networks, respectively. In appendix A, we discuss the performance of the sigmoid adjacency function as well. Figure 5 shows how the scale free topology fitting index R^2 depends on hard (τ) and soft thresholds (β). Since for the hard thresholds the R^2 curve levels off at $\tau = 0.7$ with $R^2 = 0.971$, we used $signum(s, \tau = 0.7)$ to construct the unweighted network. Since for the soft thresholds the R^2 curve levels off at $\beta = 6$ with $R^2 = 0.965$, we used $power(s, \beta = 6)$ to construct the weighted network. Different choices of these parameters are discussed below.

For module detection, we used the topological overlap node dissimilarity measure in average linkage hierarchical clustering.

The 6 major gene modules identified in the weighted network are color-coded in Figure 3. The TOM plot guided the choice of the height cut-off of the dendrogram. We found that the 6 modules were highly enriched for certain gene ontologies (functions) and will report the biological details elsewhere (Mischel *et al.*, 2005). Figure 6a) shows that these modules are highly preserved when considering an unweighted network constructed with the scale free topology criterion. Figure 6b) shows that the connectivities between

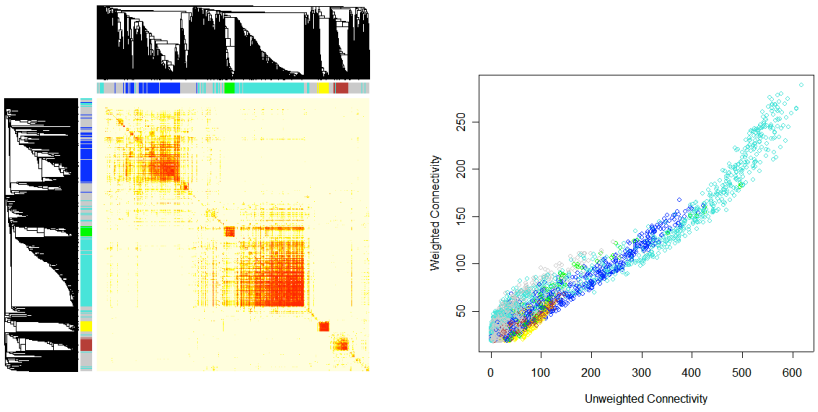


(a) $signum(s, \tau)$



(b) $power(s, \beta)$

Figure 5: *Cancer network properties for different hard and soft thresholds. For different hard thresholds (top row) and soft thresholds (bottom row), the plots visualize the scale free topology fitting index (first column), the mean connectivity (second column) and measures of biological signal (third column). Points are labelled by the corresponding adjacency function parameter. There is a trade-off between a high scale-free topology fit (R^2) and the mean number of connections. The scale free topology criterion picks adjacency function parameters that have a high biological signal (red vertical line in the third column). The biological signal is defined as the Spearman correlation between intramodular gene connectivity in the brown module and prognostic significance for patient survival. For the biological signal plot, the black and the blue curves in the first row correspond to the connectivity measures k.in and ω.in, respectively. Similarly, the red and orange curves in the second row correspond to k.in and ω.in, respectively.*



(a) (b)

Figure 6: *a) Topological overlap plot for the unweighted cancer network based on signum*$(s, \tau = 0.70)$*. Here the genes (rows and columns) are colored by the module assignment from the weighted network based on power*$(s, \beta = 6)$*. The plot shows that modules are highly robust with respect to the network construction method used. b) Scatterplot of the unweighted (hard thresholding) connectivity k versus the weighted (soft thresholding) connectivity colored by soft module colors. The connectivities are highly correlated with a Pearson correlation of 0.97.*

weighted and unweighted network are highly correlated (Pearson correlation $r = 0.97$). In appendix A, we show that the TOM-based dissimilarities that result from the signum, sigmoid and power adjacency functions are highly correlated if these networks were constructed using the scale free topology criterion. Thus, constructing the networks using the scale-free topology criterion leads to similar biological conclusions, which is reassuring.

A major goal of the analysis was to relate intramodular gene connectivity to prognostic significance for cancer survival. To define prognostic gene significance, we regressed patient survival on individual gene expressions using a univariate Cox regression model (Cox and Oakes, 1990; Klein and Moeschberger, 1997). Specifically, we defined the (prognostic) gene significance as $GS = -log_{10}(p)$ where p denotes the univariate Cox regression p-value.

Since the brown module was significantly (Pearson χ^2 p-value < 0.05) enriched with 'prognostic' genes, i.e. with genes whose univariate Cox p-value

was smaller than 0.05), we focused our attention on the brown module genes. Figure 5 shows that for brown module genes, the intramodular connectivities $k.in$ and $\omega.in$ (equation (7)) are highly correlated with gene significance, see also the third column in Table 1. In this application, we find that $\omega.in$ outperforms $k.in$. Further, connectivity measures from weighted networks consistently perform better than those from unweighted networks. Thus, we find that soft thresholding is superior to hard thresholding especially for low values of the scale free topology R^2. Soft connectivity measures are better than hard measures because they are relatively robust with respect to the parameter of the adjacency function. For soft thresholding even choosing a power of $\beta = 1$ leads to a relatively high correlation. In contrast, choosing a hard threshold of $\tau = 0.2$ leads to a network for which the biological signal of interest is reduced. Note that our proposed scale free topology criterion leads to estimates of the adjacency function parameter that are nearly optimal in this application.

For the biological signal plots in Figure 5, we compared the performance of the different intramodular connectivity measures by fixing the brown module. Since the module definition is highly variable due to its dependence on how the branches of the dendrogram are cut-off, fixing the module allows for a more direct comparison. In appendix B (Figure 16), we also report the results when the brown module definition changes 'adaptively' with different values of the adjacency function parameter. Our conclusions remain the same.

As mentioned before, we find that it is important to take a module centric view when relating connectivity to gene significance: in this application, we find no relation between whole network connectivity k and gene significance (correlation $r = 0.01$, p-value $p = 0.49$).

6 Application II: Yeast Cell-Cycle Microarray Data

The proposed framework was also used to analyze yeast cell-cycle microarray data of 44 samples. This dataset recorded gene expression levels during different stages of cell cycles in yeasts and has been widely used before to illustrate clustering methods (Spellman *et al.*, 1998). The analysis used the 4000 most varying genes (highest variance). Here our focus is to compare the performance of intramodular connectivity measures as a function of the adjacency function parameters. A discussion of the meaning of the modules

is beyond the scope of this article.

The biological goals of the analysis were a) to identify gene modules and b) to relate network connectivity to gene essentiality (gene knock-out effect) coded as 1 if the gene is essential for yeast survival and 0 otherwise. Thus, gene essentiality presents external information of gene significance, analogous to prognostic significance in the cancer microarray application.

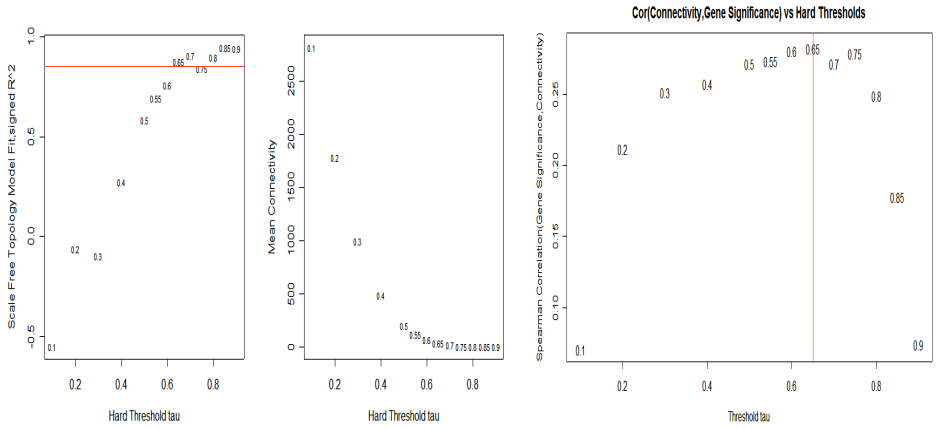
We used the scale-free topology criterion to choose the parameter values of the two adjacency functions by requiring that the scale-free topology fitting index R^2 be larger than 0.85, see Figures 7 and 4a.

For module detection, we used the TOM-based dissimilarity measure in average linkage hierarchical clustering, see Figure 8. Figure 8c) shows that the module assignment of the unweighted network is highly preserved in the weighted network. This shows that modules are robust with respect to the network construction method when the scale-free topology criterion is used.

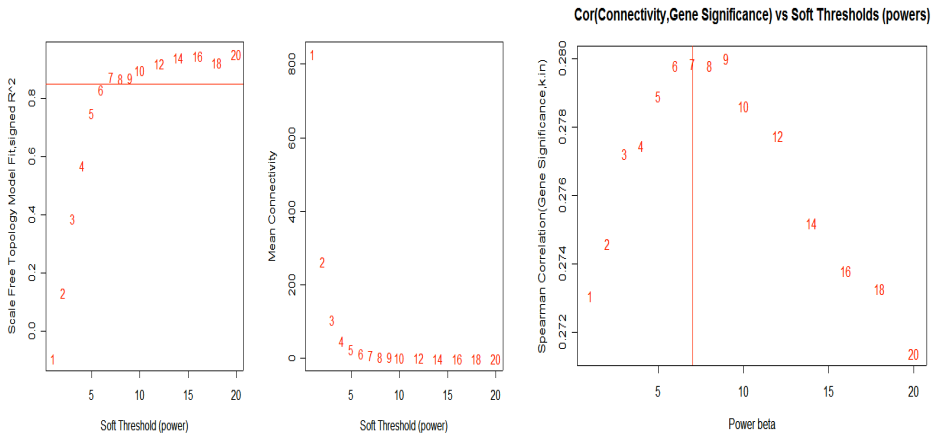
To study the relationship between connectivity and gene essentiality, we focused on the turquoise module since it was significantly enriched with essential genes (chi-square $p < 10^{-35}$). Figure 8c) related the biological signal to the scale free topology fitting index R^2 for different weighted and unweighted networks. The biological signal is defined as the Spearman correlation between intramodular connectivity (turquoise module) and gene essentiality (knock-out effect) information. Soft thresholding (weighted networks) leads to results that are much more robust with respect to the threshold chosen than those of hard thresholding.

Also we find that the intramodular connectivity has consistently higher correlations with gene essentiality than the whole network connectivity measures. For example, in the weighted network based on $power(s, \beta = 7)$, the turquoise intramodular connectivity leads to Spearman correlation of 0.28 while the whole network connectivity leads to a Spearman correlation of 0.17.

Since we do not find that the TOM-based intramodular connectivity $\omega.in$ performs better than $k.in$ in this application, we do not report it here. But details can be found in the corresponding R tutorial.

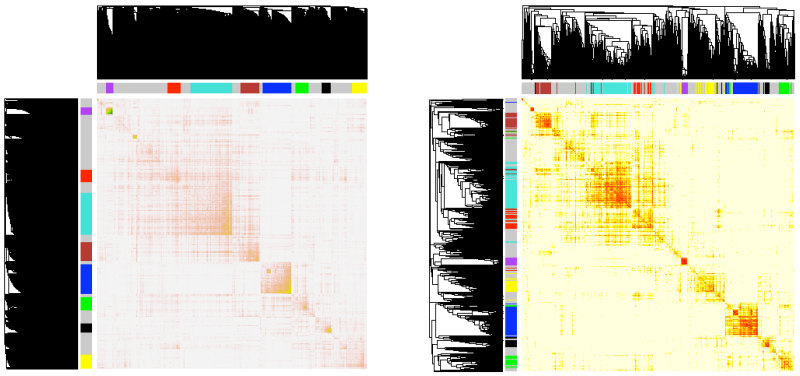


(a) $signum(s, \tau)$

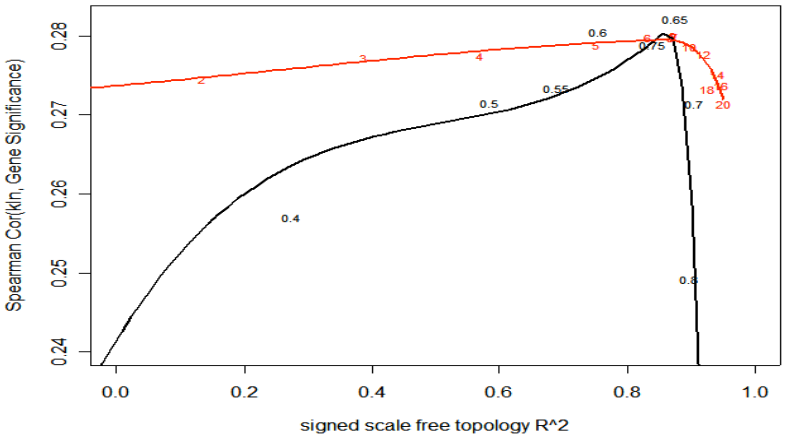


(b) $power(s, \beta)$

Figure 7: *Yeast network properties for different hard and soft thresholds. For different hard thresholds (top row) and soft thresholds (bottom row), the plots visualize the scale free topology fitting index (first column), the mean connectivity (second column) and a measure of biological signal (third column). The biological signal is defined as the Spearman correlation between intramodular gene connectivity in the turquoise module and gene essentiality. The analogous plots in the bottom row show the findings for the soft power adjacency function parameter β. Points are labelled by the corresponding adjacency function parameter. Clearly, there is a trade-off between a high scale-free topology fit (R^2) and a high mean number of connections.*



(a) (b)



(c)

Figure 8: *a) Topological overlap matrix plot of the weighted yeast network based on power$(s, \beta = 7)$. b) TOM plot of the unweighted yeast network based on signum$(s, \tau = 0.65)$. But the genes (rows and columns) are colored by the weighted network module assignment. The plot shows that modules are highly robust with respect to the network construction method used. We chose a different color coding from that in figure a) to enhance the signal. c) Scatterplot of the biological signal (intramodular correlation between connectivity and gene essentiality) versus the signed scale free topology fitting index for different hard (black) and soft thresholds (red). Points are labelled by the adjacency function parameter. Note that soft thresholding leads to results that are far more robust with respect to the choice of the adjacency function parameter.*

7 The Clustering Coefficient For Weighted Networks

The clustering coefficient of a node measures how 'cliquish' its neighbors are. The clustering coefficient C_i of node i ranges from 0 to 1. It is equal to 1 for a node at the center of a fully interlinked cluster, and it is 0 for a node whose neighbors are not connected at all. As an intuitive example consider social interaction networks: the cluster coefficient of a person is 1 (or 0) if all (or none) of their acquaintances know each other.

Let n_i be the total number of direct connections among the nodes connected to node i. For an unweighted network, n_i can be computed using the following formula

$$n_i = \frac{1}{2} \sum_{u \neq i} \sum_{\{v | v \neq i, v \neq u\}} a_{iu} a_{uv} a_{vi}. \tag{8}$$

Since the diagonal elements of the adjacency matrix equal 0 by convention, one could omit the index constraints in equation (8). But we report them explicitly to emphasize that the definition of the clustering coefficient ignores the diagonal elements (a_{ii}).

By definition, n_i is smaller than or equal to π_i, which is defined to be the maximum number of possible connections between its neighbors:

$$\pi_i = \frac{k_i(k_i - 1)}{2} \tag{9}$$

where k_i is the number of nodes directly connected to node i. As we will show below, equation (9) is only valid for unweighted networks.

Then the clustering coefficient of node i is defined as

$$C_i = \frac{n_i}{\pi_i}. \tag{10}$$

By definition, the clustering coefficient C_i of node i ranges from 0 to 1. The average clustering coefficient can be used to measure whether the network exhibits a modular organization (Ravasz *et al.*, 2002).

In the following, we generalize the n_i and π_i to the setting of weighted networks. Generalizing n_i is straightforward by using equation (8) with $0 \leq a_{ij} \leq 1$.

Generalizing π_i is more challenging. The key is to ensure that $n_i \leq \pi_i$ and $n_i = \pi_i$ for a fully interconnected network. Since $a_{ij} \leq 1$ and $a_{ii} = 0$ by

definition of the adjacency matrix, $a_{ij} \leq 1 - \delta_{ij}$ where δ_{ij} equals 1 if $i = j$ and 0 otherwise. Thus,

$$\begin{aligned} n_i &= \frac{1}{2}\sum_{u\neq i}\sum_{\{v|v\neq i, v\neq u\}} a_{iu}a_{uv}a_{vi} \\ &\leq \frac{1}{2}\sum_{u\neq i} a_{iu}(\sum_{v\neq i}(1-\delta_{uv})a_{vi}) \\ &= \frac{1}{2}\sum_{u\neq i} a_{iu}((\sum_{v\neq i} a_{vi}) - a_{ui}) \\ &= \frac{1}{2}((\sum_{u\neq i} a_{iu})^2 - \sum_{u\neq i} a_{iu}^2). \end{aligned}$$

Therefore, we define π_i for weighted networks as follows

$$\pi_i = \frac{1}{2}((\sum_{u\neq i} a_{iu})^2 - \sum_{u\neq i} a_{iu}^2). \quad (11)$$

One can show that equation (11) reduces to equation (9) in the case of unweighted networks since then $k_i = \sum_{u\neq i} a_{iu} = \sum_{u\neq i} a_{iu}^2$.

7.1 Soft Thresholding and the Clustering Coefficient

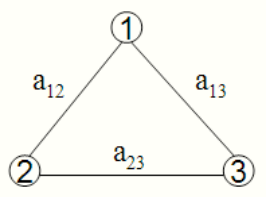


Figure 9: *A simple network with three nodes. The clustering coefficient of node 1 equals* a_{23}.

In the following, we provide a 'toy' example to illustrate the profound effect that soft thresholding has on the definition of the clustering coefficient.

Consider a network consisting of 3 nodes shown in Figure 9. Using formulas (8) and (11), one finds that $n_1 = a_{12}a_{23}a_{13}$ and $\pi_1 = \frac{1}{2}((a_{12}+a_{13})^2 - (a_{12}^2 + a_{13}^2)) = a_{12}a_{13}$. Thus, $C_1 = a_{23}$. In the case of hard thresholding C_1 takes on values 0 or 1. Unlike the case of unweighted networks, C_1 is a continuous function of the underlying correlation coefficient if soft-thresholding is used. Thus it is more robust with respect to the choice of the parameters of the adjacency function.

7.2 The Clustering Coefficient for Factorizable Adjacency Matrices

Here we show that the weighted clustering coefficient is approximately constant if the adjacency matrix A is factorizable, i.e. if $a_{ij} = a_i a_j, \forall i \neq j$. In this case, the vector $\mathbf{a} = (a_1, \ldots, a_n)^\tau$ is referred to as factorizability factor. One can show that if such a factor exists, it is unique (up to sign) when $n > 2$ (Horvath *et al.*, 2005). For factorizable networks ($=A$), the factorizability factor is highly correlated with the connectivity since

$$k_i = a_i \sum_{u \neq i} a_u. \tag{12}$$

We have found that weighted sub-networks comprised of the high connectivity genes of a particular module are approximately factorizable (Horvath *et al.*, 2005).

Using equations (8) and (11), the clustering coefficient $C_i = n_i/\pi_i$ can be rewritten with

$$n_i = \frac{1}{2}\sum_{u\neq i}\sum_{v\neq i, v\neq u} a_{iu}a_{uv}a_{vi} = \frac{1}{2}\sum_{u\neq i}\sum_{v\neq i, v\neq u} a_i^2 a_u^2 a_v^2 = \frac{a_i^2}{2}\left(\left(\sum_{u\neq i} a_v^2\right)^2 - \sum_{u\neq i} a_u^4\right)$$

and

$$\pi_i = \frac{1}{2}\left(\left(\sum_{u\neq i} a_u a_i\right)^2 - \sum_{u\neq i} a_u^2 a_i^2\right) = \frac{a_i^2}{2}\left(\left(\sum_{u\neq i} a_u\right)^2 - \sum_{u\neq i} a_u^2\right).$$

Therefore, we find that

$$C_i = n_i/\pi_i = \frac{\left(\sum_{u\neq i} a_u^2\right)^2 - \sum_{u\neq i} a_u^4}{\left(\sum_{u\neq i} a_u\right)^2 - \sum_{u\neq i} a_u^2}. \tag{13}$$

In several examples (e.g Figure 10c), we find that C_i is approximately constant:

$$C_i \approx \frac{\left(\sum_u a_u^2\right)^2 - \sum_u a_u^4}{\left(\sum_u a_u\right)^2 - \sum_u a_u^2}. \quad (14)$$

7.2.1 A Simple Simulated Network

Here we present a simple simulated network model for highlighting the differences between hard- and soft thresholding. A more realistic simulation can be found in section 8. We assume a network that is comprised of two completely unconnected modules with $n_1 = 40$ nodes (colored in blue) and $n_2 = 80$ nodes (colored in turquoise), respectively. Within each module, the similarity between nodes i and j is given by $s(i,j) = \frac{ij}{n^2}$ if $i \neq j$ and $s(i,i) = 1$. Thus the similarity matrix for the complete network is given by the following 120×120 matrix

$$S = \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix}$$

where $S_m = [\frac{ij}{n_m^2}]$.

The power adjacency function ($power(s, \beta = 2)$ and the signum adjacency function ($signum(\tau = 0.65)$ were used to define weighted and unweighted networks, respectively. Incidentally, these adjacency parameter values lead to approximate scale free topology: $R^2 \geq 0.85$. The TOM dissimilarity was used in average linkage hierarchical clustering to identify 2 modules in each network, see Figures 10a) and b).

This simulated example highlights a fundamental difference between unweighted and weighted networks when it comes to the relationship between clustering coefficient and connectivity. In the unweighted network, the clustering coefficient is anti-correlated with connectivity ($r = -0.70$) as shown in Figure 10d). In contrast, the clustering coefficient is approximately constant in the weighted network, as shown in Figure 10g). To elaborate on this difference, note that the within module similarity is factorizable. The power adjacency function preserves this property so that the cluster coefficient is approximately constant (see equation 14). In contrast, hard thresholding does not preserve the factorizability and the resulting adjacency matrix is not factorizable. This leads to a 'spurious' dependency between the cluster coefficient and the connectivity (Figure 10d)

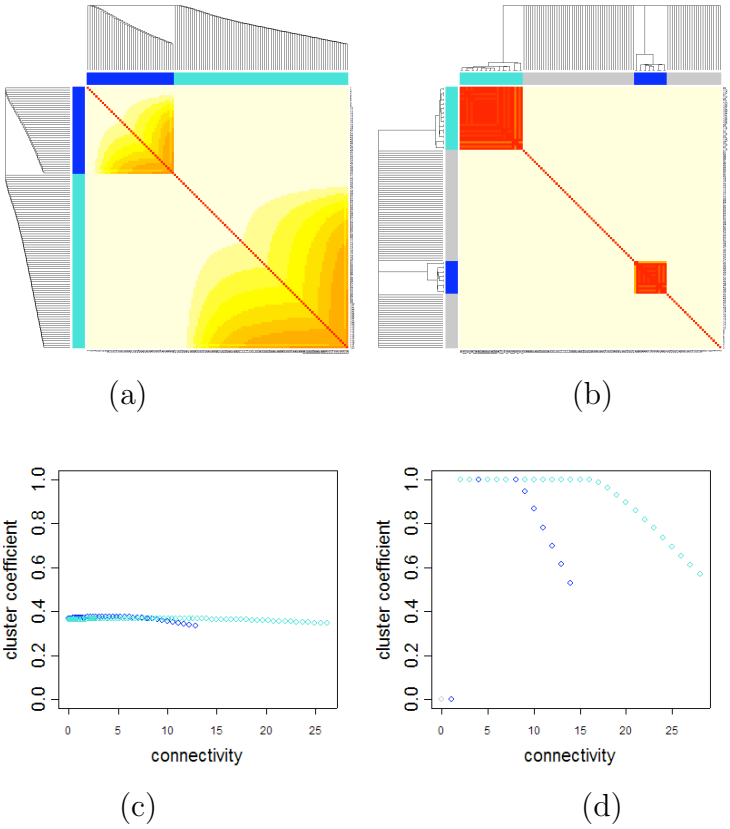


Figure 10: *Simulated example for highlighting the difference between hard and soft thresholding when it comes to the cluster coefficient. Figures (a) and (b) show the TOM plots in the weighted and unweighted network, respectively. There are two modules (turquoise and blue). Figures (c) and (d) show the relationship between cluster coefficient and connectivity colored by module membership. Note that hard thresholding gives rise to a strong inverse relationship between the cluster coefficient and the connectivity.*

Cancer Network	blue	brown	green	grey	turquoise	yellow
weighted	0.24	0.48	0.06	0.14	0.73	0.48
unweighted	-0.57	-0.76	-0.54	-0.21	-0.43	-0.91

Table 2: *Correlation coefficients between connectivity and cluster coefficient by cancer network module. The first row reports the correlations for the weighted network based on power*$(s, \beta = 6)$*. For the unweighted network based on signum*$(s, \tau = 0.70)$*, the second row reports the correlations between* k *and* C *among genes with connectivity* $k > 100$*. For the unrestricted relationship consider Figure 11.*

Yeast Network	black	blue	brown	green	grey	purple	red	turquoise	yellow
weighted	0.44	0.54	0.23	-0.05	0.06	0.61	0.57	0.45	0.29
unweighted	-0.76	-0.60	-0.58	-0.74	-0.12	-0.78	-0.56	-0.67	-0.79

Table 3: *Correlation coefficients between connectivity and cluster coefficient by yeast network module. The first row reports the correlations for the weighted network based on power*$(s, \beta = 7)$*. For the unweighted network based on signum*$(s, \tau = 0.65)$*, the second row reports the correlations between* k *and* C *among genes with connectivity* $k > 50$*. For the unrestricted relationship consider Figure 11.*

7.3 The Relationship between Cluster Coefficient and Connectivity in Real Networks

The relationship between connectivity and cluster coefficient is of interest since several authors have argued that it has implications for the overall structure of the network, (Ravasz *et al.*, 2002; Bergman *et al.*, 2004). For unweighted metabolic networks, it has been found (Ravasz *et al.*, 2002; Bergman *et al.*, 2004) that the clustering coefficient C is inversely related to the connectivity k, i.e. $C \sim k^{-1}$. To explain this relationship, Ravasz and colleagues proposed a 'hierarchical' network model that reconciles within a single framework all the observed properties of metabolic networks: their scale-free topology, high average clustering coefficient, and the power law scaling of C (Ravasz *et al.*, 2002). In contrast, non-hierarchical scale-free and modular networks predict that there is no relationship between C and k within a module.

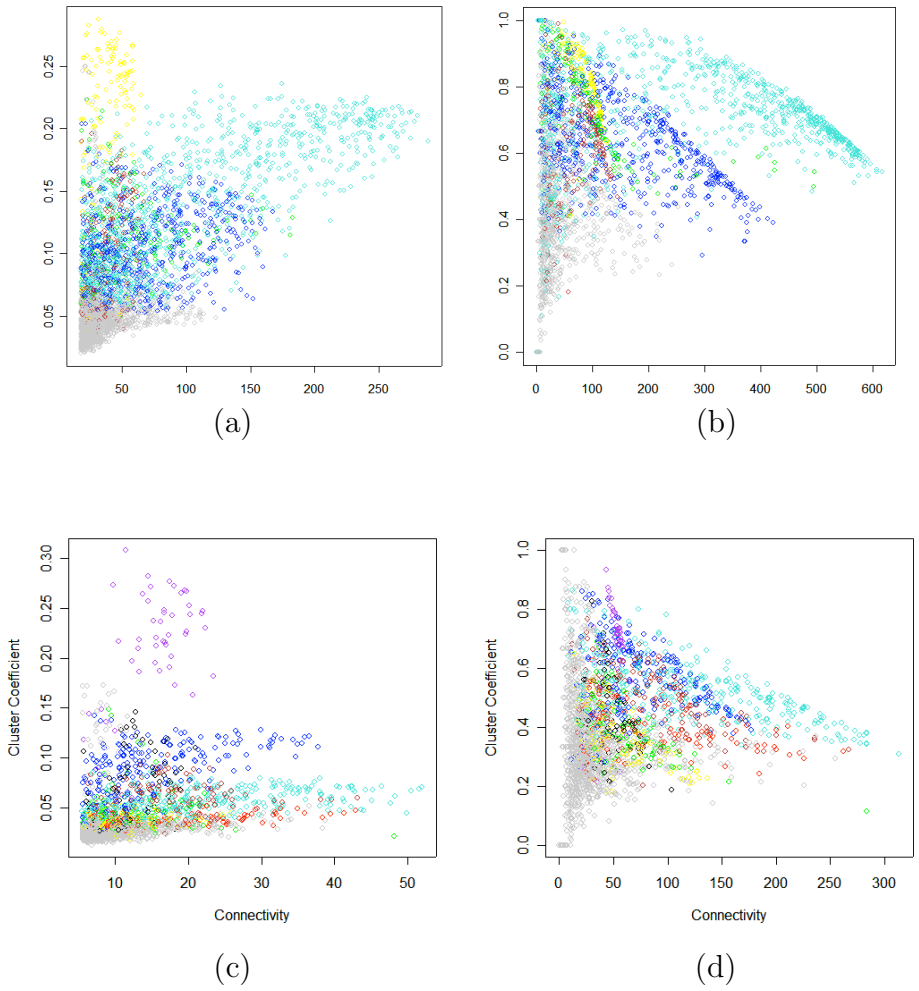


Figure 11: *Scatter plot of $C \sim k$ in different networks. Genes are colored by module membership. Figures in the the first row (a) and (b) correspond to the cancer network. Figures in the second row (c) and (d) correspond to the yeast network. Figures (a) and (c) in the first column correspond to soft thresholding with the power adjacency function. The figures in the second column correspond to hard thresholding using the signum function. Clearly, the inverse relationship between clustering coefficient and connectivity can only be observed for hard thresholding. For soft thresholding, the clustering coefficient is roughly constant for highly connected module genes.*

In *weighted* gene co-expression network, we find a positive correlation between connectivity and cluster coefficient in most modules, see Tables 2 and 3 and Figure 11. But the Figure also shows that for high connectivity nodes of a given module, the relationship between C and k is roughly constant. This approximately constant relationship can be explained using the fact that for highly connected genes inside a module, the corresponding correlation matrix is roughly factorizable, see (Horvath *et al.*, 2005). In contrast, for *unweighted* networks, we find an inverse relationship between cluster coefficient and connectivity in most modules. This is congruent with the findings reported by other groups (Bergman *et al.*, 2004; Ravasz *et al.*, 2002). The inverse relationship between C and k observed in unweighted networks may be an artefact of hard thresholding.

8 A More Realistic Simulated Example

Here we present a simulated network model that has some of the properties observed in real networks.

An R tutorial that describes this example in more detail can be found at our webpage. The example exhibits a nearly optimal (simulated) signal for weighted and unweighted networks constructed using the scale free topology criterion. The network was simulated to consist of a brown, a blue and a turquoise module with $n_1 = 100$, $n_2 = 200$, and $n_3 = 300$ genes, respectively. Further, it contained 500 grey (non-module) genes. For the genes of the brown module, we simulated an external gene significance measure $GS(i) = v_{signal}(j)$ and $v_{signal}(i) = (1 - 0.3i/n_m)^5$ where $1 \leq i \leq n_m$. One goal of the analysis is to study how the Spearman correlation between gene significance and intramodular connectivity in the brown module depends on different hard and soft thresholds.

The similarity matrix between genes of a given module contained a signal and a noise part, see Figure 13a). Specifically, for $i \leq 0.95 \times n_m$ and $j \leq 0.95 \times n_m$, we assumed that the similarity was given by $S_m(i,j) = min(v_{signal}(i), v_{signal}(j))$. The remaining 5 percent of (noise) genes were assumed to have a moderate similarity with the true signal genes. Specifically, for $i > 0.95 \times n_m$ and $j \leq 0.95 \times n_m$ (or for $i \leq 0.95 \times n_m$ and $j > 0.95 \times n_m$), we set $S_m(i,j) = v_{noise}(i) \times v_{noise}(j)$ where $v_{noise}(i) = (0.85 + 0.1 \times i/n)^5$. Further, we assumed that the noise genes had a high similarity between each other: for $i \geq 0.95 \times n_m$ and $j \geq 0.95 \times n_m$ the similarity matrix was set to

$S_m(i,j) = 0.95^5$.

The whole network similarity, which includes noise, is visualized in Figure 13b). Note that that the structure of the whole network similarity matrix is given by the 1100×1100 block-diagonal matrix

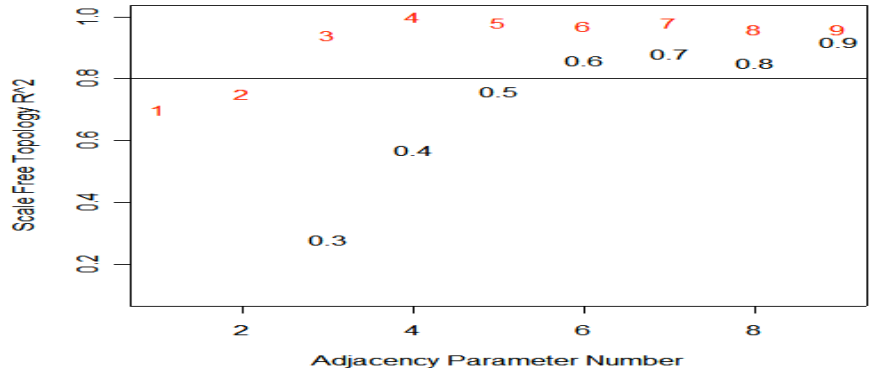
$$S = \begin{pmatrix} S_1 & 0 & 0 & 0 \\ 0 & S_2 & 0 & 0 \\ 0 & 0 & S_3 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Noise $\epsilon(i,j)$ was added to the entries $S(i,j)$ such that the result remained in the unit interval [0,1]. Specifically, $\epsilon(i,j) = (1 - S(i,j))U(i,j)^5$ where the $U(i,j)$ followed independent uniform distributions on the the unit interval $[0,1]$.

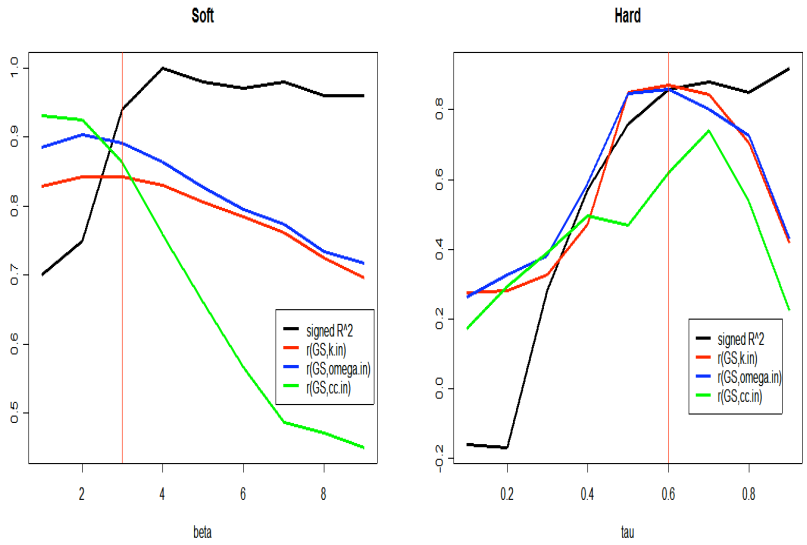
The scale free topology criterion was used to choose the hard and soft threshold parameter of the signum (unweighted network) and the power (weighted network) adjacency function, respectively. As can be seen from Figure 12a), the R^2 curve shows a kink above $R^2 > 0.80$ at a hard threshold of $\tau = 0.60$ and a soft threshold of $\beta = 3$. These adjacency function parameter values were used to define the corresponding topological overlap matrices. Average linkage hierarchical clustering was used to identify the modules, see Figures 13b) and e). The module identification method based on the weighted network recovers the underlying true module structure much better than an approach based on the unweighted network: while the weighted network module assignment misclassified 89 of the 1100 genes, the unweighted networks modules assignment misclassified 164 genes.

Figures 13e) and f) show the relationship between the cluster coefficient C and the connectivity k. Similar to the real data applications, we find a fundamental difference between the unweighted and weighted networks: there is a strong inverse relationship between C and k for high connectivity nodes in the unweighted network but not in the weighted network.

By construction, the simulated gene significance of the brown module genes correlated with the standard intramodular connectivity measure $k.in$. But for completeness, we also report the findings of correlating the TOM-based connectivity $\omega.in$ and the intramodular clustering coefficient $C.in$ with the gene significance, see Figures 12b) and c). Interestingly, we find that the TOM-based connectivity *omega.in* performs best in the weighted network. By construction, we find that the adjacency function parameters chosen by scale free topology criterion lead to nearly optimal biological signal when it



(a)



(b) (c)

Figure 12: *Simulated Network. a) Scale-free topology fitting index* R^2 *as a function of different hard (black curve) and soft (red curve) thresholds Points are labelled by the thresholds. The horizontal line corresponds to* $R^2 = 0.80$. *The scale free topology criterion leads us to choose power*$(s, \beta = 3)$ *for soft thresholding and signum*$(s, \tau = 0.60)$ *for hard thresholding. (b) Spearman correlations between simulated gene significance GS and different intramodular connectivity measures as a function of different values of the power adjacency function parameter* β. *The red curve shows the correlations for the standard intramodular connectivity measure k.in. The blue curve reports the findings for the TOM-based connectivity measure* $\omega.in$. *The green curve corresponds to the intramodular clustering coefficient* C. *The black curve reports the signed scale free topology model fitting index* R^2. *Analogous to b), figure c) shows the findings for the the hard threshold parameter* τ.

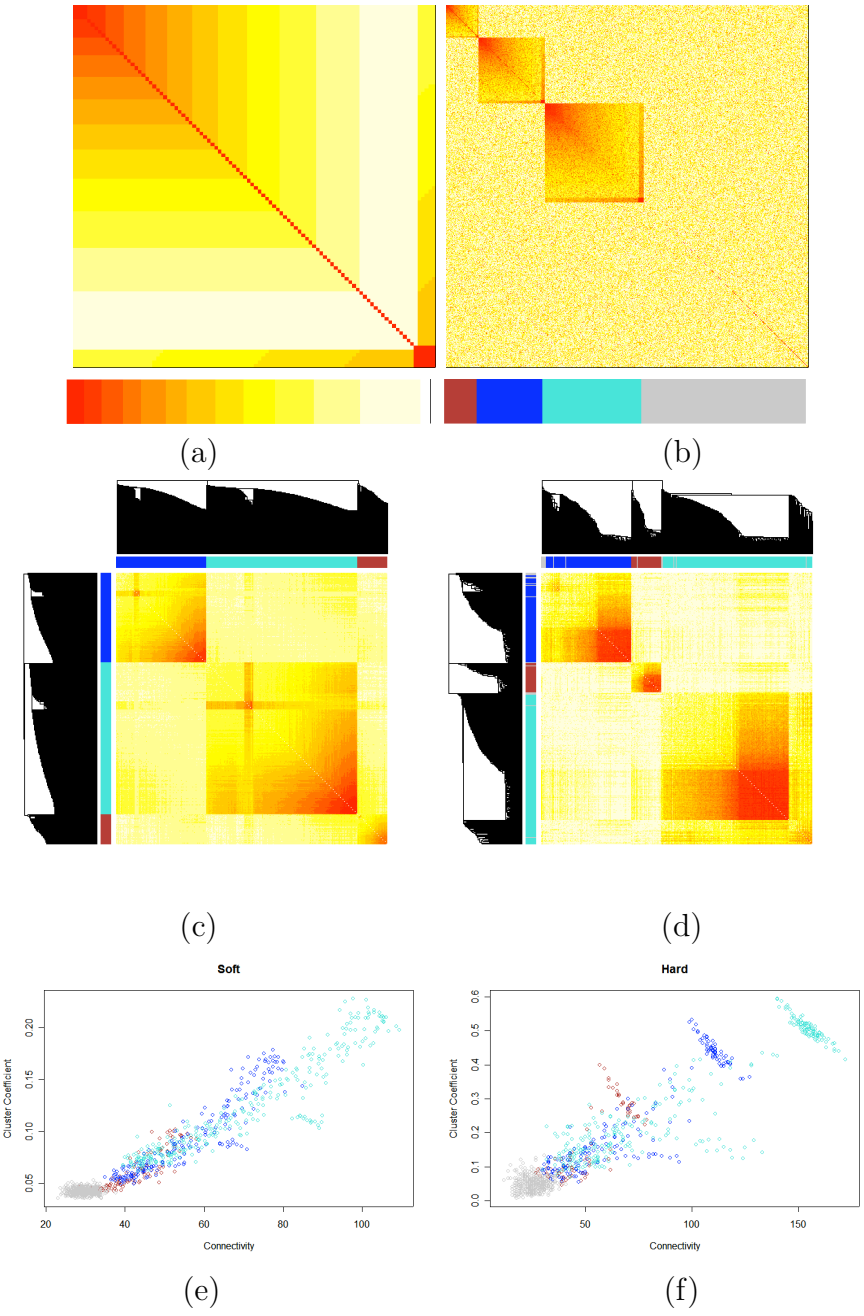


Figure 13: *(a) Color coded pictures of the brown module similarity matrix. The bottom row color-codes the simulated external gene significance measure GS for the brown module genes. (b) Whole-network similarity measure. The bottom row color codes the simulated module membership. (c) TOM plot of the weighted network restricted to nodes with connectivity* $k > 40$. *(d) Corresonding TOM plot for the unweighted network. Figures (e) and (f) depict the relationship between cluster coefficient* C *and connectivity* k *in the weighted and unweighted networks, respectively.*

comes to the Spearman correlation between $k.in$ and the gene significance. Our simulated example shows that the correlations of the weighted network are much more robust to the choice of the adjacency function parameter than those of the unweighted network.

Soft thresholding can avoid arbitrary discontinuities that sometimes result from hard thresholding. As the reader can verify using our software tutorial, we find that the results of a weighted network analysis are highly robust to the choice of the soft parameter β when it comes to module identification, connectivity definition, and the relationship between intramodular connectivity and an external gene significance measure. This is an attractive feature since it protects against (inadvertent) overfitting in a network analysis.

Software Implementation

R software tutorials for the yeast network, the cancer network, and the simulated example can be obtained from the following webpage:
http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork.

9 Conclusion

Unravelling the interactions between genes constitutes a major goal of biology. The structure of resulting gene networks is relevant to the functioning of the cell, for example, in development (Davidson *et al.*, 2003). Network analyses have shown a correlation between, on the one hand, the essentiality of a gene and, on the other hand, either the number of connections that the gene has (Jeong *et al.*, 2001; Han *et al.*, 2004; Carter *et al.*, 2004) or the topology of the metabolic network (Stelling *et al.*, 2002; Forster *et al.*, 2003). Furthermore, networks have been found useful to interpret synthetic lethal knockouts (Brummelkamp and Bernards, 2003; Sonoda, 2003). Network modules implement the classic idea that a cell can be divided into functional modules (Snel *et al.*, 2002; Yanai and DeLisi, 2002; Davidson *et al.*, 2003).

We present a general framework for constructing and analyzing gene co-expression networks. In order to construct weighted networks, we propose to use soft thresholding techniques to convert a gene co-expression similarity measure into a network connection strength. The parameters involved in various thresholding functions are estimated based on a biologically mo-

tivated criterion (referred to as scale-free topology criterion). We generalize important network concepts (connectivity, clustering coefficient, scale-free topology, topological overlap) from simple networks to weighted networks. Further, we introduce a new centrality measure based on topological overlap matrix. This new centrality measure is more predictive of gene importance than the standard measure in the cancer network and in the simulated example.

In this paper, we distinguish intramodular connectivity from whole network connectivity. Roughly speaking, the intramodular connectivity measures how connected a gene is to the genes of its module. We show that the intramodular connectivity is more strongly correlated with gene significance than the whole network connectivity. It is not clear to us whether it is meaningful to compare connectivity of genes across modules (sets of highly correlated genes): a gene that is highly connected within a small but important module may have far fewer whole-network connections than a moderately connected gene in a large but unimportant module.

We provide empirical evidence that the 'within' module clustering coefficient C has a weak positive dependence on connectivity k in weighted networks. In contrast, it is inversely related to the connectivity in unweighted networks. To understand this, we have derived an approximate formula for the relationship between k and C in factorizable networks, see section 7.2. It has been shown that the *correlation* matrix of genes from a 'tight' module is approximately factorizable (Horvath *et al.*, 2005). Soft thresholding with the power adjacency function preserves this property and the resulting adjacency matrix remains factorizable. Thus, *for highly connected* genes inside a module, one would expect a roughly constant relationship between k and C. In contrast, hard thresholding does not yield a factorizable adjacency function and an inverse relationship results as shown in our simulated example and in our real data applications. We find that the inverse relationship between clustering coefficient and intramodular connectivity derives from hard thresholding. It is worth pointing out that the cluster coefficients changes *across* modules. Genes belonging to different modules may have very different cluster coefficient as seen in our real data analysis.

Our empirical studies involving two DNA microarray data sets show that soft thresholding techniques, which result in weighted networks, lead to networks whose biological results tend to be highly robust with respect to the adjacency function parameters. In contrast, unweighted networks are less robust with respect to the threshold chosen. We have proposed a biologically

motivated criterion (referred to as the Scale-Free Topology criterion) that yields networks with high biological signal.

As we show in appendix A, when this criterion is used to estimate the parameters of the underlying adjacency functions, network concepts such as connectivity measures, hub status, and modules are quite robust with respect to the class of adjacency functions considered.

Appendix A: Comparison of the Signum-, Sigmoid- and Power Adjacency Function

Here we provide empirical evidence that the TOM-based measures d^{ω}_{ij} and the whole network connectivities k_i are highly correlated in networks that are constructed using the scale-free topology criterion. For the cancer network, the scale-free topology criterion was used to determine the parameters in the following adjacency functions: $signum(s, \tau = 0.7)$, $power(s, \beta = 6)$ and $sigmoid(0.9, 10)$. We also consider adjacency functions $power(s, 1)$ and $power(s, 2)$, which do not lead to approximate scale free topology.

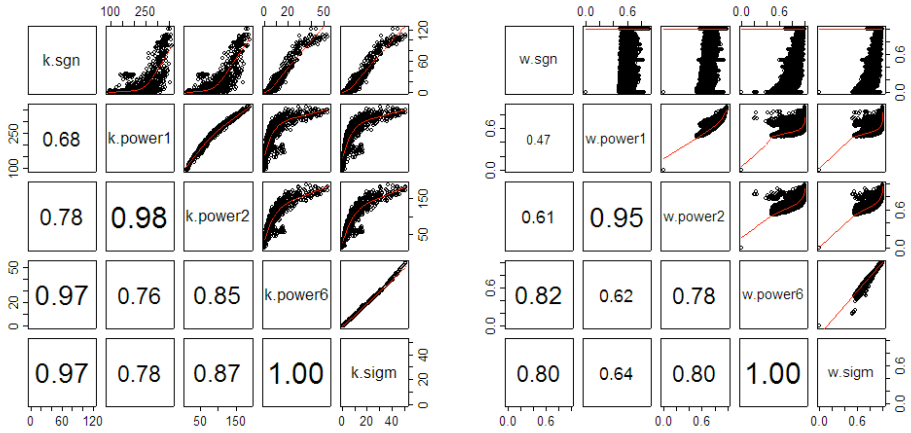
Figure 14a) shows that the whole network connectivities k_i corresponding to $signum(0.7)$, $power(s, 6)$ and $sigmoid(0.9, 10)$ are highly correlated (correlations bigger than 0.97). In contrast, these dissimilarities are less correlated with the ad-hoc dissimilarities $power(1)$ and $power(2)$. Figure 14b) shows that the TOM-based measures d^{ω}_{ij} corresponding to $signum(0.7)$, $power(s, 6)$ and $sigmoid(0.9, 10)$ are highly correlated. In contrast, these dissimilarities are less correlated with the ad-hoc dissimilarities $power(1)$ and $power(2)$.

Note also that both the connectivity and the TOM-based dissimilarity based on $power(s, \beta = 6)$ have a correlation of 1.0 with the corresponding quantities from the sigmoid adjacency function $sigmoid(0.9, 10)$.

Figure 15 shows the multidimensional scaling plots for the dissimilarity measure $1 - power(1)$ (which is widely used for clustering gene expression profiles) and the TOM-based dissimilarities resulting from different adjacency functions. The TOM-based measures lead to far more distinct modules than $(1-power(1))$. This can be used to illustrate how module identification using co-expression network analysis differs from standard gene clustering analysis.

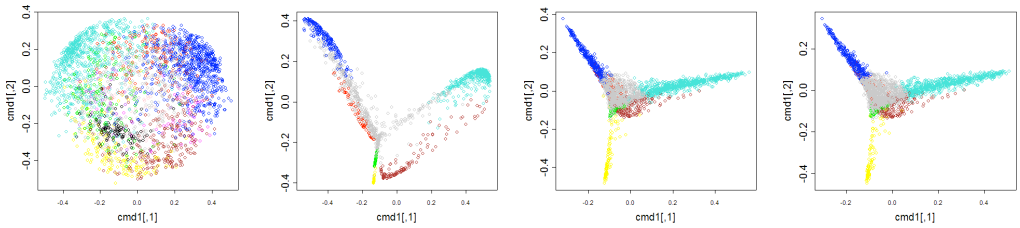
Appendix B: Comparing the Biological Signal for Adaptively Chosen Modules

In the main text, we compared the performance of the different connectivity measures by fixing the modules. Since the module definition is highly variable due to its dependence on how the branches of the dendrogram are cut-off, fixing the module allows for a more direct comparison. But here we report results when the brown module definition is changed adaptively and automatically for each value of the adjacency function parameter. Although,



(a) (b)

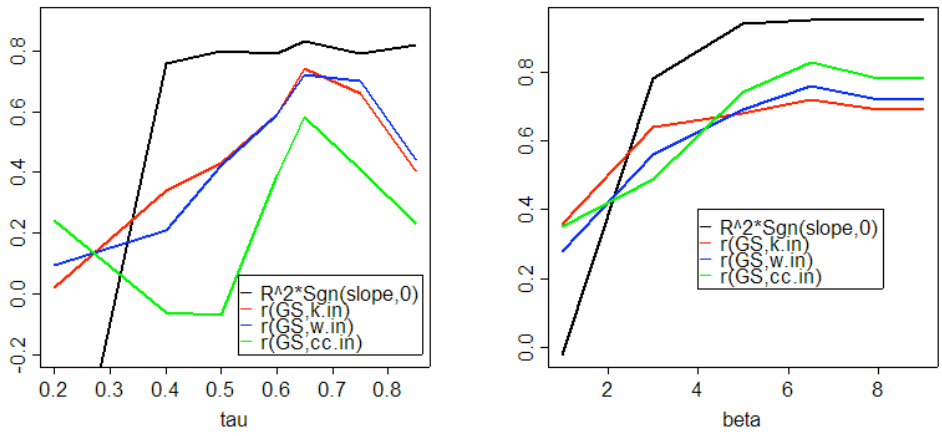
Figure 14: *The cancer microarray data are used to contrast different connectivity measures (a) and TOM-based dissimilarity measures (b) that result from different adjacency function. Above the diagonal are pairwise scatter plots and below the diagonal are the corresponding Pearson correlation coefficients. TOM-based dissimilaritys are preceded by the letter w for different adjacency functions.*



(a) (b) (c) (d)

Figure 15: *Multi-dimensional scaling plots of the genes as a function of different dissimilarity measures. (a)* $1 - power(1)$*, which is a widely used measure for clustering gene expression profiles; (b) TOM dissimilarity based on* $signum(s, \tau = 0.7)$*; (c) TOM dissimilarity based on* $power(s, \beta = 6)$*; (d) TOM dissimilarity based on* $sigmoid(s, \alpha = 10, \tau_0 = 0.9)$*.*

the results are more variable, one could argue that this comparison is more realistic. Figure 16 relates different intramodular connectivity measures of the brown module to prognostic gene significance in the cancer network. Again, we find that the biological signal is nearly optimal if the adjacency function parameter is chosen with the scale free topology criterion.



(a) (b)

Figure 16: *Spearman correlations between gene (prognostic) significance and different intramodular connectivity measures for different values of adjacency function parameters: a) correlations for the power adjacency function parameters β; b) the corresponding plot as a function of the hard threshold parameter τ. The black curve reports the signed scale free topology model fitting index R^2. The red curve shows the correlations for the standard intramodular connectivity measure k.in. The blue curve reports the findings for the TOM-based connectivity measure ω.in. The green curve corresponds to the intramodular clustering coefficient. In contrast to the results reported in the main text, the brown module definition changed adaptively for each value of the adjacency function parameter.*

References

Albert, R. and Barabasi, A. L. (2000). Topology of evolving networks: local events and universality. *Phys Rev Lett*, **85**(24), 5234–7.

Albert, R., Jeong, H. and Barabasi, A. L. (2000). Error and attack tolerance of complex networks. *Nature*, **406**(6794), 378–82.

Barabasi, A. L. and Albert, R. (1999). Emergence of scaling in random networks science. *Science*, **286**(5439), 509–512.

Barabasi, A. L. and Bonabeau, E. (2003). Scale-free networks. *Scientific American*, **288**, 60–69.

Barabasi, A L and Oltvai, Z N. (2004). Network biology: Understanding the cells's functional organization. *Nature Reviews Genetics*, **5**, 101–113.

Bergman, S, Ihmels, J and Barkai, N. (2004). Similarities and difference in genome-wide expression data of six organisms. *PLOS Biology*, **2**(1), 85–93.

Brummelkamp, T.R. and Bernards, R. (2003). New tools for functional mammalian cancer genetics. *Nat Rev Cancer*, **3**, 781–789.

Butte, A. J. and Kohane, I. S. (2000). Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*, **5**, 418–429.

Carter, Scott L., Brechbhler, Christian M., Griffin, Michael and Bond, Andrew T. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, **20**(14), 2242–2250.

Cox, D R and Oakes, D. (1990). *Analysis of survival data*. Chapman and Hall, London.

Csanyi, G. and Szendroi, B. (2004). Structure of a large social network. *Physical Review*, **69**(036131), 1–5.

Davidson, E.H., McClay, D.R. and Hood, L. (2003). Regulatory gene networks and the properties of the developmental process. *Proc Natl Acad Sci*, **100**, 1475–1480.

Davidson, G. S., Wylie, B. N. and Boyack, K. W. (2001). Cluster stability and the use of noise in interpretation of clustering. *Pages 23–30 of: IEEE Information Visualization 2001.*

Eisen, Michael B., Spellman, Paul T., Brown, Patrick O. and Botstein, David. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci*, **95**(25), 14863–14868.

Forster, J., Famili, I., Palsson, B.O. and Nielsen, J. (2003). Large-scale evaluation of in silico gene deletions in saccharomyces cerevisiae. *Omics*, **7**, 193–202.

Han, J. D., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J., Cusick, M.E., Roth, F.P. and Vidal, M. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**(6995), 88–93.

Hartwell, L., Hopefield, J., S., Leibler and Murray, A. (1999). From molecular to modular cell biology. *Nature*, **402**(6761 Suppl), C47–52.

Horvath, S., Dong, J. and Yip, A. (2005). Using the factorizability decomposition to understand connectivity in modular gene co-expression networks. *UCLA Technical Report. www.genetics.ucla.edu/labs/horvath/ModuleConformity/* .

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. and Barabasi, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, **407**(6804), 651–4.

Jeong, H., Mason, S. P., Barabsi, A. L. and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, **411**(May), 41–43.

Kaufman, Leonard and Rousseeuw, Peter J. (1990). *Finding groups in data: An introduction to cluster analysis.* New York: John Wiley Sons, Inc.

Klein, J P and Moeschberger, M L. (1997). *Survival analysis: Techniques for censored and truncated data.* Springer-Verlag, New York.

Mischel, P.s., Zhang, B., Carlson, M., Fang, Z., Freije, W., Castro, E., Scheck, A.C., Liau, L.M., Kornblum, H.I., Geschwind, D.H., Cloughesy, T.F.,

Horvath, S. and Nelson, S.F. (2005). A network approach to detecting individual prognostic genes and therapeutic targets in brain cancer. *Submitted.*

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. and Barabasi, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, **297**(Aug.), 1151–1155.

Rzhetsky, A. and Gomez, S. M. (2001). Birth of scale-free molecular networks and the number of distinct dna and protein domains per genome. *Bioinformatics*, **17**, 988–996.

Snel, B., Bork, P. and Huynen, M.A. (2002). The identification of functional modules from the genomic association of genes. *Proc Natl Acad Sci*, **99**, 5890–5895.

Sonoda, E. (2003). Multiple roles of rev3, the catalytic subunit of polzeta in maintaining genome stability in vertebrates. *EMBO*, **22**, 3188–3197.

Spellman, P., Sherlock, G., Zhang, M., Iyer, V. and Anders, K. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, **9**(12), 3273–3297.

Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S. and Gilles, E.D. (2002). Metabolic network structure determines key aspects of functionality and regulation. *Nature*, **420**, 190–193.

Stuart, J. M., Segal, E., Koller, D. and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**(5643), 249–255.

Yanai, I. and DeLisi, C. (2002). The society of genes: networks of functional links between genes from comparative genomics. *Genome Biol*, **3**, 64–68.

Ye, Y and Godzik, A. (2004). Comparative analysis of protein domain organization. *Genome Biology*, **14**(3), 343–353.

Yook, S Y, Oltvai, Z N and Barabasi, A L. (2004). Functional and topological characterization of protein interaction networks. *Proteomics*, **4**, 928–942.