


# Loss of Conservation of Graph Centralities in Reverse-engineered Transcriptional Regulatory Networks

Holger Weishaupt<sup>1</sup>  · Patrik Johansson<sup>1</sup> ·  
Christopher Engström<sup>2</sup> · Sven Nelander<sup>1</sup> ·  
Sergei Silvestrov<sup>2</sup> · Fredrik J Swartling<sup>1</sup>

Received: 30 October 2015 / Revised: 7 February 2017 /

Accepted: 1 March 2017 / Published online: 3 April 2017

© The Author(s) 2017. This article is published with open access at Springerlink.com

**Abstract** Graph centralities are commonly used to identify and prioritize disease genes in transcriptional regulatory networks. Studies on small networks of experimentally validated protein-protein interactions underpin the general validity of this approach and extensions of such findings have recently been proposed for networks inferred from gene expression data. However, it is largely unknown how well gene centralities are preserved between the underlying biological interactions and the networks inferred from gene expression data. Specifically, while previous studies have evaluated the performance of inference methods on synthetic gene expression, it has not been established how the choice of inference method affects individual centralities in the network. Here, we compare two gene centrality measures between reference networks and networks inferred from corresponding simulated gene expression data, using a number of commonly used network inference methods. The results indicate that the centrality of genes is only moderately conserved for all of the inference methods used. In conclusion, caution should be exercised when inspecting centralities in reverse-engineered networks and further work will be required to establish the use of such networks for prioritizing disease genes.

**Keywords** Transcriptional regulatory network inference · Simulated gene expression · Graph centrality

**Mathematics Subject Classification (2010)** 05Cxx · 92C42

---

This work was supported by grants from the Worldwide Cancer Research (formerly known as AICR) and the Swedish Childhood Cancer Foundation.

---

✉ Holger Weishaupt  
holger.weishaupt@igp.uu.se

<sup>1</sup> Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Dag Hammarskjöldsväg 20, 751 85 Uppsala, Sweden

<sup>2</sup> Division of Applied Mathematics, The School of Education, Culture and Communication (UKK), Mälardalen University, Box 883, 721 23 Västerås, Sweden

## 1 Introduction

With the increasing amount of ‘omics’ data made available to researchers during the last decades, biological network analysis has rapidly grown in its importance as one of the predominant methods of studying the underlying interactions and relationships between biological entities (Zhu et al. 2007). Current network based studies of topological properties of the related interactomes place particular interest on methods for clustering, pathway analysis, motif identification and graph based centrality measures on vertex or edge level (Ma and Gao 2012; Zhang et al. 2010; Aittokallio and Schwikowski 2006). Among these methods, vertex centrality can be considered the foremost technique in assigning importance to nodes in a variety of related networks (Zhang et al. 2010; Koschützki and Schreiber 2008). Importantly, it has previously been established that with respect to candidate gene identification, biological network analysis using centrality enrichment of nodes might in certain situations prove advantageous compared to a simple meta-analysis of genomic datasets (Langfelder et al. 2013).

Initial investigations of centralities in protein-protein interaction (PPI) networks of *Saccharomyces cerevisiae*, *Drosophila melanogaster* and *Caenorhabditis elegans* have suggested that developmentally and functionally essential proteins, i.e. proteins whose disruption leads to embryonal lethality, might be associated with high degree, closeness or betweenness centralities (Jeong et al. 2001; Joy et al. 2005; Hahn and Kern 2005; Estrada 2006a, b). Subsequently, graph centralities have also been investigated for the prediction of disease or cancer genes in human gene and protein networks (Özgür et al. 2008; Jonsson and Bates 2006; Wachi et al. 2005; Xu and Li 2006; Ortutay and Vihinen 2009; Siddani et al. 2013; Izudheen and Mathew 2013). It is however not yet fully established whether such an approach is meaningful. Specifically, if highly central genes are embryonically lethal or essential for development, then an early event leading to their ablation will lead to the death of the organism rather than to the development of a disease. In fact, it has been suggested that essential genes are often located at the center of hubs, while disease genes are frequently non-essential and found outside of hubcenters (Goh et al. 2007). However, while such a view is relevant especially for heritable diseases, cancer is considered a genetic disease brought on by somatic mutations. Thus cancer genes might coincide with essential genes, without causing embryonic lethality. Accordingly, genes with somatic mutations, as compared to non-essential disease genes, might still exhibit more central positions in such networks (Goh et al. 2007), especially when further considering that many cancer genes are characterized by a gain rather than loss of function, and drive abnormal proliferation and growth programs that are essential for embryonal development.

Another concern with such early studies relates to the fact that they have mainly been performed on PPI networks built from databases of validated biological interactions. Since such data usually contains only a limited number of generic interactions and biological entities, any related screen might miss important genes or might not be suitable to investigate the interactome specific for a given disease.

Alternatively, researchers often fall back on centrality based prioritization of genes from transcriptional regulatory networks (TRNs) reverse-engineered from expression data (Basso et al. 2005; Jörnsten et al. 2011; Emmert-Streib et al. 2014; Cordero et al. 2014; Knaack et al. 2014), which has become easily accessible for cells and tissues of normal and diseased conditions and can cover all known genes in the genome. Numerous methods for reverse-engineering of regulatory networks from expression data have been developed during the last decade (Margolin et al. 2006; Faith et al. 2007; Meyer et al. 2007; Langfelder and Horvath 2008; Altay and Emmert-Streib 2010a; De Matos Simoes and Emmert-Streib 2012;

Huynh-Thu et al. 2010; Yip et al. 2010; Zhang et al. 2012) (For a review of additional inference methods, we refer also to Marbach et al. 2012, 2010; Altay and Emmert-Streib 2010b; Liu 2015.) A problem with this approach however is the noise inherent to most expression datasets and the uncertainty in any network reconstructed from such data with available inference methods (Margolin and Califano 2007; Chalancon et al. 2012). Accordingly, a lot of effort has been dedicated to the evaluation of different aspects of inference accuracies and consistencies between such methods (Altay and Emmert-Streib 2010b; Schaffter et al. 2011; De Matos Simoes et al. 2013; Marbach et al. 2010; Marbach et al. 2012; Liu 2015).

However, it is still largely unexplored, how well the centralities in such inferred networks agree with the centralities of genes in the true underlying biological network. Considering that different methods are likely to make different systematic errors in inferring gene interactions (Schaffter et al. 2011; Marbach et al. 2012), it is reasonable to also expect different effects on the conservation of centrality values between inferred and true biological networks.

Here we obtain reference biological networks from a database of interactions in *Escherichia coli* (Schaffter et al. 2011) and from the *Pathway-Commons* database (Cerami et al. 2011), generate simulated gene expression for these networks using a model of stochastic differential equations, and utilize this data to reconstruct networks using a number of different inference methods. These benchmark datasets are then employed to estimate the agreement of degree and betweenness centralities between the reference and inferred networks for the different inference methods.

## 2 Generation of Benchmark Datasets

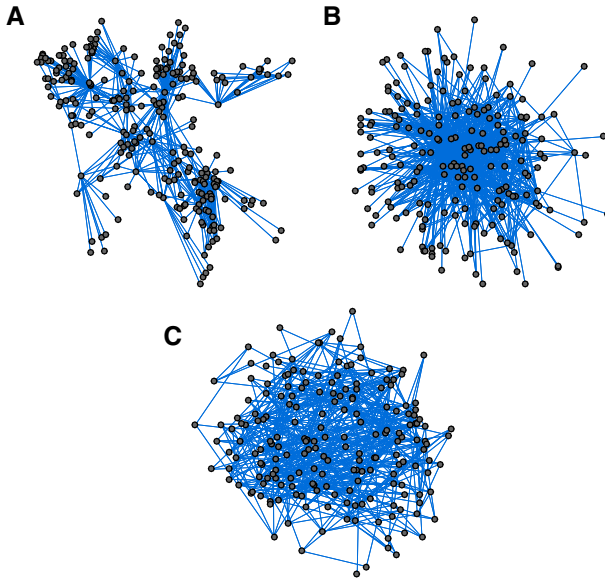
To determine how well centralities are conserved between true networks and networks inferred from gene expression data, we are here relying on benchmark datasets consisting of biological reference networks and synthetic expression data simulated from such reference networks according to Schaffter et al. (2011). Two types of reference networks are used here as described below.

**Ecoli250:** A set of 100 networks with 250 nodes each, extracted using the GeneNetWeaver (GNW) software (Schaffter et al. 2011) from the included *ecoli-regulonDB-6-7* dataset. The entire *ecoli* database contained 1565 nodes connected by 3758 edges. For each network ten random regulators (nodes with high out-degree centrality) were selected and additional nodes added from the neighborhood using a greedy search (Schaffter et al. 2011).

**PC200:** A set of 100 networks with 200 nodes each, extracted from the *Pathway-Commons* database. From the 4667832 edges (connecting 19006 nodes) with diverse interaction types in the *Pathway-Commons* database, only the subset of 105178 edges (connecting 13390 genes) of transcriptional regulatory nature, i.e. with an '*controls-expression-of*' interaction type, was chosen as a reference. Each benchmark network was then generated by selecting a random seed node and adding nodes from the neighborhood using a greedy search based on degree centralities.

The sizes of these networks was chosen from the range of network sizes investigated in Schaffter et al. (2011). A potential impact of the chosen network sizes on the following analyses will further be considered in the discussion below.

Based on the selection of seed nodes, the *Ecoli250* networks have a more nodular structure than the *PC200* networks (compare Fig. 1A-B).

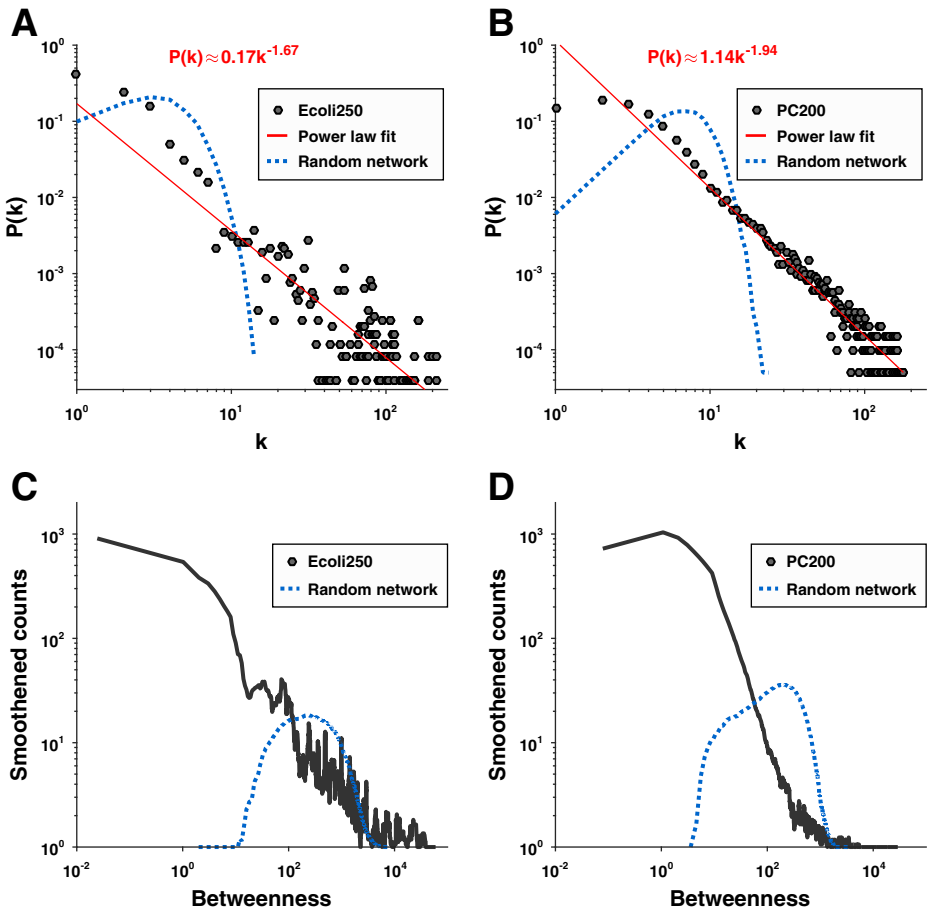


**Fig. 1** Example of *Ecoli250*, *PC200*, and random network structures. Depicted are the network structures of one example of an *Ecoli250* network, (A) and one *PC200* network (B), without displaying the direction of links. (C) A random network with a comparable number of edges was established using the Erdős-Rényi random graph model  $G(n, p)$  (Erdős and Rényi 1959), where  $n = 200$  denotes the number of nodes in the network and  $p = 0.025$  refers to the probability of placing an edge between any pair of nodes, which was chosen here to obtain a graph with a number of edges comparable to the *PC200* network

However, as compared to random networks (compare Fig. 1C), both sets of networks exhibit a roughly scale free topology as the distribution of degree centralities in these networks can be modeled using a power law of the form  $P(k) = a \cdot k^{-\lambda}$  (compare Fig. 2A-B), which is one of the characteristic properties of biological networks (Albert 2005). In addition, the investigation of betweenness centralities in these networks similarly shows a distribution with the majority of nodes exhibiting small betweenness and high betweenness values only observed in a small number of nodes, while in random networks the majority of nodes display a small range of typical betweenness values, i.e. in random networks nodes generally have very similar betweenness centralities. Thus, in the biological networks but not the random networks, there are genes with uncharacteristically high degree and betweenness values, which allows the use of these measures for centrality based gene prioritization.

Different methods have been described for the generation of *in-silico* gene expression data from an existing network structure, with the predominant method relying on the use of coupled ordinary or stochastic differential equations (Knüpfer et al. 2004; Mendes et al. 2003; Wu et al. 2014; Schaffter et al. 2011). Here, expression data for each reference network was then simulated using GNW, which makes use of stochastic differential equations (SDEs) to model transcriptional and translational processes (Schaffter et al. 2011). Specifically, the noise-free rate of concentration changes  $\frac{dY_t}{dt}$  for mRNA and proteins are simulated using an ordinary differential equation (ODE) of the form

$$\frac{dY_t}{dt} = G(Y_t) - D(Y_t),$$



**Fig. 2** Distribution of degree and betweenness centralities in the *Ecoli250* and *PC200* networks. (A–B): Degree centralities in *Ecoli250* (A) and *PC200* (B) and corresponding random networks. The scatter plot shows the number of nodes with certain degree values averaged over all biological networks in each set. The red line indicates a polynomial fit plotted according to the depicted equation in red. The blue dotted line depicts the averaged degree distribution over 100 random  $G(n, p)$  networks each, with  $n = 250$  (A) or  $n = 200$  (B) and  $p$  is individually set in accordance with the number of links observed in each of the biological networks. (C–D): Betweenness centralities in *Ecoli250* (C) and *PC200* (D) and corresponding random networks. Grey solid lines and blue-dotted lines indicate a LOESS smoothing fit to the betweenness values observed in the biological and random networks from (A–B), respectively, excluding nodes with zero betweenness

where  $G(Y_t)$  describes the contribution through production at time  $t$  and  $D(Y_t)$  describes the amount of degradation of the product at time  $t$ . The processes under biological noise are then simulated using the SDE of the form

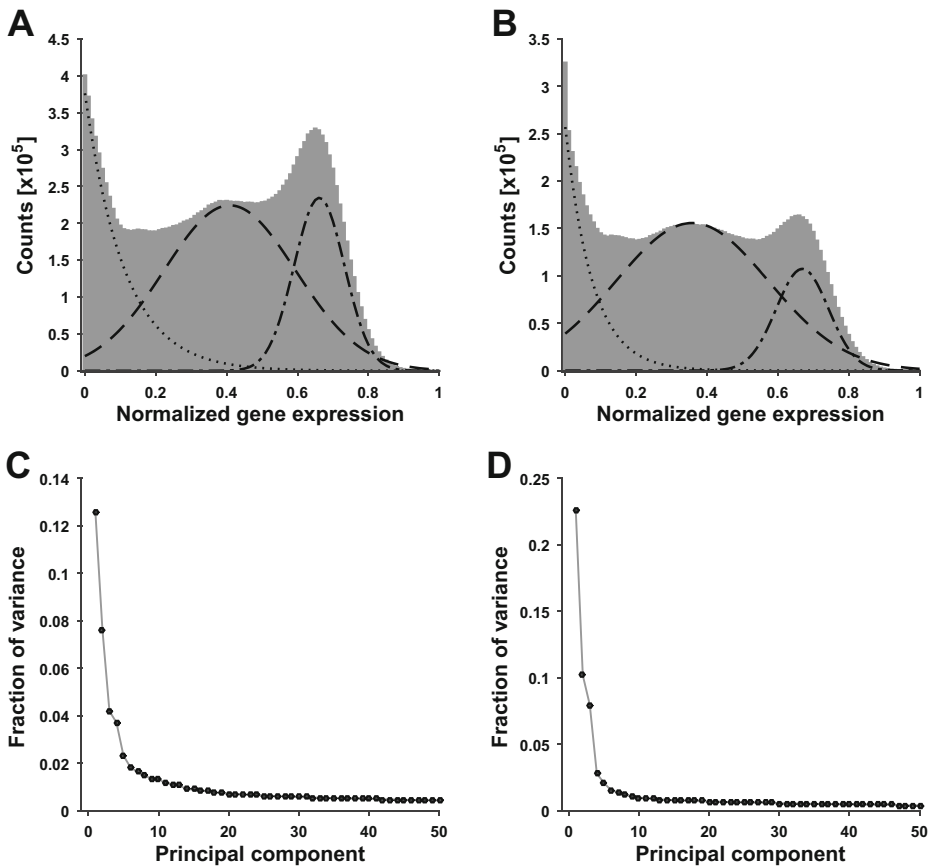
$$\frac{dY_t}{dt} = G(Y_t) - D(Y_t) + a(\sqrt{G(Y_t)}\eta_G - \sqrt{D(Y_t)}\eta_D),$$

where  $a$  is a constant and  $\eta_G$  and  $\eta_D$  are two independent white-noise signals with zero mean. In addition to molecular noise GNW also adds a model of instrumental noise introduced in typical microarray experiments (Schaffter et al. 2011). Variable expression is then

further generated through multifactorial variations in the initial activation of genes in the model (multifactorial data), through a 50% decrease in a single gene's initial activation (knockdown data) and through the complete inactivation of a single gene (knockout data).

In this study we used a white-noise term in the SDEs with standard deviation of 0.05 and generated 250 and 200 expression samples each of multifactorial, knockdown and knockout data for each *Ecoli250* and *PC200* network, respectively. These individual three datasets were then combined to obtain the final simulated expression dataset for each network.

Inspection of the simulated expression data revealed that it exhibited certain properties expected for real microarray gene expression datasets. Specifically, the distribution of simulated gene expression values over all datasets could be modeled by a mixture of three Gaussian modes (compare Fig. 3A-B). Similar observations with two to three Gaussian



**Fig. 3** Properties of the simulated expression data. (A-B): The histogram in gray shows the distribution of gene expression values over all 100 datasets simulated from the *Ecoli250* networks (A) and the *PC200* networks (B). The black dotted, dashed and dash-dotted lines indicate the different Gaussian modes obtained by modeling the distribution using a Gaussian mixture model. Specifically, the "fit" MATLAB function was utilized with the "gauss3" parameter in order to model the distribution of expression data using a mixture of 3 Gaussian modes (MATLAB, 2015). (C-D): Results of a PCA of the expression data simulated for a single *Ecoli250* network (C) and a single *PC200* network (D) showing the variance explained for by the individual principal components identified in the analysis

modes have previously been documented and utilized for microarray expression of various tissues or cell types (Wieczorek et al. 2003; Tuna and Niranjana 2009; Painter et al. 2011), where the individual modes are usually assumed to reflect the populations of lowly, (intermediary) and highly expressed genes, respectively. The skewed shape of the histogram itself, in comparison to typical microarray expression, might be due to the fact that expression was simulated from small networks and was normalized subsequently. Furthermore, performing a principal component analysis (PCA) on the obtained simulated expression data revealed that, similar to regular gene expression data sets, two to four of the principal components provided the major contribution to explaining the observed variance in the data (compare Fig. 3C-D).

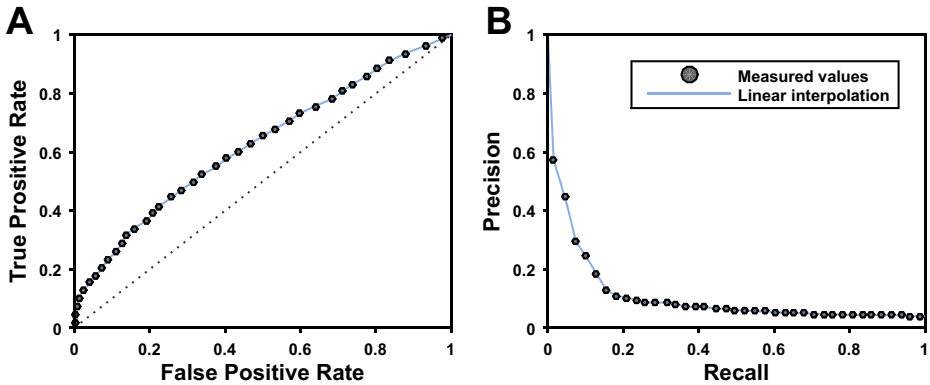
### 3 Inference of Transcriptional Networks

From the simulated expression data, TRNs were inferred using ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks; Margolin et al. 2006), CLR (Context Likelihood of Relatedness; Faith et al. 2007) and MRNET (Minimum Redundancy NETWORK; Meyer et al. 2007), all three of which are based on mutual information and are implemented in the R/Bioconductor package MINET (Meyer et al. 2008), as well as the correlation based WGCNA (Weighted Gene Co-expression Network Analysis; Langfelder and Horvath 2008) method. Importantly, since we were interested in centralities rather than the exact prediction of individual links, all links in the inferred and reference networks were considered to be undirected once the networks had been inferred.

### 4 Estimating the Accuracy of Inferred Networks

The agreement between reference and case networks with respect to existence or absence of individual links can be estimated using a number of different metrics (Schaffter et al. 2011; Marbach et al. 2012; Marbach et al. 2010; Liu et al. 2014). For comparability with previously published results, network accuracies were here estimated using the area under the Receiver-Operator-Characteristic (ROC) curve (auROC) and the area under the Precision-Recall (PR) curve (auPR) as documented in Schaffter et al. (2011). ROC curves and PR curves were obtained by first sorting the network links predicted by the inference methods based on descending absolute strength. Subsequently, false-/true-positive rates and precision/recall values were sampled at certain intervals by including increasing numbers of highly scored links (compare Fig. 4). The sampled curves were then linearly interpolated including also the additional start and end points (0,0) and (1,1) for ROC curves, as well as (0,1) and (1,Pr<sub>max</sub>) for the PR curves (compare Fig. 4), where  $\text{Pr}_{\max} = \frac{l}{(N^2 - N)/2}$  is the maximal achievable precision, when including all links, with  $l$  denoting the number of (undirected) links and  $N$  denoting the number of nodes in the reference network.

However, one concern with this approach was the difference in total number of predicted links observed for the individual methods, which led to substantial differences in the obtainable recall or true-positive rates (compare Fig. 5A-B). Due to this discrepancy in the number of predicted links between different methods and subsequent issues of comparability between methods and also in establishing networks with sufficient interactions for centrality evaluations, we restrained from including more conservative methods, such as the C3NET algorithm (Altay and Emmert-Streib 2010a), which in our hands provided in most cases fewer links than ARACNE. Instead, to remove potential bias based on different



**Fig. 4** Sampling of ROC and PR curves. ROC curve (A) and PR curve (B) for one WGCNA inferred *Ecoli250* network including sampled true-positive and false-positive rates or recall and precision values, respectively, and the interpolations over the sampled data points

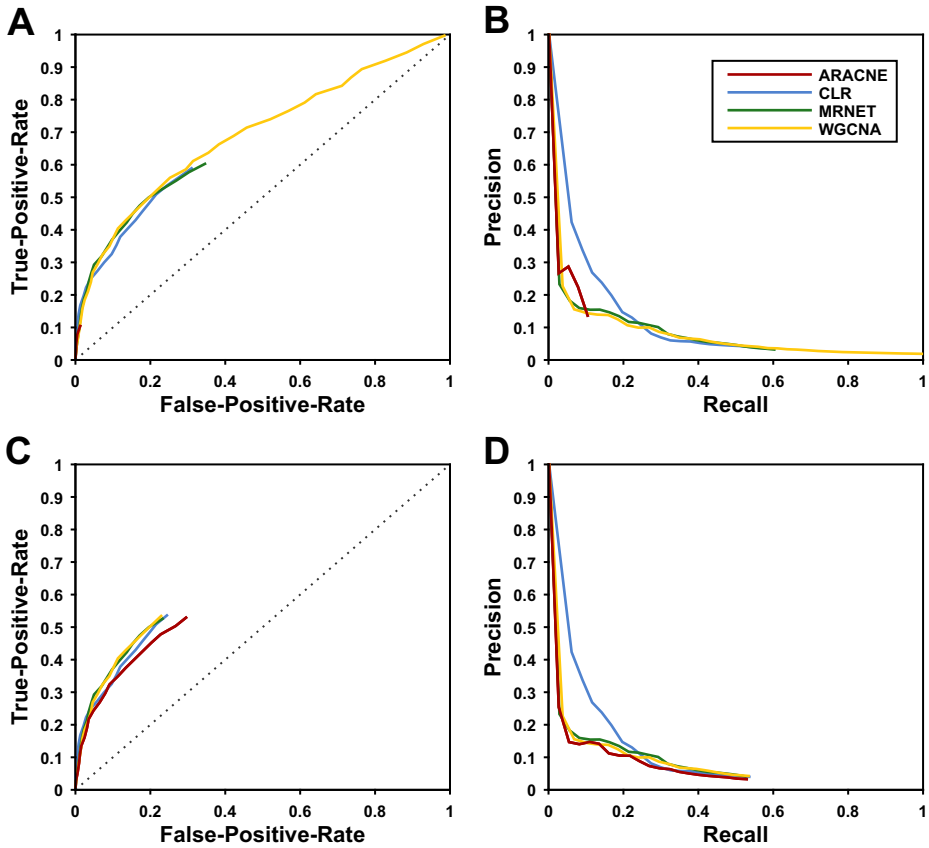
amounts of missing data between the used methods, we decided to (1) decrease the default threshold below which links in vertex-triplets are removed in ARACNE by 0.05 (Margolin et al. 2006; Meyer et al. 2008) and (2) sample all networks only up to the smallest maximum recall value obtained among all inferred networks (compare Fig. 5C-D).

The method specific auROC and auPR values were then calculated for all networks in the *Ecoli250* and *PC200* sets using the *trapz* function in MATLAB (2015). Both approaches discussed above were employed, i.e. either using all available interactions obtained by the inference methods with default parameters as illustrated in Fig. 4A-B or using the interactions filtered based on the maximum recall value as outlined in Fig. 5C-D. The results are depicted in Fig. 6A-B and C-D, respectively.

Networks derived by ARACNE had generally lower auROC values than networks inferred by the other three methods (compare Fig. 6A,C). Apart from the *PC200* comparison to WGCNA inferred networks with filtered interactions (Fig. 6C), these differences in mean auROC values were significant at the  $\alpha = 0.05$  level according to a two-sided Welch T-test (Welch 1947). Of note however, the difference of mean auROC values between ARACNE and the other methods was substantially diminished when using filtered interactions (Fig. 6C), suggesting that part of the differences in Fig. 6A can be explained due to different numbers of predicted links between methods. On the same note, while we observed comparable ranges of *Ecoli250* auPR values and a significantly higher mean *PC200* auPR for ARACNE as compared to CLR with default interactions (Fig. 6B), auPR values for ARACNE in both network sets were substantially reduced when using filtered interactions, with CLR exhibiting significantly higher mean auPR values compared to all other methods (Fig. 6D). The results in Fig. 6A appear consistent with the original results presented by Schaffter et al. (2011), who documented comparable auROC differences between the ARACNE and CLR methods also coupled to low auPR values, when tested on expression data simulated with knock-out and knock-down perturbations (Schaffter et al. 2011).

In addition, we observed that the *PC200* networks exhibited significantly lower auROC values and significantly higher auPR values than the respective *Ecoli250* networks. Considering the different structures of the networks in the *Ecoli250* and *PC200* sets, this difference points towards a potential impact of the overall network structure on the inference accuracy of the used methods.





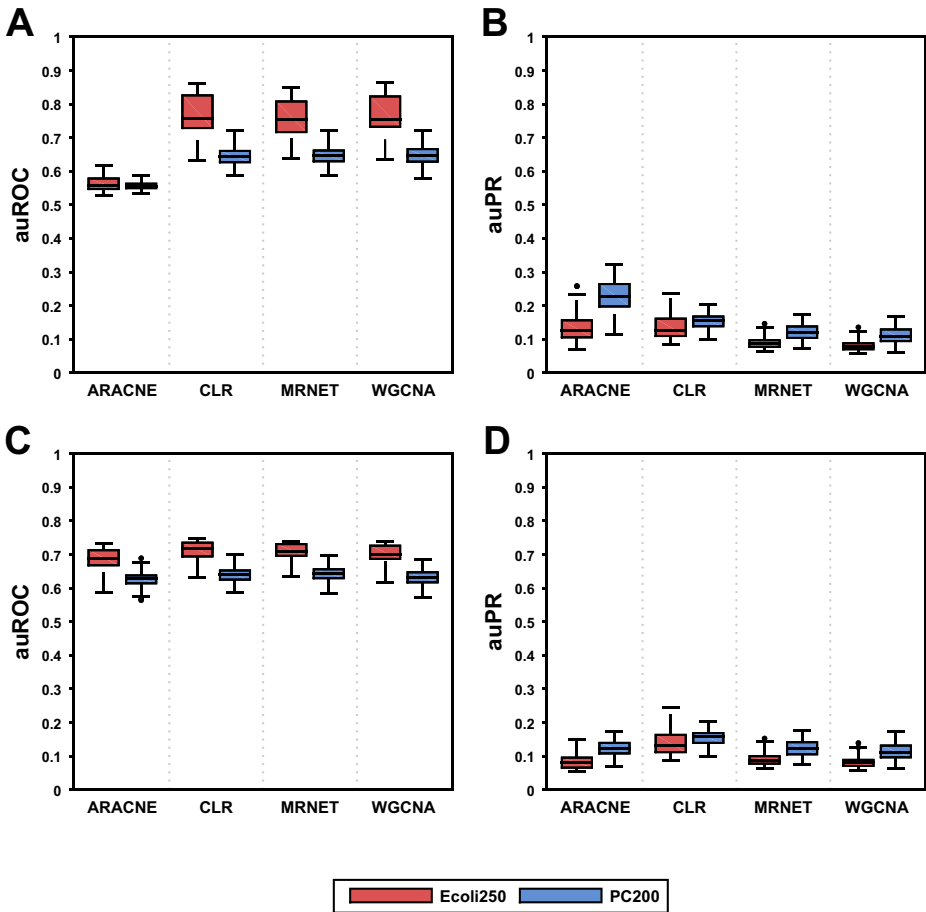
**Fig. 5** Differences in maximum recall and true-positive-rates between network inference methods. (A-B): True-positive and false-positive rates (A) as well as recall and precision values (B) of the four used inference methods calculated on one of the *E. coli*250 networks interpolated up to the maximum true-positive rate or recall value, respectively, obtained for the individual method. (C-D): True-positive and false-positive rates (C) as well as recall and precision values (D) for the four used inference methods calculated on the same network and interpolated up to the smallest maximum true-positive rate or recall value, respectively, among all inference methods and *E. coli*250 networks

### 5 Conservation of Centralities in Inferred Networks

In order to investigate the agreement of centrality measures between true and reverse-engineered networks for the four inference methods, we focused here on degree centrality and betweenness centrality (Freeman 1977).

Given the symmetric adjacency matrix  $A = \{a_{n \times n}\}$ , where  $n$  is the number of genes and  $a_{ij} = a_{ji} = 1$  if there exists a biological interaction between gene  $i$  and  $j$  and  $a_{ij} = a_{ji} = 0$  otherwise, then degree centrality is defined as

$$C_D(v_i) = \sum_{j=1}^n a_{ij},$$



**Fig. 6** Box-and-whisker plots of auROC and auPR values. The auROC (A,C) and auPR (B,D) values of the networks inferred by four different methods for the *Ecoli250* (red) and *PC200* (blue) are depicted as box-and-whisker-plots, with the median values represented by horizontal lines inside each box, lower and upper box borders indicating 25th and 75th percentiles, respectively, regions between whiskers including all non-outlier values, and dots representing outliers. The upper panel (A-B) depicts results for the first approach described above, i.e. including all available interactions obtained by the inference methods, while the lower panel (C-D) displays the results on filtered interactions

and betweenness centrality can according to Brandes (2001) be defined as

$$C_B(v_i) = \frac{1}{2} \sum_{k \neq i \neq l} \frac{\sigma_{kl}(v_i)}{\sigma_{kl}}$$

with  $v_i$  denoting the  $i$ 'th gene,  $\sigma_{kl}$  denoting the number of shortest paths between two nodes  $v_k$  and  $v_l$  and  $\sigma_{kl}(v_i)$  denoting the number of shortest paths between  $v_k$  and  $v_l$  that traverse through node  $v_i$ . Importantly, in order to allow the application of betweenness centrality on

disconnected networks, if there is no path between nodes  $v_k$  and  $v_l$  in the network then one sets

$$d(v_k, v_l) \equiv \infty, \quad \frac{\sigma_{kl}(v_i)}{\sigma_{kl}} \equiv 0.$$

Degree and betweenness measures were then computed using the MatlabBGL package (Gleich and Saunders 2009) in the reference network and three different variants of the network obtained by each of the four individual network inference methods. The three different variants were established by including the same number of top scored links as present in the reference network ( $N_1$ ), or using only a number of top scored links equal to 10% ( $N_{0.1}$ ), or 50% ( $N_{0.5}$ ) of the total number of links present in the reference network. This approach was chosen, since there might be differences in centrality conservations depending on the number and scores of links included in the generation of the reconstructed networks.

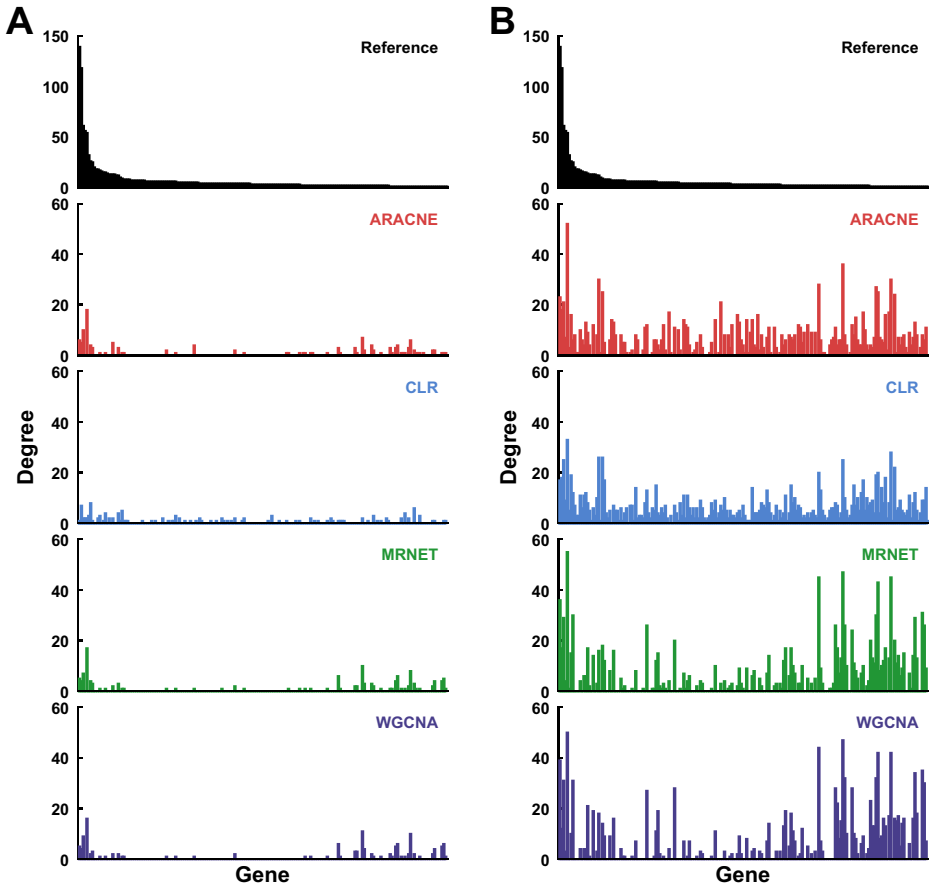
We then started with a visual inspection of the distribution of centralities between the reference networks and respective inferred networks (compare the example for degree centralities in Fig. 7). As indicated by these qualitative comparisons, nodes that showed a top degree centrality or betweenness centrality score in the reference network, would often also receive a higher centrality in the inferred networks. However, we also observed a lot of noise exemplified by a large number of nodes with low centrality score in the reference networks that would receive high centralities in the inferred network, an observation that became more clear when increasing the number of links in the inferred networks (compare Fig. 7B). Due to this spread in centralities in the inferred networks, even when including the same number of links as in the reference network, the observed maximum centrality measures were much lower than for the top scored genes in the reference network (compare Fig. 7B). Hence, potential similarities of centrality distributions between reference and inferred networks proved difficult to examine qualitatively.

In order to determine the agreement of centralities between the reference and inferred networks more quantitatively, we were then interested in comparing the ranks of nodes under the individual centralities between the two networks using a rank correlation metric. Importantly, since centralities are here understood as a method of prioritization of nodes, it is obvious that more importance should be placed on high ranked nodes, which dictates the use of a weighted rank correlation metric. During the recent years many different weighted variants of Spearman’s  $\rho$  and Kendall’s  $\tau$  have been discussed in the literature, compare (Dancelli et al. 2013; Tarsitano 2009; Pozzi et al. 2012). Since the chosen centralities in combination with the inferred network structures, which will often miss a subset of nodes thus receiving a zero centrality, implies a high number of ties in the rankings, Kendall’s  $\tau$  and specifically the  $\tau_B$  variant, designed particularly for rankings with ties, appears here to be the preferable method. Accordingly, we used here the weighted Kendall’s  $\tau_B$  measure proposed by Pozzi et al. (2012), which is computed for the two variables  $y^i$  and  $y^j$  over a running window  $\Delta t$  as

$$\tau_{ij}^w = \sum_{k=1}^{\Delta t-1} \sum_{l=k+1}^{\Delta t} w_{kl} \text{Sgn}(y_k^i - y_l^i) \text{Sgn}(y_k^j - y_l^j),$$

where we define the exponential smoothing weights as

$$w_{kl} = \frac{e^{\frac{2\Delta t-k-l}{\theta}}}{\sum_{v=1}^{\Delta t-1} \sum_{w=v+1}^{\Delta t} e^{\frac{2\Delta t-v-w}{\theta}}},$$



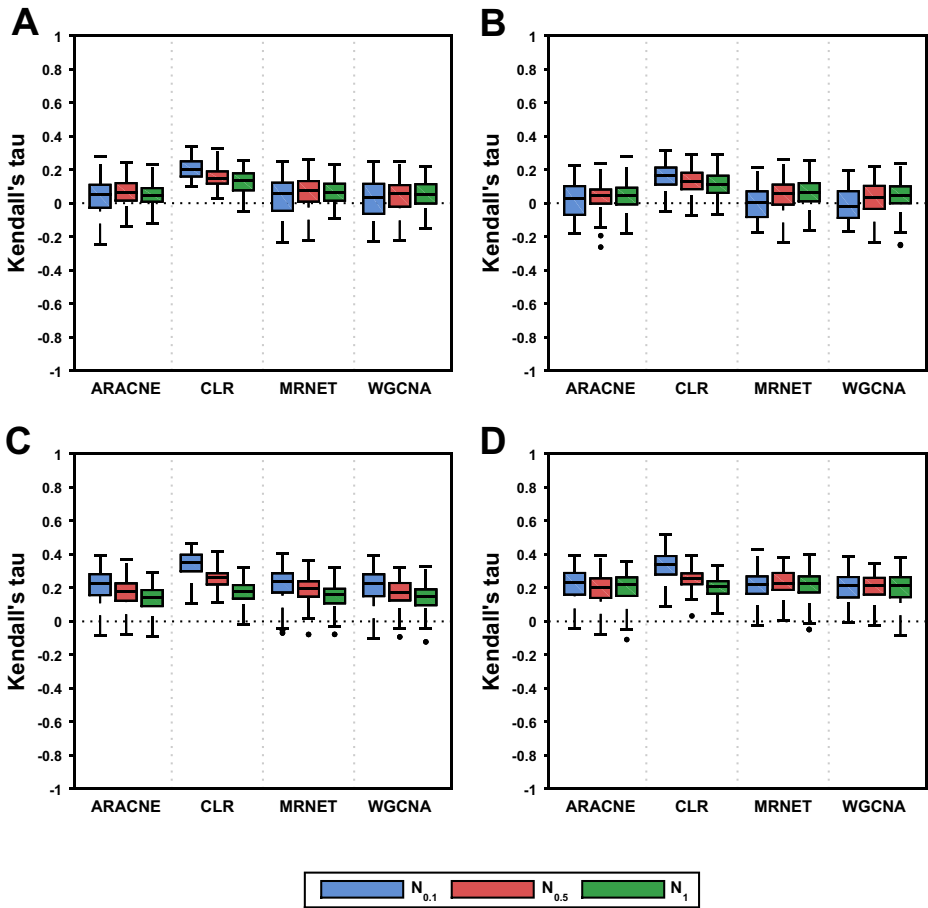
**Fig. 7 Visual comparison of degree centralities between reference and inferred networks.** The degree centrality values for one PC200 reference network sorted in descending order are shown in the top panel. The centralities obtained from the respective inferred networks with genes sorted in the same order as for the corresponding reference network are shown underneath for the  $N_{0,1}$  network type (A) and the  $N_1$  network type (B), respectively

with  $l > k$  and the constant  $\theta$  indicating the specific time of the weights, such that

$$\sum_{k=1}^{\Delta t-1} \sum_{l=k+1}^{\Delta t} w_{kl} = 1$$

Importantly, the weights here are just the weights defined by Pozzi et al. (2012) in reverse order, i.e. decreasing on both subscripts in order to place more importance on the nodes with high ranks in the reference networks.  $\Delta t$  was chosen equal to 250 and 200 for the *Ecoli250* and *PC200* datasets, respectively, and  $\theta = \frac{\Delta t}{3}$  according to the recommendations of the authors (Pozzi et al. 2012).

The results of the weighted rank correlation analysis of centralities between reference and inferred networks are shown in Fig. 8.

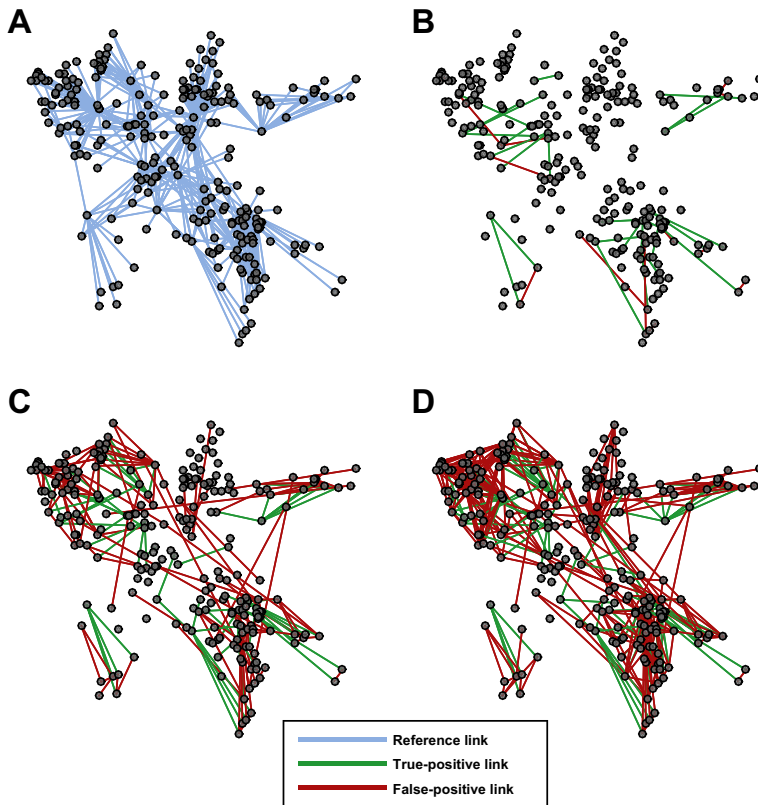


**Fig. 8** Kendall's  $\tau$  of centralities in reference and inferred networks. The Kendall's  $\tau$  values for degree (A,C) and betweenness centralities (B,D) between reference and inferred networks for *Ecoli250* (A-B) and *PC200* (C-D) are shown as box-and-whisker-plots for the four inference methods. Centralities have been compared for three variations of the inferred networks;  $N_{0.1}$  (blue),  $N_{0.5}$  (red),  $N_1$  (green)

As can be seen from Fig. 8 none of the used inference methods produced a network with a high correlation of degree or betweenness centralities between reference and inferred networks, considering the three network variants  $N_{0.1}$ ,  $N_{0.5}$  and  $N_1$ . The highest mean correlations for both centrality types in *Ecoli250* and *PC200* are seen with the CLR method across all three network variants except the case of betweenness in the *PC200*- $N_1$  variant. Interestingly, in all comparisons of degree centrality for the *PC200* networks as well as the other three CLR results, the  $N_{0.1}$  networks achieved a better mean correlation than the  $N_{0.5}$  networks with a further reduction observed in the  $N_1$  networks.

In order to explore the observed discrepancies in centralities between reference and inferred networks further, we inspected the expected and predicted links in the three network variants (compare example in Fig. 9).

From the inspection of predicted links, as exemplified by the networks in Fig. 7, it is obvious that when choosing only a small number of highly scored links from the network



**Fig. 9** Inspection of CLR inferred links. The reference *Ecoli250* network (A) from Figure 1A is shown together with the CLR inferred  $N_{0.1}$  (B),  $N_{0.5}$  (C) and  $N_1$  (D) networks, i.e. using a number of highly scored predicted links equal to 10%, 50% and 100% of the total number of links present in the reference network. Reference links in the benchmark network are shown in blue, true-positive predicted links are shown in green and false-positive predicted links are shown in red

inference results (compare Fig. 9B), one can among those links obtain a large proportion of true-positives, but the number of links does not suffice to reconstruct enough of the overall topology of the reference network to obtain a high degree of agreement between centralities. When increasing the number of links in the inferred networks, however, the topology appears to become dominated by false-positive links (compare Fig. 9C-D), which might lead to a more random distribution of centralities in the inferred networks.

## 6 Conclusion and Future Perspectives

The results presented in this study show that the inference methods are indeed able to accurately predict a certain proportion of biological links from simulated expression data. However, the number of correctly predicted links does not appear to suffice in order to also reconstruct the underlying network topology to a degree necessary to obtain a high conservation of graph centralities between reference and inferred networks. In addition, it has

previously been reported that network inference methods might show a lower prediction accuracy for links connecting to a high degree centrality node as compared to links connecting to low centrality nodes (Marbach et al. 2010). Together, these two observations might explain the revealed dissimilarities of centralities between reference and inferred networks.

As briefly mentioned in the description of the used reference networks, the choice of network sizes might have an influence on the displayed results. However, as demonstrated in Schaffter et al. (2011), networks of larger dimensions might have lower connectivity, which might affect the inference accuracy. As a consequence, the conservation of centralities in larger networks might be similarly affected. In addition, using simulated networks as well as knowledge-based networks of *Escherichia coli* and *Saccharomyces cerevisiae*, all three of which contain a substantially larger number of nodes and edges than the networks utilized in the present work, we did not find any substantial improvement on the correlation of centrality distributions between reference and inferred networks (Weishaupt et al. 2016).

It is also possible that at least part of the observed disagreements of centralities can be attributed to difficulties of inferring networks from the chosen synthetic expression data and that the tested methods might actually perform better on true microarray expression data. It would hence be important to repeat the outlined experiments using other models of expression data, compare for instance (Van den Bulcke et al. 2006; Langfelder and Horvath 2008). Additional and more comprehensive studies will also be required to evaluate the performance of other recently developed network methods with respect to the conservation of centrality methods in inferred networks. Specifically, considering the fact that methods make different systematic errors in inferring gene interactions (Schaffter et al. 2011; Marbach et al. 2012) and that accurately predicted links also differ between networks (data not shown), future studies would probably also benefit from further investigating how a combination of highly scored links from multiple network methods could be utilized to improve our ability to reconstruct relevant network topologies from gene expression data (Marbach et al. 2010).

In summary, more work will be needed to establish the usefulness of networks inferred from expression data using current methodology for the purpose of centrality based gene prioritization.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

- Aittokallio T, Schwikowski B (2006) Graph-based methods for analysing networks in cell biology. *Brief Bioinforma* 7(3):243–255
- Albert R (2005) Scale-free networks in cell biology. *J Cell Sci* 118(21):4947–4957
- Altay G, Emmert-Streib F (2010a) Inferring the conservative causal core of gene regulatory networks. *BMC Syst Biol* 4(1):132
- Altay G, Emmert-Streib F (2010b) Revealing differences in gene network inference algorithms on the network level by ensemble methods. *Bioinformatics* 26(14):1738–1744
- Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A (2005) Reverse engineering of regulatory networks in human b cells. *Nat Genet* 37(4):382–390
- Brandes U (2001) A faster algorithm for betweenness centrality\*. *J Math Sociol* 25(2):163–177

- Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur Ö, Anwar N, Schultz N, Bader GD, Sander C (2011) Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res* 39(suppl 1):D685–D690
- Chalancon G, Ravarani CN, Balaji S, Martinez-Arias A, Aravind L, Jothi R, Babu MM (2012) Interplay between gene expression noise and regulatory network architecture. *Trends Genet* 28(5):221–232
- Cordero D, Solé X, Crous-Bou M, Sanz-Pamplona R, Paré L, Guinó E, Olivares D, Berenguer A, Santos C, Salazar R et al (2014) Large differences in global transcriptional regulatory programs of normal and tumor colon cells. *BMC Cancer* 14(1):708
- Dancelli L, Manisera M, Vezzoli M (2013) On two classes of weighted rank correlation measures deriving from the spearman's  $\rho$ . In: *Statistical Models for Data Analysis*. Springer, pp 107–114
- De Matos Simoes R, Dehmer M, Emmert-Streib F (2013) B-cell lymphoma gene regulatory networks: biological consistency among inference methods. *Frontiers in genetics*, 4
- De Matos Simoes R, Emmert-Streib F (2012) Bagging statistical network inference from large-scale gene expression data. *PLoS One* 7(3):e33624
- Emmert-Streib F, De Matos Simoes R, Mullan P, Haibe-Kains B, Dehmer M (2014) The gene regulatory network for breast cancer: Integrated regulatory landscape of cancer hallmarks. *Frontiers in genetics*, 5
- Erdős P, Rényi A (1959) On random graphs, i. *Publ Math Debr* 6:290–297
- Estrada E (2006a) Protein bipartivity and essentiality in the yeast protein–protein interaction network. *J Proteome Res* 5(9):2177–2184
- Estrada E (2006b) Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics* 6(1):35–40
- Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS (2007) Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol* 5(1):e8
- Freeman LC (1977) A set of measures of centrality based on betweenness. *Sociometry*, pp 35–41
- Gleich DF, Saunders M (2009) Models and algorithms for pagerank sensitivity. Stanford University
- Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L (2007) The human disease network. *Proc Nat Acad Sci* 104(21):8685–8690
- Hahn MW, Kern AD (2005) Comparative genomics of centrality and essentiality in three eukaryotic protein–interaction networks. *Mol Biol Evol* 22(4):803–806
- Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P (2010) Inferring regulatory networks from expression data using tree-based methods. *PLoS one* 5(9):e12776
- Izudhean S, Mathew S (2013) Cancer gene identification using graph centrality. *Curr Sci* 105(8):1143
- Jeong H, Mason SP, Barabási A-L, Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature* 411(6833):41–42
- Jonsson PF, Bates PA (2006) Global topological features of cancer proteins in the human interactome. *Bioinformatics* 22(18):2291–2297
- Jörnsten R, Abenius T, Kling T, Schmidt L, Johansson E, Nordling TE, Nordlander B, Sander C, Gennemark P, Funa K et al (2011) Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. *Mol Syst Biol* 7(1):486
- Joy MP, Brock A, Ingber DE, Huang S (2005) High-betweenness proteins in the yeast protein interaction network. *BioMed Res Int* 2005(2):96–103
- Knaack SA, Siahpirani AF, Roy S (2014) A pan-cancer modular regulatory network analysis to identify common and cancer-specific network components. *Cancer Inf* 13(Suppl 5):69
- Knüpfel C, Dittrich P, Beckstein C (2004) Artificial gene regulation: A data source for validation of reverse bioengineering. In: *Proceedings of the 6th German Workshop on Artificial Life (GWAL6)*, pp 66–75
- Koschützki D, Schreiber F (2008) Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene Regul Syst Biol* 2:193
- Langfelder P, Horvath S (2008) Wgcna: an R package for weighted correlation network analysis. *BMC bioinformatics* 9(1):559
- Langfelder P, Mischel PS, Horvath S (2013) When is hub gene selection better than standard meta-analysis. *PLoS one* 8(4):e61505
- Liu Z-P (2015) Reverse engineering of genome-wide gene regulatory networks from gene expression data. *Curr Genomics* 16(1):3–22
- Liu Z-P, Wu H, Zhu J, Miao H (2014) Systematic identification of transcriptional and post-transcriptional regulations in human respiratory epithelial cells during influenza a virus infection. *BMC Bioinforma* 15(1):336
- Ma X, Gao L (2012) Biological network analysis: insights into structure and functions. *Brief Funct Genomics* 11(6):434–442



- Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Stolovitzky G et al (2012) Wisdom of crowds for robust gene network inference. *Nat Methods* 9(8):796–804
- Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, Stolovitzky G (2010) Revealing strengths and weaknesses of methods for gene network inference. *Proc Nat Acad Sci* 107(14):6286–6291
- Margolin AA, Califano A (2007) Theory and limitations of genetic network inference from microarray data. *Ann New York Acad Sci* 1115(1):51–72
- Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, Califano A (2006) Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinform* 7(Suppl 1):S7
- MATLAB (2015) version 7.10.0 (R2015a), The MathWorks Inc., Natick, Massachusetts
- Mendes P, Sha W, Ye K (2003) Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics* 19(suppl 2):ii122–ii129
- Meyer PE, Kontos K, Lafitte F, Bontempi G (2007) Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J Bioinform Syst Biol* 2007:8–8
- Meyer PE, Lafitte F, Bontempi G (2008) Minet: Ar/bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinform* 9(1):461
- Ortutay C, Vihinen M (2009) Identification of candidate disease genes by integrating gene ontologies and protein–interaction networks: case study of primary immunodeficiencies. *Nucleic Acids Res* 37(2):622–628
- Özgür A, Vu T, Erkan G, Radev DR (2008) Identifying gene–disease associations using centrality on a literature mined gene–interaction network. *Bioinformatics* 24(13):i277–i285
- Painter MW, Davis S, Hardy R, Mathis D, Benoist C, Zhou Y, Shinton S, Hardy R, Asinovski N, Ergun A et al (2011) Transcriptomes of the b and t lineages compared by multiplatform microarray profiling. *J Immunol* 186(5):3047–3057
- Pozzi F, Di Matteo T, Aste T (2012) Exponential smoothing weighted correlations. *Eur Phys J B* 85(6):1–21
- Schaffter T, Marbach D, Floreano D (2011) Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics* 27(16):2263–2270
- Siddani BR, Pochineni LP, Palanisamy M (2013) Candidate gene identification for systemic lupus erythematosus using network centrality measures and gene ontology. *PloS one* 8(12):e81766
- Tarsitano A (2009) Comparing the effectiveness of rank correlation statistics. P: Dip. di Economia e Statistica, University of della Calabria
- Tuna S, Niranjana M (2009) Cross-platform analysis with binarized gene expression data. In: *Pattern Recognition in Bioinformatics*. Springer, pp 439–449
- Van den Bulcke T, Van Leemput K, Naudts B, Van Remortel P, Ma H, Verschoren A, De Moor B, Marchal K (2006) Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinform* 7(1):43
- Wachi S, Yoneda K, Wu R (2005) Interactome–transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics* 21(23):4205–4208
- Weishaupt H, Johansson P, Engström C, Nelander S, Silvestrov S, Swartling FJ (2016) Prediction of high centrality nodes from reverseengineered transcriptional regulator networks. In: *SMTDA 2016 Proceedings: / 4th Stochastic Modeling Techniques and Data Analysis International Conference / [ed] H. Skiadas (Ed), ISAST: International Society for the Advancement of Science and Technology*, pp 5177–531
- Welch BL (1947) The generalization of student's problem when several different population variances are involved. *Biometrika*, pp 28–35
- Wieczorek G, Steinhoff C, Schulz R, Scheller M, Vingron M, Ropers H-H, Nuber UA (2003) Gene expression profile of mouse bone marrow stromal cells determined by cDNA microarray analysis. *Cell Tissue Res* 311(2):227–237
- Wu S, Liu Z-P, Qiu X, Wu H (2014) Modeling genome-wide dynamic regulatory network in mouse lungs with influenza infection using high-dimensional ordinary differential equations. *PloS one* 9(5):e95276
- Xu J, Li Y (2006) Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics* 22(22):2800–2805
- Yip KY, Alexander RP, Yan K-K, Gerstein M (2010) Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data. *PloS one* 5(1):e8121–e8121
- Zhang M, Deng J, Fang CV, Zhang X, Lu L (2010) Molecular network analysis and applications. *Knowledge-Based Bioinformatics*, pp 253
- Zhang X, Zhao X-M, He K, Lu L, Cao Y, Liu J, Hao J-K, Liu Z-P, Chen L (2012) Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics* 28(1):98–104
- Zhu X, Gerstein M, Snyder M (2007) Getting connected: analysis and principles of biological networks. *Genes Dev* 21(9):1010–1024