Gene expression

biRte: Bayesian inference of context-specific regulator activities and transcriptional networks

Holger Fröhlich

University of Bonn, Institute for Computer Science, Römerstr. 164, 53117 Bonn, Germany

Associate Editor: Igor Jurisica

Received on February 3, 2015; revised on May 13, 2015; accepted on June 15, 2015

Abstract

In the last years there has been an increasing effort to computationally model and predict the influence of regulators (transcription factors, miRNAs) on gene expression. Here we introduce biRte as a computationally attractive approach combining Bayesian inference of regulator activities with network reverse engineering. biRte integrates target gene predictions with different omics data entities (e.g. miRNA and mRNA data) into a joint probabilistic framework. The utility of our method is tested in extensive simulation studies and demonstrated with applications from prostate cancer and Escherichia coli growth control. The resulting regulatory networks generally show a good agreement with the biological literature.

Availability and implementation: biRte is available on Bioconductor (http://bioconductor.org).

Contact: frohlich@bit.uni-bonn.de

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Gene regulation is one of the most important biological processes in the living cell. Its malfunctioning is directly associated with many human diseases. Transcription factors (TFs) are one of the most relevant drivers of mRNA expression. In addition, miRNAs can degrade or inhibit translation of target mRNAs on the post-transcriptional level. Thus both, TFs and miRNAs, influence mRNA concentration jointly together in complex networks.

Based on existing TF and miRNA target gene prediction methods during the last years there has been an increasing effort to computationally model regulatory networks and draw conclusions about context specific regulator activities (Engelmann and Spang, 2012; Geeven et al., 2012; Lim et al., 2009; Setty et al., 2012). Geeven et al. (2012) also considered statistical interactions between regulators, which they depicted as networks. In our earlier publication we devised BIRTA as a method, which combines miRNA and mRNA expression data into a joint probabilistic model to infer condition specific activities of TFs and miRNAs (Zacher et al., 2012). BIRTA is formulated in a Bayesian regression framework and uses a so-called spike and slab prior (George and McCulloch, 1997) to infer probabilities of activation for transcription factors and miRNAs in a sparse manner. BIRTA does not support statistical interactions between regulators.

Independent of the question to predict context specific regulator activities, structure learning of gene regulatory networks directly from expression data has been widely studied (Friedman et al., 2000; Huynh-Thu et al., 2010; Margolin et al., 2006; Sachs et al., 2005).

In this article, we introduce biRte as a combined approach to estimate context specific regulator activities and networks between these regulators. biRte first uses a significant improvement of BIRTA to infer activities of regulators and their statistical interactions from expression data and target gene predictions. A key difference to BIRTA is a computationally efficient analytical marginalization of regression coefficients, which yields far better prediction performance and 10–15 fold less computation time. After estimation of active regulators, biRte employs Nested Effects Model (NEM) (Markowetz et al., 2005) structure learning to infer the associated transcriptional network. A key difference to many existing methods at this point is that our approach does not rely on direct mRNA measurements of regulators, but employs information of differential expression of downstream target genes.
2 Overview about biRte

The minimum input to biRte consists of (differential) gene expression data together with a genome-wide regulator-target gene network. Based on that biRte addresses two goals (Fig. 1):

1. Estimation, which regulators (miRNAs, TFs and possibly other factors) exhibit a significant influence on the expression of their target genes in a specific biological context. We call a regulator active, if such a significant influence is present and inactive otherwise. biRte also supports the inference of statistical two-way interactions between regulators.

2. Estimation of the network between active regulators based on observed nested subsets of differentially expressed target genes.

Besides gene expression data biRte can make use of direct regulator measurements (e.g. miRNA expression), if available. Gene expression data of transcription factors is treated specially: differential expression of these factors may indicate positive evidence that the respective regulator has a different activity in one compared to another experimental condition. This positive evidence can be exploited in the integrated likelihood model of biRte.

3 BiRte step I: estimation of active regulators

3.1 Likelihood model for data integration

In order to estimate the active influence of regulators on gene expression a sufficient likelihood model is required. The basic intuition behind our likelihood model is two-fold:

- If the set of active regulators is already known, then mRNA expression can be predicted from regulator activities via a sparse linear model.
- We do not assume a direct correlation between the expression level of regulators and mRNA. However, we suppose that in case a regulator is active in one and inactive in another condition, then there should be an observable shift of the expression level of the regulator itself. In other words, differential regulator expression increases the likelihood to obtain differential activity between two conditions.

Let \( n \) be the set of all measured mRNAs. With biRte we assume that these mRNAs have been measured in \( C \) conditions (in our software \( C \leq 2 \)) with \( R \) replicates. Let \( O \) be the set of all resulting mRNA expression levels (after appropriate normalization and transformation of data). Let \( A \) denote the set of all mRNA expression levels with \( R \) being the respective number of replicates per condition. Possibly further available TF expression data is denoted by \( T \) and other data (e.g. CNVs) by \( Z \). The notation for the number of measured molecular entities and replicates is done in consistency with the previously introduced notation for mRNAs and miRNAs.

In our model we distinguish between hidden regulator states (active vs. inactive) and observable measurements. We denote the set of all available experimental data as \( D \) here. Given the set of active regulators \( R = \{R_1, R_2, ..., R_C\} \) in each experimental condition (\( c = 1, 2, ..., C \)) (which is equivalent to knowing the hidden variables) together with further model parameters \( \Theta \) we decompose the likelihood of \( D \) as:

\[
L_{D, \Theta}(\mathcal{R}) = p(D|\mathcal{R}, \Theta) = \prod_{D \in \{O, A, T, X\}} p(D|\mathcal{R}, \Theta) \tag{1}
\]

\[
= \prod_{D \in \{O, A, T, X\}} \prod_{c=1}^{C} \prod_{i=1}^{R_c} p(D_{ic}|\mathcal{R}_c, \Theta) \tag{2}
\]

where \( D_{ic} \) is the \( i \)th feature and the \( r \)th replicate measured under condition \( c \) in data type \( D \).

This formulation is slightly more general than the one used in Zacher et al. (2012). It allows for including further data types (\( Z \)) and captures the situation that relative expression (log fold changes) are measured. In this case \( C = R_D = 1 \). Using relative expression data allows to apply biRte to arbitrary complex statistical designs.

Let \( T(A) \) denote the target genes of regulator \( A \). Two-way interactions between regulators \( A, B \) can be formally introduced as additional regulators \( A : B \) that only target genes lying within \( T(A) \cap T(B) \). Since the number of potential two-way interactions scales quadratically with the number of regulators, in practice we only consider \( A : B \) if \( \frac{T(A) \cap T(B)}{T(A) \cup T(B)} \) lies within a defined range (default: 0.1 to 0.8).

Next we specify, how the data specific (marginal) likelihoods \( p(D_{ic}|\mathcal{R}_c, \Theta) \) are computed.

3.2 Marginal likelihood for mRNA expression

In agreement with our previous paper each observed mRNA measurement is assumed to be determined by the set of active regulators:

\[
p(O_{ic}|\mathcal{R}_c, \omega_c; \nu^2) = N(\omega_{ic} + \sum_{q \in \mathcal{R}_c} \omega_{qc}; \nu^2) \tag{3}
\]

were \( \mathcal{R}_j \) denotes the set of active regulators of gene \( j \). The equation can be interpreted as a linear regression with coefficients \( \omega_{qc} \) and intercept \( \omega_{ic} \). The formulation assumes a constant noise variance for all genes, which can be achieved after appropriate scaling of the data.

In contrast to our previous paper (Zacher et al., 2012) regression coefficients in biRte are not sampled, but marginalized out analytically; Let \( p := |\mathcal{R}| \) and \( \omega := (\omega_{00}, \omega_{01}, ..., \omega_{0p}) \). Let \( \phi \) be the vector of expression values for a particular replicate \( r \) under condition \( c \). Then \( E[\phi|X_p, \omega_0] = X_p \omega_0 \), where \( X_p \) is a \( n \times p \) design matrix (the first column is constantly 1 for the intercept). Using conjugate priors \( \nu \sim IG(\lambda, \delta) \) and \( \omega_{0r} | \nu^2 \sim N(0, \nu^2 \lambda^{-1}) \) the marginal distribution \( p(\phi|\mathcal{R}_c) = p(\phi|X_p) \) (also called model evidence) can be written down in closed form (Gelman et al., 2004):

\[
p(\phi|X_p) = \frac{1}{(2\pi)^{n/2} \nu^{n/2}} \sqrt{\det(\nu I)} \exp\left( -\frac{1}{\nu} \phi' \nu^{-1} \phi - \frac{1}{2} \phi' (X_p' X_p + \lambda I)^{-1} \phi \right) \tag{4}
\]
Let us assume experimental data 3.3 Marginal likelihood for miRNA expression and other target genes, while the opposite can generally not assumed to be the chance that there is a measurable difference in the influence on marginal likelihood has been already derived in Zacher assertion lowers that chance.

\[ b_n = b_0 + \frac{1}{2} \left( \alpha^T o - \mu_{n} (X_n^T X_n + \lambda I) \right) \mu_n \]

(6)

\[ \mu_n = (X_n^T X_n + \lambda I)^{-1} X_n^T o \]

(7)

Let \( \xi := \det(X_n^T X_n + \lambda I) \). Then the log marginal likelihood is given as:

\[ \log p(o|X_n) = -\frac{1}{2} \log \xi + a_0 \log b_0 - a_0 \log b_n + \log \Gamma(a_n) - \log \Gamma(a_0) \]

(8)

We thus have to efficiently compute the determinant \( \xi \) and the matrix inverse of \( C_p := (X_n^T X_n + \lambda I) \) in order to make the approach practically feasible. Here we address this problem via a Cholesky decomposition \( C_p = R^T R \), where \( R \) is an upper triangular matrix: Since \( \log \xi = 2 \sum \log |R_{ij}| \) Cholesky factorization allows to efficiently calculate the log determinant. Moreover, \( C_p^{-1} = R^{-1} (R^{-1})^T \). Note that inversion of the triangular matrix \( R \) can be done quickly via back-fitting.

3.3 Marginal likelihood for miRNA expression and other experimental data

Let us assume \( C = 2 \), where—without loss of generalization—\( c = 1 \) denotes a reference condition. We distinguish two situations: In one miRNA \( i \) has switched its activity between \( c = 1 \) and \( c = 2 \). In that case we use two Gaussian distributions to model expression levels. In the other case miRNA \( i \) has not switched, and we use only one Gaussian distribution. The resulting one-way ANOVA model is thus

\[ p(A_{mv}|R_v, \sigma_i^2) = \begin{cases} N(\gamma_0 + \gamma_i \sigma_i^2) & \text{if } has switched \, \xi \text{ otherwise} \\ N(\gamma_0, \sigma_i^2) & \text{otherwise} \end{cases} \]

(9)

where \( \gamma_0 \) and \( \gamma_i \) are parameters, which describe the mean expression and log fold change of miRNA \( i \), respectively. The equation should be interpreted in a probabilistic sense: If a miRNA is differentially expressed, then the chance increases that it also exhibits a measurable influence on its target genes. Vice versa non-differential expression lowers that chance.

If we assume \( \sigma_i^2 \sim IG(\alpha, \beta) \), the marginal likelihood \( p(A_{mv}|R_v) = p(A_{mv}|X_v) \) can be calculated in analytical form:

\[ p(A_{mv}|X_v) = \frac{\Gamma(z + \frac{1}{2})}{\Gamma(z)} \frac{1}{(2\pi)^{z/2} \left( 1 + \frac{1}{2} (\mu_v - \mu_v) \right)^{z/2}} \]

(10)

where \( \Gamma(\cdot) \) is the gamma function and \( \mu_v = \gamma_0 + \gamma_i \), if the first condition in Equation (9) holds and \( \gamma_0 \), otherwise. Although this marginal likelihood has been already derived in Zacher et al. (2012) it is not used in BIRTA.

The same strategy as described here for miRNAs can be applied to experimental data of other regulatory factors, if available. More specifically, one may use miRNA data of differentially expressed TFs to support evidence for their conditions specific activity on protein level. The idea is that differential gene expression of a TF increases the chance that there is a measurable difference in the influence on target genes, while the opposite can generally not assumed to be true. Accordingly, for TFs we define a model following Equation (10) for differentially expressed TFs only.

3.4 Bayesian variable selection

In order to allow for variable selection we need to make our model sparse. There exists a large amount of literature on Bayesian variable selection methods. One possibility that we also followed in our previous BIRTA model is the spike and slab prior (George and McCulloch, 1997):

\[ \omega_{p} | \pi, \nu \sim (1 - \pi_{p}) \delta_{0} + \pi_{p} N(0, \nu^{-1}) \]

(11)

where \( \delta_{0} \) denotes a point mass at 0, \( \pi_{p} \sim \text{Bernoulli}(p_{p}) \) with \( p = 1, \ldots, N \) and \( \nu \) is the total number of possibly active regulators. One advantage of the spike and slab prior compared to a lasso type prior is a selective shrinkage of model coefficients. That means model coefficients for selected variables are only mildly pushed towards zero and thus estimates are less biased compared to a lasso prior (Hernandez-Lobato et al., 2010). Another advantage of the spike and slab prior is that sparsity can be controlled in a variable dependent manner via parameters \( p_{p} \). Here prior knowledge can be integrated: For each regulator we can encode the prior belief that it is active.

3.5 Model fitting via MCMC

3.5.1 General idea

In Section 3.2 we showed that regression coefficients can be effectively marginalized out while estimating model evidence. In biRTE we are thus left with estimating the condition specific hidden state variables \( \pi_{c} := (\pi_{p})_{p = 1, \ldots, N} \). For that purpose we use a stochastic sampling scheme based on Markov Chain Monte Carlo (MCMC). More specifically, we define two possible move types:

• switch: a regulator switches from active to inactive state within randomly picked condition \( c \)
• swap: an inactive regulator and an active regulator exchange their activity states within randomly picked condition \( c \).

In practice we restrict swap operations to regulators showing a significant overlap of regulated targets.

We formally accept an MCMC move with probability

\[ b = \min \left( 1, \frac{p(D|X, \pi_{c}^{\text{new}}) \prod_{p=1}^{n} q_{c_{p}}^{\text{old}} q_{c_{p}}^{\text{new}}}{p(D|X, \pi_{c}^{\text{old}}) \prod_{p=1}^{n} q_{c_{p}}^{\text{old}} q_{c_{p}}^{\text{new}}} \right) \]

(12)

where \( \pi_{c}^{\text{new}} \) and \( \pi_{c}^{\text{old}} \) correspond to the old and new miRNA and TF state configurations, respectively, and \( X \) is the full \( \nu_{\text{mRNA}} \times (N + 1) \) design matrix. Please note that inclusion of the proposal distribution \( q \) is necessary, because the swap operation induces non-identical probabilities to reach \( \pi_{c}^{\text{old}} \) from \( \pi_{c}^{\text{new}} \) and vice versa.

In conclusion the MCMC algorithm allows for estimating for each regulator and condition the posterior probability to influence the expression of its target genes.

3.5.2 Efficient update of marginal likelihood

One of the major differences of biRTE vs. BIRTA is the analytical marginalization of regression coefficients in Equation (3). This in turn requires updates of the design matrix \( X \) and the determinant \( \xi \) after each MCMC move. To make these updates computationally feasible we can exploit the fact that only the design matrix \( X_{p} \) restricted to active regulators needs to be updated. When a switch or swap operation is performed a column in \( X_{p} \) is added, removed or
exchanged. Let \( X_p \) denote our updated design matrix, then the difference in the marginal log likelihood for mRNA data is given by:

\[
\log p(o|X_p) - \log p(o|X_p) = \frac{1}{2} \left( \log \xi - \log \xi' \right) + a_e (\log b_a - \log b_a')
\]

Based on Section 3.2 it is sufficient to update the Cholesky factor \( R \) of \( C_p = X_p^T X_p + \lambda I \) in order to compute this formula efficiently: Suppose we add a column \( v \) to \( X_p \), i.e. \( X_p = X_{p+1} = [X_p, v] \). We are interested in the updated Cholesky factor \( R' \). Note that

\[
C_p' = \begin{pmatrix}
C_p & X_p'v \\
X_p v^T & v^T v + \lambda
\end{pmatrix} = \begin{pmatrix}
R^R & X_p'v \\
X_p v^T & v^T v + \lambda
\end{pmatrix}
\]

According to the Schur complement and block-wise matrix inversion lemmas we can decompose \( (X^T X + \delta I)^{-1} \) as:

\[
\begin{pmatrix}
R^R & X_p'v \\
X_p v^T & v^T v + \lambda
\end{pmatrix}^{-1} = \begin{pmatrix}
I & -R^R(X^R)^{-1}X_p v \\
0 & I
\end{pmatrix}
\]

\[
= \begin{pmatrix}
I & -R^R(X^R)^{-1}X_p v \\
0 & 1/\delta
\end{pmatrix}
\]

\[
= \begin{pmatrix}
R^R & X_p'v \\
X_p v^T & v^T v + \lambda
\end{pmatrix}^{-1}
\]

where \( \delta := v^T v + \lambda - v^T X_p R^{-1}(X^R)^{-1}X_p v = v^T v + \lambda - v^T U U^T v \). Hence, \( R^{-1} \) can be updated efficiently according to the block-wise matrix decomposition indicated in the third line, and \( R' \) is computed from \( R^{-1} \) via back-fitting. If the last column is deleted in \( X_p \), the whole procedure has just to be done in reverse order. However, if an arbitrary column \( 1 \leq \ell < p \) is removed, one would have to re-triangulate \( R' \), e.g. via Householder reflections. This is, however, computationally not beneficial compared to just re-calculating the Cholesky decomposition of matrix \( C_p \) after removal of column \( \ell \) and row \( \ell \).

3.6 Posterior inference

Inference about active regulators based on MCMC samples can be done in different ways. One method is to return the configuration with highest posterior probability among all sampled ones. Another possibility is to look at the marginal selection frequencies during sampling and filter those above a defined cutoff \( \tau \). The arising question is, how such a cutoff could be chosen. We here implemented a method that we recently proposed in the context of Bayesian Network learning (Frohlich and Klaau, 2013). Briefly the idea is to fit a beta mixture model to the observed regulator frequency profile. Based on this mixture model it is possible to select \( \tau \) such that the expected false positive detection rate is below a defined cutoff (e.g. 0.1%).

3.7 Hyperparameter settings

Our model contains a few hyper-parameters that need to be specified. These include parameters for the variance priors \( (x, \beta, a_0, b_0) \), which \textit{biRte} estimates in an empirical Bayes sense via maximum likelihood from the global distribution of empirical variances (separately for each data type). The method is taken from Venables and Ripley (2002) and implemented in the R-package MASS.

Following (Hernandez-Lobato et al., 2010) we set \( \lambda = 1 \) in the spike and slab prior. Moreover, \( \gamma_0 \) is set to the average expression level in the reference condition, and \( \gamma_1 \) to the median log fold change of differentially expressed regulators (separately for each regulator type and for up- and down-regulation). This procedure was chosen in order to reduce false positive predictions of regulator activities.

4 BiRte step II: network inference

After having determined active regulators one may ask, in which way these regulators influence each other. Established methods, such as Bayesian Networks, would usually require direct measurements of regulators, which are at least difficult to obtain for TFs. Moreover, the typically small sample size imposes a principal limitation. In \textit{biRte} we thus restrict ourselves to subset relationships between differentially expressed target genes. These subset relationships can express biologically interesting properties: Suppose regulator \( A \) to act upstream of regulator \( B \). There is a certain subset of target genes \( E_{AB} \), which is only affected by \( A \). However, because \( A \) is upstream of \( B \), \( A \) also influences target genes \( E_B \) of \( B \) (Fig. 2). This might happen, because \( A \) activates or deactivates \( B \) on protein level and only \( B \) itself influences \( E_B \). Another possible mechanism is that \( A \) and \( B \) together modulate expression of \( E_B \) in a combinatorial fashion. Both mechanisms are not distinguishable without further experiments and thus the schema in Figure 2 comprises both of these situations. Our task is to infer the correct wiring of the network from observable differential expression of target genes as well as target gene predictions for individual regulators.

The whole idea has striking similarities to Nested Effects Models (NEMs) (e.g. Frohlich et al., 2011; Markowitz et al., 2005), which have been introduced for causal network inference from perturbation data. Although in our case we do not have targeted perturbations of individual regulators, probabilistic inference of subset relationships between differentially expressed targets of regulator pairs can be done effectively via existing NEM structure learning algorithms. For the sake of simplicity we here explain the idea of the pairwise NEM inference algorithm discussed in Markowitz et al. (2007): Let \( n = |E_A| + |E_B| \) and \( E \in \{0, 1\}^{n \times 2} \). We set \( E_{ik} = 1 \), if gene \( i \) is differentially expressed and a predicted target of regulator \( k \), otherwise \( E_{ik} = 0 \). Let us further assume we know certain type I and type II error rates \( \alpha, \beta \) that occur while declaring differential gene expression. Let \( H \) be one of the four possible hypotheses \{\( A \rightarrow B, A \leftarrow B, A \rightarrow B, A \leftarrow B \)\}, where \( A.B \) means no edge / subset relationship. In order to score \( H \) we are interested in the marginal likelihood of our observations \( E \):

\[\begin{array}{ccc}
E_A & & E_B \\
& A & \\
& B & \\
& &\end{array}\]

Fig. 2. Idea behind regulatory network inference: Regulator \( A \) influences genes \( E_A \) and \( E_B \), whereas \( B \) only targets \( E_B \). The model thus represents a subset relationship between target genes of \( A \) and \( B \).

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/31/20/3290/196051 by guest on 25 October 2018
Overall the results indicate a significant improvement of biRte compared to all competing methods. For miRNAs pAUCs are typically close to 1 and Brier scores close to 0, even for the highest levels of false positives and negatives. TF predictions react more sensitive, because TFs typically have fewer target genes than miRNAs in our network. As expected, prediction of regulator-regulator interactions is most challenging for all compared methods. We observed 10–15 fold speed-up of biRte compared to BIRTA and GEMULA (Supplementary Fig. S11).

6.2 Network inference

In a second round of simulations we compared the network reconstruction used in our biRte approach to Bayesian Networks (BN) (Friedman et al., 2000), ARACNE (Margolin et al., 2006), GENIE3 (Huynh-Thu et al., 2010) and GeneNet (Opgen-Rhein and Strimmer, 2007) on 10 randomly selected regulatory sub-networks of Escherichia coli and 20 correspondingly simulated datasets per network (details in Supplements). For BN structure learning a greedy hill climber using the BGe score was employed. For ARACNE we investigated additionally the effect of a data discretization into three bins using the information theoretic approach by Hartemink (2001). Accordingly, for BN the BDe was used to score candidate networks for discretized data. For biRte the assumed type I and type II error rates for differential expression detection were fixed to 0.05 and 0.1, respectively. GENIE3 returns an edge weight matrix. We transformed these edge weights into z-scores and then selected all edges with $P < 0.1$. For GeneNet we considered the presence of an edge, regardless of direction. This was appropriately taken into account while counting false positive predictions, when comparing the inferred network against the true one. Likewise, for BNs the CPDAG and for biRte the transitive closure of the original network was considered as reference. That means a predicted edge was only counted as false positive, if it was not contained into the respective equivalence class of the true network.

We conducted our comparison for different fractions of false positive and false negative target genes (Supplementary Figs S13–S16), number of replicates $r$ (Supplementary Figs S13 and S17–S19), total number of downstream targets $m$ (Supplementary Figs S13, S20 and S21) as well as number of network nodes $n$ (Supplementary Figs S13 and S22–S24). For $n > 15$ nodes ARACNE and BN on discretized data became infeasible slow, so they had to be omitted.

The boxplots in all cases show a clear advantage of biRte’s network reconstruction, which uses subset relationships between differentially expressed target genes. The computation time varied between less than 1 second (15 nodes, Supplementary Fig. S24) and 30 seconds (60 nodes, Supplementary Fig. S27), which makes the approach practically highly feasible. biRte showed a highly robust behavior against false positive and false negative target gene predictions, given a sufficient amount of target genes. Sensitivities and specificities stayed in all cases close to 1. As expected, increasing the number of replicates from 5 to 10, 20 and 30 yielded a small improvement of all competing methods, because they directly use expression data of regulators. Decreasing the number $m$ of total target genes from 450 to 150 on average led to a decrease of sensitivity to around 80% for biRte, but in single cases (e.g. network 4) the effect was stronger. Increasing the number $n$ of network nodes from 15 to 30, 40 and 60 lead to a decrease of sensitivity for all methods. However, even for $n = 60$ biRte with 5 replicates still achieved a sensitivity of around 60% at a specificity close to 1.

$P(E_i|H) = \sum_{\theta_i \in \{A,B,\text{null}\}} P(E_i|H, \theta_i = j)p(\theta_i = j)$

(15)

where $\theta_i$ is a parameter that indicates whether $E_i$ is attached to regulator $j$. In Equation (15) we marginalize over all possible attachment positions, possibly weighted by a prior (which is set uniform in the default case). While doing so we consider an additional dummy node null which is unconnected to $A$, $B$. This trick allows for effective removal of target genes that react unspecifically (Tresh and Markowetz, 2008).

The local likelihoods $P(E_i|H, \theta_i = j)$ are defined as:

$$P(E_{ik} = 1|H, \theta_i = j) = \begin{cases} 1 - \beta & \text{if edge } k \rightarrow j \text{ exists} \\ \frac{\beta}{2} & \text{otherwise} \end{cases}$$

(16)

The algorithm selects for each regulator pair $A$, $B$ the one with maximum marginal likelihood. Afterwards the transitive closure of the resulting graph is returned. We refer the reader e.g. to Fröhlich et al. (2009) for a detailed overview and empirical study of different NEM structure learning approaches. In the biRte software package the user has full access to all methods implemented in R-package nem (Fröhlich et al., 2008).

5 Software

biRte is available as a Bioconductor R-package. To ease the analysis of relative expression levels the software offers appropriate interface functions to limma (Smyth, 2004) and thus allows to be applied for arbitrary complex statistical designs. Since biRte relies on the assumption of multivariate normally distributed data we recommend the limma-voom mechanism for RNA-seq data (Law et al., 2014).

Notably biRte optionally allows for grouping regulators with highly similar target gene sets (being unlikely to be distinguishable) into clusters in a pre-processing step (see details in Supplements). In the subsequent network analysis clusters are automatically decomposed into individual regulators.

6 Simulation results

6.1 Active regulators and regulator interactions

In order to gain insights into the principal behavior of our model we conducted several simulations. In a first simulation study we compared biRte to our previous BIRTA method (Zacher et al., 2012), to GEMULA (Geeven et al., 2012) as well as to a hypergeometric test (Lim et al., 2009) with multiple testing correction via the method by Benjamini and Yekutieli (2001). This was done on the basis of a human regulatory sub-network and accordingly simulated expression data of 900 target genes (details in Supplements). In order to have a fair comparison we here applied GEMULA, biRte and the hypergeometric test such that they used the same set of potentially relevant two-way interaction terms (see Supplements). BIRTA does not support inference of regulator-regulator interactions.

Supplementary Figures S1–S10 depict the quality of regulator activity predictions for different fractions of false positive and false negative target gene predictions, i.e. erroneous edges in our regulatory network. Results are shown for 20 simulation repeats in terms of partial area under ROC curve (for a specificity cutoff of 90%) and Brier scores (i.e. average squared differences between probabilistic predictions and indicators of true regulator activities).
7 E.coli growth conditions

We investigated a microarray dataset by Covert et al. (2004) together with a filtered TF-target graph (Castelo and Roverato, 2009). The gene expression data comprises three replicates from E.coli during aerobic growth and four replicates during anaerobic growth. The TF-target graph contains annotations for 160 transcription factors. The data are also contained in the vignette of the biRte R package.

Potential two-way interaction terms between regulators were selected such that the overlap of target genes was between 10 and 80%.

Limma analysis yielded 224 differentially expressed genes (FDR < 5%, log FC > 1). Moreover, 9 TFs showed differential expression on mRNA level. For these 9 TFs we set a prior activation probability of 50%, whereas for the others a uniform prior was used.

We conducted a biRte analysis with a burn-in and sampling phase of 20 000 iterations. The calculation took around 2 min on a 2.5 GHz AMD Opteron 6380 machine. Convergence can be checked according to the trace plot of the marginal log-likelihood (Supplementary Fig. S28). Active regulators were selected according to their marginal selection probability (cutoff: expected FPR < 0.001).

Application of biRte’s network inference to all active regulators resulted in a network of 28 unique regulators (Fig. 3), which have all at least one differentially expressed target gene. The network consists of one connected component with 25 regulators and three singletons. The master regulator of the connected network component, fumarate and nitrate reductase (fnr), is well known to be essential for the switch from aerobic to anaerobic metabolism (Unden and Schirawski, 1997). Fnr regulates arcA in anaerobic conditions (Compan and Touati, 1994) and has an effect on the narX/narL system (Lamberg and Kiley, 2000), as also reflected by our network. Furthermore, arcA represses betI under anaerobic growth conditions (Lamark et al., 1996). The activity of gadE, gadX, gadW, appY, fhlA and marA under anaerobic growth conditions have also been described previously (Barbosa and Levy, 2002; Oshima et al., 2002; Takahata et al., 2008; Zhou et al., 2011). Likewise, ompR, iscR, soxS, iclR and dnaA activities have been associated to aerobic growth and oxidative stress (Chodavarapu et al., 2008; Demple, 1996; Giel et al., 2006; Lin et al., 2005; Matsubara et al., 2000). The TF crp (cAMP receptor protein) regulates several TFs in our inferred network. It is known to regulate more than 100 genes involved in energy metabolism (Bott, 1997). The singleton argP is an arginine and lysine responsive transcriptional regulator. It competes with dnaA for iteron binding (Rudd, 2000).

When comparing the estimated against the original, literature based TF-target gene network, direct TF-TF interactions could explain the edges fnr \rightarrow arcA, gadE \rightarrow gadX, crp \rightarrow galS, crp \rightarrow fur, arcA \rightarrow betI as well as fnr \rightarrow soxS, fnr \rightarrow pdhR, fnr \rightarrow narL, fnr \rightarrow gadX, fnr \rightarrow gadW, fnr \rightarrow fhlA, crp \rightarrow pdhR, crp \rightarrow gadX, crp \rightarrow gadE, arcA \rightarrow gadX, arcA \rightarrow gadE. Several of these edges are contained in the equivalence class of our estimated network (which is transitively closed), but removed for better visualization purposes in Figure 3.

8 Prostate cancer

We downloaded a dataset comprising gene expression from 48 normal and 47 prostate cancer tissue samples (GSE29079, Affymetrix
Human Exon 1.0 ST). Accompanying qPCR-based array data (2 plates) from 464 human miRNAs is available for 48 normal and 51 cancer samples (GSE54516). MiRNAs in each individual sample were normalized to plate median signals. Gene expression data was normalized via RMA (Irizarry et al., 2003). For the following analysis we restricted ourselves to the set of 44 normal and 47 cancer samples that were matching between mRNA and miRNA data. These data were mapped to a human regulatory network (see Supplements). Only regulators targeting more than 5 probesets were considered. Potential two-way interaction terms between regulators were selected such that the overlap of target probesets was between 25 and 50%. This was done in order to ensure significant effects that were clearly distinguishable from main effects of regulators.

Limma analysis yielded 262 differentially expressed probesets and 134 differentially expressed, network map-able miRNAs (FDR < 5%, |log FC| > 1). Moreover, on mRNA level there were four differentially expressed, network map-able TFs. Differentially expressed miRNAs were given a prior activation probability of 80%, and for differentially expressed TFs we set this number to 50%. TF expression is expected to be less indicative of TF protein activity. All other regulators were given a uniform prior probability per respective regulator class (TF, miRNA, interactions).

We ran biRte with 1 000 000 iterations with an additional burn-in phase of 1 000 000 iterations. The whole calculation took around 16.2 hours. Convergence can be checked according to the trace plot of the marginal log-likelihood (Supplementary Fig. S29). Active regulators were selected according to their marginal selection probability (cutoff: FDR < 0.001).

We applied biRte’s network inference to all active regulators, resulting in a hierarchical network structure (Fig. 4). All 26 regulators in this network have at least one differentially expressed target gene. Several are known to be involved in cancer and prostate cancer specifically: For example, HLTF encodes a member of the SWI/SNF family. Members of this family have helicase and ATPase activities and are thought to regulate transcription of certain genes by altering the chromatin structure around those genes. They have been associated to cancer progression (Debauve et al., 2008). EP300 encodes the adenovirus E1A-associated cellular p300 transcriptional co-activator protein. It functions as histone acetyltransferase that regulates transcription via chromatin remodeling and is important in the processes of cell proliferation and differentiation. Several studies suggest EP300 to play a role in the development of prostate cancer (Zhong et al., 2014). ATF4 is a member of the activating TF family and discussed as a cancer drug target (Singleton and Harris, 2012). E2F1 is a member of the E2F family and plays a crucial role in the control of cell cycle and action of tumor suppressor proteins. It has been associated to prostate cancer (Ma et al., 2014). FOXC1 belongs to the forkhead box TF family. These TFs play important roles in the regulation of tissue- and cell type-specific gene transcription and have been associated to prostate cancer (Long et al., 2012).

Comparison of the network against the BioGRID database (Chatr-Aryamontri et al., 2013) could verify the edges $EP300 \rightarrow TAL1$. $ATF4$
biRte combines inference of regulator activities with network reverse engineering. For inference of regulator activities we rely on an extension of our previous BIRTA method (Zacher et al., 2012). The most significant difference is the analytical and computationally efficient marginalization of regression coefficients, which results in 10–15 fold faster computation time. Furthermore, our simulations indicate that biRte not only outperforms BIRTA, but also a hyper-geometric test (Lim et al., 2009) and the lasso based GEMULA (Geeven et al., 2012) with respect to prediction of active regulators and regulator-regulator interactions. biRte performs post-hoc inference of networks between active regulators via NEM structure learning, which infers subset relationships of differentially expressed target genes. According to our simulations the approach yields far more reliable network reconstructions than typical methods that use observational expression data of regulators.

Application of biRte to a gene expression data measuring response to different growth conditions in E.coli as well as to a combined mRNA and miRNA dataset from prostate cancer patients demonstrated the principal agreement with the biological literature. We thus think that biRte could be a useful contribution to decipher biological context specific regulatory networks.

Conflict of Interest: none declared.

References

Frohlich,H. et al. (2011) Fast and efficient dynamic nested effects models. Bioinformatics, 27, 238–244.
Ma,Y. et al. (2014) Tfdp3 was expressed in coordination with e2f1 to inhibit c2h1-mediated apoptosis in prostate cancer. Gene, 537, 253–259.


