Bagging statistical network inference from large-scale gene expression data.

de Matos Simoes, R., & Emmert-Streib, F. (2012). Bagging statistical network inference from large-scale gene expression data. PLoS ONE, 7(3), 1-11. [e33624]. DOI: 10.1371/journal.pone.0033624

Published in:
PLoS ONE

Document Version:
Publisher's PDF, also known as Version of record

OPEN ACCESS Freely available online

PLoS one

Bagging Statistical Network Inference from Large-Scale Gene Expression Data

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

Modern biology and medicine aim at hunting molecular and cellular causes of biological functions and diseases. Gene regulatory networks (GRN) inferred from gene expression data are considered an important aid for this research by providing a map of molecular interactions. Hence, GRNs have the potential enabling and enhancing basic as well as applied research in the life sciences. In this paper, we introduce a new method called BC3NET for inferring causal gene regulatory networks from large-scale gene expression data. BC3NET is an ensemble method that is based on *bagging* the C3NET algorithm, which means it corresponds to a Bayesian approach with noninformative priors. In this study we demonstrate for a variety of simulated and biological gene expression data from *S. cerevisiae* that BC3NET is an important enhancement over other inference methods that is capable of capturing biochemical interactions from transcription regulation and protein-protein interaction sensibly. An implementation of BC3NET is freely available as an R package from the CRAN repository.

Citation: de Matos Simoes R, Emmert-Streib F (2012) Bagging Statistical Network Inference from Large-Scale Gene Expression Data. PLoS ONE 7(3): e33624. doi:10.1371/journal.pone.0033624

Editor: Matteo Pellegrini, UCLA-DOE Institute for Genomics and Proteomics, United States of America

Received October 24, 2011; **Accepted** February 14, 2012; **Published** March 30, 2012

Funding: This project is partly supported by the Department for Employment and Learning through its "Strengthening the all-Island Research Base" initiative and the Engineering and Physical Sciences Research Council (EPSRC). 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

Gene networks represent the *blueprint* of the causal interplay between genes and their products on all molecular levels [1–6]. Gene regulatory networks (GRN) inferred from large-scale gene expression data aim to represent signals from these different levels of the gene network. The inference, analysis and interpretation of a GRN is a daunting task due to the fact that the concentrations of mRNAs provide only indirect information about interactions occurring between genes and their gene products (e.g., protein interactions). The reason for this is that DNA microarrays measure only the concentration of mRNAs rather than the binding, e.g., between proteins or between a transcription factor and the DNA. Despite the increased community effort in recent years [7,8] and a considerable number of suggested inference methods [9–19] there is an urgent need to further advance our current methods to provide reliable and efficient procedures for analyzing the increasing amount of data from biological, biomedical and clinical studies [20–22]. For this reason, this field is currently vastly expanding. A detailed review for many of the most widely used methods can be found in [15,18,23–26].

A major problem for the inference of regulatory networks are the intricate characteristics of gene expression data. These data are high-dimensional, in the order of the genome size of the studied organism, and nonlinear due to the intertwined connection of the underlying complex regulatory machinery including the multilevel regulation structures (DNA, mRNA, protein, protein complexes, pathways) and turnover rates of the measured mRNAs, products and proteins. Further, gene expression data for network inference are large-scale, although, the "Large p Small n" [27] problem holds, because the number of explanatory variables (p genes) exceeds the number of observations (n microarray samples). In addition, technical noise and outliers can make it difficult to gain access to the true biological signal of the expression measurement itself.

The main contribution of this paper is to introduce a new network inference method for gene expression data. The principle idea of our method is based on *bootstrap aggregation* [28,29], briefly called *bagging*, in order to create an ensemble version of the network inference method C3NET [9]. For this reason we call our new method bagging C3NET (BC3NET). The underlying procedure of BC3NET is to generate an ensemble of bootstrap datasets from which an ensemble of networks is inferred by using C3NET. Then the obtained inferred networks are aggregated resulting in the final network. For the last step we employ statistical hypotheses tests removing the need to select a threshold parameter manually. Instead, a significance level with a clear statistical interpretation needs to be selected. This is in contrast with other studies, e.g., [30].

Given the challenging properties of gene expression data, briefly outlined above, BC3NET is designed to target these in the following way. First, BC3NET is based on statistical estimators for mutual information values capable of capturing nonlinearities in the data. Second, in order to cope with noise and outliers in expression data, we employ bagging because it has the desirable ability to reduce the variance of estimates [28]. Computationally, this introduces an additional burden, and a necessary prerequisite for any method to be used in combination with bagging is its tractability to be applicable to a bootstrap ensemble. C3NET is computationally efficient to enable this, even for high-dimensional massive data.

There are a few network inference methods that are similar to BC3NET. The method GENIE3, which was best performer in the DREAM4 *In Silico Multifactorial challenge* [31], employs also an ensemble approach, however, in combination with regression trees, e.g., in form of *Random Forests* [32]. In [30] a bootstrap approach has been used in combination with Bayesian networks to estimate confidence levels for features. However, we want to emphasize that, in contrast to BC3NET, both methods [30,31] do not provide a statistical procedure for determining an optimal confidence threshold parameter. Finally, we note that also for ARACNe a bootstrap version has been introduced [33], which has so far been used for inferring subnetworks around selected transcription factors [34,35].

Methods

The BC3NET approach for GRN inference

In general, mutual information based gene regulatory network inference methods consists of three major steps. In the first step, a mutual information matrix is obtained based on mutual information estimates for all possible gene pairs in a gene expression data set. In the second step, a hypothesis test is performed for each mutual information value estimate. Finally, in the third step, a gene regulatory network is inferred from the significant mutual information values, according to a method specific procedure.

The basic idea of BC3NET is to generate from one dataset $D(s)$, consisting of s samples, an ensemble of B independent bootstrap datasets $\{D_k^b\}_{k=1}^B$ by sampling from $D(s)$ with replacement by using a non-parametric bootstrap [36] with $B=1000$. Then, for each generated data set D_k^b in the ensemble, a network G_k^b is inferred by using C3NET [9]. From the ensemble of networks $\{G_k^b\}_{k=1}^B$ we construct one weighted network

$$G_w^b \xleftarrow{\text{aggregate}} \{G_k^b\}_{k=1}^B \tag{1}$$

which is used to determine the statistical significance of the connection between gene pairs. This results in the final binary, undirected network G. Fig. 1 shows a schematic visualization of this procedure.

A base component of BC3NET is the inference method C3NET introduced in [9], which we present in the following in a modified form to obtain a more efficient implementation. Briefly, C3NET consists of three main steps. First, mutual information values among all gene pairs are estimated. Second, an extremal selection strategy is applied allowing each of the p genes in a given dataset to contribute *at most* one edge to the inferred network. That means we need to test only p different hypotheses and not $p(p-1)/2$. This potential edge corresponds to the hypothesis test that needs to be conducted for each of the p genes. Third, a multiple testing procedure is applied to control the type one error. In the above described context, this results in a network G_k^b.

In order to test the statistical significance of the connection between gene pairs BC3NET utilizes the edge weights of the aggregated network G_w^b as test statistics. The edge weights of G_w^b are componentwise defined by

$$G_w^b(i,j)=\sum_{k=1}^{B} I_1(G_k^b(i,j))=\#\{G_k^b(i,j)=1|\{G_k^b\}_{k=1}^B\}. \tag{2}$$

Here $I()$ is the indicator function which is 1 if its argument is $G_k^b(i,j)=1$ and 0 otherwise. This expression corresponds to the number of networks in $\{G_k^b\}_{k=1}^B$ which have an edge between gene i and j. For brevity, we write in the following $n_{ij}=G_w^b(i,j)$. From Eqn. 2 follows that n_{ij} assumes integer values in $\{0,\ldots,B\}$. Based on the test statistic n_{ij}, we formulate the following null hypothesis which we test for each gene pair (i,j).

$H_0^{n_{ij}}$: The number of networks n_{ij} in the ensemble $\{G_k^b\}_{k=1}^B$ with an edge between gene i and j is less than $n_0(\alpha)$.

Here the cut-off value n_0 depends on the significance level α. Due to the independence of the bootstrap datasets we assume the null distribution of n_{ij} to follow a binomially distributed $Bin(B,p_c)$, whereas B corresponds to the size of the bootstrap ensemble and p_c is the probability that two genes are connected by chance. The parameter p_c relates to a population of networks, estimated from randomized data by using BC3NET, and corresponds to the fraction of randomly inferred edges in the bootstrap population ($\mathbb{E}[E_b(B,D(s))]$) divided by the total number of possible edges in this population ($E_t(B)$) that means

$$p_c=\frac{\mathbb{E}[E_b(B,D(s))]}{E_t(B)}. \tag{3}$$

The maximal number of gene pairs that can be formed from p genes in B bootstrap datasets is given by

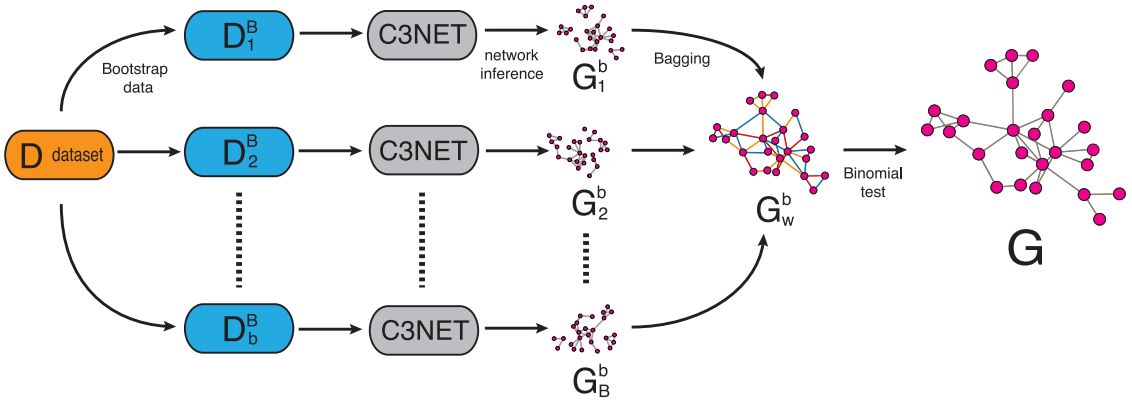

Figure 1. BC3NET algorithm: The gene regulatory network G is inferred from a bootstrap ensemble generated from a single gene expression dataset D. For each generated dataset in the ensemble, D_k^b, a network, G_k^b, is inferred using C3NET. From $\{G_k^b\}_{k=1}^B$ an aggregated network G_w^b is obtained whose edges are used as test statistics to obtain the final network G.
doi:10.1371/journal.pone.0033624.g001

$$E_t(B) = \frac{p(p-1)}{2} \cdot B \tag{4}$$

This value is independent of the sample size. $\mathbb{E}[E_b(B,D(s))]$ corresponds to the expectation value of the number of randomly inferred edges for a population of an ensemble of bootstrap datasets of size B. Because $E_b(B,D(s))$ is a random variable it is necessary to average over all possible bootstrap datasets of size B with sample size s. On a theoretical note we remark that these bootstrap datasets constitute a population that specifies a probability mass function (pmf) for which the expectation of $E_b(B,D(s))$ needs to be evaluated. Due to the fact that this pmf is unknown the value of $\mathbb{E}[E_b(B,D(s))]$ needs to be estimated.

In order to estimate $\mathbb{E}[E_b(B,D(s))]$ we randomize the data to estimate the number of edges randomly inferred in an bootstrap ensemble of size B, $\{\tilde{G}_k^b\}_{k=1}^B$

$$E_b = \#\text{edges randomly inferred in } \{\tilde{G}_k^b\}_{k=1}^B \tag{5}$$

Using $E_b \approx \mathbb{E}[E_b(B,D(s))]$ as plug-in estimator for Eqn. 3 we obtain an estimate for p_c. This allows us to calculate a p-value for each gene pair (i,j) and a given test statistic n_{ij}, given by Eqn. 2, from the null distribution of n_{ij} by

$$p(i,j) = Pr(n \geq n_{ij}) = \sum_{n=n_{ij}}^{B} \binom{B}{n} p_c^n (1-p_c)^{B-n} \tag{6}$$

Here $p(i,j)$ is the probability to observe n_{ij} or more edges by chance in a bootstrap ensemble of size B and sample size s.

Because we need to test $p(1-p)/2$ hypotheses simultaneously (one for each gene pair) we need to apply a multiple testing correction (MTC) [37,38]. For our analysis we are using a Bonferroni procedure for a strong control of the family-wise error rate (FWER). Typically, procedures controlling the FWER are more conservative than procedures controlling, e.g., the false discovery rate (FDR) by making only mild assumptions about the underlying data [39,40]. Based on these hypotheses tests the final network G is componentwise defined by

$$G(i,j) = \begin{cases} 1 & \text{if } p(i,j) \leq \alpha; \\ 0 & \text{otherwise}. \end{cases} \tag{7}$$

That means if the connection between a gene pair is statistically significant they are connected by and edge, otherwise there is no connection.

Null-distribution of mutual information values

In order to determine the statistical significance of the mutual information values between genes we test for each pair of genes the following null hypothesis.

H_0^I : The mutual information between gene i and j is zero.

Because we are using a nonparametric test we need to obtain the corresponding null distribution for H_0^I from a randomization of the data. Principally, there are several ways to perform such a randomization which conform with the formulated null hypothesis. For this reason, we perform 3 different randomizations and compare the obtained results with respect to the performance of the inference method to select the most appropriate one. Two randomization schemes (RM1 and RM2) permute the expression profiles for *each gene pair* separately. RM1 permutes *only* the sample labels and RM2 permutes the sample *and* the gene labels. In contrast, the randomization scheme RM3 permutes the sample and gene labels *for all genes* of the entire expression matrix at once.

Mutual Information Estimators

Due to the expected nonlinearities in the data we use mutual information estimators to assess the similarity between gene profiles instead of correlation coefficients. In a previous study, we found that for normalized microarray data the distribution among individual gene pairs can strongly deviate from a normal distribution [41]. This makes it challenging to judge by theoretical considerations only which statistical estimator is most appropriate for gene expression data because most estimators were designed assuming normal data. For this reason we compare eight different estimators and investigate their influence on the performance of C3NET.

Mutual Information is frequently estimated from the marginal and joint entropy H of two discretized random variables X and Y [42],

$$I(X,Y) = H(X) + H(Y) - H(X,Y). \tag{8}$$

In our study, we use four MI estimators based on continuous data and four MI estimators based on discretized data. The MI estimators for discretized data are the empirical estimator [42], Miller-Madow [42], shrinkage [43] and the Schürmann-Grassberger [44] mutual information estimator. For the emipirical estimator, the entropy H^{emp} is estimated from the observed cell frequencies for each bin k of a random variable discretized into p bins, i.e.,

$$H^{emp} = -\sum_{k=1}^{p} n_k \log(n_k). \tag{9}$$

With an increasing number of bins, the empirical estimator underestimates the true entropy H due to undersampling of the cell frequencies n_k. The different estimators attemp to adjust the undersampling bias by a constant factor [42], estimate cell frequencies by a shrinkage function between two models [43] or add a pseudo count from a probability distribution to the cell frequencies [44].

Mutual information can also be estimated from continuous random variables. The B-spline estimator considers the bias induced by the discretization for values falling close to the boundaries of a bin. For each bin, weights are estimated for the corresponding values from overlapping polynomial B-spline functions [45]. Hence, this method allows to map values to more than one bin.

For normal data, there is an analytical correspondence between a correlation coefficient and the mutual information [25],

$$I(X_i,X_j) = -\frac{1}{2}\log(1-\rho^2). \tag{10}$$

In this equation, the coefficient ρ could be the Pearson correlation coefficient ρ, Spearman rank correlation coefficient ρ or the Kendal rank correlation coefficient τ.

Yeast gene expression data

We use the *S. cerevisiae* Affymetrix ygs98 RMA normalized gene expression compendium available from the Many Microbe Microarrays Database M3D [46]. The yeast compendium dataset comprises 9335 probesets and 904 samples from experimental and

observational data from anaerobic and aerobic growth conditions, gene knockout and drug perturbation experiments. We map the yeast affymetrix probeset IDs to gene symbols using the annotation of the *ygs98.db* Bioconductor package. Multiple probesets for the same gene are summarized by the median expression value. The resulting expression matrix comprises a total of 9163 features for 4837 gene symbols and 4326 probesets that cannot be assigned to a gene symbol.

Simulated gene expression data

We simulate a variety of different gene expression datasets for Erdös-Rényi networks [47] with an edge density of $=\{0.003,0.006,0.008,0.010\}$. An Erdös-Rényi network is generated by starting with n unconnected vertices. Then, between each vertex pair an edge is included with a pre-selected probability. The generated networks contain 150 genes of which $\{60,22,19,10\}$ genes are unconnected. For each network, simulated gene expression datasets were created for various sample sizes of $\{50,100,200,500,1000\}$ by using Syntren [48] including biological noise. We generate also simulated gene expression datasets for different subnetworks from the *E.coli* transcriptional regulatory network obtained from RegulonDB. The giant connected component (GCC) of the transcriptional regulatory network of *E.coli* consists of 1192 genes. We sample seven connected subnetworks from the GCC of sizes $\{50,100,150,200,300,400,500\}$. Again, using Syntren we simulate 100 different expression datasets including biological noise with sample size $n=50$ for each of these seven networks.

Gene pair enrichment analysis (GPEA)

To test the enrichment of GO-terms in the inferred yeast BC3NET network we adopt a hypergeometric test (one-sided Fisher exact test) for edges (gene pairs) instead of genes in the following way. For p genes there is a total of $N=p(p-1)/2$ different gene pairs. If there are p_{GO} genes for a given GO-term then the total number of gene pairs is $m=p_{GO}(p_{GO}-1)/2$. Suppose the inferred yeast BC3NET network contains n edges of which k are among genes from the given GO-term, then a p-value for the enrichment of this GO-term can be calculated from a hypergeometric distribution by

$$p=\sum_{i=k}^{m} P(X=i)=\sum_{i=k}^{m}\frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}} \qquad (11)$$

Here the p-value estimates the probability to observe k or more edges between genes from the given GO-term. For all GO:0032991 (macromolecular complex) offspring terms from Cellular Component that correspond to protein complexes, the above null hypothesis reflects the expected connection in a protein complex which is a clique (fully connected). For all other GO categories that we test, e.g., from the category Biological Process, the above is a very conservative assumption.

Results

Influence of the randomization and MTC

The influence of the randomization scheme on the performance of BC3NET is shown in Fig. 2. Here we use simulated data from a Erdös-Rényi network consisting of 150 genes, of which 60 are unconnected. The figure shows results for RM1-RM3 with and without MTC for five different sample sizes, shown in the legend of the figure. As one can see, all three randomization schemes with a Bonferroni correction perform similarly good. Also RM1-RM3 without MTC perform similarly, however, significantly worse indicating the importance to correct for multiple hypotheses testing. Due to the fact that RM3 is from a computational point of view more efficient than RM1 or RM2 we use this randomization scheme for our following investigations.

Fig. 2 includes also the F-scores obtained from the randomization of the expression data itself (right-hand side) to obtain baseline values for a comparison with the results from RM1-RM3. This is interesting because, e.g., in contrast to the AU-ROC [49], the F-score for data containing only noise is not 0.5 as for the AU-ROC. From this perspective, one can see that even the results without MTC are significantly better than expected by chance.

Influence of the mutual information estimator

To study the influence of the statistical estimators of the mutual information values, we use simulated data for several different network topologies. Fig. 3 shows results for eight different estimators and different sample sizes for a Erdös-Rényi network with an edge density of $\varepsilon=0.006$. The three continuous estimators, Pearson, Spearman and Kendall as well as B-spline, perform better for smaller sample sizes. For large sample sizes the empirical, Miller-Madow, shrinkage and Schürmann-Grassberger perform slightly better. We want to note that for different parameters of the Erdös-Rényi network and different network types we obtain qualitatively similar results (not shown). Considering the size of the studied networks we used for our analysis, which contain 150 genes, sample sizes up to 100 lead to a realistic ratio of $n/p \lesssim 0.66$ which one can also find for real microarray data. Larger ratios are currently and the near future hard to achieve. For this reason, we assess the results for smaller sample sizes as more important, due to their increased relevance for practical applications. Based on these results we use for the following studies the B-spline estimator.

Comparative analysis of BC3NET

Computational complexity. In [9] the computational complexity of C3NET has been estimated as $O(n^2)$, where n corresponds to the number of genes. For BC3NET this means that its computational complexity is $O(B\times n^2)$. Here B is the number of bootstraps. In order to provide a practical impression for the meaning of these numbers, we compare the computational complexity between the ARACNe bootstrap network approach, described in [33], and BC3NET. We performed an analysis for a gene expression data set with 5000 genes and 200 samples. The ARACNe algorithm needed 22 hours for a single run that means to analysis one bootstrap data set. This results in a total time of 2200 hours (100×22 hours) for 100 bootstraps, which are about $\sim$92 days. In contrast, the BC3NET algorithm completed this task in only 28 minutes for all 100 bootstraps.

Comparative analysis using simulated data. In order to gain insight into the quality of BC3NET we study it comparatively by contrasting its performance with GENIE3 and C3NET. In Fig. 4 we show results for three different Erdös-Rényi networks each with 150 genes, of which $\{22,19,10\}$ genes are unconnected. The edge density of these networks is $\in\{0.003,0.006,0.008\}$. We use these edge densities because regulatory networks are known to be sparsely connected [50]. The F-score distributions for all studied conditions are larger for BC3NET. We repeated the above simulations for subnetworks from the transcriptional regulatory network of *E. coli* and obtained qualitatively similar results. This demonstrates the robustness of the results with respect to different network types and network parameters.

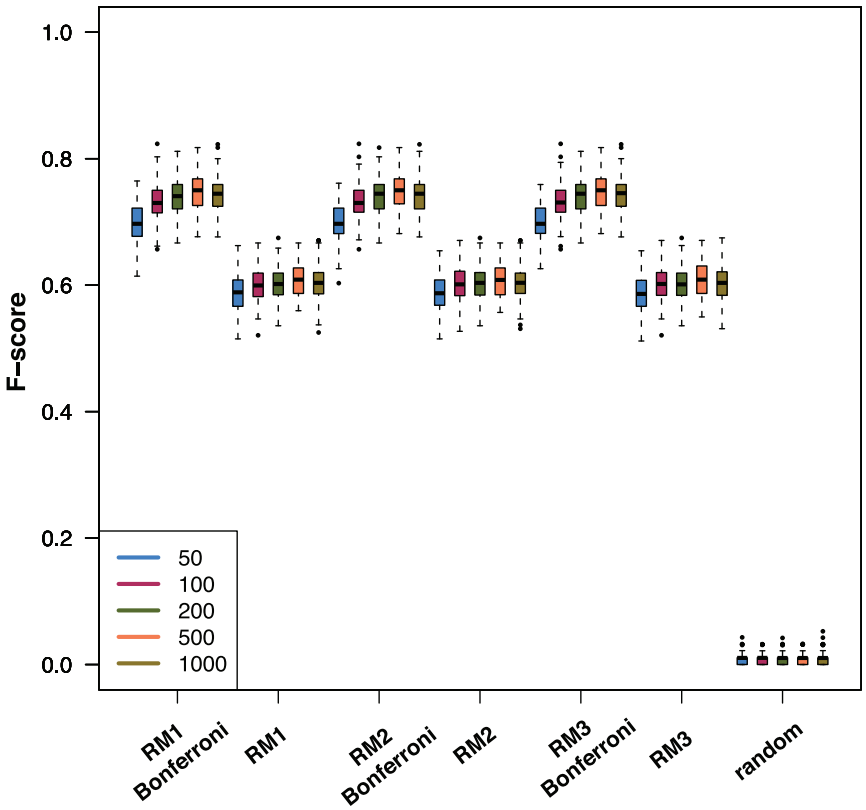

Figure 2. Influence of different randomization schemes (*RM1*, *RM1* and *RM3*) and the multiple hypothesis testing correction on the network inference performance, measured by the F-score. The legend shows the used sample sizes. Each randomization scheme is used with and without a Bonferroni correction. The boxplots labeled 'random' correspond to randomly permuted data to get an impression for random F-scores.
doi:10.1371/journal.pone.0033624.g002

To emphasize the actual gain in the number of true positive edges with respect to C3NET, on which BC3NET is based, we present in Fig. 5 the percentage of the increase of inferred true positive edges for various network sizes ranging from 50 to 500 genes for subnetworks from *E. coli*. For the results shown in the left figure, we use a fixed sample size of $n=50$ and for the right figure the sample size equals the number of genes, i.e., $n=p$. For a fixed sample size ($n=50$) the BC3NET networks show an increase of true positives edges $>45\%$, with a more prominent increase for the larger networks. Quantitatively, this observation is confirmed by a linear regression which gives a none vanishing positive slope of 0.039 and an intercept of 44.24%. Both parameters are highly significant with p-values $\lesssim 10^{-16}$. For the datasets with variable sample sizes ($n=p$) the percentage of inferred true positive edges remains constant with an increasing network size and is around 30%. We want to note that the results for $n=p$ assess the asymptotic behavior of BC3NET because the number of samples n increases linearly with the number of genes p. That means, asymptotically, the gain of BC3NET over C3NET is expected to be $\sim 30\%$. On the other hand, for real data for which $p>n$ holds, the expected gain is much larger, as one can see from the left figure, reaching 70%.

Analysis of the regulatory network of yeast

Using BC3NET, we infer a regulatory network for a large-scale gene expression dataset of *Saccharomyces cerevisiae*. Due to the fact that for *Saccharomyces cerevisiae* no gold standard reference network is available to assess the quality of the inferred GRN we evaluate the resulting network by using functional gene annotations and experimentally validated protein interactions.

The yeast network inferred by BC3NET is a connected network that contains 9,163 genes and 27,493 edges with an edge density of $=0.00065$. The degree distribution of this network follows a power-law distribution, $\sim k^{-\alpha}$, with an exponent of $\alpha=4.38$. We tested a total of 2159 GO-terms from the category Biological Process, whereas each GO-term contains less than 1000 annotated genes. From these, 525 (24.31%) test significant using a Bonferroni procedure indicating an enrichment of gene pairs for the corresponding GO-terms. The strongest enrichment of gene pairs we find in our analysis are for ribosome biogenesis, ncRNA and rRNA processing, mitochondirial organization, metabolic and catabolic processes and cell cycle. See Table 1 for an overview of the top 30 results.

One of the most reliable to detect (biochemical) interaction types that can be experimentally tested and that correspond to *causal* interactions, are protein-protein interactions from protein complexes. The reason therefore is that protein-protein interactions establish a direct connection between the proteins by forming physical bonds. Therefore we study the extend of protein complexes, as defined in the GO database [51], that are present in the yeast BC3NET network. We perform GPEA for 377 GO-terms, which correspond to different protein complexes. From these we identify 94 protein complex terms with significantly enriched gene-pairs. The top 30 GO-terms of protein complexes we find are listed in Table 2. Some of the largest protein complexes detected in the BC3NET network are ribonucleopro-

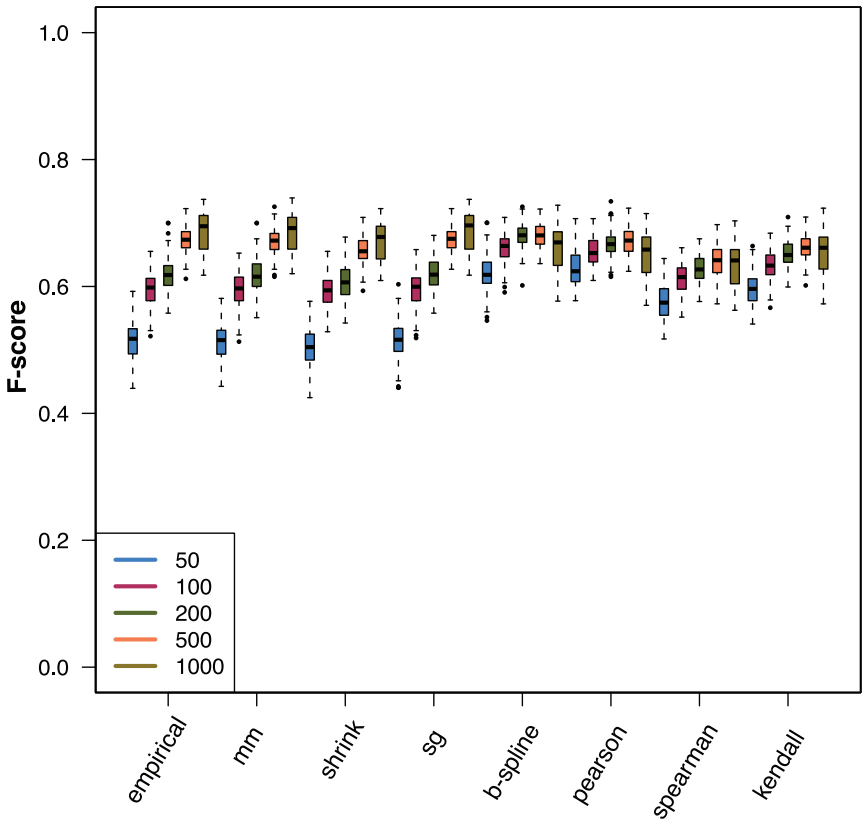

Figure 3. Influence of the statistical mutual information estimators (x-axis) on the network inference performance, measured by the F-score. The legend shows the used sample sizes. Gene expression data were simulated for an Erdös-Rényi network with $=0.006$.
doi:10.1371/journal.pone.0033624.g003

tein complexes (789 edges) including the cytosolic ribosome (315 edges) and mitochondrial ribosome (142 edges). Further protein complexes present in the yeast BC3NET network are the proteasome complex (77 edges), proton-transporting ATP synthase complex (13 edges) and DNA-directed RNA polymerase complex (16 edges).

Finally, we study experimentally evaluated protein-protein interactions extracted from the BioGrid database (release 3.1.77) [52] and compare them with our yeast BC3NET network. First, we find that the yeast PPI network from BioGrid and our yeast BC3NET network have 4,723 genes in common. Further, we find a total of 878 BioGrid interactions among 1,043 genes that are present in the yeast BC3NET network. These interactions are distributed over a total of 282 separate network components, each consisting of 2 or more genes. Among these, we find 11 network components with a significant component size, where the largest significant component includes 147 genes and the smallest significant component includes 9 genes. Significance was identified from gene-label randomized data generating a null distribution for the size of connected network components of the 1,043 genes. The resulting p-values were Bonferroni corrected. For each BioGRID component that is nested in the yeast BC3NET network, we conduct a GO enrichment analysis. From this analysis we use the GO-term with the highest enrichment value to annotate the individual network components, see Table 3.

One of the most extensively studied biological processes in *Saccharomyces cerevisiae* is the cell cycle. For cell cycle the GPEA gives a gene-pair enrichment p-value of $2.6e-55$, see Table 1. In Fig. 6 we show the largest network component of the cell cycle inferred by BC3NET that includes 304 genes and 423 edges. From this network, 57 edges are confirmed in BioGrid (violet edges), 25 edges are from protein complex units (GO) (green edges) and 7 edges are present in both databases (orange edges).

Discussion

From the analysis of BC3NET for the gene expression data set from *S. cerevisiae*, we find in addition to a significant enrichment of over 500 GO-terms in the category Biological Process, the significance of 94 GO-terms in Cellular Component for protein complexes. The largest complexes we identified are the ribosome ($p=1.9e-310$) and proteasome protein complex ($p=1.3-104$). There are two main reasons why edges of these protein complexes are highly abundant in the yeast BC3NET network. First, the ribosome and proteasome protein complexes are well annotated because they have been extensively studied in yeast [53]. Second, the ribosome and proteasome protein complex are mainly regulated on the gene expression level and, where observed, having highly dependent gene expression patterns [53]. Therefore, it is plausible that GRN inference methods can also pick-up signals from physical interactions between protein subunits of protein complexes.

We want to note that we are not the first to recognize that gene expression data contain information about protein-protein interactions. For example, [53,54] provide evidence that proteins from the same complex show a significant coexpression of their corresponding genes. Also in [55] it is mentioned that inferred interactions from gene expression data 'may represent an expanded class of interactions' [55]. However, when it comes to

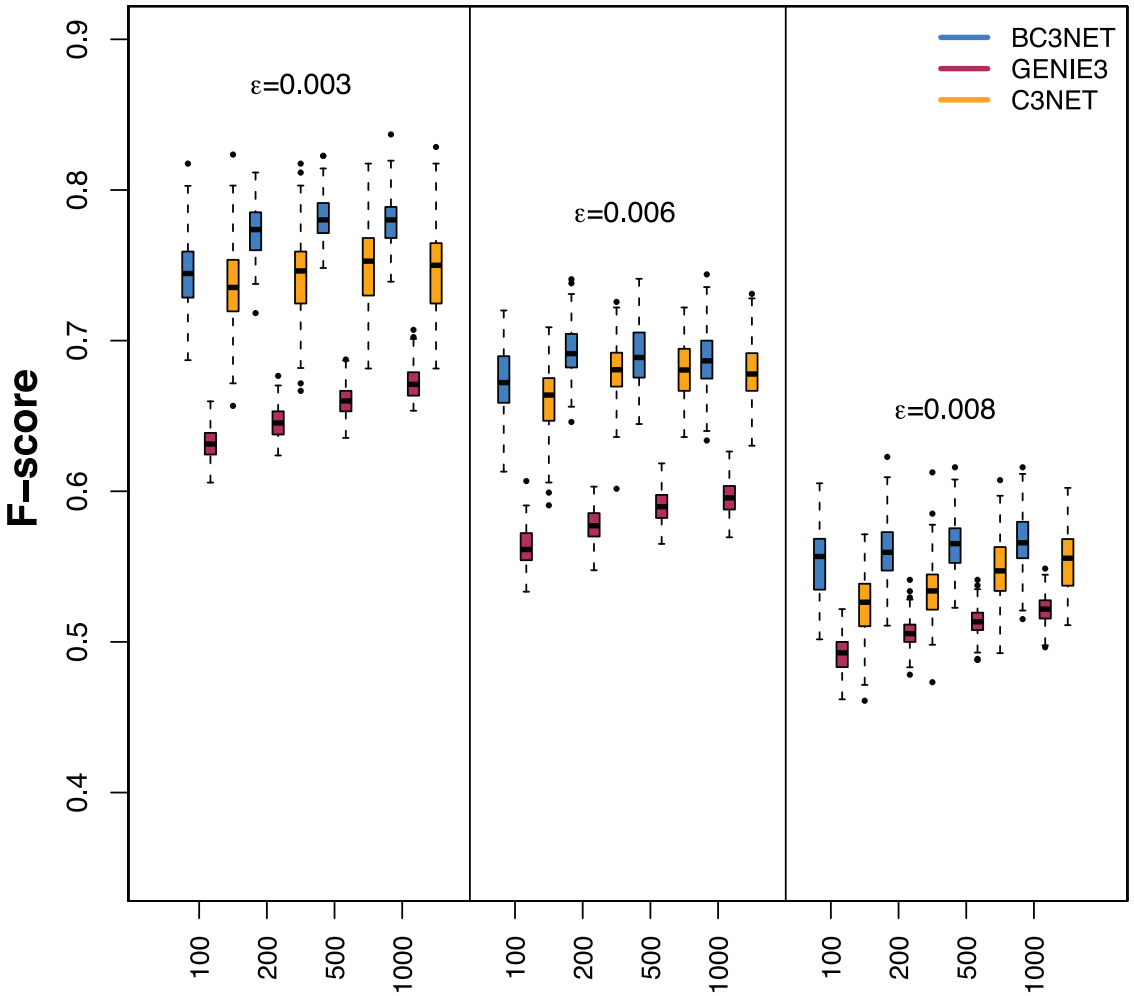

Figure 4. Comparative analysis of BC3NET, GENIE3 and C3NET for Erdös-Rényi networks with edge density. The x-axis shows the sample size *n*.
doi:10.1371/journal.pone.0033624.g004

the experimental assessment of the inferred networks, usually, only interactions related to the transcriptional regulation are studied, e.g., with ChIP-chip experiments [11,16]. To our knowledge we are the first to provide a large-scale analysis of an inferred GRN from gene expression data with respect to the presence of protein-protein interactions.

BC3NET is an ensemble method that uses as base network inference algorithm C3NET [9,56]. As for other ensemble methods based on bagging, e.g., random forests, the interpretability and characteristics of the base method does usually not translate to the resulting ensemble method [28,32]. In our case this means that the inferred network can actually have more than p edges, despite the fact that networks inferred by C3NET can not. However, in our case this is a desirable property because it improves BC3NET leading ultimately to a richer connectivity structure of the inferred network. Specifically, our numerical results demonstrate that BC3NET gains in average more than 40% true positive edges compared to C3NET (see Fig. 5). Another more general advantage of an ensemble approach is that it is straight forward to use on a computer cluster because a parallelization is naturally given by the base inference methods. Given the increasing availability of computer clusters this appears to be a conceptual advantage over none ensemble methods, likely to gain even more importance in the future. In this paper we pursued a conservative approach by using a Bonferroni procedure for MTC to demonstrate that even in this setting our method is capable of inferring many significant interactions that can be confirmed biologically. However, there is certainly potential to use more adopted MTC procedures that are less conservative. For example, procedures controlling the *false discovery rate* (FDR) could be investigated [39,57].

Further, we want to note that despite the fact that the network inference method C3NET is no Bayesian method [58,59], BC3NET is. The reason for this is that it is known for the bootstrap distribution of a parameter to correspond approximately to the Bayesian posterior distribution for a noninformative prior, and the bagged estimate thereof is the approximate mean of the Bayesian posterior [60]. Hence, BC3NET can be considered as a Bayesian method with noninformative priors for the connectivity structure among the genes. Given the problem to define informative priors for a Bayesian approach in a genomics context, either because not enough reliable information about a specific organism is available or because it is difficult to select this information in an uncontroversial manner, a noninformative prior is in the current state of genomics research still a prevalent choice. From a theoretical point of view, a bootstrap implementation is

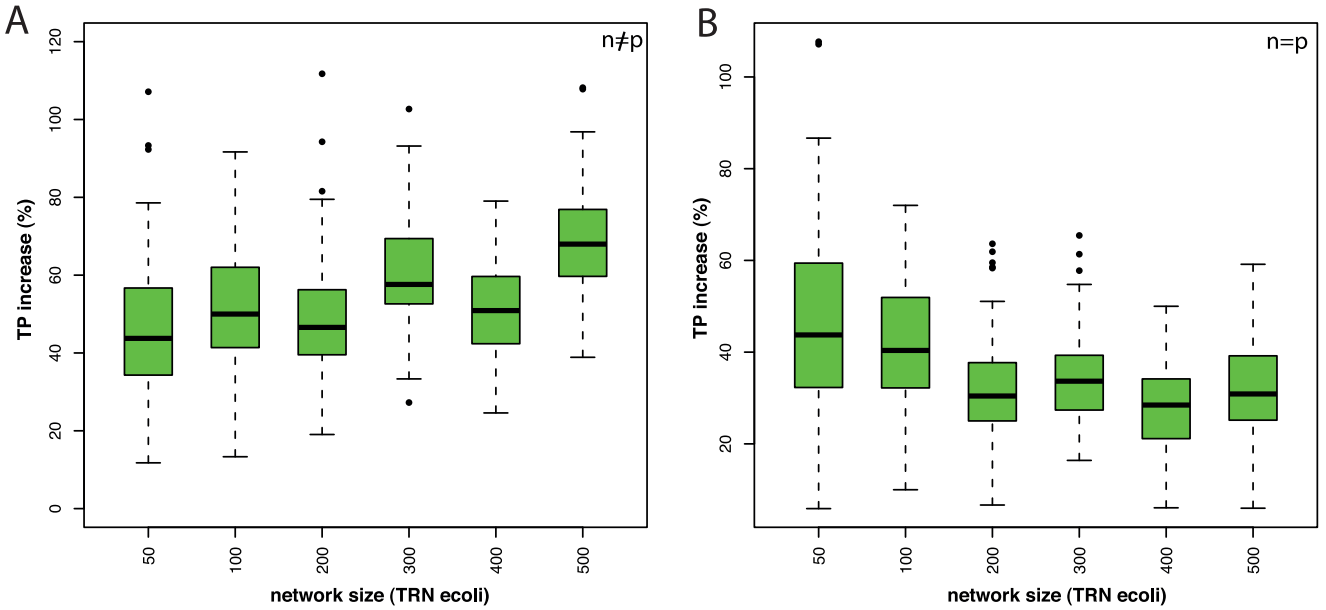

Figure 5. Gain in the number of true positive edges in BC3NET compared with C3NET. The x-axis shows the size (number of genes p) of the used subnetwork of *E.coli*. A: Influence of network size on TP gain with constant sample size $n=50$ ($n \neq p$). B: Influence of network size on TP gain with sample size $n=p$.
doi:10.1371/journal.pone.0033624.g005

Table 1. Top 30 GO-terms for a GPEA for the category Biological Process.

GOID	Term	Genes	Edges	Exp	p_{Bonf}
GO:0042254	ribosome biogenesis	349	546	66	$1.5e-297$
GO:0022613	ribonucleoprotein complex biogenesis	398	581	86	$3.3e-273$
GO:0034470	ncRNA processing	332	472	60	$4.1e-248$
GO:0006364	rRNA processing	237	355	31	$1.6e-239$
GO:0016072	rRNA metabolic process	246	364	33	$1.7e-238$
GO:0034660	ncRNA metabolic process	386	519	81	$2.2e-233$
GO:0006412	translation	699	822	267	$6.1e-176$
GO:0006396	RNA processing	506	581	140	$4.0e-175$
GO:0007005	mitochondrion organization	282	330	43	$3.0e-168$
GO:0032543	mitochondrial translation	100	159	5	$4.2e-167$
GO:0044085	cellular component biogenesis	841	905	386	$1.1e-127$
GO:0044281	small molecule metabolic process	890	847	432	$2.9e-82$
GO:0044257	cellular protein catabolic process	347	271	66	$2.0e-78$
GO:0030163	protein catabolic process	369	288	74	$1.1e-77$
GO:0006082	organic acid metabolic process	388	303	82	$4.5e-77$
GO:0044248	cellular catabolic process	720	612	283	$1.2e-69$
GO:0006519	cellular amino acid and derivative metabolic process	296	216	48	$1.4e-68$
GO:0009056	catabolic process	810	709	358	$4.5e-68$
GO:0019752	carboxylic acid metabolic process	370	271	75	$3.5e-67$
GO:0043436	oxoacid metabolic process	370	271	75	$3.5e-67$

All terms contain <1000 and >2 genes. 'Exp' denotes the expected number of edges for a GO-term. A total of 2159 terms were tested of which 525 (24.31%) tested significant.
doi:10.1371/journal.pone.0033624.t001

Table 2. Top 30 GO-terms for a GPEA for protein complexes.

GOID	Term	Genes	Edges	Exp	p_{Bonf}
GO:0033279	ribosomal subunit	210	442	24	0
GO:0022626	cytosolic ribosome	151	315	12	$1.2e-314$
GO:0005840	ribosome	291	485	46	$1.9e-310$
GO:0030529	ribonucleoprotein complex	568	789	176	$3.3e-262$
GO:0000313	organellar ribosome	78	142	3	$2.1e-173$
GO:0005761	mitochondrial ribosome	78	142	3	$2.1e-173$
GO:0015934	large ribosomal subunit	124	154	8	$2.1e-132$
GO:0030684	preribosome	130	155	9	$1.3e-127$
GO:0000502	proteasome complex	49	77	1	$1.3e-104$
GO:0022625	cytosolic large ribosomal subunit	82	96	4	$1.2e-96$
GO:0000315	organellar large ribosomal subunit	42	58	1	$2.6e-79$
GO:0005762	mitochondrial large ribosomal subunit	42	58	1	$2.6e-79$
GO:0031597	cytosolic proteasome complex	30	45	0	$5.6e-70$
GO:0034515	proteasome storage granule	30	45	0	$5.6e-70$
GO:0015935	small ribosomal subunit	86	69	4	$6.7e-57$
GO:0022627	cytosolic small ribosomal subunit	54	48	2	$6.2e-51$
GO:0030686	90S preribosome	80	55	3	$2.4e-43$
GO:0005838	proteasome regulatory particle	24	27	0	$6.3e-41$
GO:0022624	proteasome accessory complex	24	27	0	$6.3e-41$
GO:0005839	proteasome core complex	15	21	0	$1.4e-38$

All terms contain more than 2 genes. A total of 377 different terms were tested of which 94 protein complexes (24.93%) were significant.
doi:10.1371/journal.pone.0033624.t002

easier to accomplish than the corresponding (full) Bayesian method. Hence, our approach is more elementary [60]. Employing a similar argument as above, one can also see that BC3NET performs a model averaging of the individual networks inferred by C3NET.

From a conceptual point of view, one may wonder if an inferred GRN using BC3NET corresponds to a causal or an association network [19,61]. Here, by *causal* we denote an edge that corresponds to a *direct* interaction between gene products, e.g., the binding of a transcription factor to the promoter region on the DNA for regulating the expression of this genes. The quantitative evaluation of our simulated data, provide actually a quantification of the *causal content* of the inferred networks in the form of F-scores. It is clear that due to the statistical nature of the data, any inference is accompanied by a certain amount of uncertainty leading to an inferred GRN that contains false positive as well as

Table 3. Shown are 11 significant BC3NET network components nested in the BioGrid PPI yeast network.

Component	Genes	Edges	p_{Bonf}	GO
c_1	147	210	$2.01e^{-3}$	ribosome biogenesis ($p=3.16e-27$)
c_2	49	50	$2.01e^{-3}$	protein amino acid glycosylation ($p=5.05e-13$)
c_3	41	69	$2.01e^{-3}$	ubiquitin-dependent protein catabolic process ($p=3.16e-27$)
c_4	22	21	$2.01e^{-3}$	actin cytoskeleton organization ($p=4.74e-8$)
c_5	22	25	$2.01e^{-3}$	DNA replication ($p=3.13e-9$)
c_6	19	19	$2.01e^{-3}$	mitochondrial translation ($p=1.39e-21$)
c_7	15	17	$2.01e^{-3}$	ergosterol biosynthetic process ($p=5.05e-20$)
c_8	10	10	$1.00e^{-2}$	cytokinesis ($p=6.32e-03$)
c_9	10	9	$1.00e^{-2}$	DNA replication initiation ($p=4.42e-10$)
c_{10}	10	12	$1.00e^{-2}$	response to pheromone ($p=1.67e-11$)
c_{11}	9	9	$1.61e^{-2}$	microtubule-based process ($p=7.26e-05$)

Shown are the number of genes and concordant edges for each BC3NET network component. The p-values were adjusted using a Bonferroni procedure. We annotated these network components by using the most enriched GO term from the category Biological Process.
doi:10.1371/journal.pone.0033624.t003

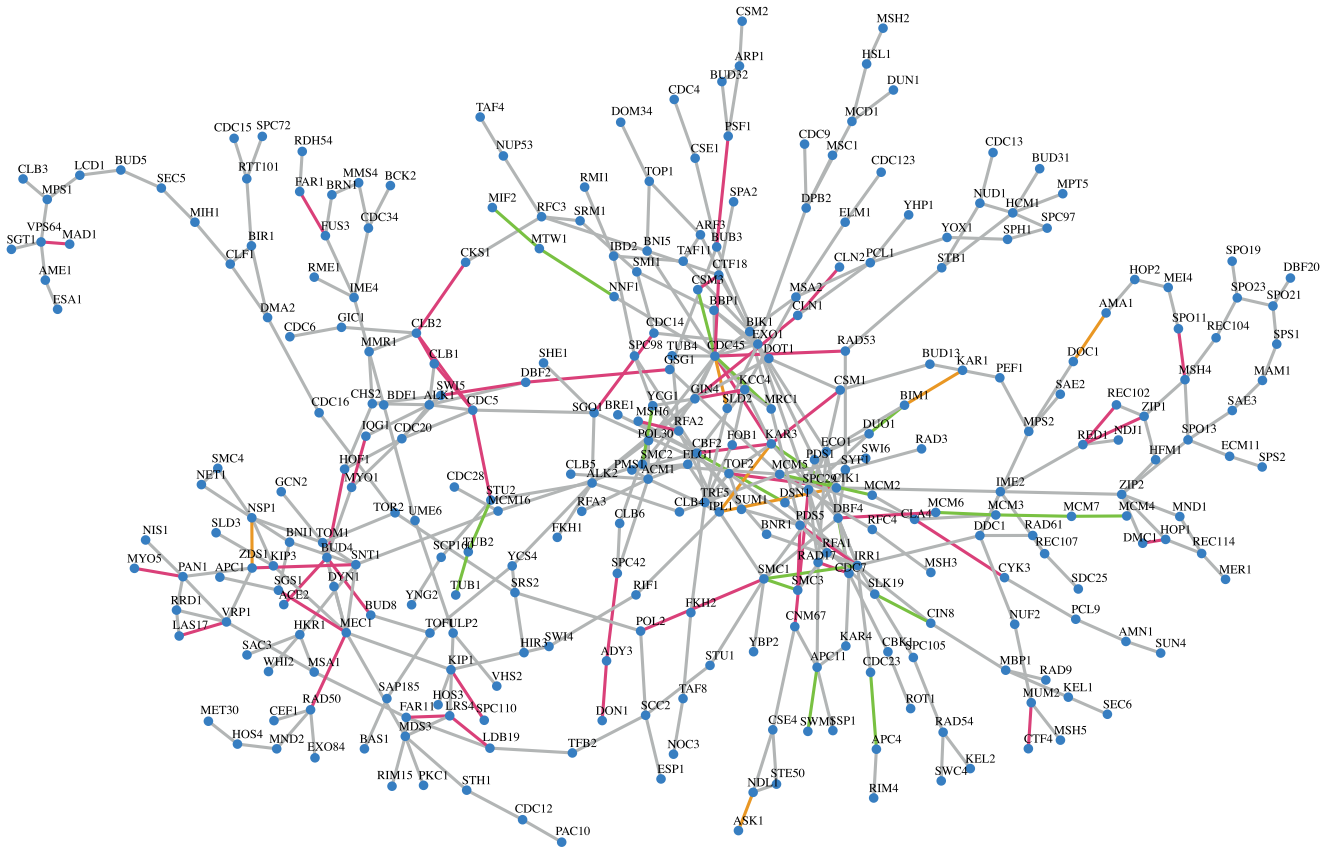

Figure 6. The largest network component of the yeast BC3NET network nested in *cell cycle* (GO category Biological Process: GO:0007049). The network component comprises 304 genes and 423 edges (GPEA *cell cycle* $p_{Bonferroni} \leq 2.6e-55$). The 57 violet edges correspond to interactions present in BioGrid, 25 green edges correspond to protein-protein interactions in protein complexes (GO) and the 7 orange edges are present in both databases.
doi:10.1371/journal.pone.0033624.g006

false negative edges. However, as demonstrated by our numerical analysis, BC3NET is an important improvement toward the inference of causal gene regulatory networks.

Despite the fact that the presented inference method BC3NET was introduced by using gene expression data from DNA microarray experiments, it can also be used in connection with data from RNA-seq experiments. Given the rapidly increasing importance of this new technology we expect that within the next few years datasets with sufficient large sample size are available to infer GRN.

Acknowledgments

We would like to thank Gökmen Altay, Dirk Husmeier and Shailesh Tripathi for fruitful discussions. For our simulations we used R [62] and the network was visualized with igraph [63].

Author Contributions

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

References

1. Barabási AL, Oltvai ZN (2004) Network biology: Understanding the cell's functional organization. Nature Reviews 5: 101–113.
2. Emmert-Streib F, Glazko G (2011) Network Biology: A direct approach to study biological function. Wiley Interdiscip Rev Syst Biol Med 3: 379–391.
3. Kauffman S (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. Journal of Theoretical Biology 22: 437–467.
4. Palsson B (2006) Systems Biology. Cambridge; New York: Cambridge University Press.
5. Vidal M (2009) A unifying view of 21st century systems biology. FEBS Letters 583: 3891–3894.
6. Waddington C (1957) The strategy of the genes. Geo, Allen & Unwin, London.
7. Stolovitzky G, Califano A, eds (2007) Reverse Engineering Biological Networks: Opportunities and Challenges in Computational Methods for Pathway Inference. Wiley-Blackwell.
8. Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, et al. (2010) Revealing strengths and weaknesses of methods for gene network inference. Proceedings of the National Academy of Sciences 107: 6286–6291.
9. Altay G, Emmert-Streib F (2010) Inferring the conservative causal core of gene regulatory networks. BMC Systems Biology 4: 132.
10. Bulashevska S, Eils R (2005) Inferring genetic regulatory logic from expression data. Bioinformatics 21: 2706–2713.
11. Faith J, Hayete B, Thaden J, 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: e8.
12. de la Fuente A, Bing N, Hoeschele I, Mendes P (2004) Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics 20: 3565–3574.
13. Hache H, Lehrach H, Herwig R (2009) Reverse engineering of gene regulatory networks: A comparative study. EURASIP J Bioinform Syst Biol 2009: 617281.
14. 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.
15. Markowetz F, Spang R (2007) Inferring cellular networks–a review. BMC Bioinformatics 8: S5.
16. Margolin A, Nemenman I, Basso K, Wiggins C, Stolovitzky, et al. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7: S7.

17. Meyer P, Kontos K, Bontempi G (2007) Information-theoretic inference of large transcriptional regulatory networks. EURASIP journal on bioinformatics and systems biology 2007: 79879.
18. Werhli A, Grzegorczyk M, Husmeier D (2006) Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. Bioinformatics 22: 2523–31.
19. Xing B, van der Laan M (2005) A causal inference approach for constructing transcriptional regulatory networks. Bioinformatics 21: 4007–4013.
20. Barabási AL (2007) Network Medicine – From Obesity to the "Diseasome". N Engl J Med 357: 404–407.
21. Emmert-Streib F, Dehmer M, eds. Medical Biostatistics for Complex Diseases. Weinheim: Wiley-Blackwell.
22. Zanzoni A, Soler-Lopez M, Aloy P (2009) A network medicine approach to human disease. FEBS Letters 583: 1759–1765.
23. De Smet R, Marchal K (2010) Advantages and limitations of current network inference methods. Nature Reviews Microbiology 8: 717–729.
24. Emmert-Streib F, Glazko G, Altay G, de Matos Simoes R (2012) Statistical inference and reverse engineering of gene regulatory networks from observational expression data. Frontiers in Genetics 3: 8.
25. Olsen C, Meyer P, Bontempi G (2009) On the impact of entropy estimator in transcriptional regulatory network inference. EURASIP Journal on Bioinformatics and Systems Biology 2009: 308959.
26. Penfold CA, Wild DL (2011) How to infer gene networks from expression profiles, revisited. Interface Focus 1: 857–870.
27. West M (2003) Bayesian factor regression models in the "large p, small n" paradigm. In: Bayesian Statistics 7 Oxford University Press. pp 723–732.
28. Breiman L (1996) Bagging Predictors. Machine Learning 24: 123–140.
29. Zhang H, Singer BH (2010) Recursive partitioning and applications. Springer; New York: Springer.
30. Friedman N, Goldszmidt M, Wyner A (1999) Data Analysis with Bayesian Networks: A Bootstrap Approach. In: Proc Fifteenth Conf on Uncertainty in Artificial Intelligence (UAI) Society for Artificial Intelligence in Statistics. pp 196–205.
31. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P (2010) Inferring regulatory networks from expression data using tree-based methods. PLoS ONE 5: e12776.
32. Breiman L (2001) Random Forests. Machine Learning 45: 5–32.
33. Margolin A, Wang K, Lim W, Kustagi M, Nemenman I, et al. (2006) Reverse engineering cellular networks. Nat Protoc 1: 662–71.
34. Lefebvre C, Rajbhandari P, Alvarez MJ, Bandaru P, Lim WK, et al. (2010) A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. Molecular Systems Biology 6: 377.
35. Zhao X, D Arca D, Lim WK, Brahmachary M, Carro MS, et al. (2009) The N-Myc-DLL3 cascade is suppressed by the ubiquitin ligase Huwe1 to inhibit proliferation and promote neurogenesis in the developing brain. Developmental Cell 17: 210–221.
36. Efron B, Tibshirani RJ (1993) An Introduction to the Bootstrap. Chapman et Hall.
37. Dudoit S, van der Laan M (2007) Multiple Testing Procedures with Applications to Genomics. New York; London: Springer.
38. Farcomeni A (2008) A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. Stat Methods Med Res 17: 347–88.
39. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B (Methodological) 57: 125–133.
40. Ge Y, Dudoit S, Speed T (2003) Resampling-based multiple testing for microarray data analysis. TEST 12: 1–77.
41. Emmert-Streib F, Altay G (2010) Local network-based measures to assess the inferability of different regulatory networks. IET Systems Biology 4: 277–288.
42. Paninski L (2003) Estimation of entropy and mutual information. Neural Computation 15: 1191–1253.
43. 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.
44. Schürmann T, Grassberger P (1996) Entropy estimation of symbol sequences. Chaos 6: 414427.
45. Daub C, Steuer R, Selbig J, Kloska S (2004) Estimating mutual information using B-spline functions–an improved similarity measure for analysing gene expression data. BMC Bioinformatics 5: 118.
46. Faith J, Driscoll M, Fusaro V, Cosgrove E, Hayete B, et al. (2008) Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. Nucleic Acids Res 36: D866–70.
47. Erdos P, Renyi A (1960) On the evolution of random graphs. Publ Math Inst Hungary Acad Sci 5: 17–61.
48. 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.
49. Husmeier D (2003) Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic bayesian networks. Bioinformatics 19: 2271–82.
50. Leclerc RD (2008) Survival of the sparsest: robust gene networks are parsimonious. Mol Syst Biol 4: 213.
51. Ashburner M, Ball C, Blake J, Botstein D, Butler, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics 25: 25–29.
52. Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, et al. (2008) The BioGRID Interaction Database: 2008 update. Nucl Acids Res 36: D637–640.
53. Jansen R, Greenbaum D, Gerstein M (2002) Relating whole-genome expression data with proteinprotein interactions. Genome Res 12: 37–46.
54. Grigoriev A (2001) A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae. Nucleic Acids Res 29: 3513–9.
55. Margolin A, Califano A (2007) Theory and limitations of genetic network inference from microarray data. Ann N Y Acad Sci 1115: 51–72.
56. Altay G, Emmert-Streib F (2011) Structural Influence of gene networks on their inference: Analysis of C3NET. Biology Direct 6: 31.
57. Storey J (2002) A direct approach to false discovery rates. Journal of the Royal Statistical Society, Series B 64: 479–498.
58. Bernardo JM, Smith AFM (1994) Bayesian Theory Wiley.
59. Gelman A, Carlin JB, Stern HS, Rubin DB (2003) Bayesian Data Analysis Chapman & Hall/CRC.
60. Haste T, Tibshirani R, Friedman J (2009) The elements of statistical learning: Data mining, inference and prediction. New York: Springer.
61. Opgen-Rhein R, Strimmer K (2007) Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. BMC Bioinformatics 8: S3.
62. 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.
63. Csardi G, Nepusz T (2006) The igraph software package for complex network research. InterJournal Complex Systems. 1695 p.