BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 10 2012, pages 1376–1382
doi:10.1093/bioinformatics/bts143

Systems biology

Advance Access publication March 30, 2012

Inferring gene regulatory networks by ANOVA

Robert Küffner*, Tobias Petri, Pegah Tavakkolkhah, Lukas Windhager and Ralf Zimmer

Department of Informatics, Ludwig-Maximilians University, Amalienstr. 17, 80333 Munich, Germany

Associate Editor: Martin Bishop

ABSTRACT

Motivation: To improve the understanding of molecular regulation events, various approaches have been developed for deducing gene regulatory networks from mRNA expression data.

Results: We present a new score for network inference, η^2, that is derived from an analysis of variance. Candidate transcription factor:target gene (TF:TG) relationships are assumed more likely if the expression of TF and TG are mutually dependent in at least a subset of the examined experiments. We evaluate this dependency by η^2, a non-parametric, non-linear correlation coefficient. It is fast, easy to apply and does not require the discretization of the input data. In the recent DREAM5 blind assessment, the arguably most comprehensive evaluation of inference methods, our approach based on η^2 was rated the best performer on real expression compendia. It also performs better than methods tested in other recently published comparative assessments. About half of our predicted novel predictions are true interactions as estimated from qPCR experiments performed for DREAM5.

Conclusions: The score η^2 has a number of interesting features that enable the efficient detection of gene regulatory interactions. For most experimental setups, it is an interesting alternative to other measures of dependency such as Pearson's correlation or mutual information.

Availability: See http://www2.bio.ifi.lmu.de/~kueffner/anova.tar.gz for code and example data.

Contact: kueffner@bio.ifi.lmu.de

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

Received on July 11, 2011; revised on March 18, 2012; accepted on March 22, 2012

1 INTRODUCTION

The reconstruction of gene regulatory networks (GRNs) from expression data can help to improve our understanding of molecular regulation events. A variety of algorithms have been devised to predict gene regulatory interactions, frequently based on mutual dependencies between the expression of regulators and their targets (see related work).

We propose to evaluate transcription factor:target gene (TF:TG) interactions by the measure η^2 [Cohen (1973)], a non-linear correlation coefficient derived from an analysis of variance (ANOVA). Although η^2 has a number of interesting features it has, to our knowledge, not been applied to network inference or to bioinformatics in general.

*To whom correspondence should be addressed.

A high proportion of our predicted novel interactions were confirmed by small-scale qPCR experiments performed by the DREAM5 organizers. In addition, our approach was evaluated as the best performer for the inference of real datasets in the recent DREAM5 blind assessment [http://wiki.c2b2.columbia.edu/dream/index.php; Marbach *et al.* (2010); Prill *et al.* (2010)]. Here, 29 participating teams applied a diverse set of inference methods to a variety of large real (*Escherichia coli*, *Saccharomyces cerevisiae*) and artificial expression compendia with thousands of genes from several hundreds of microarray measurements. The microarray experiments consisted of various gene, drug or environmental perturbations that were in some cases carried out as time courses.

After a brief summary of related work we describe the GRN inference setting, introduce our inference approach based on the score η^2 and evaluate its properties and performance.

1.1 Related work

The inference of large GRNs of 500+ nodes is frequently tackled by unsupervised, data-driven approaches that aim to resolve dependencies from expression data alone. We briefly review some commonly used techniques in the following and refer the reader to review papers [e.g. by Altay and Emmert-Streib (2010b), Lee and Tzou (2009) and Markowetz and Spang (2007)] for a more comprehensive overview of methods.

Unparameterized topologies can be approximated even for large networks by measures of pairwise gene dependencies, e.g. using Pearson's linear correlation coefficient [Butte and Kohane (1999)]. To take non-linear correlations into account, information theoretic approaches can be employed such as Bayes conditional probability tables or mutual information [Butte and Kohane (2000); Ding and Peng (2005); Faith *et al.* (2007); Margolin *et al.* (2006); Meyer *et al.* (2007); Zhao *et al.* (2006)]. The latter techniques require a very careful discretization of the expression data to avoid the loss of signal [Altay and Emmert-Streib (2010a); Mukherjee and Speed (2008); Zhu *et al.* (2008)].

One source of false positive predictions are indirect effects, i.e. in a cascade $A \rightarrow B \rightarrow C$ methods are likely to also predict the additional effect $A \rightarrow C$. Extensions like the data processing inequality [ARACNe, Margolin *et al.* (2006)] and gene dependent background distributions [CLR, Faith *et al.* (2007)] have been proposed to overcome this problem. The minimum redundancy maximum relevance concept [Ding and Peng (2005); Meyer *et al.* (2007)] offers another way to select important edges. Indirect effects might also be identified and removed by partial correlations [Castelo and Roverato (2009)], elastic net or lasso, a L1-penalized estimation of the inverse covariance matrix [Friedman *et al.* (2008); Kabir *et al.* (2010); Wang *et al.* (2009)]. All of the mentioned approaches measure *global* dependencies, i.e. dependencies that are visible

across the majority of measured experimental conditions. *Local* dependencies that are only apparent in a subset of the conditions [Kwon *et al.* (2003)] might thus be missed.

Models like Boolean, (probabilistic) Bayesian networks, ordinary differential equations (ODE) or Petri Nets are generative i.e. they allow the generation of the original training datasets by simulation. Optimization approaches minimize the deviation from given data by parameterizing models [Guthke *et al.* (2005); Küffner *et al.* (2010); Wang *et al.* (2006)]. Due to the huge parameter space these algorithms may not scale well to large networks.

The assessment of the multitude of reconstruction algorithms is quite difficult. Comparative studies [Hache *et al.* (2009); Michoel *et al.* (2009); Narendra *et al.* (2011); Soranzo *et al.* (2007); Zou and Feng (2009)] evaluate only subsets of approaches. More comprehensive assessments are facilitated through community-wide challenges conducted by the DREAM consortium.

2 METHODS

2.1 Inference setting, data sources and evaluation

Problem statement and evaluation. GRN inference aims at the detection of gene regulatory relationships from mRNA expression datasets. The task is to reverse engineer the directed topology of one network for each of the available expression datasets (Table 1). In the following, we describe a setup for the evaluation of inference methods that has been adopted by many comparative assessment studies including DREAM5 (http://wiki.c2b2.columbia.edu/dream/index.php/D5c4) and Narendra *et al.* (2011).

For each dataset, potential TFs are given. Only these TFs should be included as regulators in the network predictions as the used gold standards do not contain gene regulatory interactions for other regulators such as sigma factors or miRNAs. The list of TFs was available to all participants of the challenge. Approaches were then required to check and rank $|\text{TF}| \times |\text{Genes}|$ candidate relationships. Lists of ranked candidate interactions are evaluated against the true topology (in case of the artificial dataset) or against experimentally determined TF:TG interactions. Candidate lists are evaluated against gold standard networks (see below) based on the area under the precision-recall curve (AUPR) and the area under the receiver-operator characteristics curve [AUROC; see Prill *et al.* (2010)]. In DREAM5, only the top 100 000 interactions were considered for this analysis. The resulting AUPR and AUROC will be lower if only a subset of the interactions is considered. Although this difference has only little effect on the ranking of the approaches, we will report AUROC values for the top 100 000 predictions as well as for all predictions in order to enable the comparison to other studies [e.g. Narendra *et al.* (2011)]. The performance evaluation in this article focuses on the AUROC, but additional evaluation and scores can be found in the Supplementary Material (part 4).

Table 1. DREAM5 and M3D datasets used in this study

Dataset	|TF|	|Genes|	|TF| pert.	|Genes| pert.	|Chips|
Artificial^{D5}	195	1643	38	38	805
E. coli^{D5}	334	4511	20	43	805
E. coli^{M3D}	167	4297	17	67	907
S. cerevisiae^{D5}	333	5950	5	14	536
S. cerevisiae^{M3D}	156	6572	11	37	904

Shown is the size of the examined datasets as well as the number of measurements subject to gene specific perturbations (gene over-expressions and deletions).

Expression compendia. In this study, we used three datasets provided by DREAM5 and two additional datasets from M3D [Faith *et al.* (2008)]. All datasets consisted of several thousand genes and several hundred microarray measurements (Table 1). In comparison to data repositories such as GEO [Barrett *et al.* (2010)], DREAM5 and M3D provide fewer but uniformly preprocessed and normalized datasets. Measurements as well as annotations are rendered comparable across different experiments and are thus suited to automated network inference.

In case of the real DREAM5 datasets, organism, experiment and gene names are replaced by random IDs to enable the evaluation of the inferred networks against experimentally confirmed interactions unknown to the participants. Thus, no prior knowledge could be utilized for the inference.

Datasets in the expression compendia are subdivided into *experiments* that consist of all microarrays described in a single publication or conducted by the same experimenter. Besides wild-type measurements, experimental *conditions* represent (combinations of) drug, environmental and gene perturbations. Some of the drug or environmental perturbations are provided as time course measurements. We considered each time point as a separate condition. A condition may contain multiple *replicates*. In case of gene perturbations (deletion or over-expression), the annotations provide the IDs of the perturbed genes. The artificial dataset was generated by the tool GeneNetWeaver [Marbach *et al.* (2009), see Section 6 of the challenge description at http://wiki.c2b2.columbia.edu/dream/index.php/D5c4] and mimicked the *E. coli* dataset in the composition of the perturbations and time courses.

For additional validation, we obtained *E. coli* and *S. cerevisiae* expression data from the M3D database [see Table 1 and Faith *et al.* (2008)]. Similar to the chip annotations provided in DREAM5, M3D provided manually curated metadata for their chip measurements.

Gold standard networks. Predicted *E. coli* interactions were validated based on RegulonDB [Gama-Castro *et al.* (2011); Huerta *et al.* (1998)], a database of gene regulatory relationships that are both experimentally validated and manually curated. The *S. cerevisiae* gold standard [MacIsaac *et al.* (2006)] was automatically derived by large-scale chromatin immunoprecipitation (ChIP) binding assays. Physical binding is not a sufficient evidence, as noted by Hu *et al.* (2007) and Boulesteix and Strimmer (2005). Thus, ChIP will lead to many false positive interactions. MacIsaac *et al.* (2006) aimed to overcome this problem by complementing ChIP assays with conservation-based motif discovery algorithms. Due to the more reliable small-scale assays and the manual curation, the *E. coli* gold standard should be regarded as more reliable than the one for *S. cerevisiae*. This is supported by a recent review by Narendra *et al.* (2011) where even otherwise accurate methods fail to predict this gold standard. The true network topology is known for the artificial dataset and was used for evaluation.

Combined with the above mentioned M3D datasets, these gold standard networks have been used for evaluating inference methods by DREAM5 as well as Faith *et al.* (2007) and Narendra *et al.* (2011). Because the same gold standards are used and because the majority of the experiments in the DREAM5 datasets on *E. coli* were taken from the M3D database (Daniel Marbach, personal communication) results of DREAM5 and Narendra *et al.* (2011) are approximately comparable. This also applies to assessments based on artificial datasets that have been generated by the tool GeneNetWeaver [Marbach *et al.* (2009)] in both studies.

2.2 Network inference

Fold changes. Basal gene levels can be very different between experiments. To compensate for this, we transformed the absolute expression values into expression fold changes (see also Supplementary Material, part 1). Fold changes are computed by mapping the measurements $m_i, i=1...|m|$ of each condition m onto one or more valid control conditions (Fig. 1). Each m is subject to a combination of gene, drug or environmental perturbations P. A condition m^c measured at time $t(m^c)$ under the set of perturbations P^c is called a valid control condition for m if $P^c \subset P$ and $t(m)=t(m^c)$, where $P-P^c$ represents the differential treatment between two conditions.

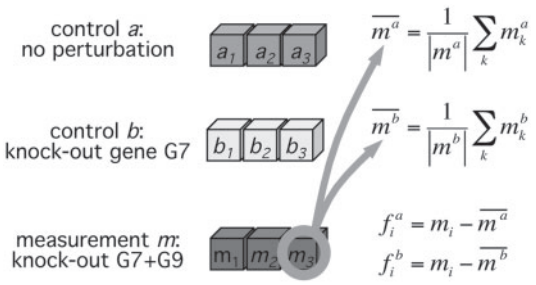

Fig. 1. Transformation of absolute expression values into fold changes. We compute log-fold changes by mapping each measured condition m to one or more control conditions (replicated measurements) from the same experiment. A control may have fewer drug or gene perturbations than the corresponding measurements, but not more. In the example shown, both conditions a and b have fewer perturbations than m and are valid control conditions. Then, the replicates of the controls are averaged and (here: a total of 6) fold changes are computed for the replicates in m against the means of the selected controls. Log-fold changes are computed as differences as measurements are already log-transformed. For the given application, the resulting geometric mean performs similar to an average (not shown).

For instance, the DREAM5 *E. coli* dataset consisted of 805 chip measurements of 487 different experimental conditions. Among the 805 chips we selected controls for 599 chips (corresponding to 379 conditions). Due to the multiplicity of measurement-control combinations, 935-fold changes were computed from the 599 chips.

Relevance networks. Our network inference approach is based on the estimation of the relevance of candidate interactions [Butte and Kohane (1999); Butte and Kohane (2000)]. Candidate interactions, i.e. pairs of a TF and a TG, are ranked by a score s. The score s can be any measure of dependency between the expression of the TF and its TG. Frequently used measures of dependency are based on Pearson's or Spearman's correlation coefficients or mutual information. In this article, we propose to use the score η^2 that is introduced below. The application of η^2 in the relevance network framework will be referred to as the η^2 approach. In addition to relevance networks, we also evaluate η^2 in the context of the C3NET [Altay and Emmert-Streib (2010a)] and CLR [Faith *et al.* (2007)] frameworks (Supplementary Material, part 4).

A measure of association derived from two-way ANOVA. Our inference approach is based on a two-way ANOVA. A two-way ANOVA can be used to model experimental observations Y_{ijk} as responses to two factors C and G as well as the measurement error,

$$Y_{ijk} = \mu + \tau_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \quad (1)$$

where μ is the average response, τ_i is the effect from the i-th level of the factor C, β_j is the effect from the j-th level of factor G, γ_{ij} is the joint effect from the interaction between factors C and G and ϵ_{ijk} represents the remaining unexplained error in replicate k. In our application of ANOVA, C models the effect of differential expression across $i \in [1 \ldots c]$ different experimental conditions and G models whether the expression profiles of the genes $j \in [g, t]$ (as we consider exactly one TF t and one TG g) differ. Thus, we apply ANOVA to a matrix of conditions, genes and replicates as depicted in Figure 2A. A two-way ANOVA tests three null hypotheses: (i) no differences in means of factor C; (ii) no differences in means of factor G; and (iii) no interaction between C and G, by partitioning the total sum of squares SS_T into four components (Fig. 2):

$$\mathrm{SS}_T = \mathrm{SS}_C + \mathrm{SS}_G + \mathrm{SS}_{CG} + \mathrm{SS}_{\mathrm{err}} \quad (2)$$

A sum of squares (SS) is a sum of squared deviations from a mean [Miller (1997)] and can be regarded as an unadjusted measure of dispersion.

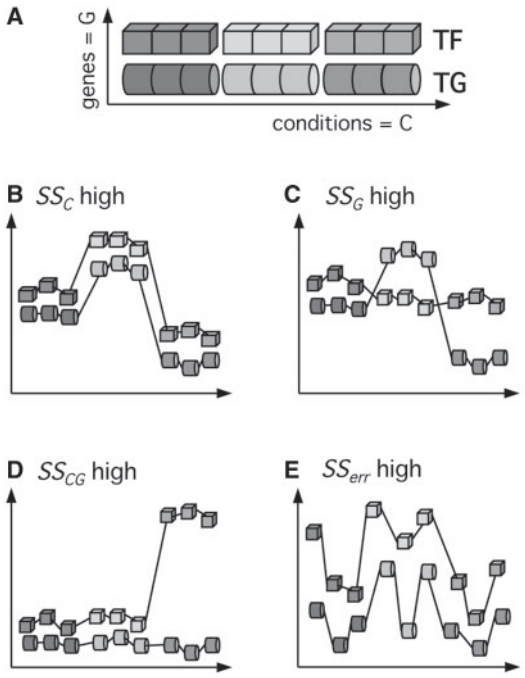

Fig. 2. Sum of squares and the two-way ANOVA. A two-way ANOVA analyzes two dimensions or effects (here: C for conditions and G for genes) by partitioning the SS into four components: $\mathrm{SS}_T = \mathrm{SS}_C + \mathrm{SS}_G + \mathrm{SS}_{CG} + \mathrm{SS}_{\mathrm{err}}$. The first example (panel B) exhibits strong associations between TF and TG. Here, SS_C is high as there is strong differential expression between the conditions. In panel C, the genes exhibit strong differences so SS_G will be high. If the two effects are linked (panel D), i.e. differential expression across conditions occurs only if strong differences are exhibited between both genes, SS_{CG} will be high. A high replicate variance leads to a high $\mathrm{SS}_{\mathrm{err}}$ (panel E).

A variance V_x is computed by adjusting the SS_x for the degree of freedom df_x, where df_x is the number of data points under consideration minus 1, and $x \in [C, G, CG, \mathrm{err}, T]$. An F-value is computed by weighting the effect variance against the error variance [Equation (3)]. F-values follow the F-statistic, which can be used to derive the statistical significance of the involved factors as p-values. For instance, to estimate the significance of differential expression across conditions we compute F_C by:

$$V_x = \frac{\mathrm{SS}_x}{\mathrm{df}_x}, F_C = \frac{V_C}{V_{\mathrm{err}}} \quad (3)$$

Effects so far describe differences, but ANOVA can also be used to detect specific similarities or associations between TF and TG. Phrased in terms of the two-way ANOVA, the strength of an association is proportional to the fraction of SS_C in the total sum of squares SS_T:

$$\eta^2_+ = \frac{\mathrm{SS}_C}{\mathrm{SS}_T}, F_{\eta+} = \frac{V_C}{V_T} \quad (4)$$

Thus, $\eta^2_+ \in [0 \ldots 1]$ measures association as the fraction of the total variance that is explained by the differential expression across experimental conditions. Cohen (1973) refers to η^2_+ as the non-parametric non-linear correlation coefficient. Its statistical significance can be estimated via $F_{\eta+}$. A more detailed description of the algorithm including pseudo code can be found in the Supplementary Material (parts 2 and 3).

Adjusting for negative correlation. In contrast to Pearson's ρ^2, η^2 does not directly test for negative correlations. We therefore propose to reverse the signs of the TF-fold changes to compute an additional η^2_-. The final ranking of candidate interactions is performed using $\eta^2 = \max(\eta^2_+, \eta^2_-)$.

Incorporation of gene perturbation experiments. We extend the basic approach to incorporate measurements on gene specific perturbations. A candidate interaction between a TF and a putative TG should be considered more likely if the TG shows a response to the knock-out or over-expression of the TF. In the calculation of η^2 for such an interaction this is taken into account by increasing the weight of conditions that involve gene perturbations affecting the TF by a user specified weight parameter w_{gp}. Values may range between 10 and 1000 (see Supplementary Material, part 4) and have been estimated based on M3D data [Faith *et al.* (2008)] prior to our participation in DREAM. The weight of such a condition is increased by inserting $w_{gp}-1$ additional copies into the ANOVA matrix (Fig. 2A). Conditions where non-TFs or TFs other than the currently tested TF are perturbed receive the default weight of 1.

3 RESULTS

Differences between datasets. Prokaryotes organize the regulation of expression by operons. A promotor region controls several structural genes which show similar expression patterns. We analyzed expression patterns by Markov clustering [mcl; Enright *et al.* (2002)]. Figure 3 shows the resulting expression patterns from the largest clusters of the *E. coli* and the artificial data. Co-regulation patterns are easy to detect in all real datasets but are virtually absent in the artificial data. We also noticed that ∼50% of the regulation in the artificial data is due to inhibition. Inhibition, i.e. negative correlations between TF and TG is comparatively rare in the real datasets (Fig. 4). Real networks are more complex than the artificial networks created by Marbach *et al.* (2009) as they do not account for interactions involving proteins or other molecules.

The real *E. coli* and *S. cerevisiae* datasets were also markedly different. While correlation between TF and TG is a good predictor of a gene regulatory relationship in artificial and *E. coli* data, this is not the case in *S. cerevisiae* (Fig. 4). This finding is reproducible across different yeast datasets, gold standards and measures of dependency (not shown). That network inference is more difficult in *S.cerevisiae* as compared with *E. coli* is consistent with the work of Hu *et al.* (2007) and Narendra *et al.* (2011).

Run time complexity. The run time complexity of our network inference approach is $O(|TF| \times |genes| \times |chips|)$ (see Supplementary Material, part 3), i.e. the complexity of the ANOVA estimator is linear in the number of chips. The datasets (Table 1) required between 280 k and 2 M evaluations of η^2 across 160 to 907 chips. The largest datasets required a run time of 2 min on a single processor core. The complexity of other inference approaches has been discussed previously [Narendra *et al.* (2011)].

Performance evaluation. Table 2 and Figure 5 show the performance of our η^2 method in comparison to other approaches. For a broad comparison of methods we combined our own evaluation results for some of the publicly available methods (for additional evaluation and scores see Supplementary Material, part 4) with the results of the DREAM5 network inference challenge as well as the large comparative assessment study of Narendra *et al.* (2011). We selected methods that performed best on one of the three DREAM5 datasets (methods 1–3) and the best performing methods (with respect to AUROC) as determined by Narendra *et al.* (2011) (methods 6–9). For comparison, we also applied the methods 2 as well as 5–7 as end-user ready tools were available.

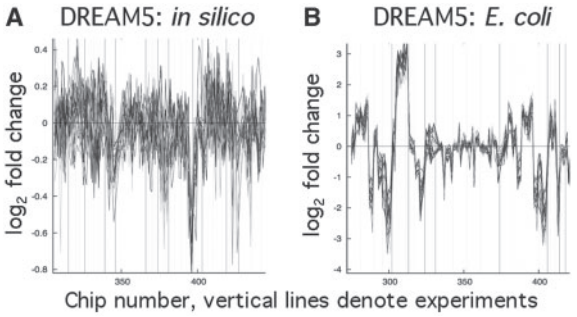

Fig. 3. Co-regulation patterns in artificial and real datasets. In contrast to the artificial data (panel A), all real datasets (panel B) show strong co-regulation patterns. In *E. coli*, clustered patterns largely overlap with operons. Shown are subsets of 20 genes selected from the largest clusters derived by Markov clustering.

In order to render the DREAM evaluation (that considered only the top 100 k predictions) comparable to the evaluation by Narendra *et al.* (2011), we re-computed the performance based on all predictions (i.e. not only the top 100 k) and re-applied publicly available methods. Considering all predictions usually increases the resulting AUROC by up to a few percentage points, which usually has only little effect on the performance ranking of the methods.

Interestingly, some of the participants in the DREAM5 challenge outperformed existing inference methods significantly, particularly in case of the artificial and *E. coli* datasets (Fig. 5). For the inference of yeast interactions such a clear statement was not possible as all methods performed rather poorly.

The presented η^2 method was the best performer for inferring *E. coli* interactions and was also competitive for the artificial dataset, outperforming the previously published methods analyzed in this article or by Narendra *et al.* (2011). The performance on artificial and *E. coli* interactions is depicted in Figure 5 for all methods that participated in the DREAM5 network inference challenge.

The prediction performance apparently also depends on the number of chip experiments. All methods yielded a better prediction performance for the M3D *E. coli* dataset (Table 2), which contains more measurements than the DREAM5 *E. coli* dataset. The same reasons also contribute to the lower performance observed in the *S. cerevisiae* dataset.

Properties of η^2 exemplified via selected interactions. We analyzed properties of η^2 on the *E. coli* dataset from M3D. A strong linear correlation is exhibited for instance by the fis:dusB interaction (fis: *organization and maintenance of nucleoid structure*; dusB: *tRNA-dihydrouridine synthase B*, Fig. 6A). The observed linear correlation might be due to the fact that both genes are part of the same operon. Non-linear correlations such as gadE:hdeA are also detected by η^2 (gadE: *acid-induced positive regulator of glutamate-dependent acid resistance*; hdeA: *stress response acid resistance protein*). hdeA is already activated by low gadE concentrations (Fig. 6B). In contrast, mdtE (*multidrug transporter component*) is activated only at high concentrations of gadX (*regulator of acid resistance*) resulting in an upwardly-curved scatterplot (not shown).

η^2 also enables the detection of correlations that are only apparent in a subset of the measured conditions, i.e. it detects local correlations. This increased sensitivity is due to the

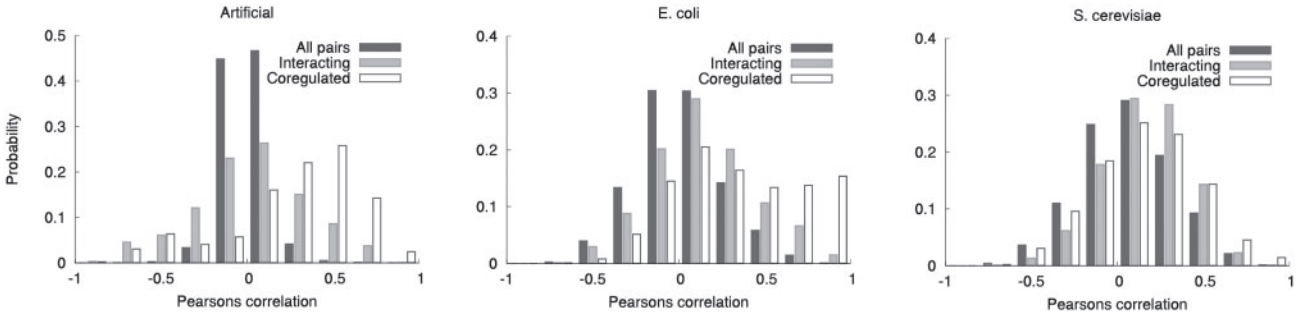

Fig. 4. Differences between artificial and real datasets. The correlation distributions of artificial, *E. coli* and *S. cerevisiae* data expression data are quite different. Shown are histograms of the correlation of non-interacting and interacting gene pairs as well as gene pairs regulated by the same set of TFs. In contrast to artificial and *E. coli* data, correlation between a TF and a TG is not a good indicator of a true regulatory relationship in *S. cerevisiae*.

Table 2. Performance of selected methods on the DREAM5 and M3D datasets.

		Artificial			*E. coli*				*S. cerevisiae*			
Methods	References (abbrev.)	D5:100k	D5	Nar2011	D5:100k	D5	M3D	Nar2011	D5:100k	D5	M3D	Nar2011
ANOVA η^2	This article	78.0[a]	81.6[c]		**67.1**[a]	**74.6**[c]	**79.8**[d]		51.8[a]	57.8[c]	55.0[d]	
Genie3	Huynh-Thu *et al.* (2010)	**81.5**[a]	**83.4**[c]		61.7[a]	69.0[c]	67.3[d]		51.8[a]	54.5[c]	51.3[d]	
Team 395	unpublished	69.5[a]			60.2[a]				**53.9**[a]			
Pearson's ρ^2	Butte and Kohane (1999)	75.7[b]	76.5[c]		57.2[b]	61.2[c]	64.6[d]		51.0[b]	56.9[c]	53.8[d]	
MRNet	Meyer *et al.* (2007)	71.5[b]	73.0[c]		58.1[b]	66.2[c]	64.5[d]		50.9[b]	52.2[c]	52.3[d]	
CLR	Faith *et al.* (2007)	76.2[b]	77.4[c]	76.2[e]	59.1[b]	66.1[c]	64.2[d]	64.0[e]	51.6[b]	52.6[c]	52.4[d]	50.9[e]
ARACNe	Margolin *et al.* (2006)	76.3[b]	77.5[c]	76.7[e]	57.2[b]	64.2[c]	63.5[d]	64.4[e]	50.4[b]	51.3[c]	49.9[d]	49.1[e]
qp graphs	Castelo and Roverato (2009)			69.6[e]				63.5[e]				54.5[e]
GeneNet	Opgen-Rhein and Strimmer (2007)			52.4[e]				59.9[e]				**55.2**[e]

The area under the ROC curve (AUROC) curve is used for the evaluation of inference methods performed by the DREAM5 organizers[a], by the authors of the present article[b,c,d] and from the paper of Narendra *et al.* (2011)[e]. To render the DREAM5 protocol[a,b] (gray background, considering the top 100 k predictions only) comparable to other studies, the performance on the DREAM datasets has been re-calculated[c] with all predictions. Also, publicly available methods have been re-applied[d] to M3D datasets. Methods show similar performance between the DREAM5[c] and M3D[d,e] real (because of their large overlaps) as well as artificial datasets (because they were generated by the same tool, GeneNetWeaver). The best predictions are shown in bold. See Supplementary Material (part 4) for additional scores. All methods were invoked with the designated options to utilize the preselected lists of TFs.

effective utilization of the replicated measurements to quantify the measurement error. The expression profiles of an interaction between the multiple antibiotic resistance (mar, GeneOntology GO:0046677—response to antibiotic) genes marA and marB are a good example of a local correlation between TF and TG. While Pearson's ρ^2 would argue against this edge, it is considered as relevant by η^2 (Fig. 6, panels C and E). The interaction between marA and marB is active in *E. coli* treated with the antibiotic norfloxacin (a gyrase inhibitor). The co-regulation across various gene over-expression experiments in the presence of norfloxacin is depicted in Figure 6E (left side). The treatment with other antibiotics such as ampicillin and kanamycin also triggers the co-regulation of the two mar genes (not shown). The interaction is not active in the experiments on biofilm formation and growth phases as marA and marB exhibit virtually no co-regulation here (Fig. 6E, right side).

In vivo confirmation of novel interactions. Novel candidate interactions in *E. coli* were preselected by applying a 50% precision cutoff to the predictions, i.e. we stop iterating over the list of predictions from most to least confident when the precision evaluated against RegulonDB drops <50%.

Predicted TF:TG interactions were tested by quantifying the presence of the TG mRNA through qPCR amplification in *E. coli* gene knockouts of the corresponding TF. The mRNA levels for the same TG were quantified in non-mutant, wild-type *E. coli* to measure gene expression differences. Expression differences >2-fold for TGs are considered evidence for a true regulatory relationship between the predicted TF:TG pair.

In total, 5 TFs were sampled and 53 interactions not contained in the gold standard were tested. Here, 21 TF:TG pairs showed greater than a 2-fold change corresponding to a confirmation rate of 39.6%. Relaxing the fold change cutoff to 1.8, 26 pairs are reported (precision of 49.1%). This approximately confirms the 50% precision cutoff from the computational analysis. At a precision cutoff of 50% we predict 1995 novel interactions thus expecting $\sim$1000 ($1995 \times 49.1\% = 979$) additional true interactions not contained in RegulonDB. The qPCR experiments were performed by the lab of James J. Collins at the Boston University in the context of the DREAM5 challenge. The full description and analysis of these interactions as well as the participating inference approaches will be the subject of a future paper (Marbach,D., Costello,J. and Küffner,R. *et al.*, The wisdom of crowds for gene network inference, submitted).

4 DISCUSSION

To infer gene regulatory networks (GRNs), we rank the relevance of candidate relationships consisting of a TF and a TG by measuring the dependency between their respective expression profiles.

For the detection of dependencies we proposed the measure η^2 that is derived from an analysis of variance (ANOVA). To our knowledge, η^2 has not been applied to network inference or to other problems in bioinformatics although it has a number of interesting features (Fig. 6). Like Pearson's ρ^2, but in contrast to Bayes conditional probability tables or mutual information, η^2 does not require the discretization of the input data. This increases the robustness of our method as inappropriate discretization might lead to loss of signal. In contrast to Pearson's linear correlation coefficient, η^2 is a non-parametric, non-linear correlation coefficient. It also detects local correlations that are only apparent in a subset of the measured conditions. This increased sensitivity is due to the effective utilization of replicated measurements to model the measurement error.

The recent DREAM5 blind assessment solicited the prediction of GRNs with thousands of genes from two real datasets (*E. coli* and *S. cerevisiae*) and one artificial dataset. The 29 participating teams employed a variety of methods based on regression (Lasso, random forests), Bayesian networks, mutual information and correlation. In DREAM5, our approach was rated the best performer on the inference of real networks and the second best performer on real and artificial networks combined. Especially for the inference of *E. coli* interactions, our approach performed significantly better than the methods evaluated in DREAM5 (Fig. 5) as well as in the large assessment of Narendra *et al.* (2011) (Table 2).

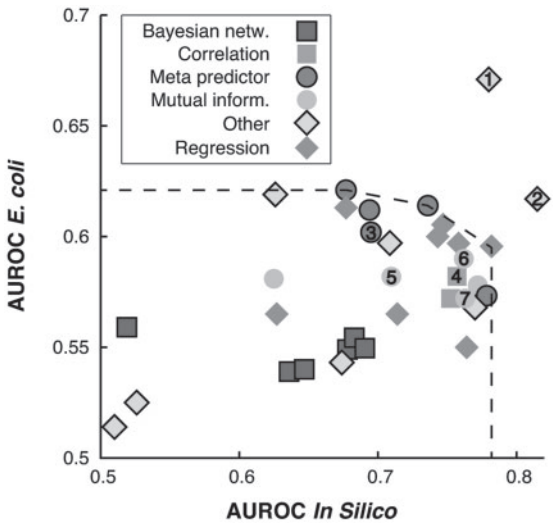

Fig. 5. Comparative prediction performance: artificial versus *E. coli*. DREAM5 participants applied a range of network inference approaches including meta predictors (=combining different approaches) and others (=methods eluding categorization). Some of the participants, particularly η^2 (denoted as 1) and Genie3 (=2) significantly outperformed previously published inference approaches (=4–7, compare Table 2, first two gray columns). The pareto cover of the remaining DREAM5 participants is depicted as dashed line.

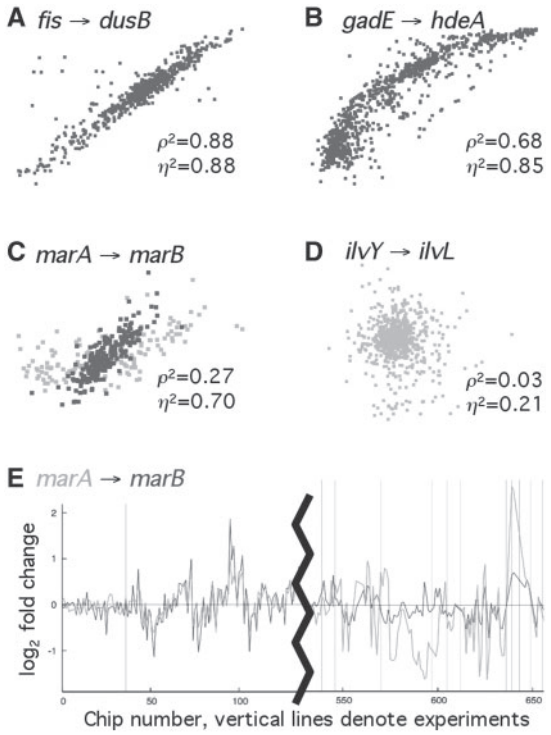

Fig. 6. Correlations between TFs and TGs in *E. coli*. As only mRNA expression data is available, inference depends on the accurate detection of correlations. The plots in panels (**A**)–(**D**) (expression of TF/abscissa scattered against TG/ordinate, data obtained from M3D) depict a series of interactions from RegulonDB that are increasingly difficult to detect. Global linear correlations (A) are easier to detect than non-linear correlations (B). The expression profiles of the multiple antibiotic resistance (mar) genes marA and marB are a good example of a local correlation that is detected by η^2 but not by ρ^2 (C). The correlation becomes visible if *E. coli* is treated with the antibiotic norfloxacin (C: blue dots, E: left side) but not in the experiments on growth phases (C: orange dots, E: right side). No correlation (panel D) might result if the TF itself is not regulated at the level of transcription.

In contrast to *E. coli*, predictions for *S. cerevisiae* received significantly lower scores because the yeast gold standard network is less reliable. Compared with *E. coli* and artifical networks, inference is substantially more difficult in *S. cerevisiae* as here the expression of TF and their regulated genes is hardly correlated (Fig. 4). Indeed, with an AUC between 0.49 and 0.54, predictions were hardly better than guessing. The difficulty of network inference in *S.cerevisiae* has also been recognized by Hu *et al.* (2007) and Narendra *et al.* (2011). Many publications on network inference approaches solely focus on *E. coli* [Faith *et al.* (2007); Mordelet and Vert (2008)].

Some of the known *E. coli* interactions identified by our approach were quite interesting biologically. For instance, an interaction between multiple antibiotic resistance genes was active after antibiotic treatment (local correlation) but not in growth phase experiments. According to qPCR experiments that were performed as part of the DREAM5 conference >50% of our novel predictions represent true interactions. At a precision of 50% we thus expect that our predictions contain 1000 previously unobserved true interactions.

Funding: P.T. and T.P. are partially funded by the DFG (IRTG 1563/1 RECESS and Z.I. 616/3 CLA, respectively). L.W. is partially funded by the Helmholtz Alliance on Systems Biology, Project CoReNe.

Conflict of Interest: none declared.

REFERENCES

Altay,G. and Emmert-Streib,F. (2010a) Inferring the conservative causal core of gene regulatory networks. *BMC Syst. Biol.*, **4**, 132.

Altay,G. and Emmert-Streib,F. (2010b) Revealing differences in gene network inference algorithms on the network level by ensemble methods. *Bioinformatics*, **26**, 1738.

Barrett,T. *et al*. (2010) NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.

Boulesteix,A.-L. and Strimmer,K. (2005) Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach. *Theor. Biol. Med. Model*, **2**.

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

Butte,A. and Kohane,I. (1999) Unsupervised knowledge discovery in medical databases using relevance networks. In *Proceedings of the AMIA Symposium*. American Medical Informatics Association, **7**, 11–15

Castelo,R. and Roverato,A. (2009) Reverse engineering molecular regulatory networks from microarray data with qp-graphs. *J. Comput. Biol.*, **16**, 213–227.

Cohen,J. (1973) Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educ. Psychol. Meas.*, **33**, 107.

Ding,C. and Peng,H. (2005) Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol.*, **3**: 185–205.

Enright,A.J. *et al*. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.

Faith,J.J. *et al*. (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, **5**, e8.

Faith,J.J. *et al*. (2008) Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res.*, **36**, D866–D870.

Friedman,J. *et al*. (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432–441.

Gama-Castro,S. *et al*. (2011) RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res.*, **39**, D98–D105.

Guthke,R. *et al*. (2005) Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection. *Bioinformatics*, **21**, 1626–1634.

Hache,H. *et al*. (2009) Reverse engineering of gene regulatory networks: a comparative study. *EURASIP J. Bioinform. Syst. Biol.*, **2009**, 1–12.

Huerta,A.M. *et al*. (1998) RegulonDB: a database on transcriptional regulation in Escherichia coli. *Nucleic Acids Res.*, **26**, 55–59.

Huynh-Thu,V.A. *et al*. (2010) Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, **5**, e12776.

Hu,Z. *et al*. (2007) Genetic reconstruction of a functional transcriptional regulatory network. *Nat. Genet.*, **39**, 683–687.

Kabir,M. *et al*. (2010) Reverse engineering gene regulatory network from microarray data using linear time-variant model. *BMC Bioinformatics*, **11** (Suppl. 1), S56.

Küffner,R. *et al*. (2010) Petri Nets with Fuzzy Logic (PNFL): reverse engineering and parametrization. *PLoS One*, **5**, e12807.

Kwon,A.T. *et al*. (2003) Inference of transcriptional regulation relationships from gene expression data. *Bioinformatics*, **19**, 905–912.

Lee,W.-P. and Tzou,W.-S. (2009) Computational methods for discovering gene networks from expression data. *Brief. Bioinform.*, **10**, 408–423.

MacIsaac,K.D. *et al*. (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, **7**, 113.

Marbach,D. *et al*. (2009) Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *J. Comput. Biol.*, **16**, 229–239.

Marbach,D. *et al*. (2010) Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl Acad. Sci. USA*, **107**, 6286–6291.

Margolin,A.A. *et al*. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7** (Suppl. 1), S7.

Markowetz,F. and Spang,R. (2007) Inferring cellular networks–a review. *BMC Bioinformatics*, **8** (Suppl. 6), S5.

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

Michoel,T. *et al*. (2009) Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks. *BMC Syst. Biol.*, **3**, 49.

Miller,R. (1997) *Beyond ANOVA: Basics of Applied Statistics*. Chapman & Hall/CRC.

Mordelet,F. and Vert,J.-P. (2008) SIRENE: supervised inference of regulatory networks. *Bioinformatics*, **24**, i76–i82.

Mukherjee,S. and Speed,T.P. (2008) Network inference using informative priors. *Proc. Natl Acad. Sci. USA*, **105**, 14313–14318.

Narendra,V. *et al*. (2011) A comprehensive assessment of methods for de-novo reverse-engineering of genome-scale regulatory networks. *Genomics*, **97**, 7–18.

Opgen-Rhein,R. and Strimmer,K. (2007) From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Syst. Biol.*, **1**, 37.

Prill,R.J. *et al*. (2010) Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PLoS One*, **5**, e9202.

Soranzo,N. *et al*. (2007) Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: synthetic versus real data. *Bioinformatics*, **23**, 1640–1647.

Wang,Y. *et al*. (2006) Inferring gene regulatory networks from multiple microarray datasets. *Bioinformatics*, **22**, 2413–2420.

Wang,Z. *et al*. (2009) An extended Kalman filtering approach to modeling nonlinear dynamic gene regulatory networks via short gene expression time series. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **6**, 410–419.

Zhao,W. *et al*. (2006) Inferring gene regulatory networks from time series data using the minimum description length principle. *Bioinformatics*, **22**, 2129–2135.

Zhu,J. *et al*. (2008) Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat. Genet.*, **40**, 854–861.

Zou,C. and Feng,J. (2009) Granger causality vs. dynamic Bayesian network inference: a comparative study. *BMC Bioinformatics*, **10**, 122.