BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 14 2010, pages 1738–1744
doi:10.1093/bioinformatics/btq259

Systems biology

Advance Access publication May 25, 2010

Revealing differences in gene network inference algorithms on the network level by ensemble methods

Gökmen Altay and Frank Emmert-Streib*

Computational Biology and Machine Learning, Center for Cancer Research and Cell Biology, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, 97 Lisburn Road, Belfast BT9 7BL, UK

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: The inference of regulatory networks from large-scale expression data holds great promise because of the potentially causal interpretation of these networks. However, due to the difficulty to establish reliable methods based on observational data there is so far only incomplete knowledge about possibilities and limitations of such inference methods in this context.

Results: In this article, we conduct a statistical analysis investigating differences and similarities of four network inference algorithms, ARACNE, CLR, MRNET and RN, with respect to local network-based measures. We employ ensemble methods allowing to assess the inferability down to the level of individual edges. Our analysis reveals the bias of these inference methods with respect to the inference of various network components and, hence, provides guidance in the interpretation of inferred regulatory networks from expression data. Further, as application we predict the total number of regulatory interactions in human B cells and hypothesize about the role of Myc and its targets regarding molecular information processing.

Contact: v@bio-complexity.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 12, 2009; revised on April 26, 2010; accepted on May 16, 2010

1 INTRODUCTION

The orchestral behavior of genes is coordinated by different types of gene networks, e.g. the transcriptional regulatory, signaling or protein network and interactions among them (Barabasi and Oltvai, 2004; Palsson, 2006). As these networks represent biochemical interactions among gene products, they form causal (Pearl, 2000; Shipley, 2000) instead of merely association networks (Zhang and Horvath, 2005). For this reason the reconstruction of gene networks from experimental data on a genomic scale is considered as one of the most important goals in systems biology. A reflection of this importance can be seen in the establishment of the DREAM (Dialogue for Reverse Engineering Assessments and Methods) project and its conference series that is devoted to all aspects of this endeavor providing a forum for the community (Stolovitzky and Califano, 2007; Stolovitzky *et al*., 2009).

If expression data from DNA microarray experiments are used, the inferred networks are called gene regulatory networks (GRNs) because a large portion of estimated interactions may come from transcription regulation directly, but not only (Emmert-Streib and Dehmer, 2008; Levine *et al*., 2005). In general, GRN inference (GRNI) is important for a better understanding of normal cell physiology and for the explanation of pathological phenotypes (Almudevar *et al*., 2006; Emmert-Streib, 2007). The difficulty of the problem (Hache *et al*., 2009; Holland, 1986; Margolin and Califano, 2007; Werhli *et al*., 2006) stems from the fact that expression data are noisy, sample size is small relative to the number of covariates and the type of data is frequently limited to observational data only, e.g. because of ethical reasons.

Recently, several mutual information (MI; Cover and Thomas, 1991) based methods for the inference of genome-scale regulatory networks have been suggested to approach this problem. Four of these methods are ARACNE (Margolin *et al*., 2006), CLR (Faith *et al*., 2007), MRNET (Meyer *et al*., 2007) and Relevance Networks (RN; Butte and Kohane, 2000). These methods have been investigated using global performance measures such as the F-score (Meyer *et al*., 2007) or area under the receiver operator curve (AUROC) (Husmeier, 2003). Further, such studies are frequently based on an individual dataset rather than an ensemble of datasets drawn under the same condition. In this article, we analyze network inference methods statistically employing local network-based measures to assess their performance, as proposed in Emmert-Streib and Altay (2010), in combination with ensemble data. This introduces a graph-theoretical perspective (Chartrand and Oellermann, 1993) on the problem that allows to study arbitrary *network components* instead of the entire network only.

2 APPROACH

In this article, we aim to compare four different network inference algorithms (ARACNE, CLR, MRNET and RN) by using local network-based measures for their assessment. In Emmert-Streib and Altay (2010), several general local network-based measures have been introduced studying various synthetic and one real biological network from *Escherichia coli*. It was demonstrated that these measures allow an exploratory analysis of inference algorithms on the level of network components, e.g. edges, motifs or subnetworks. This means, in the finest resolution, e.g. true positive rates (TPRs) of individual edges, instead of entire networks, can be investigated. In this study, we apply the local network-based measures introduced in Emmert-Streib and Altay (2010) and complement them with additional ones. As a practical application of the results obtained by our approach, in combination with results from a previous study (Basso *et al*., 2005), we estimate the total number of regulatory interactions in human B cells and hypothesize about the role of Myc

*To whom correspondence should be addressed.

with respect to the hierarchical organization (Ma *et al*., 2004; Yu and Gerstein, 2006) of this regulatory network.

3 METHODS

Our comparative analysis of network inference algorithms utilizes local network-based measures (Emmert-Streib and Altay, 2010). These measures assume the availability of an ensemble of datasets $\mathcal{D}=\{D_1(G),\ldots,D_E(G)\}$, instead of just a single one, belonging to the same underlying structure of a gene network G. More precisely, for the following analysis we use a subnetwork of the transcriptional regulatory network of Yeast (Guelzim *et al*., 2002), G, consisting of 100 genes. This subnetwork was randomly sampled from the entire transcriptional regulatory network by using SynTReN (Van den Bulcke *et al*., 2006). From this network, expression data in steady-state condition are obtained by using dynamic equations with Michaelis–Menten and Hill enzyme kinetics. These data are generated with SynTReN (Van den Bulcke *et al*., 2006). In total, we generate $E=300$ different datasets for sample size 200 and another $E=300$ datasets for sample size 20 using the same subnetwork G of the transcriptional regulatory network of yeast but with different kinetic parameters for each dataset. Biologically, the data $\mathcal{D}$ may correspond to a population of one species spanning the whole dynamic range that different individual organisms from the same species can exhibit. The reason for this variability comes from the fact that molecular systems behave unlike a clockwork utilizing parallel pathways for inter- and intra-cell communication. Statistically, using one network structure G underlying the ensemble data allows the statistical assessment of network components down to the level of individual edges. If different networks, $G_i\neq G_j$ for $i\neq j$, would be used for different datasets $D_i(G_i)$ for $i\in\{1,\ldots,E\}$, the identification of such network components would be no longer possible and, hence, an averaging over different datasets would become meaningless. For example, given two networks of the same size, there may be an edge connecting gene m and n in the first network but no edge in the second network. This demonstrates the problem to identify common parts in these networks. If instead of mathematical labels, m and n, the nodes in these networks would be labeled with gene names, this problem would become more apparent.

The network inference algorithms that we compare in this study are ARACNE, CLR, MRNET and RNs. For detailed information about these GRNI algorithms considered in this study, we refer the readers to Butte and Kohane (2000), Faith *et al*. (2007), Margolin *et al*. (2006) and Meyer *et al*. (2007, 2008). For ARACNE, we set the data processing inequality (DPI; Cover and Thomas, 1991) tolerance parameter $\epsilon=0.1$, as in Basso *et al*. (2005). The MI values are estimated using non-parametric Gaussian estimator as described in Meyer *et al*. (2008) and Olsen *et al*. (2009). The optimal cutoff value for each dataset, D_i, used to declare edges significant is obtained by maximizing the F-score,

$$F(I_0')=\frac{2p(I_0')r(I_0')}{p(I_0')+r(I_0')}. \quad (1)$$

Here the F-score, $F(I_0')$, precision, $p(I_0')=\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$ and recall, $r(I_0')=\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, are a function of the MI threshold I_0' [CLR uses z scores instead (Faith *et al*., 2007)] and so are the number of true positive (TP), false positive (FP) and false negative (FN) edges (see the Supplementary Material for more details). This results in E different F-scores, correspondingly E inferred networks, for the ensemble $\mathcal{D}=\{D_1(G),\ldots,D_E(G)\}$ for a given sample size. For the following simulations, we use data of sample sizes 20 and 200. The overview block diagram of our approach is shown in the Supplementary Figure 1.

Despite the fact that the main objective of this article is to investigate local network-based measures, we use one global measure, the F-score, not only to find a common ground to traditional approaches to this problem, but also to investigate the variability of this measure by using an ensemble of datasets. In conventional studies, only one dataset is used resulting in one, e.g. F-score. However, as we will see in Section 4 one value is usually no good representation of the underlying population. From many simulations, also using the AUROC, we found that the quality of this statement is not sensitive to the actual global measure. For this reason, we present only results for the F-score because it seems sufficient for our argument.

The conceptional idea that motivated the introduction of local network-based measures is to categorize network *components* according to a given rule. For example, we might categorize edges in different classes according to a graph-theoretical description that is solely based on the network topology of the regulatory network. Or we could categorize edges according to their effect sign as given by the dynamical equations generating the expression data. An example for the first measure, in the following called D_s, is a function of the degrees of the edge enclosing genes whereas an example for the second is the categorization in activator and repressor edges. It is apparent that there are many more measures that can be constructed based on your guiding idea. One further measure we study in the following, besides individual edges, is based on three-gene motifs because it has been recognized that such motifs are functional building blocks of larger networks (Milo *et al*., 2002). All these network-based measures emphasize one specific aspect, either of the network topology or of the dynamical equations. However, collectively they allow to gather multidimensional information that can be used to evaluate inference methods. Formal details of our measures and their defining rules can be found in the Supplementary Material or in Emmert-Streib and Altay (2010).

4 RESULTS

We begin our analysis of the four inference algorithms by comparing three global measures. The results are shown in Figure 1. In this figure, the number 20 (respectively 200) indicates the sample size. By maximizing the F-score for each of the E datasets, we obtain the optimal MI threshold (cutoff) values I_0, respectively z_0, illustrated in Figure 1A. From this figure, one can see that CLR results in much higher cutoff values, z_0, than the other algorithms. That is why we plotted these results separately. The average MI (z for CLR) values for each edge of the true network G (averaged over E datasets) are illustrated in Figure 1C. Figure 1B shows the maximal F-scores for the E datasets. This figure shows that, statistically, on a global basis and sample size 200, MRNET having a median F-score of $\sim$0.45 is significantly better than the other methods: ARACNE with a median F-score $\sim$0.38, CLR $\sim$0.32 and RN $\sim$0.28. For sample size 20, the median F-scores are: CLR $\sim$0.33, MRNET $\sim$0.32, ARACNE $\sim$0.29 and RN $\sim$0.28. When the sample size is smaller, the performance of the GRNI algorithms becomes comparable. It is also worth mentioning that MRNET and ARACNE increase their performance significantly by increasing the sample size, while the other methods are only moderately affected with respect to the median value of the F-score.

In the remainder of this section, we study local network-based performance measures. The first measure we use investigates the influence of activator (positive effect) and repressor (negative effect) edges. In Figure 2, we show histograms, for sample size 200, to visualize the effect of activator (red) and repressor (blue) edges on the TPR of edges for all GRNI algorithms under investigation. The TPR of an edge is the number of times a specific edge is inferred correctly divided by the total number of datasets (E). A brief overview of this measure is given in the Supplementary Material. To investigate the results in Figure 2 quantitatively, we apply a two-sample Kolmogorov–Smirnov test (Sheskin, 2004), to each GRNI algorithm, for testing differences in the cumulative distribution function (CDF) of activator and repressor edges. For sample size 200, we obtain P-values of 0.0009688, 0.001554, 0.0001145 and 3.432×10^{-6} for ARACNE, CLR, MRNET and RN, respectively. The results suggest that, for a significance level of $\alpha=0.01$, the

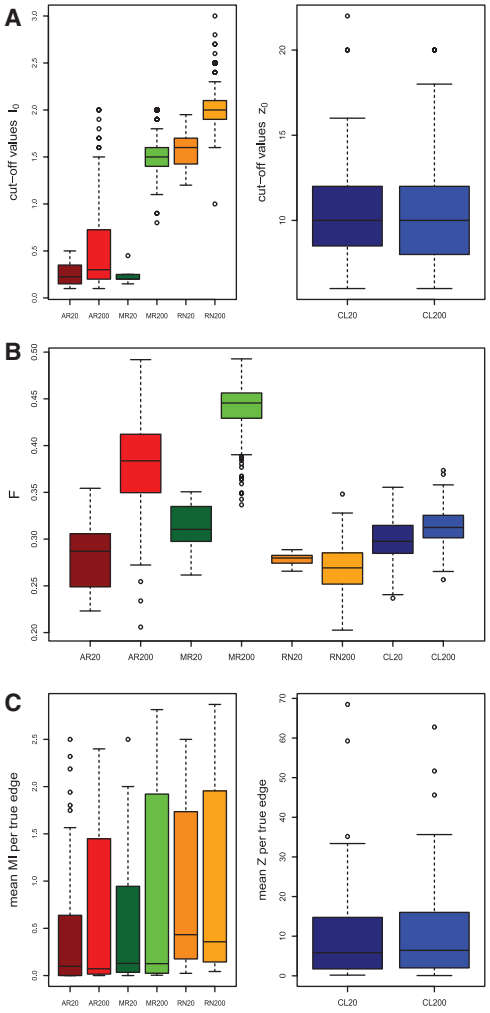

Fig. 1. Boxplots for ARACNE (red), MRNET (green), RN (orange) and CLR (blue). Dark color (left boxplot) corresponds to sample size 20, light color (right boxplot) to sample size 200. (**A**) Optimal cutoff values. (**B**) F-scores. (**C**) Average MI values, respectively, z scores per edge.

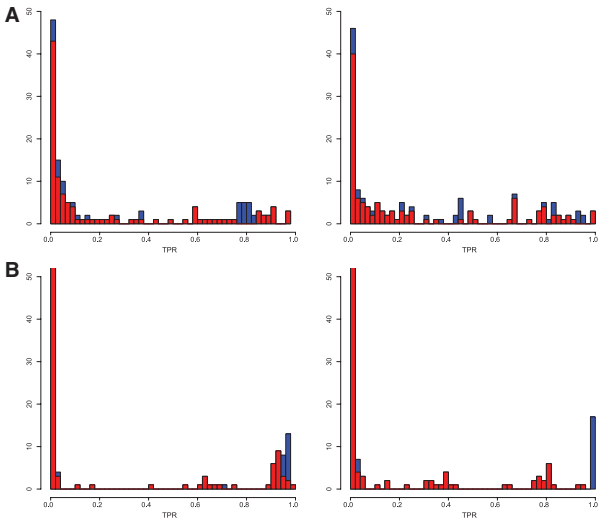

Fig. 2. Histogram of TPRs for edges in the true network. (**A**) ARACNE (left) and CLR (right). (**B**) MRNET (left) and RN (right). Red indicates the contribution from activator and blue from repressor edges. Sample size is 200 for all figures.

edge type has a systematic effect on all four inference algorithms. Qualitatively, this can be seen from the histograms in Figure 2 where the repressor edges have a higher TPR, which means that they are easier to infer. For sample size 20, we obtain P-values of 0.1097, 0.00823, 0.08638 and 4.726×10^{-5} for ARACNE, CLR, MRNET and RN, respectively. These results indicate that only CLR and RN are systematically affected by the edge type. In summary, this means that not only the used inference algorithm may introduce a bias in this context but also the sample size.

The next network-based measure, D_s, that we use allows us to categorize edges in more than two classes as for the activator and repressor edges. The measure Ds_{ij} is defined as the sum of the out-degree of node i plus the in-degree of node j (Emmert-Streib and Altay, 2010) (for a more detailed description see the Supplementary Material). In Figure 3, we illustrate the relation between mean TPR ($\overline{\text{TPR}}$) and D_s for sample size 200. In order to quantitatively investigate whether there is a systematic effect of D_s on $\overline{\text{TPR}}$, we apply an one-factor ANOVA test to test for equal means of $\overline{\text{TPR}}$ for each GRNI algorithm. ANOVA tests give P-values of 0.002016, 1.56×10^{-7}, 0.4053 and 0.874 for ARACNE, CLR, MRNET and RN, respectively. For a significance level of $\alpha=0.01$, CLR and ARACNE show a significant heterogeneous behavior of $\overline{\text{TPR}}$ with respect to D_s. This suggests that $\overline{\text{TPR}}$ is significantly affected by the value of D_s. On the other hand, the results do not suggest a systematic dependence of $\overline{\text{TPR}}$ on D_s for MRNET and RN. For a sample size of 20 (Supplementary Fig. 2), the ANOVA tests give P-values of 5.182×10^{-6}, 1.814×10^{-8}, 6.714×10^{-5} and 0.811 for ARACNE, CLR, MRNET and RN, respectively. This suggests that, for $\alpha=0.01$, ARACNE, CLR and MRNET show a significant influence of D_s on $\overline{\text{TPR}}$ whereas RN does not. For an appropriate interpretation of the shown error bars in Figure 3, the reader is referred to the histogram of D_s counts provided in Supplementary Figure 3.

The next network-based measure evaluates the inferability of basic motif types that consists of four different three-gene motifs, as shown in Supplementary Figure 5. In Table 1, we present the results for these four network motifs for sample size 200, providing their mean true reconstruction rate $\bar{p}$ and its SD $\sigma(\bar{p})$. The mean true reconstruction rate, e.g. for motif of types 1,2 and 3 is given by averaging

$$p=\frac{1}{3}\big(\text{TPR}(\text{A}\leftrightarrow\text{B})+\text{TPR}(\text{B}\leftrightarrow\text{C})+\text{TNR}(\text{A}\not\leftrightarrow\text{C})\big). \qquad (2)$$

over all motifs of the same type within the network (Emmert-Streib and Altay, 2010). More details about this measure are given in the Supplementary Material. From the table we can observe that all algorithms behave similarly with respect to the inferability for these network motifs, favoring motif types 1 and 3. The mean true reconstruction rate for motif type 4 is consistently the worst for all four methods. A general explanation for this behavior (different $\bar{p}$ values for different motif types) seems not to be straight forward. However, assuming a homogeneous TPR for an edge of 0.17 (as implied by motif type 4) and an idealized true negative rate (TNR) of 1.0 for a non-edge would result in, e.g. for motif type 1 or 2,

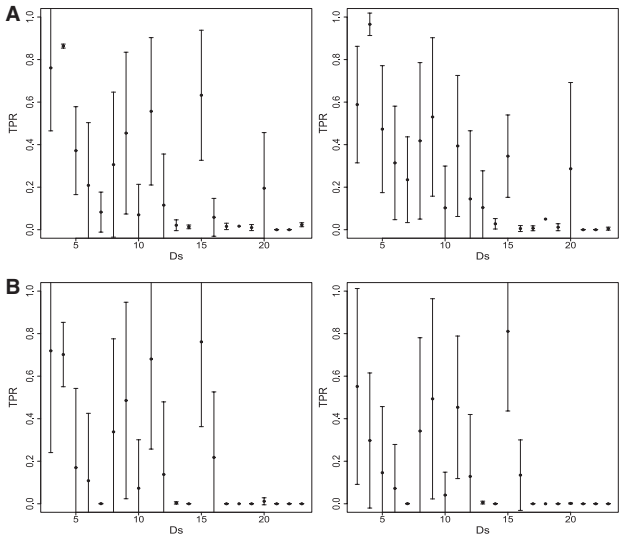

Fig. 3. Mean TPRs for various edge types, D_s, in the true network for sample size 200. The error bars correspond to SDs. (**A**) ARACNE (left) and CLR (right). (**B**) MRNET (left) and RN (right).

$\bar{p} = 0.45 = (2 \times 1.0 + 0.17)/3$. This value would be an upper bound that cannot be exceeded because a TNR of 1.0 is the highest possible value. However, as one can see from Table 1 all $\bar{p}$ values for motif types 1 and 3 are larger than 0.45. Hence, the actual situation is more intricate implying a complex dependency structure. For sample size 20, we observe results that are qualitatively similar to the results discussed above (Supplementary Table 1). These results suggest that there is no bias introduced by the inference algorithms regarding the inferability of individual motif types (compare columns in Table 1); however, each algorithm is biased toward motif types 1 and 2 (compare rows in Table 1).

In order to compare the estimations of motif types with real biological data, we apply our measure to one of the DREAM datasets (DREAM2, Challange 5, *Genome-Scale Network Inference*). This dataset consists of 520 Affymetrix arrays from *E.coli* (Faith *et al.*, 2007). Comparison of the results for CLR and CLR (EC) (Table 1) reveals that the results from our simulations are well confirmed. Specifically, motif type 4 is by far the most difficult one to infer. Interestingly, from the number of motifs present, #*m*, one can see that these are quite different to the subnetwork of yeast we used for our simulations; still, the general inferability of different motif types remains similar. This indicates that for simulations to provide meaningful results it is not necessary to know the underlying network with arbitrary precision but meeting basic characteristics seem to be sufficient. This demonstrates that our approach results in good estimates of the inferability of motif types with a GRNI algorithm for real biological data.

Finally, we investigate the behavior of individual edges, which is the finest resolution one can achieve with any measure. We start with an overview of these results in Figure 4 showing the boxplots of the MI values for TP, FN and FP, TN edges with respect to the true network G.[1] From these boxplots in Figure 4, one sees that for all inference methods the highest MI values of TP edges are in the same order of magnitude as the MI values of FP edges. That means indirect interactions may result in MI values as high as MI values for direct interactions. This does not only hold for the maximum values but also the median values.

[1]We want to note that the terminus, e.g. TP is only possible because we know the true network structure.

Table 1. Summary of motif statistics for ARACNE, CLR, MRNET and RN

Measure/motif type	1	2	3	4
ARA				
#*m*	40	171	446	10
$\bar{p}$	0.591	0.352	0.530	0.156
$\sigma(\bar{p})$	0.15	0.04	0.18	0.12
CLR				
#*m*	40	171	446	10
$\bar{p}$	0.506	0.378	0.480	0.171
$\sigma(\bar{p})$	0.131	0.072	0.137	0.194
MR				
#*m*	40	171	446	10
$\bar{p}$	0.568	0.326	0.582	0.176
$\sigma(\bar{p})$	0.1576	0.0083	0.2237	0.1434
RN				
#*m*	40	171	446	10
$\bar{p}$	0.511	0.321	0.515	0.151
$\sigma(\bar{p})$	0.121	0.010	0.152	0.112
CLR (EC)				
#*m*	2105	321 896	1315	997
$\bar{p}$	0.3625	0.3353	0.3355	0.0558
$\sigma(\bar{p})$	0.103	0.026	0.031	0.168

#*m* is the number of motifs, $\bar{p}$ is the mean true reconstruction rate for a motif and $\sigma(p)$ is its SD. The sample size is 200 for all cases. CLR (EC): Summary of motifs for the *E.coli* dataset from the DREAM conference (Faith *et al.*, 2007).

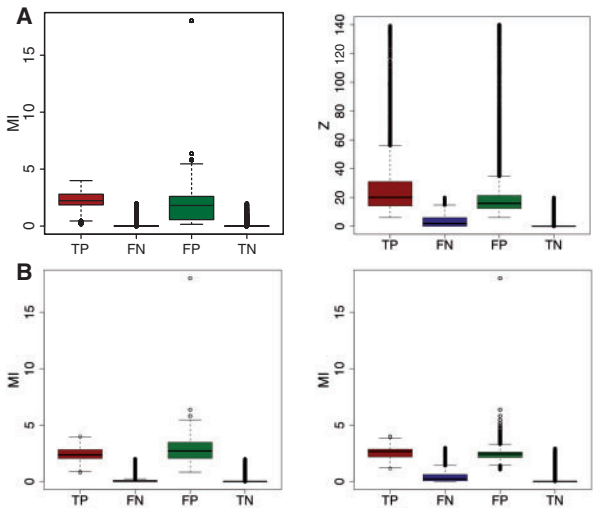

Fig. 4. MI values for TP and FN edges (with respect to the true network) and for FP and TN for sample size 200. (**A**) ARACNE (left) and CLR (right). (**B**) MRNET (left) and RN (right).

In Figure 5, we show a visualization of the mean TPR of edges in the true network for MRNET (Fig. 5A) and CLR (Fig. 5B)—corresponding results for ARACNE and RN can be in found in Supplementary Figure 4. The color code of the edges corresponds to their mean TPR. Specifically, for black edges, $1 \geq \overline{\text{TPR}} > 0.75$; for blue edges, $0.75 \geq \overline{\text{TPR}} > 0.5$; for green edges, $0.5 \geq \overline{\text{TPR}} > 0.25$; and for red edges, $0.25 \geq \overline{\text{TPR}} \geq 0.0$. A visual inspection of these figures suggests that there might be a systematic influence of *in-hubs* and *leafs* on the inferability as reflected by the color of edges. Here, an in-hub is defined as a gene that has more than three incoming edges. We term these incoming edges as in-hub edges. A leaf node is a terminal gene that has exactly one incoming edge. We call this edge leaf edge. In order to quantify this observation, we provide a summary in Table 2. For each of the four algorithms, we count the number of leaf edges (leaf #) and in-hub edges (in-hub #) with respect to the occurrence of red, green, blue and black edges (as defined above). For example, for RN we find a total of $E_L = 41$ leaf edges of which six are green corresponding to 14.6% with respect to E_L. Further, these six green edges represent 50% of all 12 green edges found with respect to the entire network (N_{et}). One can see from this table that in general the probability to observe blue or black leaf edges is much higher than to observe red or green leaf edges, whereas for the in-hub edges this situation is reversed. This implies a systematic bias for all four algorithms.

From our results presented in Table 2 follows an immediate application if combined with results for experimental expression data supposing that the assumptions made for our analysis extrapolate to real data. ARACNE has been applied to expression data from human B cells (Basso *et al*., 2005). The inferred regulatory network consisted of ∼129.000 edges (Basso *et al*., 2005). Using our simulation results, we predict from this that the inferred network constitutes only of about 29.9% of the total number of interactions present in the regulatory network of a B cell. That means the predicted total number of all interactions is about 431.000.[2] This is a conservative estimate that should be considered as a lower bound of the total number of regulatory interactions because of two reasons. First, we did not only consider black (21.1%) but also blue (8.8%) edges. Second, in our simulations we did not estimate a cutoff value I_0 for the declaration of significant MI values from the data but by using the true network structure. This gives an optimal cutoff value resulting in the best performance possible. For estimated cutoff values, the percentages for blue and black edges can only be smaller and, hence, the predicted number of total interaction would increase.

We conduct a similar investigation for *non-edges*. This term refers to the absence of a regulatory interaction in the true network G. A summary of our results is shown in Table 3. From this table, one can see that the TNR is not only highest for MRNET but close to perfect because only three non-edges belong to I_2. ARACNE is slightly worse but also gives very good results. CLR and RN perform worst meaning that their TNR is lowest and the number of non-edges belonging to I_2, I_3 or I_4 is highest. This indicates that there are several non-edges that may very frequently lead to wrong declarations (FP edges) regardless of the used dataset. Hence, even if one would utilize an ensemble of datasets, this would not reduce the risk for declaring several non-edges falsely, a true positive finding.

[2]The number of predicted regulatory interactions holds with respect to the number of genes considered, which was ∼6000 (Basso *et al*., 2005).

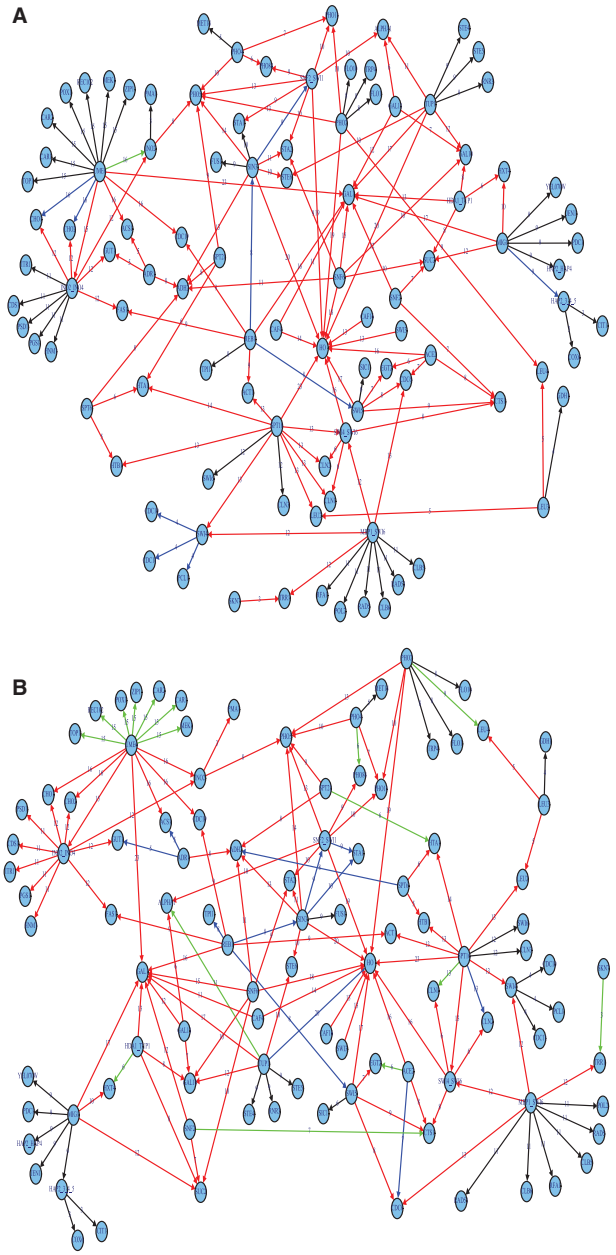

Fig. 5. Visualization of the results for MRNET (**A**) and CLR (**B**) for sample size 200 inferring a subnetwork of Yeast consisting of 100 genes. The color of each edge reflects its mean TPR. Specifically, for black edges, $1 \geq \overline{\text{TPR}} > 0.75$; for blue edges, $0.75 \geq \overline{\text{TPR}} > 0.5$; for green edges, $0.5 \geq \overline{\text{TPR}} > 0.25$; and for red edges, $0.25 \geq \overline{\text{TPR}} \geq 0.0$. The integer numbers at the edges correspond to the value of D_s.

5 DISCUSSION

In this article, we compared four different network inference algorithms—ARACNE, CLR, MRNET and RN—with respect to their performance. For this comparison, we used several local

Table 2. Summary statistics for leaf and hub edges for ARACNE, CLR, MRNET and RN

Measure/edge type	Red	Green	Blue	Black
ARACNE				
True network (#)	93	10	13	31
True network (%)	63.3	6.8	8.8	21.1
Leafs (#)	1	0	10	30
Leafs (% wrt E_L)	2.4	0.0	24.4	73.2
Leafs (% wrt N_{et})	1.1	0.0	77.0	96.8
In-hubs (#)	42	2	0	0
In-hubs (% wrt E_H)	95.5	4.5	0.0	0.0
In-hubs (% wrt N_{et})	45.2	20.0	0.0	0.0
CLR				
True network (#)	91	16	12	28
True network (%)	62.0	10.9	8.2	19.0
Leafs (#)	6	7	1	27
Leafs (% wrt E_L)	14.6	17.1	2.4	65.9
Leafs (% wrt N_{et})	6.6	43.8	8.3	96.4
In-hubs (#)	41	1	2	0
In-hubs (% wrt E_H)	93.2	2.3	4.5	0.0
In-hubs (% wrt N_{et})	45.1	6.3	16.7	0.0
MRNET				
True network (#)	98	1	9	39
True network (%)	66.7	0.7	6.1	26.5
Leafs (#)	0	0	3	38
Leafs (% wrt E_L)	0.0	0.0	7.3	92.7
Leafs (% wrt N_{et})	0.0	0.0	33.3	97.4
In-hubs (#)	44	0	0	0
In-hubs (% wrt E_H)	100.0	0.0	0.0	0.0
In-hubs (% wrt N_{et})	44.9	0.0	0.0	0.0
RN				
True network (#)	100	12	2	33
True network (%)	68.0	8.2	1.4	22.4
Leafs (#)	3	6	2	30
Leafs (% wrt E_L)	7.3	14.6	4.9	73.2
Leafs (% wrt N_{et})	3.0	50.0	100.0	91.0
In-hubs (#)	44	0	0	0
In-hubs (% wrt E_H)	100.0	0.0	0.0	0.0
In-hubs (% wrt N_{et})	44.0	0.0	0.0	0.0

Leafs (% wrt E_L) refers to the percentage of leaf edges of a certain color with respect to the total number of leaf edges, and leafs (% wrt N_{et}) refers to the percentage with respect to the entire network. Correspondingly for in-hubs.

Table 3. Summary statistics of non-edges for ARACNE, CLR, MRNET and RN

Measure/edge type	I1	I2	I3	I4
ARACNE				
(#)	4796	3	2	2
(%)	99.8542	0.06	0.04	0.04
CLR				
(#)	4709	28	47	19
(%)	98.0428	0.58	0.97	0.39
MRNET				
(#)	4800	3	0	0
(%)	99.9375	0.06	0	0
RN				
(#)	4734	21	2	46
(%)	98.5634	0.43	0.04	0.95

I1, I2, I3 and I4 represent the intervals, $1.0 \geq \overline{\text{TNR}} > 0.75$, $0.75 \geq \overline{\text{TNR}} > 0.50$, $0.50 \geq \overline{\text{TNR}} > 0.25$ and $0.25 \geq \overline{\text{TNR}} \geq 0.00$, respectively.

network-based measures in combination with ensemble simulations allowing a detailed analysis down to the level of individual edges. This is the highest resolution achievable. The main purpose of our investigation was not only to reveal the general performance of these methods with respect to the studied novel measures but also to gain insights into a possible bias of these methods. For example, our finding that repressor edges are easier to infer for all four algorithms than activator edges means that the regulatory networks inferred by these methods do systematically discriminate activating interactions. Hence, an interpretation of the inferred networks in biological terms should take this bias into account to avoid spurious conclusions that are in fact induced by the working mechanism of the employed method. We found similar results for network motifs consisting of three genes. Also for these measures all four algorithms behaved largely the same. This is different for the measure D_s. Only ARACNE and CLR showed a significant dependence on D_s.

Application of our simulation results for ARACNE allowed to make a prediction about the expected number of regulatory interactions in human B cells. Extending this discussion we can also formulate a hypothesis about direct interaction partners of Myc as presented in Basso *et al*. (2005). Based on our results presented in Table 2, we hypothesize that these targets are likely to form leaf nodes in the underlying regulatory network. This means that many of these targets may only interact with Myc but no other gene products. This would make these genes peripheral in this network with respect to information processing because they represent so to say one-way streets. More interestingly, because they directly connect with Myc, this transcription factor may also not form a central component of the information processing because it is generally assumed that gene networks are organized hierarchically. Hence, either regulatory networks are hierarchically organized, then the closeness of Myc to leaf genes suggests its decentral character, or Myc is central suggesting either a non-hierarchical organization of the network or the existence of genes that behave non-hierarchical in an otherwise hierarchically organized system. If the latter case is true this might be an indicator of a network characteristics that remained so far covert.

A further prediction that we can make relates to the direction of interactions. Again, based on our results in Table 2, we predict that edges should be oriented toward leaf genes making, e.g. Myc the source of outgoing edges. Due to the fact that Myc is a transcription factor, this appears compelling. However, we want to emphasize that our methods employed in this article are not familiar with the semantics of *transcription factor*, making such a prediction not straight forward for a theoretical method.

For the experimental design of the inference of regulatory networks from expression data (Margolin and Califano, 2007) follow at least two suggestions from our results. First, despite the fact that we studied four different inference methods that have been introduced separately, we demonstrated, by usage of

motif statistics (Table 1), that they behave quantitatively similar. A possible explanation may be that these methods have in common to be based on *bivariate* estimates of MI, not considering higher orders thereof. For this reason, it seems sensible to study multivariate informations in combination with our measures to improve certain regulatory combinations that are by current methods systematically discriminated. Although the extension to motifs involving more than three genes is straight forward, the interpretation of these results guiding the design of n-variate information may be intricate. For this reason, we suggest focusing first on three-gene motifs and corresponding information measures. Second, it would be interesting to study differences between observational and interventional data with respect to, e.g. motif statistics. Specifically, it would be beneficial for future experiments to understand what *parts* of the regulatory network can or cannot be inferred well from observations data only, respectively, from interventional data, to identify an optimal combination of both data types balancing performance and economic constraints. Hence, our study may contribute complementing recent results from DREAM (Stolovitzky and Califano, 2007; Stolovitzky *et al*., 2009) by introducing a novel network-based perspective that may not only help to evaluate methods but also to guide the design of novel statistical estimators.

This discussion underlines the importance of sound simulation studies in order to obtain meaningful interpretations of the inferred networks. Also, as demonstrated with our discussion about Myc and human B cells, such studies enable practical predictions and intriguing hypotheses about the intricate working mechanism of biological pathways and their underlying information processing.

6 CONCLUSION

Despite the wealth of literature already existing studying network inference algorithms, our study is the first comparing such algorithms with respect to local network-based measures in combination with ensemble methods. By emphasizing *network based* and *ensemble*, we want to point out, first, measures should be problem specific, in contrast to general measures such as the F-score, allowing to gain domain-specific insights in the underlying problem. Second, the overall problem of network inference should be put in the context of statistical inference to gain reliable information about the performance and especially errors of such methods under well-defined conditions.

ACKNOWLEDGEMENTS

We would like to thank Andrea Califano and his group for helpful discussions about ARACNE and Shailesh Tripathi for help with the presented figures.

Funding: Department for Employment and Learning through its 'Strengthening the all-Island Research Base' initiative.

Conflict of Interest: none declared.

REFERENCES

Almudevar,A. *et al.* (2006) Utility of correlation measures in analysis of gene expression. *NeuroRx*, **3**, 384–395.

Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: Understanding the cell's functional organization. *Nat. Rev.*, **5**, 101–113.

Basso,K. *et al.* (2005) Reverse engineering of regulatory networks in human B cells. *Nat. Genet.*, **37**, 382–390.

Butte,A.J. and Kohane,I.S. (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.*, **5**, 415–426.

Chartrand,G. and Oellermann,O. (1993) *Applied and Algorithmic Graph Theory*. McGraw-Hill, New York.

Cover,T.M. and Thomas,J.A. (1991) *Information Theory*. John Wiley & Sons, Inc., New York.

Emmert-Streib,F. (2007) The chronic fatigue syndrome: a comparative pathway analysis. *J. Comput. Biol.*, **14**, 961–972.

Emmert-Streib,F. and Altay,G. (2010) Local network-based measures to assess the inferability of different regulatory networks. *IET Systems Biol.*, in press.

Emmert-Streib,F. and Dehmer, M. (eds) (2008) *Analysis of Microarray Data: A Network Based Approach*. Wiley-VCH, Weinheim.

Faith,J.J. *et al.* (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, **5**, 54–66.

Guelzim,N. *et al.* (2002) Topological and causal structure of the yeast transcriptional regulatory network. *Nat. Genet.*, **30**, 60–63.

Hache,H. *et al.* (2009) Reverse engineering of gene regulatory networks: a comparative study. *EURASIP J. Bioinform. Syst. Biol.*, **2009**, 617281.

Holland,P. (1986) Statistics and causal inference. *J. Am. Stat. Assoc.*, **81**, 945–960.

Husmeier,D. (2003) Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, **19**, 2271–2282.

Levine,M. and Davidson,E.H. (2005) Gene regulatory networks for development. *Proc. Natl Acad. Sci. USA* , **102**, 4936–4942.

Ma,H.W. *et al.* (2004) Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach. *BMC Bioinformatics*, **5**, 199.

Margolin,A.A. and Califano,A. (2007) Theory and limitations of genetic network inference from microarray data. *Ann. N. Y. Acad. Sci.*, **1115**, 51–72.

Margolin,A.A. *et al.* (2006) Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7**, S7.

Meyer,P.E. *et al.* (2007) Information-theoretic inference of large transcriptional regulatory networks. *EUROSIP J. Bioinform. Syst. Biol.*, **2007**, 79879.

Meyer,P.E. *et al.* (2008) minet: r/bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, **9**, 461.

Milo,R. *et al.* (2002) Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–827.

Olsen,C. *et al.* (2009) On the impact of entropy estimator in transcriptional regulatory network inference. *EURASIP J. Bioinform. Syst. Biol.*, **2009**, 308959.

Palsson,B.O. (2006) *Systems Biology*. Cambridge University Press, Cambridge, New York .

Pearl,J. (2000) *Causality: Models, Reasoning, and Inference.* Cambridge University Press, Cambridge, UK; New York, USA.

Sheskin,D.J. (2004) *Handbook of Parametric and Nonparametric Statistical Procedures*, 3rd edn. RC Press, Boca Raton, FL.

Shipley,B. (2000) *Cause and Correlation in Biology*. Cambridge University Press, Cambridge, UK; New York, NY, USA.

Stolovitzky,G. and Califano,A. (2007) *Reverse Engineering Biological Networks: Opportunities and Challenges in Computational Methods for Pathway Inference*, Wiley-Blackwell, Boston, Mass.

Stolovitzky,G. *et al.* (2009) *The Challenges of Systems Biology: Community Efforts to Harness Biological Complexity*, Wiley-Blackwell, Hoboken, NJ.

Van den Bulcke,T. *et al.* (2006) SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics*, **7**, 43.

Werhli,A.V. *et al.* (2006) Comparative evaluation of reverse engineering gene regulatory networks with relevance netorks, graphical Gaussian models and Bayesian networks, *Bioinformatics*, **22**, 2523–2531.

Yu,H. and Gerstein,M. (2006) Genomic analysis of the hierarchical structure of regulatory networks. *Proc. Natl Acad. Sci. USA*, **103**, 14724–14731.

Zhang,B. and Horvath,S. (2005) A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.*, **4**, 17.