Influence of Statistical Estimators of Mutual Information and Data Heterogeneity on the Inference of Gene Regulatory Networks

Ricardo de Matos Simoes, Frank Emmert-Streib*

Computational Biology and Machine Learning Lab, Center for Cancer Research and Cell Biology, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, Belfast, United Kingdom

Abstract

The inference of gene regulatory networks from gene expression data is a difficult problem because the performance of the inference algorithms depends on a multitude of different factors. In this paper we study two of these. First, we investigate the influence of discrete mutual information (MI) estimators on the global and local network inference performance of the C3NET algorithm. More precisely, we study 4 different MI estimators (Empirical, Miller-Madow, Shrink and Schürmann-Grassberger) in combination with 3 discretization methods (*equal frequency*, *equal width* and *global equal width* discretization). We observe the best global and local inference performance of C3NET for the Miller-Madow estimator with an *equal width* discretization. Second, our numerical analysis can be considered as a systems approach because we simulate gene expression data from an underlying gene regulatory network, instead of making a distributional assumption to sample thereof. We demonstrate that despite the popularity of the latter approach, which is the traditional way of studying MI estimators, this is in fact not supported by simulated and biological expression data because of their heterogeneity. Hence, our study provides guidance for an efficient design of a simulation study in the context of network inference, supporting a systems approach.

Citation: de Matos Simoes R, Emmert-Streib F (2011) Influence of Statistical Estimators of Mutual Information and Data Heterogeneity on the Inference of Gene Regulatory Networks. PLoS ONE 6(12): e29279. doi:10.1371/journal.pone.0029279

Editor: Vladimir Brusic, Dana-Farber Cancer Institute, United States of America

Received September 21, 2011; **Accepted** November 23, 2011; **Published** December 29, 2011

Funding: Part of this work has been supported by the Engineering and Physical Sciences Research Council (EPSRC). No additional external funding was received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: v@bio-complexity.com

Introduction

The mutual information (MI) is a measure to quantify the non-linear dependency between two random variables [1,2]. The most popular strategies for estimating mutual information values are based on a discretized model for continuous data [3]. This strategy is widely known as the *histogram approach* that approximates the joint probability distribution by their empirical joint frequencies in bins of the two discretized random variables [1]. A variety of different mutual information estimators were developed in order to obtain statistical estimates for data sampled from an underlying distribution. The simplest estimator is the Empirical estimator [3] that is computed from the observed cell frequencies of a discretized distribution. However, it has been shown that the Empirical estimator underestimates the entropy due to an undersampling of cell frequencies and of zero cell frequencies which increase with the number of bins [4]. This is a major problem for practical applications due to finite data and the requirement for a large number of bins for accurate estimates. To account for the induced bias of the MI estimate, a variety of methods were developed that adjust the estimate by a constant factor [3], use a shrinkage regularization [5] or employ a Bayesian approach to estimate the joint frequencies for the bins from a Dirichlet distribution [6] to gain more accurate estimates.

In systems biology [7–10], many gene regulatory network (GRN) inference methods use mutual information as an estimator to unveil the interaction structure and the relations among genes in a cellular system from gene expression data [11–16]. One of the first methods based on mutual information for GRN inference was introduced in [17]. The underlying method, called *relevance network* (RN), assigns edges to gene pairs if the corresponding MI value is above a given threshold. The networks that result from application of RN are association networks because an edge between two genes indicates merely their association but not necessarily a causal effect [18–20]. A different type of GRN methods are network inference methods that aim to inferring causal interactions among genes and their products which can be experimentally validated [21–26]. The networks that result from such an inference are called *gene regulatory networks* [27]. So far, there is no generally agreed gold standard to conduct and include the routine of gene regulatory network inference and their analysis for molecular studies. However, necessary preprocessing steps of the data to prepare them for the subsequent inference of a gene regulatory network involve standardized procedures for the normalization of gene expression distributions within and between samples and a summarization step to obtain gene-centric values of the gene expression [28–30].

The major purpose of this paper is to investigate the influence of the MI estimator and the choice of the discretization method on the inference of gene regulatory networks. Despite the enormous popularity of the inference of GRNs this topic has so far only been addressed by a few studies. A notable exception is [31]. There gene regulatory network algorithms were evaluated for different mutual information estimators demonstrating that the choice of the estimator influences the inference performance in a significant manner. It has also been shown that the choice of the discretization and MI estimator is specific for a gene regulatory network inference method. In contrast to this study, there are many investigations that analyze statistical estimators of MI values directly without considering the MI estimator as part of a larger model, like the inference algorithm for a GRN. Usually, such studies assess either the estimates of the probability distribution respectively of the cell frequencies of an estimator, or study the resulting entropies [32–35]. The first approach is based on the *plug-in* usage of the estimated probabilities whereas the second utilizes the fact that the mutual information can be expressed in terms of entropies [36]. This can be seen as traditional approach because it places the statistical estimator itself in the focus of the investigation. In addition, another characteristics of these traditional studies is that they make assumptions about the distribution of the data and then assess the statistical MI estimator by means of simulated data, generated in accordance with these assumptions. This allows a thorough statistical analysis because each model parameter can be controlled appropriately, possibly, up to computational limitations.

The major purpose of this paper is to analyze the influence of statistical estimators of the mutual information on the inference of regulatory networks from large-scale gene expression data for the C3NET algorithm using global and local network-based error measures. More specifically, we investigate 3 different discretization methods in combination with 4 different statistical estimators for mutual information values. We infer networks from *in silico* (simulated) gene expression datasets generated for three Erdös-Rényi networks for various sample sizes and assess the influence of the MI estimators by global and local network-based error measures. In addition, we investigate and define data heterogeneity and discuss general consequences thereof on the simulation approach for network inference methods. This will provide us with general insights and reveals a problem of traditional studies of MI estimators which do not place the MI estimator into a model for which it is intended, but study it in isolation. For this reason, nonlinear effects that are only present for the larger model may lead to unexpected performance results which does not reflect the performance of the estimator in the isolated study. One reason therefore is the violation of assumptions. It is clear that assumptions that reflect real data *appropriately* lead to comparable results for simulated and real data. However, the more these assumptions are violated the more the results may deviate. This is well known, but does not reveal the only problem one encounters by using a statistical estimator within a larger model. The additional problem with the latter is in the specific context we are interested in, namely the inference of GRN from expression data, that the correlation structure within the data is only poorly understood, and, hence, there is no simple way known to simulate the expression of individual genes without simulating the entire gene network [37,38]. In the results section of this paper we will show that as a consequence of these correlations among genes, there are many different types of probability distributions present within one gene regulatory network caused by the heterogeneity of the expression data. This makes it practically impossible to reduce the test of a MI estimator to a single probability distribution, but one would need to consider a *distribution* of probability distributions to test a MI estimator. From the presence of the data heterogeneity we conclude that a MI estimator can only be meaningfully studied within a network inference model and genome-wide data. This can be seen as a systems approach [39,40] because reducing either the model to the MI estimator solely or the data to a *fixed* joint probability distribution does not lead to a realistic testing of the biological system.

This paper is organized as follows. In the next section we introduce the methods used for our analysis. Then we present numerical results of the influence different MI estimators have on the inference performance of C3NET by using global and local error measures. In addition, we analyze the effect of data heterogeneity on the estimation of MI values. The paper finishes with a discussion and conclusions.

Methods

The network inference method C3NET

The C3NET (conservative causal core) algorithm consists of three main steps [21]. The first step is for estimating mutual information values for all gene pairs. In the second step, the most significant link for each gene is selected. In the third step, non-significant links, according to a chosen significance level α, between gene pairs are eliminated by application of a multiple testing correction procedure. The complexity for multiple hypothesis testing (MHT) for C3NET is $O(N)$, whereas N corresponds to the number of genes. The C3NET algorithm selects at most N edges and therefore at most N multiple tests are required. In comparison, other gene regulatory network inference approaches, e.g., RN [41], ARACNE [25] or CLR [42] eliminate non-significant links between *all* possible gene pairs in the first step. This leads to a complexity for multiple hypothesis testing (MHT) of $O(N^2)$. Some methods often circumvent the more extensive computational effort by applying arbitrarily chosen fixed significance thresholds.

The inferred edges in a C3NET gene regulatory network correspond to the highest MI value among the neighbor edges for each gene. This implies that the highest possible number of edges that can be inferred by C3NET is equal to the number of genes under consideration. This number can decrease for several reasons. For example, when two genes have the same edge with maximum MI value. In this case, the same edge would be chosen by both genes to be included in the network. However, if an edge is already present another inclusion does not lead to an additional edge. Another case corresponds to the situation when a gene does not have significant edges at all. In this case, apparently, no edge can be included in the network. Since C3NET employs MI values as test statistics among genes, there is no directional information that can be inferred thereof. Hence, the resulting network is undirected and unweighted. Figure 1 shows the principle working mechanism of the network inference method C3NET.

Estimating mutual information

In the following we investigate 4 different types of estimators that are based on the so called histogram approach. In the first step the expression values of two genes are discretized into defined intervals, denoted as bins. The mutual information is a measure for the nonlinear dependence of the two random variables. Mutual information is defined by the marginal probability $P(X)$ and $P(Y)$ and joint probability $P(X,Y)$ of two random variables X and Y [36]:

$$I(X,Y)=\sum_{x_i\in X}\sum_{y_j\in Y}P(X=x_i,Y=y_j)\cdot\log\frac{P(X=x_i,Y=y_j)}{P(X=x_i)\cdot P(Y=y_j)} \quad (1)$$

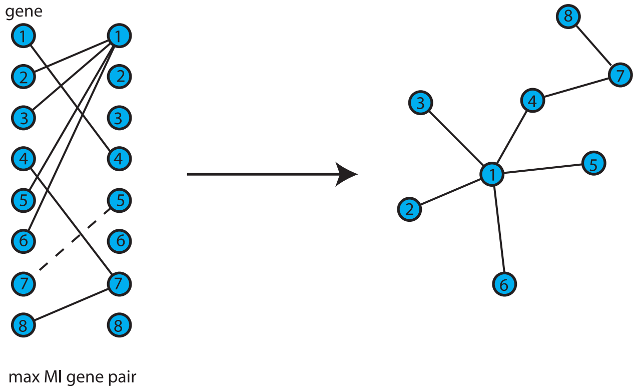

Figure 1. Principle working mechanism of C3NET. For each gene an edge is assigned to the gene neighbor with maximal MI value. The MI value between gene 7 and 5 (dashed) is not significant. The resulting network describes the core of a gene regulatory network.
doi:10.1371/journal.pone.0029279.g001

Here $\log$ means the logarithm to the base of 2. $I(X,Y)$ is always ≥ 0. For example if the two random variables are independent from each other the mutual information is 0, because $P(x,y)=P(y)P(x)$.

The mutual information can also be expressed in terms of entropies [36],

$$I(X,Y)=H(X)+H(Y)-H(X,Y) \qquad (2)$$

Here the entropy for a random variable X is defined by:

$$H(X)=-\sum_{x_i \in X} P(X=x_i)\cdot\log(P(X=x_i)) \qquad (3)$$

and the joint entropy $H(X,Y)$ is given by

$$H(X,Y)=-\sum_{x_i \in X}\sum_{y_j \in Y} P(X=x_i,Y=x_j,)\cdot\log(P(X=x_i,Y=x_j))(4)$$

We describe four different strategies for estimating mutual information for a discretized model. The simplest estimator is the empirical MI estimator [3] that estimates entropy from the observed joint frequencies for each bin. The empirical entropy H_{emp} is estimated from the observed probability distribution with n_k number of samples in bin k, total number of samples N and the total number of bins b. Note that the entropy formulas shown in the following are for a single random variable.

$$H_{emp}=-\sum_{k=1}^{b}\left(\frac{n_k}{N}\right)\log\left(\frac{n_k}{N}\right) \qquad (5)$$

The Empirical estimator gives the maximum-likelihood entropy estimate for a discretized random variable. A main problem of the empirical approach is the underestimation of the true entropy H due to undersampling of the cell frequencies with increasing number of bins. A variety of approaches were developed to account for the induced bias that range from correcting the estimate by a constant factor or using a multinomial distribution to model the extend of missing information. In the following we show 4 different MI estimators that are based on a discretized model.

The Miller-Madow estimator [3] accounts for the undersampling bias by adjusting the estimate by a constant factor that is proportional to the bin size and sample size:

$$H_{mm}=H_{emp}+\frac{b-1}{2\cdot N} \qquad (6)$$

for b number of bins and N number of samples.

The following two estimators consider the correction of the probability distributions directly. The shrinkage estimator [5] combines two models defining a model with cell frequency of $\frac{1}{b}$ and a model defining the empirical cell frequency for each bin k.

$$\hat{p}_{\lambda}(n_k)=\lambda\frac{1}{b}+(1-\lambda)\frac{n_k}{N}$$

The weighting parameter λ is estimated by minimizing the mean squared error for the two models for each k of b bins (Equation 7).

$$\lambda^{*}=\underset{\lambda\in[0,1]}{\text{argmin}}\; E[\sum_{k\in b}(p_{\lambda}(n_k)-p(n_k))^2] \qquad (7)$$

The entropy for the shrinkage optimized probability distribution is computed by:

$$\hat{H}^{shrink}=-\sum_{k=1}^{b}\hat{p}_{\lambda}(n_k)\log\hat{p}_{\lambda}(n_k) \qquad (8)$$

The Schürmann-Grassberger estimator [6] is based on a Bayesian approach that uses the Dirichlet probability distribution as conjugate prior for the likelihood given by the Empirical estimator. The Dirichlet distribution describes the distribution of probability distributions with mean θ_k.

$$f(\chi;\theta)=\frac{\Pi_{k\in\chi}\Gamma(\theta_k)}{\Gamma(\sum_{k\in\chi}\theta_k)}\prod_{k\in\chi}x_k^{\theta_k-1} \qquad (9)$$

The mean θ_k probability for each bin k is estimated from the posterior using the Schürmann-Grassberger parameter $\frac{1}{b}$ that equals:

$$\hat{\theta}_k=\frac{n_k+\frac{1}{b}}{N+1} \qquad (10)$$

In overall one pseudocount is added to the total sample count N. The entropy is estimated by:

$$\hat{H}^{dir}=-\sum_{k=1}^{p}\hat{\theta}_k\,\log\hat{\theta}_k \qquad (11)$$

The MI estimators are used in combination with three different discretization methods. The first, *equal frequency* method assigns the same frequency of values to each bin k. The *equal width* method uses for each interval width for each bin the same value. However, this is done independently for each of the two random variables X and Y. In contrast, the *global equal width* uses the same interval width for both random variables. The number of bins are defined by the proportional k-interval discretization method with $\sqrt{N}$ [43], where the number of bins is dependent on the number of samples.

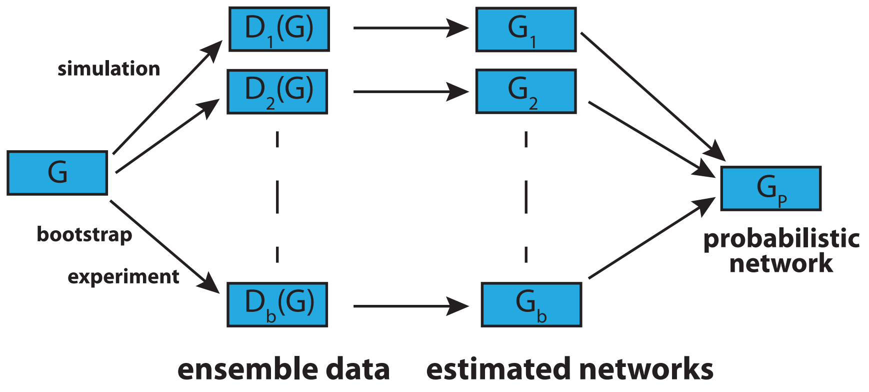

Figure 2. Schematic visualization of ensemble data from which networks are inferred and subsequently aggregated to estimate a probabilistic network.
doi:10.1371/journal.pone.0029279.g002

Global error measure

In order to measure the influence of the MI estimators on the inference performance of the C3NET algorithm we use the area under the precision-recall curve for the *receiver operator characteristics* (AUC-PR) [44]. The precision and recall measures [45] are obtained by comparison of an inferred (predicted) network with the true network used to simulate the underlying data. The recall, also known as the sensitivity, denotes the proportion of true positive edges relative to all edges in the reference network.

$$\text{recall (sensitivity)} \qquad \mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{12}$$

The precision gives the proportion of correctly inferred edges relative to all inferred edges.

$$\text{precision} \qquad \mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{13}$$

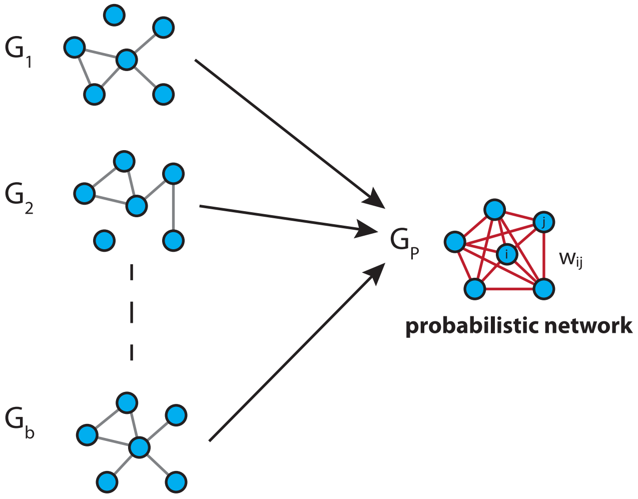

Figure 3. The ensemble of networks $\mathcal{G}^e = \{G_i\}_{i=1}^{b}$ is used to obtain a weighted network G_P.
doi:10.1371/journal.pone.0029279.g003

The predicted edges of gene regulatory network are ranked, e.g., by their respective MI estimate or alternative statistical measures. For a given threshold θ a confusion matrix can be defined when the true underlying network is known such as for simulated data. A confusion matrix tabulates the number of true positive, false positives, true negatives and false negative predictions. The PR-curve describes the precision (predicted true positives) as function of the sensitivity (recall, true positive rate) obtained by using various threshold values $\theta \in \Theta$ for the rank measures of the predicted edges. The AUC-PR area under the curve value is computed by a numerical integration along each point of the curve.

Ensemble data and local network-based measures

In contrast to the above measure, which is a global error measure, we use also local network-based measures to assess the influence of the MI estimators. The principle idea of local *network-based measures* was introduced in [38,46]. These local network-based error measures are based on ensemble data and the availability of a reference network G that represents the true regulatory network. Ensemble data means that there is more than one dataset available from the biological phenomenon under investigation. This ensemble of data could be either obtained by bootstrapping from one large data set, or from a simulation study, or from multiple experiments.

After obtaining the ensemble of data $\mathcal{D} = \{D_1(G), \ldots, D_b(G)\}$, the inference algorithm is applied to $\mathcal{D}$ resulting in an ensemble of estimated networks $\mathcal{G}^e = \{G_i\}_{i=1}^{b}$ (see Fig. 2 for a visualization). Here we emphasize that each dataset depends on the underlying network structure, G, that governs the coupling among the genes by writing, e.g, $D_i(G)$. Further, this indicates that always the same network G is used.

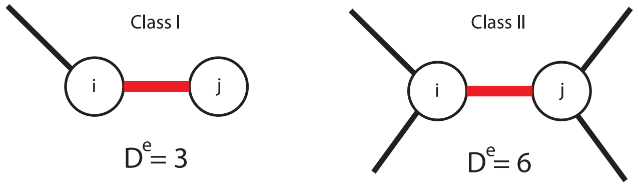

Figure 4. An edge (red) is scored according to the degree of the parental nodes of an undirected network (sum of degree *i* and degree *j*) and assigned to Class I if $D^e \leq 4$ and to Class II if $D^e > 4$.
doi:10.1371/journal.pone.0029279.g004

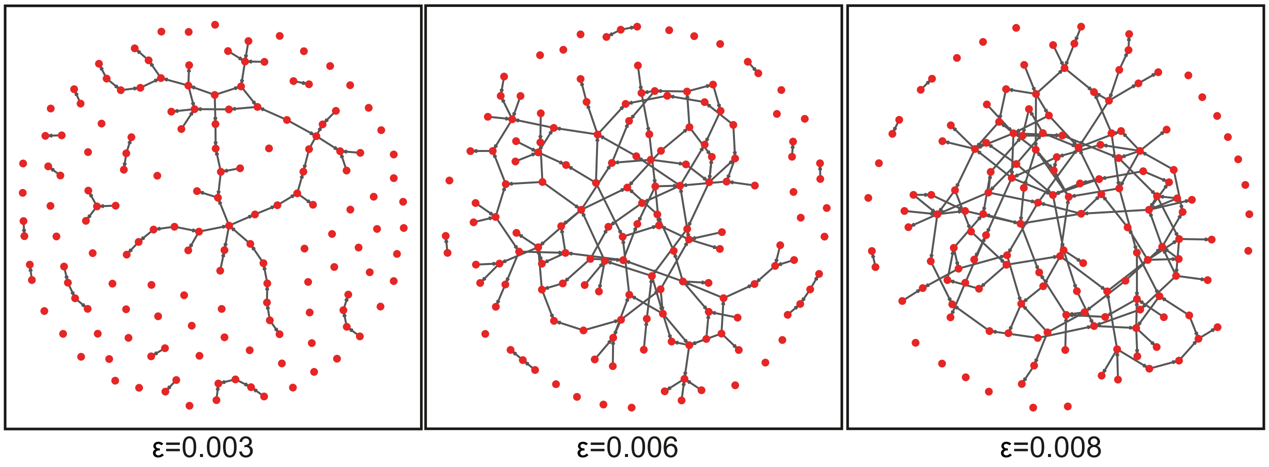

Figure 5. Shown are three Erdös-Rényi networks with edge density $\epsilon = \{0.003, 0.006, 0.008\}$**.**
doi:10.1371/journal.pone.0029279.g005

If the network structure of the underlying network G is available it is possible to obtain estimates of the TPR (true positive rate) of edges and non-edges in G. From the ensemble of estimated networks $\mathcal{G}^e = \{G_i\}_{i=1}^b$ one obtains a probabilistic network G_P. The edge weights of G_P give the TPR for each edge which quantifies how often an edge was observed in the ensemble. The edge weights between gene i and j of G_P are defined by

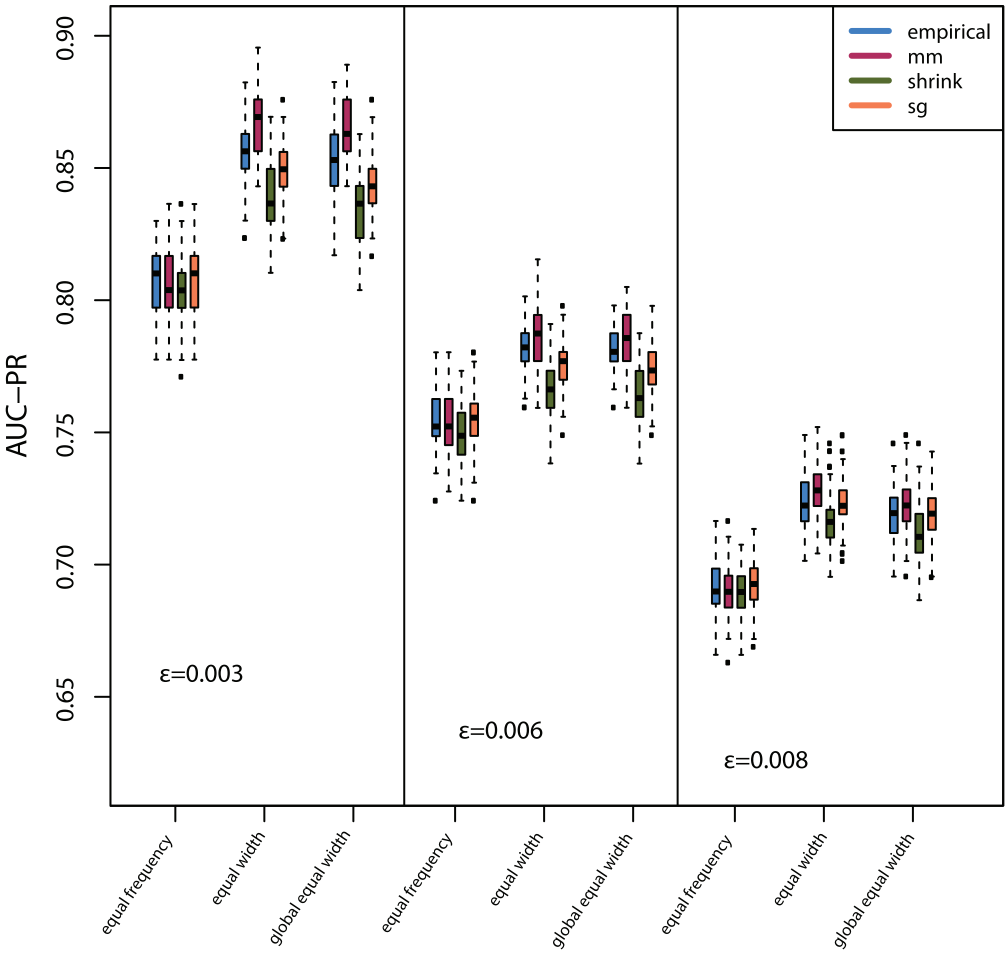

Figure 6. The influence of different discretization methods on the global network inference performance (AUC-PR) for three Erdös-Rényi networks and 4 **MI estimators.** The simulated gene expression datasets have sample size 200.
doi:10.1371/journal.pone.0029279.g006

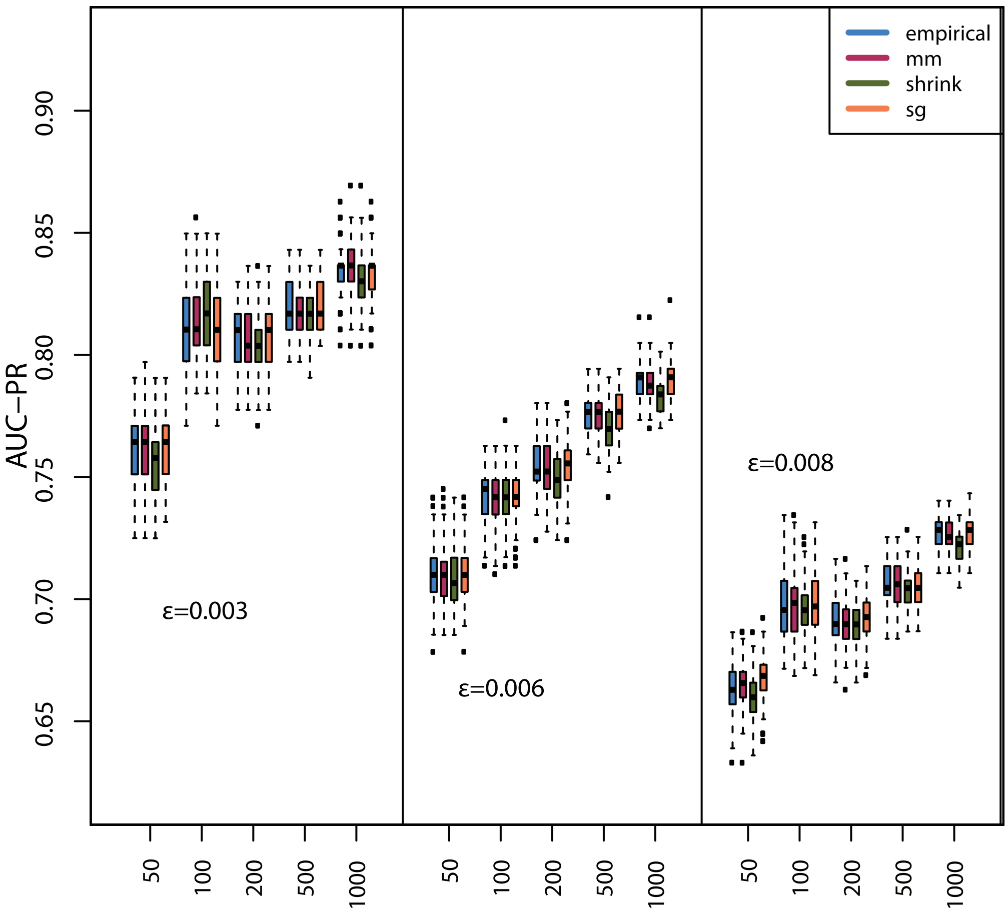

Figure 7. The influence of the *equal frequency* discretization method on the global network inference performance (AUC-PR) for three Erdös-Rényi networks and 4 MI estimators.
doi:10.1371/journal.pone.0029279.g007

$$G_P(i,j)=\frac{1}{b}\sum_{k=1}^{b}I_1(G_k(i,j))=\frac{1}{b}\#\{G_k(i,j)=1|\{G_k\}_{k=1}^{b}\}. \quad (14)$$

The indicator function $I()$ is 1 if an edge between gene i and j is observed in a network $G_k(i,j)=1$ and 0 otherwise. From Eqn. 14 follows that $G_P(i,j)$ corresponds to the probability that an edge is present in $\mathcal{G}^e$ connecting gene i with j. The combination of the networks in $\mathcal{G}^e=\{G_i\}_{i=1}^{b}$ leading to the probabilistic network G_P is visualized in Fig. 3.

Local network-based error measure

We study the influence of the MI estimator on the inference performance for two classes of edges. The first class (Class I) of edges corresponds to linearly connected nodes and the second class (Class II) of edges corresponds to nodes with a high degree. The two edge classes are defined via *local network-based* measures. We define the *local network-based* measure D^e for each edge in an undirected graph, by the sum of the degrees of node i and the degrees of node j:

$$D^e_{ij}=deg(v_i)+deg(v_j) \quad (15)$$

Based on the values of D^e we define a binary classification for the edges by:

- Class I: edges with $D^e\leq 4$ (corresponds to a chain-like structure)
- Class II: all other edges

In order to visualize our definition, we present in Fig. 4 two examples for D^e. The left side shows an example for a Class I edge (with a score ≤ 4) and the right side an example for a Class II edge (with a score >4). As described in the section 'Ensemble data and local network-based measures', a TPR is obtained for each edge. From the ensemble of datasets the distribution of the average TPRs is obtained for the two edge classes which we use for their comparison.

Simulation of gene expression data

In order to study the influence of the network connectivity on the MI value estimators we are using random networks with different values of ϵ. Here ϵ is the probability for the presence of an edge between two nodes [47]. Because real gene networks, e.g., the transcriptional regulatory network or the protein interaction network, are sparsely connected, the value of ϵ needs to be chosen to fall within a realistic interval. Typically, gene networks have an edge density of about $\sim 10^{-3}$ [48].

We generate three Erdös-Rényi graphs [47,49] with 150 genes and an edge density of $\epsilon\in\{0.003,\ 0.006,\ 0.008\}$. The resulting networks have $\{60,22,19\}$ unconnected genes to model non-expressed genes. These three networks are shown in Figure 5. For

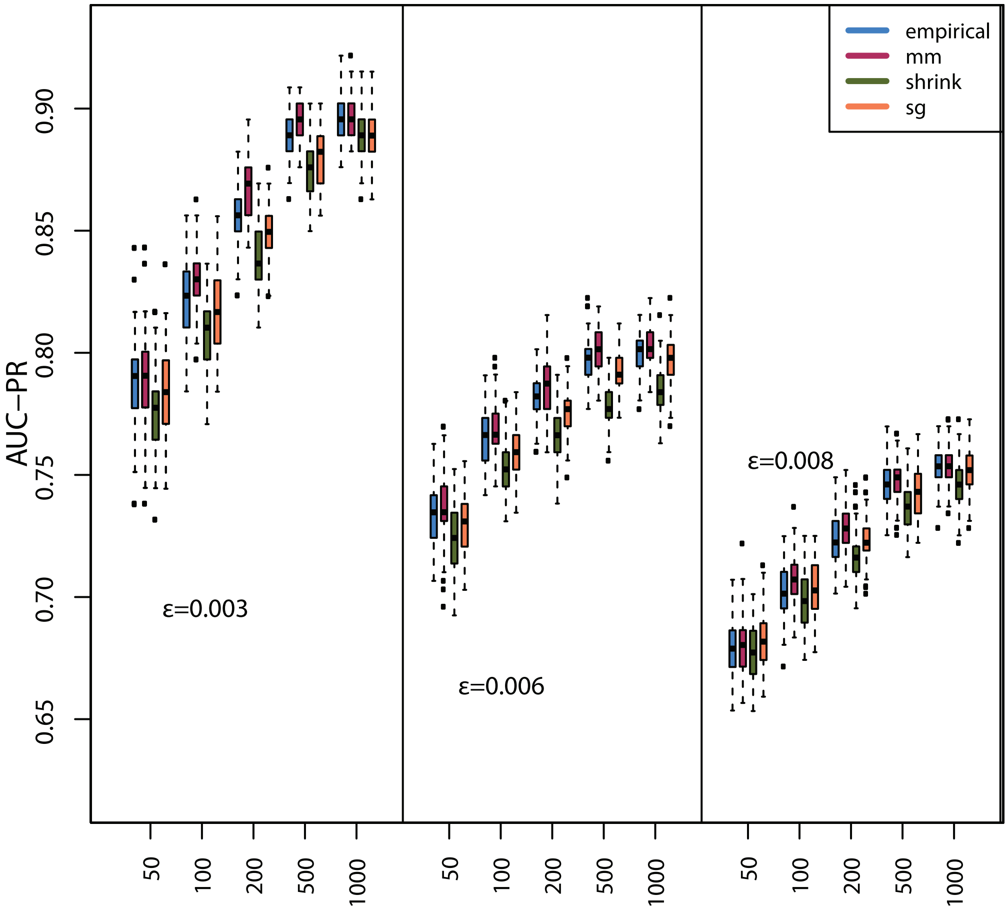

Figure 8. The influence of the *equal width* discretization method on the global network inference performance (AUC-PR) for three Erdös-Rényi networks and 4 MI estimators.
doi:10.1371/journal.pone.0029279.g008

our study we generate *steady-state* gene expression data using SYNTREN [50]. For each Erdös-Rényi network and sample size $s \in \{50,100,200,500,1000\}$ we generate $M = 100$ datasets.

Results

In the following we study 4 MI estimators in combination with 3 discretization methods. We assess the influence of these estimators on the inference of regulatory networks obtained with the C3NET algorithm by application of global and local network-based error measures. To ensure that our results are statistically robust, we base our study on ensemble data by simulating M datasets for each studied condition. Using simulated data enables the comparison of the inferred networks with the true reference networks and also the control of important parameters. In the last results section we study the influence of data heterogeneity on the MI estimates.

Influence of MI estimators on the global error measure AUC-PR

First, we compare the impact of the discretization method on the AUC-PR (Figure 6). For all three Erdös-Rényi networks, the *equal width* and the *global equal width* discretization lead to a better inference performance of C3NET compared to the *equal frequency* discretization (Figure 6). Further, the *equal width* and *global equal width* discretization in combination with the Miller-Madow estimator is better than any other combination of MI estimator and discretization method. The second best MI estimator is the Empirical estimator. The Schürmann-Grassberger estimator and the Shrink estimator perform worse, whereas the Schürmann-Grassberger estimator performs better than the Shrink estimator (Figure 6). When using the *equal frequency* discretization, all MI estimators perform equally and show no substantial difference, see Fig. 6. Further, one can see that the network inference performance is decreasing with an increasing edge density. This is reasonable and related to the working mechanism of C3NET. Due to the fact the C3NET allows each gene to contribute at most one edge to the resulting network, the inference of networks that have a higher edge density is systematically disfavored.

In addition, we study also the dependency of the MI estimators, for each of the 3 discretization methods, on the sample size. The corresponding results are shown in Fig. 7 (*equal frequency* discretization) Fig. 8 (*equal width* discretization) and Fig. 9 (*global equal width* discretization). For all investigated network types the inference performance increases with the sample size, as expected. However, independent of the influence of the sample size, the Miller-Madow estimator in combination with the *equal width* or the *global equal width* discretization show the best performance with respect to the inference performance of C3NET.

Influence of MI estimators on the local error measure D^e

In the previous section we studied MI estimators by using a global error measure. When a global error measure is used, we actually do

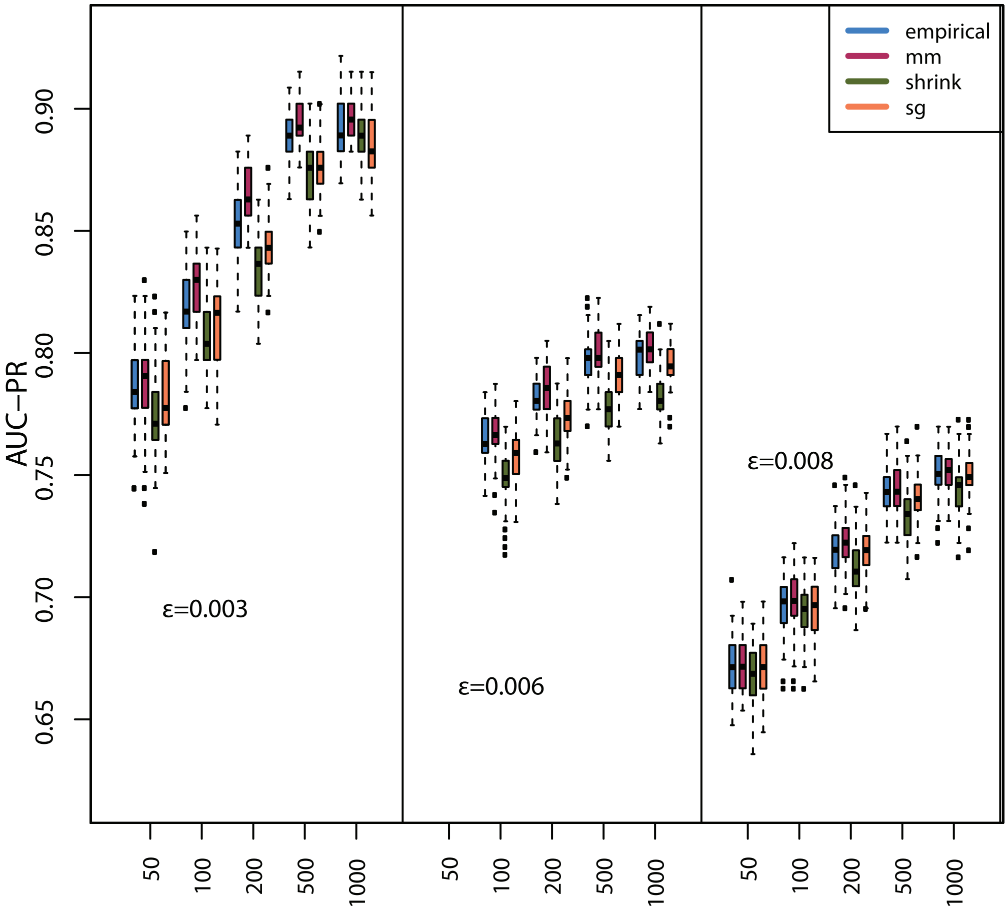

Figure 9. The influence of the *global equal width* discretization method on the global network inference performance (AUC-PR) for three Erdös-Rényi networks and 4 MI estimators.
doi:10.1371/journal.pone.0029279.g009

not obtain information about the performance of individual edges, but only of the average performance of all edges in the network. In order to *zoom in* to local properties of the network, we study in this section the local network-based measure D^e. This measure allows to divide the edges in a network into two classes, according to the structural neighborhood of an edge, as defined in section 'Local network-based error measure'. Specifically, this will allow us to gain information about the influence the MI estimators have on edges with a certain structural property. We use the local network-based measure D^e to distinguish between two edge classes. The first class represents edges from linearly connected genes (Class I) and the second class represents edges that belong to genes with multiple edges (Class II), see Fig. 4 for a visualization. In the following, we study the influence of the MI estimators on these two edge classes separately. We expect that edges from linear connected genes have only few dependencies that affect the underlying gene expression patterns and thus are more easier to infer. In contrast, edges connected to genes that are influenced by multiple other genes are expected to show more complex gene expression patterns and are therefore more difficult to estimate.

For the following simulations we use the *equal width* discretization in combination of the 4 MI estimators for gene regulatory network inference with C3NET. In Figure 10 and 11 we show the distributions of true positive rates for the three random networks and different sample sizes for the two edge classes. In general, the Class I edges show a much better inference performance compared to the Class II edges. Also, for the Class I edges the estimators do not have a substantial influence on the inference performance. However, for the Class II edges we observe that the Miller-Madow estimator performs best, followed by the Empirical estimator. The Schürmann-Grassberger estimator and the shrinkage estimator rank last. In contrast to the global error measure (AUR-PR), for edges of Class II the shrinkage estimator performs better than the Schürmann-Grassberger estimator (Figure 11).

Influence of data heterogeneity

Traditionally, when MI estimators are investigated, they are studied by making an assumption about the joint probability distribution $p(x,y)$ of the two random variables the MI estimator depends on. From this, the marginal distributions $p(x) = \int p(x,y)dy$ and $p(y) = \int p(x,y)dx$ are obtained. Frequently, normality is assumed but also other probability distributions have been studied [32–34]. Regardless of the actual probability distribution selected, all studies have in common to make an implicate assumption which translates into a *homogeneity* of the data. That mean a MI estimator is investigated with respect to a *fixed* probability distribution from which data are sampled. This may be repeated for several different but *fixed* probability distributions. The crucial point here is that this investigation is done for each probability distribution separately which means that a MI estimator is assessed by using datasets that come from only one underlying distribution. With respect to this distribution, the

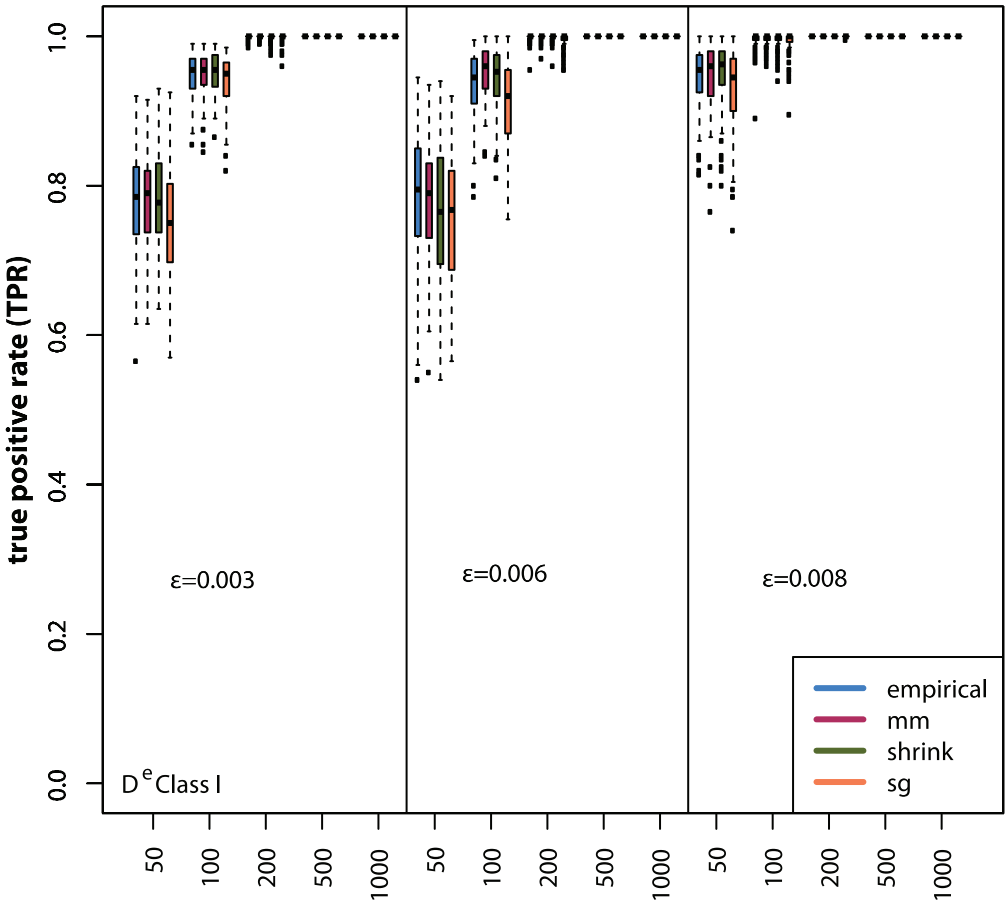

Figure 10. Local network inference performance for Class I edges. Simulated gene expression datasets for Erdös-Rényi networks with edge density ϵ for sample sizes ranging from 50 to 1000 samples.
doi:10.1371/journal.pone.0029279.g010

sampled datasets are homogeneous because no otherwise distributed data are present. In this section we will study different aspects of the homogeneity assumption and consequences thereof. We proceed by, first, investigating various tests about the normality of the data and then consider arbitrarily distributed data.

We start by testing the null hypothesis if a gene expression profile is normally distributed, by using the Anderson-Darling test [51]. For the $M=100$ simulated gene expression datasets, we tested the null hypothesis of normality for the expression profiles. For the three Erdös-Rényi networks with edge density $\epsilon=\{0.003,0.006,0.008\}$ containing 150 genes, 74%, 80% and 88% of the genes reject in average the null hypothesis testing for normality. These values are obtained for a significance level of $\alpha=0.001$ and a Bonferroni correction [52]. We repeated the same analysis for a normalized gene expression dataset from *S. cerevisiae* [42] containing 4837 genes and 904 samples to see if our simulated data represent realistic aspects of biological data. Testing each of the 4837 genes for normality leads to a rejection of 35% of the tests ($\alpha=0.001$, Bonferroni corrected). This demonstrates, first, that the characteristics of our simulated data is comparable to that of biological data and, second, that there is a non-negligible fraction of genes whose expression is not normally distributed, even after appropriate normalization of the data.

Next, we investigate the relation between the occurrences of TP and FP edges and the normality of the gene expression values. From testing the null-hypothesis that a gene expression profile follows a normal distribution, using the Anderson-Darling normality test, we obtain p-values for all genes. Combining pairs of p-values with Stouffer's method [53] gives us a p-value we assign to all gene pairs which correspond to edges and non-edges in the reference network. More precisely, we test the null hypothesis for normality of the gene expression profile of a gene using the Anderson-Darling test [51], as before. Then for each gene pair, their p-values are combined using the Stouffer method [53]. This method, first, transforms individual p-values into z-scores, $z_i=\Phi^{-1}(1-p_i)$. Here Φ is the cumulative distribution function of the standard normal distribution. After this transformation the resulting z-scores are aggregated into a combined z-score, $z=\sum_i^2 z_i/\sqrt(2)$, from which the combined p-value is obtained. These combined p-values reflect the normality of the genes that enclose an edge (or a non-edge). We estimate these p-values for each of the M datasets. This results in M vectors of length $L_p=11{,}175(=(p^2-p)/2)$ of p-values, which we rank in ascending order. We call these vectors v_i, for $i\in\{1,\ldots,M\}$. From the comparison of the inferred network with the reference network we obtain for each dataset a categorization of these p-values into the four categories TP, FP, TN and FN with respect to correctly/falsely identified edges. Considering TP and FP edges only, we obtain two categories which we use in the following. Due to the fact that, usually, the number of $TP(i)$ edges is not equal to the number of $FP(i)$ edges for a given dataset i, we identify their common length,

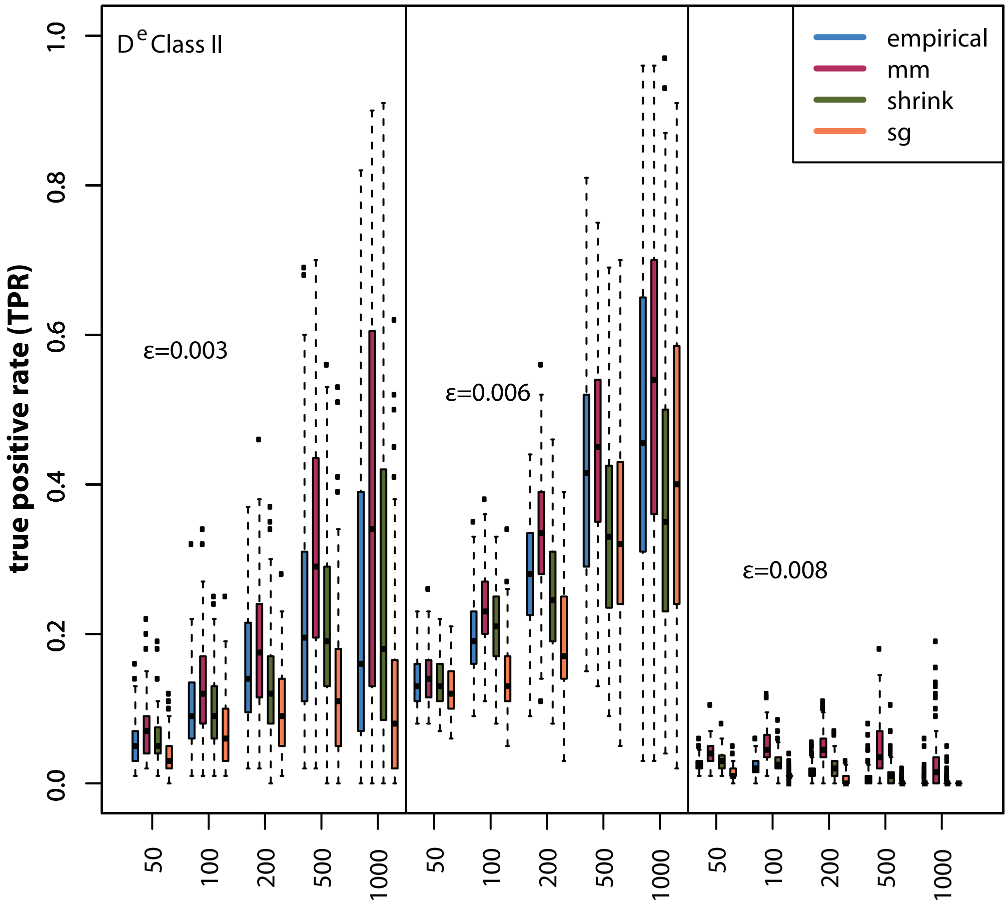

Figure 11. Local network inference performance for Class II edges. Simulated gene expression datasets for Erdös-Rényi networks with edge density ϵ for sample sizes ranging from 50 to 1000 samples.
doi:10.1371/journal.pone.0029279.g011

$$L_i = \text{argmin } (\#(TP(i)),\#(FP(i))), \quad (16)$$

and use exactly L_i true positive and false positive hits from this dataset only. This information is used to define two rank vectors v_t and v_f, each of length L_p, one for TP and one for FP edges, by extracting the ranks for the first L_i TP and FP edges for the i-th dataset from the vector v_i. This results in two L_i dimensional vectors I_i^t and I_i^f containing the ranks of the TP and FP edges in v_i. Starting from $v_t(k)=0$ and $v_f(k)=0$ for $k\in\{1,\ldots,L_p\}$, I_i^t and I_i^f are used to update these vectors for every dataset by $v_t(I_i^t(j))=v_t(I_i^t(j))+1$ and $v_f(I_i^f(j))=v_f(I_i^f(j))+1$ for $j\in\{1,\ldots,L_i\}$. Repeating this procedure for all M datasets provides us with two vectors, v_t and v_f, whose components reflect the frequency with which they occurred in all datasets. For example, $v_t(17)=6$ would mean that the edge ranked at position 17 appeared 6 times in the set of TP edges among the first L_i hits in the M datasets that were considered. The interpretation for v_f for the FP edges is analogously. For each set of datasets with sample size 200 of the Erdös-Rényi network with $\epsilon=\{0.003,0.006,0.008\}$, we compare the ranks of the p-values between TP edges, denoted by v_t, and FP edges denoted by v_f.

In Figure 12 we show the empirical cumulative distribution functions (ecdf) for v_t and v_f calculated for a Erdös-Rényi network with $\epsilon=0.003$. One can see that the TP edges are more likely to occupy lower ranked p-values compared to the FP edges. Hence, the distributions corresponding to TP edges have the tendency to show a stronger deviation from normality. In contrast, the curves of the FP edges follow a straight line which indicates the absence of a systematic relation between FP edges and the normality of the data. For the networks with $\epsilon=\{0.006,0.008\}$ we obtain qualitatively similar results.

Finally, we generalize the above analysis by testing for general probability distributions, to see if there is a systematic relation between the occurrences of TP and FP edges and the distribution of the gene expressions. For this analysis we use a Kolmogorov-Smirnov test to compare the gene expression distributions for all edge and non-edge gene pairs. For a significance level of $\alpha=0.05$ (Bonferroni corrected) the equality of the expression distribution was rejected for 39.9%, 59.8% and 76.5% ($\epsilon=0.003,0.006,0.008$) of all gene pairs. For reasons of comparison, we performed a similar analysis for the gene expression dataset from yeast [42] and found that 96% of the tests were rejected. Then we compared the ranks obtained from the Kolmogorov-Smirnov tests between TP and FP edges, as explained above for the normality test. Again, this procedures was performed for the Empirical, Miller-Madow, Shrinkage and Schürmann-Grassberger estimator for each expression dataset with sample size 200 of the Erdös Rényi networks with $\epsilon=\{0.003,0.006,0.008\}$. Similar to the results shown in Figure 12, we find that gene pairs of TP edges are likely

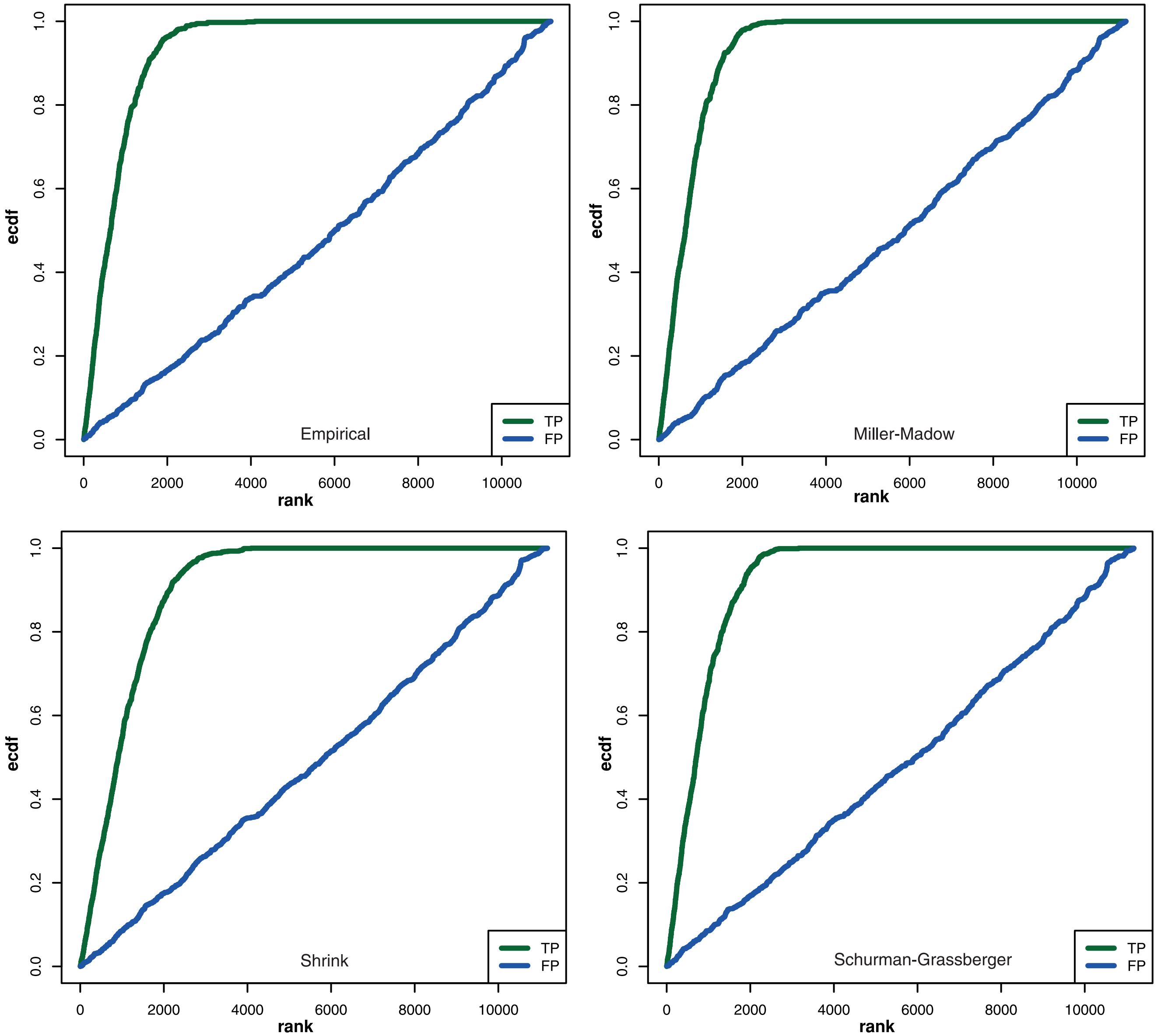

Figure 12. Empirical cumulative distribution functions (ecdf) for the four MI estimators. The green curves correspond to TP edges and the blue curves to FP edges as a function of their rank.
doi:10.1371/journal.pone.0029279.g012

to show a more prominent difference between their underlying expression profile (not shown). Hence, this observation is in agreement with the rank comparisons of TP and FP edges obtained from the Anderson-Darling normality test, and demonstrate that TP and FP edges behave quantitatively different, independent of distributional assumptions.

In Fig. 13 we summarize our findings about the data heterogeneity graphically. We found that for a given gene regulatory network one can find a multitude of different joint probability distributions, visualized by the different edge colors. Similarly, for gene pairs that are not directly connected by an edge (non-edges) one can also find many different probability distributions. From this one can obtain (discrete) probability distributions of the occurrence of probability distributions, visualized in the bottom part of Fig. 13. In these two diagrams, each color bar represents one specific probability distribution that can be found for the edges or non-edges. These discrete probability distributions P_e and P_n can even be different from each other.

Mathematically, the distributions P_e and P_n can be used to define the terms data heterogeneity respectively data homogeneity. More precisely, if we would observe $P_e(i) = 1 = P_n(i)$ for one distribution $i = i_0$, and $P_e(i) = 0 = P_n(i)$ for all other distributions ($i \neq i_0$), the underlying data would be homogeneous because they can be described by the probability distribution i. For all other distributions of P_e and P_n the data show, at least to some degree, a heterogeneity. It is easy to see that the case

$$P_e(i) = P_n(i) = \begin{cases} 1, & \text{for } i = i_0 \\ 0, & \text{for all other } i \end{cases} \qquad (17)$$

corresponds to the conventional (implicit) assumption made when MI estimators are studied in isolation.

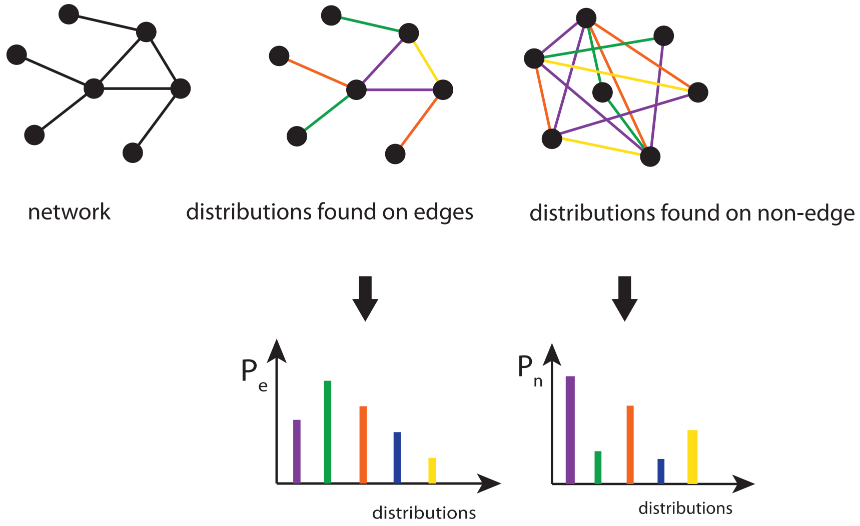

Figure 13. Schematic summary of the effect of data heterogeneity. Top row: Shown is a gene regulatory network (left) and the occurrence of different probability distributions observed on edges (middle) and non-edges (right). Bottom row: From the occurrence frequency of different probability distributions the two discrete distributions P_e and P_n are obtained.
doi:10.1371/journal.pone.0029279.g013

Discussion

In this study we presented a comprehensive investigation of the influence the MI estimators have on the inference performance of C3NET. We observed a strong influence of the MI estimators and the discretization methods on the inference performance, revealed by global and local network-based error measures. In summary, we found that the Miller-Madow estimator in combination with the *equal width* and the *global equal width* discretization methods provided the best performance. However, the major influence on the C3NET inference performance was observed for the discretization method itself, whereas the *equal width* and the *global equal width* lead to significantly better results than the *equal frequency* discretization method. Hence, using the *equal frequency* discretization in applications is likely to lead to a reduced performance of C3NET because the inference performance is prominently lower compared to the other two discretization methods. A potential explanation why the Miller-Madow estimator performs better than the other estimators is that the Miller-Madow estimator is the only estimator among the used MI estimators that considers a bias correction, which depends on the sample size and the number of bins that are introduced by the discretization.

In [31] it was shown that the influence of the discretization method and a MI estimator is method specific, e.g., for ARACNE, CLR and MRNET. In their study, MRNET and ARACNE performed best with the *equal frequency* discretization in combination with the Empirical or Miller-Madow estimator while CLR was observed to be less influenced by different estimators. In contrast to our results for C3NET, the *equal width* discretization method was observed to outperform the *equal frequency* discretization method if used in combination with the Miller-Madow estimator. A major factor that explains the differences of the discretization methods on the network inference performance of our study compared to the study performed by [31] are likely due to the different working mechanisms and characteristics of the statistical principles, employed by the different network inference methods. For this reason it is necessary to identify the optimal combination of a statistical MI estimator and a discretization method for each network inference algorithm individually, as pointed out in [31].

Interestingly, the influences of the different MI estimators found in [31] are less pronounced than the one we found for the C3NET algorithm. A reason for this may be that the ensemble data used in [31] consist of only 10 expression datasets for each setting. A large ensemble size of 100 datasets, as used in our study, allows to capture finer variations among the simulated datasets and thus allows more robust comparisons.

Another factor that could lead to differences, is the simulation strategy used for the simulation of gene expression data. For example, the effect of noise and missing data were studied in [31] and it was shown that this is an important influence that needs to be taken into account when assessing the inference performance. However, future studies are necessary to investigate this influence in more detail and also to provide guidance with respect to the selection of a simulation setting.

Global error measures make the implicit assumption to observe an approximately equal inference performance for all edges in the network. However, for C3NET it was shown that edges of linearly connected nodes, e.g., edges of leaf nodes, have a higher inference performance than edges of highly connected nodes [21]. Hence, it is likely that edges of highly connected genes are more difficult to infer due to the more complex expression patterns of the corresponding genes. For this reason, we used in addition to global error measures also local network-based error measures to study the inference performance for edges of linearly connected genes (Class I) and all other edges (Class II). We found that edges from genes with a high degree (Class II) are likely to be underrepresented in the inferred networks because they have a lower (median) true positive rate than edges from Class I. Further, we compared the influence of the MI estimator on the edge classes Class I and Class II. Among the tested combinations of discretization methods and estimators, the Miller-Madow estimator with *equal width* or *global equal width* discretization showed the best performance on the inference performance of C3NET, independent from the edge density and the sample size.

In order to obtain the above results we simulated gene expression data from underlying regulatory networks, instead of making a distributional assumption to sample data from such a distribution. The latter is the traditional approach to study MI estimators [32–34]. Despite its popularity and simplicity we demonstrated that in the context of the inference of GRNs this reductionistic approach is not appropriate because of the data heterogeneity. Specifically, we showed that one can find a multitude of different probability distributions in simulated as well as biological expression data which can be represented as a (discrete) distribution of probability distributions, P_e, for edges and, P_n, for non-edges. Hence, assuming the presence of merely one probability distribution is not supported by data. For this reason, if one would like to study MI estimators in isolation, one would need to make assumptions not only about the usage of one probability distribution, but of the distribution of probability distributions (P_e, for edges and, P_n for non-edges). In order to avoid this complication, we recommend to simulate expression data from an underlying gene regulatory network because this provides naturally such a distribution of probability distributions.

In summary, we studied the influence of discrete MI estimators and discretization methods on the inference performance of C3NET and provided suggestions for the most beneficial combination. However, our study may be also useful for the development of novel MI estimators that take the various underlying probability distributions for different edge classes into consideration. Future studies are required to evaluate the vast catalogue of existing and novel MI estimators, the impact of different network structures and the simulation procedures in order to explore the particular factors that are required to understand the influence of different MI estimators on individual inference algorithms to enable an efficient analysis of real biological gene expression datasets.

Acknowledgments

We would like to thank Shailesh Tripathi for fruitful discussions. We also thank the reviewers for their comments and suggestions on the manuscript. We used *minet* [54], *Hmisc* [55], *igraph* [56] and R [57] for our numerical analysis.

Author Contributions

Conceived and designed the experiments: FES. Performed the experiments: RDMS FES. Analyzed the data: RDMS FES. Contributed reagents/materials/analysis tools: RDMS FES. Wrote the paper: RDMS FES.

References

1. Shannon C, Weaver W (1949) The Mathematical Theory of Communication University of Illinois Press.
2. Li W (1990) Mutual information functions versus correlation functions. Journal of Statistical Physics 60: 823–837.
3. Paninski L (2003) Estimation of entropy and mutual information. Neural Computation 15: 1191–1253.
4. Miller G (1955) Note on the bias of information estimates. Information Theory in Psychology: Problems and Methods. pp 95–100.
5. Schäfer J, Strimmer K (2005) A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. Statistical Applications in Genetics and Molecular Biology 4: 32.
6. Schurmann T, Grassberger P (1996) Entropy estimation of symbol sequences. Chaos 6: 414–427.
7. Alon U (2006) An Introduction to Systems Biology: Design Principles of Biological Circuits. Boca RatonFL: Chapman & Hall/CRC.
8. Dehmer M, Emmert-Streib F, Graber A, Salvador A, eds. Applied Statistics for Network Biology: Methods for Systems Biology Wiley-Blackwell.
9. Palsson B (2006) Systems Biology. Cambridge; New York: Cambridge University Press.
10. Vidal M (2009) A unifying view of 21st century systems biology. FEBS Letters 583: 3891–3894.
11. Liang S, Fuhrman S, Somogyi R (1998) Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In: Pac Symp Biocomput. pp 18–29.
12. Liang K, Wang X (2008) Gene regulatory network reconstruction using conditional mutual information. EURASIP J Bioinform Syst Biol 2008: 253894.
13. Luo W, Hankenson K, Woolf P (2008) Learning transcriptional regulatory networks from high throughput gene expression data using continuous three-way mutual information. BMC Bioinformatics 9: 467.
14. Stolovitzky G, Califano A, eds. Reverse Engineering Biological Networks: Opportunities and Challenges in Computational Methods for Pathway Inference Wiley-Blackwell.
15. Stolovitzky G, Prill R, Califano A (2009) Lessons from the DREAM 2 Challenges. Ann N Y Acad Sci 1158: 159–195.
16. Watkinson J, Liang K, Wang X, Zheng T, Anastassiou D (2009) Inference of regulatory gene interactions from expression data using three-way mutual information. Ann N Y Acad Sci 1158: 302–13.
17. Butte A, Kohane I (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput 2000: 418–29.
18. Pearl J (2000) Causality: Models, Reasoning, and Inference. Cambridge; New York: Cambridge.
19. Spirtes P, Glymour C, Scheines R (1993) Causation, Prediction, and Search. New York: Springer.
20. Verma T, Pearl J (1990) Causal networks: semantics and expressiveness. In: Proceedings of the 4th workshop on uncertainly in artificial intelligence. Mountain View CA. pp 352–359.
21. Altay G, Emmert-Streib F (2010) Inferring the conservative causal core of gene regulatory networks. BMC Syst Biol 4: 132.
22. Bulashevska S, Eils R (2005) Inferring genetic regulatory logic from expression data. Bioinformatics 21: 2706–2713.
23. Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian Network to Analyze Expression Data. Journal of Computational Biology 7: 601–620.
24. Husmeier D (2003) Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. Bioinformatics 19: 2271–82.
25. Margolin A, Nemenman I, Basso K, Wiggins C, Stolovitzky G, et al. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7: S7.
26. Xing B, van der Laan M (2005) A causal inference approach for constructing transcriptional regulatory networks. Bioinformatics 21: 4007–4013.
27. Emmert-Streib F, Glazko G (2011) Network biology: A direct approach to study biological function. Wiley Interdiscip Rev Syst Biol Med 3: 379–391.
28. Bolstad B, Irizarry R, Astrand M, Speed T (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185–193.
29. Irizarry RA, Wu Z, Jaffee HA (2006) Comparison of Affymetrix GeneChip expression measures. Bioinformatics 22: 789–794.
30. Steinhoff C, Vingron M (2006) Normalization and quantification of differential expression in gene expression microarrays. Brief Bioinform 7: 166–177.
31. Olsen C, Meyer P, Bontempi G (2009) On the impact of entropy estimation on transcriptional regulatory network inference based on mutual information. EURASIP J Bioinform Syst Biol 2009(1): 308959.
32. Hausser J, Strimmer K (2009) Entropy Inference and the James-Stein Estimator, with Application to Nonlinear Gene Association Networks. Journal of Machine Learning Research 10: 1469–1484.
33. Khan S, Bandyopadhyay S, Ganguly A, Saigal S, Erickson D, et al. (2007) Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. Physical Review E 76: 026209.
34. Kraskov A, Stögbauer H, Grassberger P (2004) Estimating mutual information. Phys Rev E 69: 066138.
35. Nemenman I, Shafee F, Bialek W (2002) Entropy and inference, revisited. In: Dietterich TG, Becker S, Ghahramani Z, eds. Advances in Neural Information Processing Systems 14. CambridgeMA: MIT Press. pp 569–595.
36. Cover T, Thomas J (1991) Elements of Information Theory John Wiley & Sons, Inc.
37. Almudevar A, Klebanov L, Qiu X, Salzman P, Yakovlev A (2006) Utility of correlation measures in analysis of gene expression. NeuroRx 3: 384–395.
38. Emmert-Streib F, Altay G (2010) Local network-based measures to assess the inferability of different regulatory networks. IET Syst Biol 4: 277–88.
39. von Bertalanffy L (1950) The theory of open systems in physics and biology. Science 111: 23–29.
40. von Bertalanffy L (1950) An Outline of General Systems Theory. British Journal for the Philosophy of Science 1: 134–165.
41. Butte A, Kohane I (2000) Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. Pacific Symposioum on Biocomputing 5: 415–426.
42. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, et al. (2007) Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles. PLoS Biol 5.
43. Yang Y, Webb GI (2001) Proportional k-Interval Discretization for Naive-Bayes Classifiers. Proceeding EMCL '01 Proceedings of the 12th European Conference on Machine Learning.
44. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognition Letters 27: 861–874.

45. Kent A, Berry MM, Luehrs Jr. FU, Perry JW (1955) Machine literature searching VIII. Operational criteria for designing information retrieval systems. American Documentation 6: 93–101.
46. Altay G, Emmert-Streib F (2011) Structural influence of gene networks on their inference: analysis of C3NET. Biol Direct 6: 31.
47. Gilbert EN (1959) Random graphs. Annals of Mathematical Statistics 20: 1141–1144.
48. Leclerc R (2008) Survival of the sparsest: robust gene networks are parsimonious. Mol Syst Biol 4: 213.
49. Erdös P, Rényi A (1959) On random graphs. I Publicationes Mathematicae 6: 290–297.
50. Van den Bulcke T, Van Leemput K, Naudts B, van Remortel P, Ma H, et al. (2006) SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. BMC Bioinformatics 7: 43.
51. Sheskin DJ (2004) Handbook of Parametric and Nonparametric Statistical Procedures Chapman and Hall/CRC, 3rd edition.
52. Dudoit S, van der Laan M (2007) Multiple Testing Procedures with Applications to Genomics. New York; London: Springer.
53. Stouffer S, Suchman E, DeVinnery L, Star S, Williams R (1949) The American Soldier, Vol.1: Adjustment during Army Life. Princeton: Princeton University Press.
54. Meyer P, Lafitte F, Bontempi G (2008) minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. BMC Bioinformatics 9: 461.
55. Harrell F (2010) Hmisc: Harrell Miscellaneous. R package version 3.8–3.
56. Csardi G, Nepusz T (2006) The igraph software package for complex network research. Inter Journal Complex Systems. 1695.
57. R Development Core Team (2008) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.