Bioinformatics, 31(17), 2015, 2836–2843
doi: 10.1093/bioinformatics/btv215
Advance Access Publication Date: 24 April 2015
Original Paper

Systems biology

Addressing false discoveries in network inference

Tobias Petri[1], Stefan Altmann[1], Ludwig Geistlinger[1], Ralf Zimmer[1] and Robert Küffner[1,2,*]

[1]Ludwig-Maximilians-Universität München, Institut für Informatik, Munich, Germany and [2]Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

*To whom correspondence should be addressed.

Associate Editor: Igor Jurisica

Received on December 11, 2014; revised on March 11, 2015; accepted on April 7, 2015

Abstract

Motivation: Experimentally determined gene regulatory networks can be enriched by computational inference from high-throughput expression profiles. However, the prediction of regulatory interactions is severely impaired by indirect and spurious effects, particularly for eukaryotes. Recently, published methods report improved predictions by exploiting the a priori known targets of a regulator (its local topology) in addition to expression profiles.

Results: We find that methods exploiting known targets show an unexpectedly high rate of false discoveries. This leads to inflated performance estimates and the prediction of an excessive number of new interactions for regulators with many known targets. These issues are hidden from common evaluation and cross-validation setups, which is due to Simpson's paradox. We suggest a confidence score recalibration method (CoRe) that reduces the false discovery rate and enables a reliable performance estimation.

Conclusions: CoRe considerably improves the results of network inference methods that exploit known targets. Predictions then display the biological process specificity of regulators more correctly and enable the inference of accurate genome-wide regulatory networks in eukaryotes. For yeast, we propose a network with more than 22 000 confident interactions. We point out that machine learning approaches outside of the area of network inference may be affected as well.

Availability and implementation: Results, executable code and networks are available via our website http://www.bio.ifi.lmu.de/forschung/CoRe.

Contact: robert.kueffner@helmholtz-muenchen.de

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

1 Introduction

Gene regulatory networks (GRNs) consist of interactions of regulators such as transcription factors (TFs) that physically bind to specific nucleotide sequences to regulate the expression of target genes. GRNs can be experimentally derived from TF-binding studies (Gerstein *et al.*, 2012) such as Chromatin Immuno-Precipitation [ChIP, Zheng *et al.* (2010)] or DNase footprinting (Neph *et al.*, 2012). A large fraction of the interactions reported by these approaches are not associated with changes in target expression (Wu *et al.*, 2007). On the other hand, expression changes in potential TF targets can be detected from TF-knockout profiles (Chua *et al.*, 2006). This approach, however, is prone to indirect or spurious effects (Hu *et al.*, 2007).

Although the number of conducted TF-binding and TF-knockout studies is growing (Petricka and Benfey, 2011) the discovery of novel regulations detected with each additional study decreases.

Thus, a combination of experimental results and computational inference approaches is likely to provide more comprehensive networks.

Many inference methods use expression data exclusively. An interaction is predicted if a TF and its putative target are co-expressed. Such *expression-based* approaches infer prokaryotic networks successfully (Faith *et al.*, 2007; Greenfield *et al.*, 2013; Margolin *et al.*, 2006; Michoel *et al.*, 2009). However, they perform hardly better than random for inference of eukaryotic networks (Hu *et al.*, 2007; Küffner *et al.*, 2012; Marbach *et al.*, 2012; Michoel *et al.*, 2009; Narendra *et al.*, 2011; Soranzo *et al.*, 2007; Wu and Chan, 2012), although they can achieve useful results in special cases [e.g. for respiratory genes, Michoel *et al.* (2009)]. Interactions in eukaryotes are difficult to infer as observable dependencies between the expression of regulator and target are weaker and context-dependent. One reason is the increased level of complexity and the combinatorial nature of the eukaryotic regulation of transcription (Neph *et al.*, 2012).

The prediction of novel interactions can be improved for prokaryotic and *in silico* data by exploiting a priori known interactions [local topology priors, Greenfield *et al.* (2013)]. This allows to determine whether a given TF is active based on the expression of its known targets (Ciofani *et al.*, 2012; Naeem *et al.*, 2012) enabling a more reliable prediction of novel targets (De Smet and Marchal, 2010; Mordelet and Vert, 2008; Qian *et al.*, 2003). See Supplementary, Section S5 for an overview on related methods.

Here, we investigate whether eukaryotic networks are accurately inferred by methods exploiting topology priors. First, we demonstrate that many existing performance evaluations are misleading. They are not adequate for local topology methods and overestimate network quality substantially. This effect is due to Simpson's Paradox, well-known in causal theory (Pearl, 2009; Simpson, 1951). Second, this also strongly influences the quality and composition of inferred networks. We develop a simple recalibration strategy and demonstrate how it can be applied for the inference of a confident genome-scale regulatory network in yeast.

2 Materials and methods

Network inference methods score all pairs of regulators and putative target genes to quantify the confidence that a given pair represents a true interaction. For both types of inference methods discussed here, namely expression-based methods and local topology methods, confident predictions are selected by applying a unified cutoff. Expression-based methods are based exclusively on expression data and ignore known interactions. Local topology methods use expression data and known interactions (topology priors) to train a so-called local model per regulator (Fig. 1).

2.1 Data

We obtained five yeast expression compendia (for details see Supplementary, Section S2.1) from (i) the 5th DREAM Challenge [challenge 4, Marbach *et al.* (2012)], (ii) the Many Microbe Microarray Database [M3D, Faith *et al.* (2008)], (iii) the study of Hu *et al.* (2007), (iv) the study of Chua *et al.* (2006) and (v) the Gene Expression Omnibus [GEO, Barrett *et al.* (2011)]. Case-control pairs were selected from 2442 yeast microarrays as described by Küffner *et al.* (2012) to compute $\log_2$ fold-changes. Thereby, we obtained a matrix $M \in \mathbb{R}^{p\times n}$ with $p = 1829$ microarray pairs and $n = 5402$ genes. We normalize M by two successive z-score transformations of rows and columns, respectively.

We then collected experimentally supported interactions from the Yeastract database (Abdulrehman *et al.*, 2011), augmented by a study of MacIsaac *et al.* (2006). We filtered genes that were not contained in the expression data. We excluded TFs regulating less than six known targets to enable training and cross-validation (see later). The resulting reference standard contains 153 TFs, 4870 target genes and 24 462 interactions derived from 356 TF-target binding assays.

2.2 Training and assessment of local topologies

A regulatory interaction network of n genes G is a directed graph $N = (G, I)$, $G = R \cup T$, where R is the set of regulators, T is the set of targets and $I \subseteq R \times T$ are regulatory interactions. Each instance of I is a regulator-target pair $(r, t) \in R \times T$ that is labeled with a weight w_{rt} denoting the number of TF-binding studies that support interaction (r, t).

Machine learning models are trained to predict novel regulations $(r, t) \in R \times T$. Based on the known interactions (Supplementary, Section S2.3.2), each putative regulation is labeled by l_{rt}, where l_{rt} is 1 if $w_{rt} \geq 1$ and 0 otherwise. The matrix of fold-changes $M = (m_{ij}) \in \mathbb{R}^{p\times n}$ represents the feature vectors. The value m_{ij} is the fold-change for gene $j \in G$ in array pair i, and we denote row i by $M_{i\cdot}$ and column j by $M_{\cdot j}$. Then, the feature vector for (r, t) is given by $M_{\cdot t}$ (Supplementary, Fig. S1).

We train $|R|$ local models, each predicting confidence estimates $\hat{c}_{rt}$ specific for a single regulator r of putative regulations (r, t):

$$s_r : \mathbb{R}^p \to \mathbb{R},\ s_r(M_{\cdot t}) = \hat{c}_{rt}. \tag{1}$$

Alternatively, a single *global* model is trained for all regulators using combined feature vectors, i.e. feature vectors of regulator and target are concatenated to represent an interaction (Supplementary, Section S2.3)

$$s : \mathbb{R}^{p+p} \to \mathbb{R},\ s(M_{\cdot r} \oplus M_{\cdot t}) = \hat{c}_{rt}. \tag{2}$$

From all regulations, we build k splits for each model stratified with respect to their label distribution. Cross-validation (here: 3-CV, Supplementary, Section S2.4) is performed by retaining one split at a time and training a model on the remaining $k-1$ splits, so that interactions are either used in evaluation or training, but not both. Every split results in $|R| * |T|$ confidence values $\hat{c}_{rt}$ that score all regulations $(r, t) \in R \times T$. For regulator r, we denote the distribution of these confidence values as D_r (Fig. 1b and c).

The quality of inferred networks is assessed after integrating the predictions across all regulators. Assessment compares predictions to a reference standard of a priori known interactions, for instance by the area under the receiver operator characteristics curve (AUC). An AUC of 1.0 indicates that the confidence scores for the true interactions are higher than those for false positives, while an AUC of 0.5 would indicate random predictions. Such a cross-validated AUC analysis is a standard approach for the assessment of inference methods (Mordelet and Vert, 2008).

2.3 Confidence recalibration

Randomized topologies are generated to share key statistics with the reference standard of known interactions (Fig. 1a and d). We remove all regulations from the network and randomly introduce new regulations until each node k has regained its original in- and out-degree [compare Dorogovtsev and Mendes (2003), p.12]. Further, the association of expression data and genes is shuffled by gene label permutation. For each of the q randomized networks $N^{(1)}, \ldots, N^{(q)}$ we perform a CV prediction to obtain confidence values $\hat{c}_{rt}^{(i)}$ as

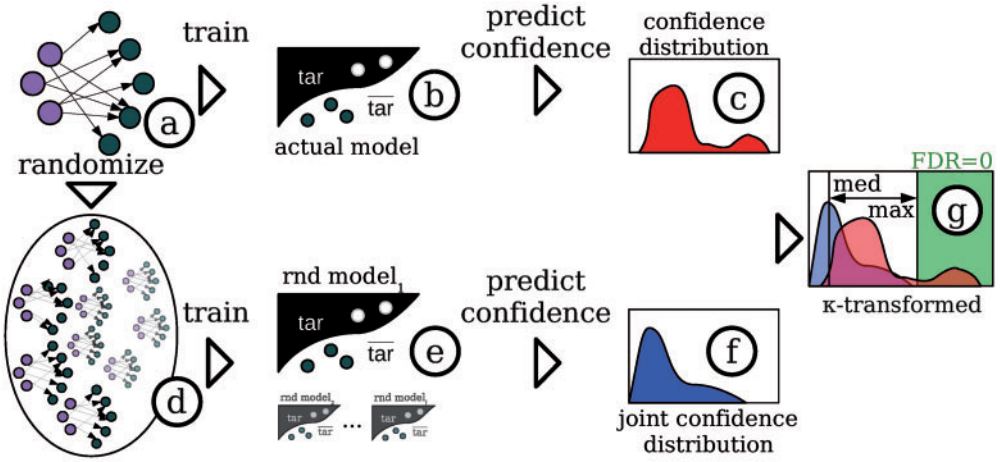

Fig.1. Outline of the recalibration approach. Based on the known network (**a**), a regulator-specific model (**b**) is trained to predict potential targets for this regulator. This results in a confidence score distribution for each regulator (**c**). Additionally, we generate random networks (**d**) maintaining in- and out-degrees from the original network and train models (**e**) for each random topology in the same way as for the original network. For each TF out-degree, we combine resulting random confidence scores into a joint distribution (**f**). Finally, we compare the two distributions c and f based on their respective medians (med) and maxima (max). We minimize false discoveries by selecting regulations [shaded area in (**g**)] that exceed values observed for random networks

described earlier (Fig. 1). Let $D_r^{(i)}$ the distribution of confidence values specific to a regulator r computed using the random prior $N^{(i)}$. We then compute a joint distribution D'_r that encompasses all confidence values derived from random networks that are associated to regulators of the same out-degree (Fig. 1f).

D'_r denotes the *randomized complement* of D_r. By comparing these two distributions we select interactions with scores higher than those observed in the randomized case. Each regulation's confidence $\hat{c}_{rt}$ is replaced by its complement κ_{rt} (Fig. 1c and g):

$$\kappa_{rt} = \frac{\hat{c}_{rt} - \text{med}(D'_r)}{\max(D'_r) - \text{med}(D'_r)}. \quad (3)$$

Scores are thus recalibrated based on the median confidence med (D'_r) and the distribution scale $(\max(D'_r) - \text{med}(D'_r))$. A κ value above 1.0 corresponds to a false discovery rate (FDR) of 0, i.e. to confidence estimates not achieved in random topologies.

3 Results

3.1 Simpson's paradox

We followed the SIRENE approach [Mordelet and Vert (2008); Section 2] and trained local models based on support vector machines (SVMs) to predict confidence values for potential regulations. On a large expression dataset of 2442 yeast microarrays and a regulatory network of 24 462 interactions (Section 2.1) the cross-validated predictions achieved a network-wide AUC of 0.784.

However, we found this standard, cross-validated AUC analysis misleading in case of methods integrating topology priors. We demonstrated this by training the methods on randomized networks (random re-assignment of targets to regulators). The confidence scores for individual regulators are random, resulting in regulator-specific AUC values of 0.5 (Supplementary, Section S2.3.3). Strikingly, an evaluation across all regulators yielded an AUC of 0.798, a score above the AUC achieved by SIRENE.

These two results seem to be in conflict: a method that performs randomly for each regulator induced subnetwork should yield random overall performance as well. This effect resembles the Simpson's or 'amalgamation' paradox (Pearl, 2009; Simpson, 1951): each of the regulator-specific distributions achieves an AUC of 0.5, while the AUC of the joint distribution suggests non-random performance (compare Supplementary, Section S3.1).

Here, the paradox results from the fact that predicted confidence score distributions are heterogeneous across regulators and are characterized by different scale and location parameters (Fig. 2a, light gray boxes). In particular, score distributions for regulators with many known targets (high out-degree) such as *ste12* are wider and systematically above average. We refer to this effect as High Degree Preference (HDP). These regulators contribute many true positives, i.e. after the integration higher scores become enriched for true positives. This in turn leads to non-random AUC values. Selected high-scoring predictions thus remain unspecific while biologically more specific signals are likely being missed (Pavlidis and Gillis, 2013). Following this line of argument, the regulator out-degree confounds the integration of confidence values. This is consistent with results demonstrated for the prediction of genes involved in biological processes (Gillis and Pavlidis, 2011).

To examine whether the paradox is an artifact of SVMs we trained further model classes (e.g. decision trees and logistic regression; Supplementary, Sections S2.6 and S3). We observed similar effects across all examined techniques, suggesting that regulator-specific methods using topology priors are generally affected by an HDP.

Besides the confounding of network quality measures, the composition of predicted networks is also affected. We predicted networks by selecting high-scoring interactions using a threshold determined from the estimated size of the complete yeast network (Supplementary, Sections S2.2 and S3.2), which should be twice as large as the known network. A score threshold was chosen so that selected regulations contain 50% previously confirmed ones (the Precision-50, or P50 network).

For a regulator with out-degree d we obtained two types of score distributions: (i) from the model trained on its known targets and (ii) from models trained on the targets of randomized regulators with out-degree d (Fig. 2a). A unified cutoff selects an excessive number of predictions for high-degree TFs that overlap with random scores. To quantify this, we computed the FDR based on the number of interactions scored above the P50 threshold in distribution (ii) divided by the total number of interactions above that threshold in (i) and (ii).

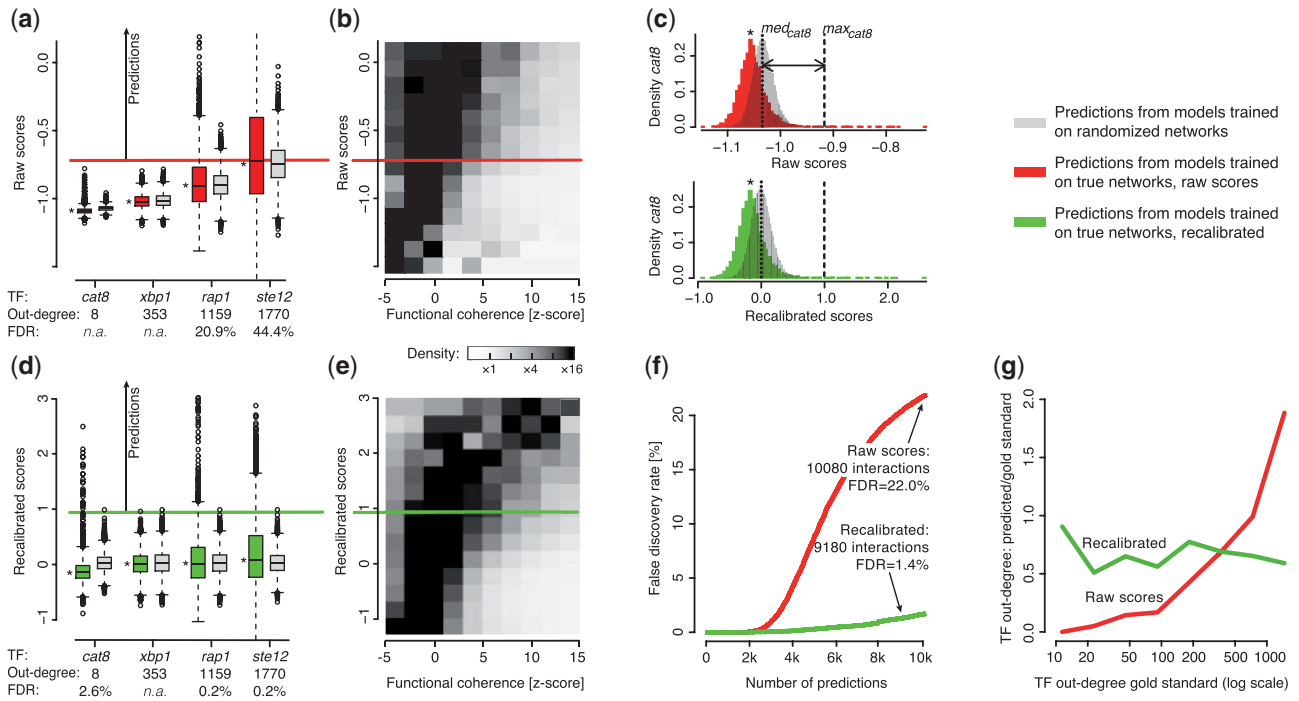

Fig. 2. Score recalibration in network predictions. (**a**) We trained SVMs for each TF (compare Supplementary, Fig. S1). Putative target genes were selected by a threshold (horizontal line) on the resulting TF-specific scores (boxplots marked by asterics). Additional SVMs were trained on random networks (light gray) and FDRs were computed for all regulators but those such as *xbp1* where no predictions were made. (**b**) The density map displays whether predicted and known targets of the same TF overlap in their biological function. Positive *z*-scores (abscissa) indicate significant function overlaps for corresponding scores (ordinate). (**c**) Score distributions (marked by asterics) were recalibrated via randomized distributions (light gray): for each TF, the median *med* (dotted line) and maximum *max* (dashed line) are mapped to 0.0 and 1.0, respectively. Panels (**d**) and (**e**) show boxplots, threshold, and a density map of function overlap after recalibration. Panel (**f**) plots the FDR as a function of the number of predicted interactions. Arrows indicate the number of interactions achieving a precision of at least 50% (P50). In (**g**), the ratio of predicted to gold standard targets (ordinate) is depicted across the range of TF out-degrees (=number of gold standard targets, abscissa) before and after recalibration

For example, the FDR is 44.4% for high-degree *ste12* and 22% across all TFs (Fig. 2f), which is unacceptably high. In contrast to *ste12*, all predictions are rejected in case of low-degree TFs such as *cat8*, even if they substantially exceed random scores (Fig. 2a). Only 81 of 153 TFs (53%) receive predictions. We concluded that neither cross-validation nor AUC analysis are sufficient to ensure the overall quality of networks inferred using structural priors.

We also assessed whether TFs frequently regulate the expression of targets that share similar biological functions (Segal *et al.*, 2003). We therefore tested whether known and predicted targets of the same TF exhibit substantial functional overlaps (Supplementary, Section S2.8). We observed that the high proportion of random scores (e.g. for *ste12*) concealed most of the signal as interactions with higher scores hardly showed an increased functional coherence (Fig. 2b).

3.2 Correction through score recalibration

We introduce a confidence recalibration (CoRe) as a wrapper for existing methods (Section 2.3). Based on the random networks, we derived expected location (median score) and scale (maximum score) properties for each out-degree *d* and used them to transform the predicted confidences into topology-corrected scores. Scores for each regulator are recalibrated by scaling the median and maximum scores to 0 and 1, respectively (Fig. 2c). This renders score distributions comparable so that they can be integrated across TFs. The FDR is then 0 for predictions with scores above 1 as they appear only for the true but not for the randomized networks. Thus, interactions for each regulator selected after CoRe are scored above the random level.

To obtain a P50 network, we select interactions that achieve a corrected score of > 0.92. The FDR for this network was reduced to 1.4% (as compared with 22.0% without recalibration). We observed that predictions are now balanced across TF degrees (Fig. 2g), predicting interactions for 138 TFs versus 81 without recalibration.

To gain further insight in the nature of the corrected network, we estimated the functional relationship between known and novel predicted targets (Supplementary, Section S2.8). Regulatory patterns were more coherent for the corrected network (compare Figs 2b and e).

3.3 Application of CoRe to network inference

For all subsequent methods and analyses we report corrected results. To evaluate the yeast regulatory network obtained, we conducted a comparative assessment of frequently used inference approaches and a consensus approach (Supplementary, Sections S2.3.4 and S2.6). The approaches are roughly classified by five attributes (Fig. 3a and Supplementary, Section S5):

1. **method:** unsupervised expression-based (Faith *et al.*, 2007) versus supervised using a structure prior;
2. **formulation:** one-class (Mordelet and Vert, 2010) versus two-class that treat unknown interactions as informative;
3. **strategy:** lazy (Supplementary, Section S5) versus parameterized models;
4. **data handling:** non-integrative versus integrative e.g. using TF-binding site preferences (Ernst *et al.*, 2008);
5. **models:** global (Yip *et al.*, 2009) versus local (regulator-specific).

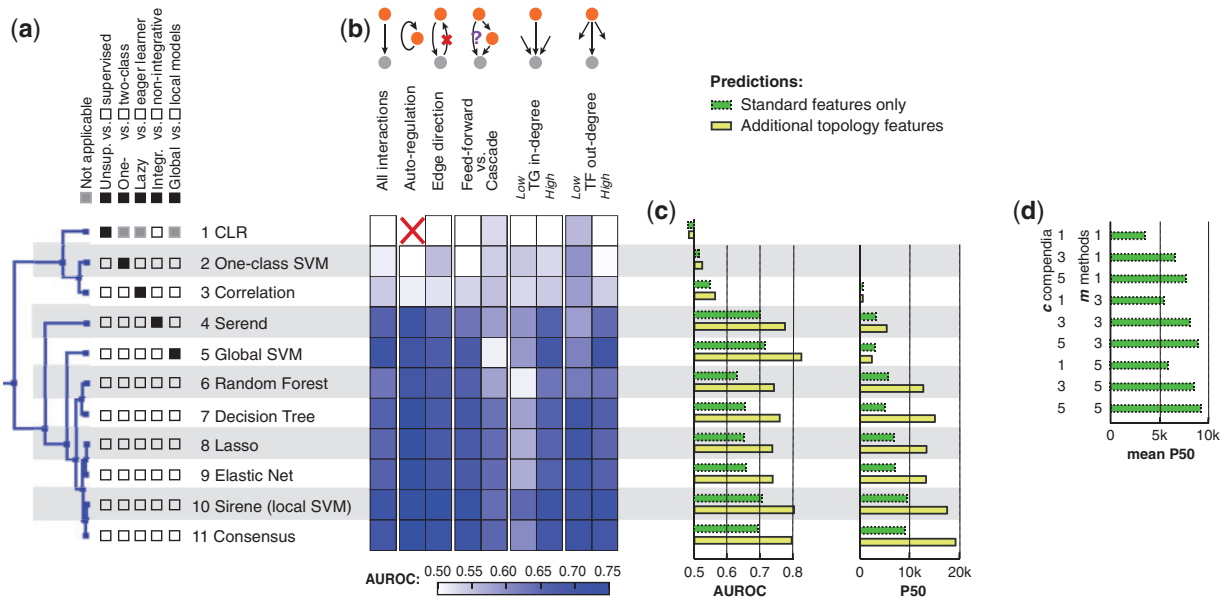

Fig. 3. Assessment of predicted interactions. We analyzed the predictions of 11 different inference methods across five yeast gene expression compendia. (**a**) The dendrogram groups methods according to the similarity of their predictions. Properties that discriminate between different classes of methods are indicated by the check boxes. Panel (**b**) shows if interactions in particular network motifs are easier (dark) or harder (light) to detect in comparison to all interactions. In (**c**) method performance (AUC) and the number of interactions predicted at a precision of 50% or better (P50) (bars with dotted borders) are assessed. Furthermore, we encoded experimentally determined targets of TFs into additional features (bars with solid borders, Supplementary, Sections S2.6.8 and S3.9) and integrated methods 6–10 into a consensus approach (method 11). Panel (**d**) illustrates mean results from integrating all subsets of $c = 1...5$ compendia and $m = 1...5$ methods. All results are based on recalibrated scores

SIRENE (Mordelet and Vert, 2008) is a supervised, two-class, parameterized, non-integrative, local approach. For all methods, we predicted confidence scores in a 3-CV scheme and recalibrated them as described earlier.

Subsequently, we analyzed network motifs (Supplementary, Section S2.7) to capture method- and topology-specific preferences (Fig. 3b). Unsupervised, expression-based approaches do not use topology priors but infer interactions if expression profiles of TFs and putative targets are mutually dependent. An example is CLR (context likelihood of relatedness, Faith *et al.*, 2007). These methods are unable to detect auto-regulation as in this case both expression profiles would be identical. Confirming previous findings (Marbach *et al.*, 2012), expression-based approaches could hardly detect feed-forward motifs or the correct direction of interactions. In contrast, regulator-specific approaches were less affected by such difficult cases and exhibited a consistently higher performance. For cascades and low in-degree targets, a slight decrease in performance was observed. Potentially, the latter indicated the prediction of novel regulators for genes that were less well-studied previously.

Next, we evaluated the performance of approaches across all interactions. Expression-based, one-class, and lazy learners performed substantially worse than the remaining methods (Fig. 3c). We observed that integrative methods like Serend (Ernst *et al.*, 2008) suffered from false positive predictions. This is likely due to the low specificity of positional weight matrices [PWMs, Holloway *et al.* (2008)] predicting targets only for 6.5% of all TFs (Supplementary, Section S3). These methods were not further analyzed. Of the remaining five methods (methods 5–12 in Fig. 3), the best results were obtained from regulator-specific SVMs and decision trees trained on bootstrap samples (bagging). See Supplementary, Section S3.9 for methodological extensions such as the integration of multiple predictors to perform a consensus prediction.

3.4 A comprehensive yeast network

Our final yeast network includes 22 231 interactions with 153 TFs and 3747 target genes. Of all predicted regulations, 12 869 are contained in the reference standard, while 9362 are novel predictions. The remaining $24,462 - 12,869 = 11,593$ reference standard interactions (Supplementary, Fig. S4a) lacked an observable effect on expression and were thus not included.

The visualization and interpretation of organism-wide networks is challenging due to their size and complexity. Instead of fully depicting each regulator, target and their interactions, we employed a modular visualization. We derived regulatory modules by grouping TFs with overlapping target sets and, vice versa, target modules by grouping genes regulated by overlapping sets of TFs. We connected regulator and target modules via *meta-interactions* if >40% of all induced regulator-target pairs were connected. This reduced representation featured 13 meta-interactions among 9 target and 9 regulatory modules, capturing half of the final interactions (11 232 interactions, 50.5% of all predicted). See Figure 4 for an excerpt (full details are in Supplementary, Sections S2.9 and S6, accessible through clickable maps, see availability).

This modular view enables an integrated display of the network as well as module-associated expression profiles. Given current data and knowledge, the respective TF-modules likely control the forming of transcriptional response patterns in the regulated target modules. Some key aspects of module-associated expression profiles are summarized below (for a comprehensive literature review on all network modules see Supplementary, Section S6). A representative gene was selected manually for each module.

The *hxt2* module features the most versatile regulation in our network, regulated by three different TF clusters comprising the highest total number of TFs (Fig. 4). According to GO (The Gene Ontology Consortium, 2010), most of the 190 genes of the *hxt2*

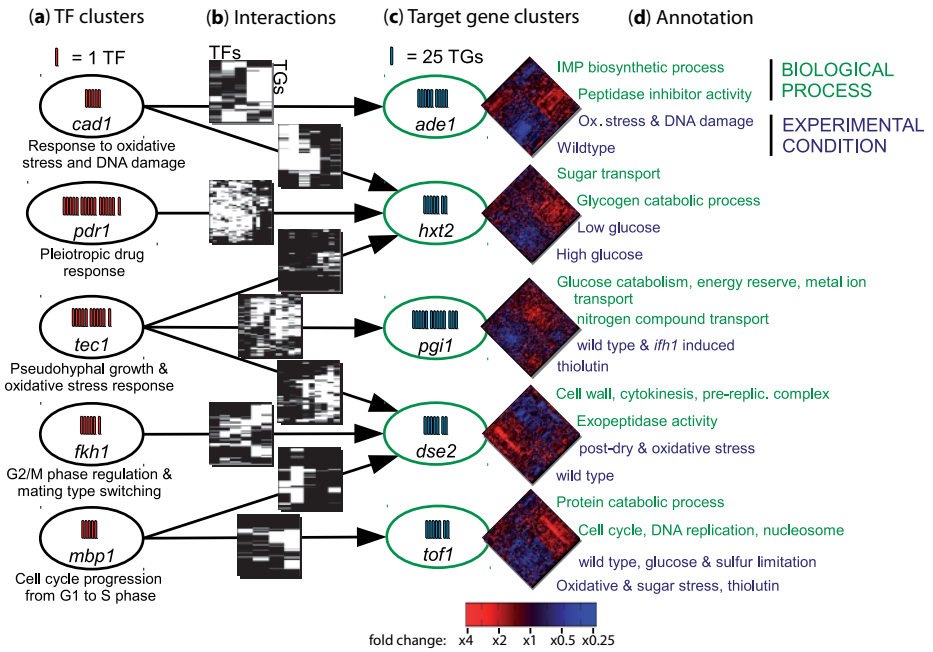

Fig. 4. Interactions and expression profiles. We partitioned our network of 22 231 gene regulatory interactions for visualization and identification of network modules. We derived (**a**) 9 clusters of 61 TFs that, via (**b**) 13 interactions between clusters (arrows), regulate (**c**) 9 clusters of 1758 target genes (excerpt of five TF and five target clusters connected by nine interactions shown here, Supplementary, Section S6 for full figure). A representative gene is displayed for each TF and target cluster. Cluster interaction maps (black = interaction, white = no interaction) comprise a total of 11 232 (50.5%) interactions. (**d**) Thus, depicted TF-modules are likely to trigger expression responses in respective target modules and associated processes (top part of heatmap). The heatmaps display the differential expression of these target modules under the indicated knockout (KO) and other experimental conditions (bottom part) (Color version of this figure is available at *Bioinformatics* online.)

cluster belong either to sugar transport (*hxt* genes) or glycogen metabolic process (e.g. *gac1*). Consequently, we observe differential expression of these genes under low- versus high-glucose growth conditions. When glucose is available, the sugar transporters are abundantly expressed (Ozcan and Johnston, 1995), whereas under glucose starvation glycogen storage is catabolized to produce glucose preferably for fermentation (François and Parrou, 2001).

The *pdr1* (pleiotropic drug response) cluster comprised the largest number of *hxt2* regulators. It consisted of 16 TFs, all tightly connected to the cellular response to drug and nutrition stress such as differing glucose concentrations. Despite this general response mediated by the *pdr* TFs (*stb5* and *msn1*), much of the regulation was performed by pseudohyphal growth TFs (*nrg1*, *mga1* and *ash1*) in conditions of nitrogen limitation and abundant fermentable carbon sources like glucose (Lorenz and Heitman, 1998).

Interestingly, a strong regulatory impact on the *hxt2* module was also observed for regulators of the oxidative stress response—on the one hand from the *cad1* cluster (5 TFs, also responding to resulting DNA damage), and, on the other hand, from the *tec1* cluster (11 TFs, also driving pseudohyphal growth). Oxidative stress results in cellular protection mechanisms, e.g. DNA repair and targeted protein degradation, which is associated with increased energy consumption (Morano *et al.*, 2012), initiated by the *hxt2* cluster via increased glucose uptake.

4 Discussion

GRNs are crucial to understand how regulators like TFs affect their target genes on the expression level. Experimentally derived networks are typically incomplete as the number of available experiments is limited. To complement them, computational inference of networks has been introduced. We revealed critical aspects but also demonstrated that inference is necessary and feasible in eukaryotes.

Even in well-studied eukaryotes such as yeast, where ~900 publications on experimental TF-binding studies are available, current networks are far from complete and benefit from computational predictions. We found that only about half of all regulations that induce detectable expression changes ('active' interactions) are currently known. In addition, experimental techniques are prone to discover regulations without effect on the expression level. We applied computational inference both for the detection of novel active and the pruning of inactive regulations.

We reported three crucial findings based on the analysis of a wide spectrum of data-driven inference methods (for reviews see De Smet and Marchal 2010; Myers *et al.*, 2006). First, we demonstrated that methods incorporating experimentally derived interactions as topology priors possess sufficient predictive power for the inference of eukaryotic networks. Methods using expression data alone fail here (Marbach *et al.*, 2012; Narendra *et al.*, 2011). We also showed that topology priors lead to Simpson's paradox (Pearl, 2009; Simpson, 1951) distorting prediction and assessment of regulatory interactions. Finally, we showed how to avoid the occurrence of the paradox.

Generally, network inference methods that exploit the local topology assign an excessive number of predictions to TFs with many known targets (Ambroise *et al.*, 2012; De Smet and Marchal, 2010), and it has been doubted whether a correction is possible or sensible (Gillis and Pavlidis, 2011; Myers *et al.*, 2006). Our analysis revealed that the number of known targets for a regulator is a confounder of regulator-target predictions. This effect is not detected by common

cross-validation routines: surprisingly, the same performance reported for published network inference approaches can be achieved by guessing random regulations. We developed a CoRe approach wrapping existing methods and showed that it corrected for both the over-estimation of performance and the distortion of the topology toward TFs with many known targets (HDP).

We conducted a comprehensive assessment of methods integrating topology priors and identified methods suitable to derive a corrected, accurate yeast regulatory network of active regulations. We describe disadvantages of several methods, which we excluded due to prediction performance, or the inadequate scale-up for large expression datasets. Our evaluation suggested that the selected methods detect several types of interactions successfully that are difficult to predict. For instance, auto-regulatory interactions and the assignment of directions are handled accurately, and immediate and indirect interactions could be distinguished. We then integrated the predictions from the selected methods to construct a network consisting of half novel and half experimentally determined regulations. This choice was based on our extrapolation of the size of the complete yeast network. Our final yeast network (see availability) contains 153 TFs that regulate 3747 target genes via 22 231 interactions. These include many novel and confident hypotheses of regulatory relationships, while we expect less than 150 false positives in total. At the same time, we reject more than half of the experimentally determined interactions as they appear to be without observable regulatory effect.

To gain an overview of the network, we derived modules of target genes that were jointly regulated by sets of TFs. The resulting modular structure was strikingly simple featuring 13 meta-regulations that represent an index for inspecting the expression effects of interactions. A thorough literature review confirmed that the modules and their expression patterns correspond well to biological processes such as respiration, sulfate/energy metabolism, transport, stress response and cell division.

We conclude that methods integrating local topology can extend known networks substantially and at a high reliability, even in well-studied model organisms. These methods, in contrast to those using expression data alone, are well-suited for the prediction of interactions in yeast and presumably other eukaryotes. Due to Simpson's paradox, however, their application was more difficult than previously acknowledged and required a correction approach. We emphasize that topology, structural priors and parameterized models are widely applied beyond network inference (Supplementary, Section S5 for an overview) and encourage a review of fields that may benefit from confidence recalibration strategies such as *CoRe*.

Funding

Parts of this work have been funded by the BMBF under project AgroClustEr/Phaenomics FKZ 0315536B and the Bavarian Research Network of Molecular Biosystems (BioSysNet). LG has been supported by the DFG, GRK 1563. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest: none declared.

References

Abdulrehman,D. *et al.* (2011) YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in Saccharomyces cerevisiae through a web services interface. *Nucleic Acids Res.*, **39**(Database issue), D136–D140.

Ambroise,J. *et al.* (2012) Transcriptional network inference from functional similarity and expression data: a global supervised approach. *Stat. Appl. Genet. Mol. Biol.*, **11**, 1–24.

Barrett,T. *et al.* (2011) NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Res.*, **39**(Database issue), D1005–D1010.

Chua,G. *et al.* (2006) Identifying transcription factor functions and targets by phenotypic activation. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 12045–12050.

Ciofani,M. *et al.* (2012) A validated regulatory network for th17 cell specification. *Cell*, **151**, 289–303.

De Smet,R. and Marchal,K. (2010) Advantages and limitations of current network inference methods. *Nat. Rev. Microbiol.*, **8**, 717–729.

Dorogovtsev,S.N. and Mendes,J.F. (2003) *Evolution of Networks: From Biological Nets to the Internet and WWW.* Oxford University Press, Oxford, UK.

Ernst,J. *et al.* (2008) A semi-supervised method for predicting transcription factor-gene interactions in Escherichia coli. *PLoS Comput. Biol.*, **4**, e1000044.

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**(Database issue), D866–D870.

François,J. and Parrou,J.L. (2001) Reserve carbohydrates metabolism in the yeast Saccharomyces cerevisiae. *FEMS Microbiol. Rev.*, **25**, 125–145.

Gerstein,M.B. *et al.* (2012) Architecture of the human regulatory network derived from encode data. *Nature*, **489**, 91–100.

Gillis,J. and Pavlidis,P. (2011) The impact of multifunctional genes on "guilt by association" analysis. *PLoS One*, **6**, e17258.

Greenfield,A. *et al.* (2013) Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics*, **29**, 1060–1067.

Holloway,D.T. *et al.* (2008) Classifying transcription factor targets and discovering relevant biological features. *Biol. Direct*, **3**, 22.

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

Küffner,R. *et al.* (2012) Inferring gene regulatory networks by ANOVA. *Bioinformatics*, **28**, 1376–1382.

Lorenz,M.C. and Heitman,J. (1998) Regulators of pseudohyphal differentiation in Saccharomyces cerevisiae identified through multicopy suppressor analysis in ammonium permease mutant strains. *Genetics*, **150**, 1443–1457.

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

Marbach,D. *et al.* (2012) Wisdom of crowds for robust gene network inference. *Nat. Methods*, **9**, 796–804.

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.

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

Morano,K.A. *et al.* (2012) The response to heat shock and oxidative stress in Saccharomyces cerevisiae. *Genetics*, **190**, 1157–1195.

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

Mordelet,F. and Vert,J.-P. (2010) A bagging SVM to learn from positive and unlabeled examples. *Technical report arXiv:1010.0772*. Cornell University Library.

Myers,C.L. *et al.* (2006) Finding function: evaluation methods for functional genomic data. *BMC Genomics*, **7**, 187.

Naeem,H. *et al.* (2012) Rigorous assessment of gene set enrichment tests. *Bioinformatics*, **28**, 1480–1486.

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

Neph,S. *et al.* (2012) Circuitry and dynamics of human transcription factor regulatory networks. *Cell*, **150**, 1274–1286.

Ozcan,S. and Johnston,M. (1995) Three different regulatory mechanisms enable yeast hexose transporter (HXT) genes to be induced by different levels of glucose. *Mol. Cell. Biol.*, **15**, 1564–1572.

Pavlidis,P. and Gillis,J. (2013) Progress and challenges in the computational prediction of gene function using networks: 2012–2013 update. *F1000Res*, **2**, 230.

Pearl,J. (2009) *Causality.* 2nd edn. Cambridge University Press, New York, NY, USA.

Petricka,J.J. and Benfey,P.N. (2011) Reconstructing regulatory network transitions. *Trends Cell Biol.*, **21**, 442–451.

Qian,J. *et al.* (2003) Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics*, **19**, 1917–1926.

Segal,E. *et al.* (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, **34**, 166–176.

Simpson,E. (1951) The interpretation of interaction in contingency tables. *J. R. Stat. Soc. Ser. B (Methodol.)*, **13**, 238–241.

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.

The Gene Ontology Consortium. (2010) The gene ontology in 2010: extensions and refinements. *Nucleic Acids Res.*, **38**(Database issue), D331–D335.

Wu,M. and Chan,C. (2012) Learning transcriptional regulation on a genome scale: a theoretical analysis based on gene expression data. *Brief Bioinform.*, **13**, 150–161.

Wu,W.-S. *et al.* (2007) Identifying regulatory targets of cell cycle transcription factors using gene expression and ChIP-chip data. *BMC Bioinformatics*, **8**, 188.

Yip,K.Y. *et al.* (2009) Multi-level learning: improving the prediction of protein, domain and residue interactions by allowing information flow between levels. *BMC Bioinformatics*, **10**, 241.

Zheng,W. *et al.* (2010) Genetic analysis of variation in transcription factor binding in yeast. *Nature*, **464**, 1187–1191.