Bioinformatics Advance Access published August 21, 2015

Bioinformatics, 2015, 1–8
doi: 10.1093/bioinformatics/btv461
Advance Access Publication Date: 6 August 2015
Original Paper

Systems biology

A statistical approach to virtual cellular experiments: improved causal discovery using accumulation IDA (aIDA)

Franziska Taruttis*, Rainer Spang and Julia C. Engelmann*

Department of Statistical Bioinformatics, University of Regensburg, 93053 Regensburg, Germany

*To whom correspondence should be addressed.

Associate Editor: Igor Jurisica

Received on March 31, 2015; revised on July 9, 2015; accepted on August 1, 2015

Abstract

Motivation: We address the following question: Does inhibition of the expression of a gene X in a cellular assay affect the expression of another gene Y? Rather than inhibiting gene X experimentally, we aim at answering this question computationally using as the only input observational gene expression data. Recently, a new statistical algorithm called Intervention calculus when the Directed acyclic graph is Absent (IDA), has been proposed for this problem. For several biological systems, IDA has been shown to outcompete regression-based methods with respect to the number of true positives versus the number of false positives for the top 5000 predicted effects. Further improvements in the performance of IDA have been realized by stability selection, a resampling method wrapped around IDA that enhances the discovery of true causal effects. Nevertheless, the rate of false positive and false negative predictions is still unsatisfactorily high.
Results: We introduce a new resampling approach for causal discovery called accumulation IDA (aIDA). We show that aIDA improves the performance of causal discoveries compared to existing variants of IDA on both simulated and real yeast data. The higher reliability of top causal effect predictions achieved by aIDA promises to increase the rate of success of wet lab intervention experiments for functional studies.
Availability and implementation: R code for aIDA is available in the Supplementary material.
Contact: franziska.taruttis@ur.de, julia.engelmann@ur.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Understanding of the complex molecular networks underlying cellular processes is the key challenge of systems biology. Here, we focus on gene regulatory networks. Interpreting these networks causally requires a direction of information flow. Given a causal network, the most basic question is: If I perturb gene X, what happens to gene Y? Of course, this question can be answered experimentally by inhibiting X, e.g. by PCR-based gene deletion strategy (Baudin *et al.*, 1993; Wach, 1996), RNA interference (Fire *et al.*, 1998), or CRISPRi (Larson *et al.*, 2013) screening, and observing Y. We here discuss to what extend the same question may be answered in a virtual way, i.e. without the need to perform a cellular perturbation experiment. A high-throughput version of this problem would be the following: Find all pairs of genes ($X \rightarrow Y$), where perturbing X affects Y. We refer to this causal data mining approach as causal discovery.

Statistically, the challenge is to estimate causal effects between random variables from purely observational data. For gene expression data the causal effect of gene X on gene Y can be defined as the number of units gene Y changes in expression, if the expression of gene X is experimentally altered by one unit. Causal Networks are a well established statistical framework for analyzing this problem (Pearl, 2000).

In causal transcriptional networks the nodes of a directed acyclic graph (DAG) are random variables representing genes and their expression values. A directed edge from X to Y denotes that the expression of X causally and directly influences the expression of Y. Consequently, the expression of Y is determined by the expression of its parents in the network. If the causal network is known, Pearl's do-calculus allows estimating the strength of the causal effects between pairs of random variables from observational data (Goldszmidt and Pearl, 1992; Pearl, 1995). However, causal transcriptional networks are largely unknown. Multiple statistical algorithms have been established to estimate the undirected skeleton of a network from conditional independence tests (Huynh-Thu *et al.*, 2010; Margolin *et al.*, 2006). The PC algorithm (Kalisch and Bühlmann, 2007; Spirtes *et al.*, 2000) allows to define the direction of some edges allowing for a causal interpretation of the network. The network cannot be fully directed because different networks with identical skeletons can encode the same conditional independence assumptions and thus can not be distinguished on observational data only (Verma and Pearl, 1991).

Recently, Maathuis *et al.* (2009) introduced the Intervention calculus when the Directed acyclic graph is Absent (IDA) algorithm, which combines partial estimation of the network with the estimation of lower bounds for causal effects between genes. They prove consistency of the bounds (Maathuis *et al.*, 2009) and show in applications that IDA outperforms regression-based methods in terms of number of true positives versus number of false positives for the top 5000 predicted effects on the transcriptome of yeast gene deletion strains from a large dataset of expression profiles of wild type yeast (Maathuis *et al.*, 2010). However, in simulations with large networks and medium sized datasets, which are typical for many biological applications, networks often cannot be reconstructed correctly. Stekhoven *et al.* (2012) suggest using stability selection (Meinshausen and Bühlmann, 2010), a subsampling strategy that is wrapped around IDA as a remedy. In fact, their CStaR algorithm outperforms plain IDA with respect to true positive selections versus false positive selections (Stekhoven *et al.*, 2012).

Here we propose an alternative subsampling strategy in combination with a modification of the IDA algorithm called accumulation IDA (aIDA). Our method represents an improvement over CStaR, IDA, and regression based methods as demonstrated for both simulated and two yeast datasets.

2 Causal discovery

2.1 Causal transcriptional networks

We represent transcriptional causal networks by DAGs consisting of nodes $\mathbf{X} = \mathbf{X}_1, ..., \mathbf{X}_p$ and edges $\mathbf{E} = \mathbf{E}_1, ..., \mathbf{E}_s$. Every gene is represented by one node in the network. Statistically, every node is a random variable whose values indicate the expression levels of a gene. We further assume that the variables follow a multivariate Gaussian distribution. Edges represent conditional dependencies; nodes that are not connected represent variables that are conditionally independent. Moreover, we interpret the edges causally. If there is a directed edge $X \rightarrow Y$, we assume that an experimental perturbation of the expression level of X will affect Y, but not vice versa. A DAG fully specifies the conditional dependencies of all nodes (Pearl, 1988), but not vice versa. Several DAGs with the same skeleton of undirected edges and the same v-structures (Pearl, 2000) encode the same conditional dependency structure (Verma and Pearl, 1991). These equivalence classes of DAGs can be represented by completed partially directed acyclic graphs (CPDAGs) that consist of the joint skeleton and the directed edges which are common to all DAGs in the equivalence class (Chickering, 2002).

2.2 Causal effects

In a hypothetical perturbation experiment we may increase the random variable X by one unit ($X \mapsto X+1$) and observe the effect on a second variable Y. If Y is changed by $Y \mapsto Y + \beta$, we call β the causal effect of X on Y. Given a causal network and a dataset of joint observations of the nodes X, Pearl's do-calculus (Goldszmidt and Pearl, 1992; Pearl, 1995) provides guidelines how to estimate causal effects β for all pairs of nodes. In the Gaussian case one can use a linear regression of Y on X and a set of additional nodes Z that fulfill the Back-door criterion (Pearl, 2000), a criterion that only depends on the topology of the underlying causal network. A set of variables Z satisfies the Back-door criterion relative to an ordered pair (X, Y) in a DAG G, if

i. no node in Z is a descendant of X.
ii. Z blocks (see Supplementary Section 1) every path between X and Y that contains an arrow into X

(Pearl, 2000). It can be shown, that the true parents of a node X do always fulfill the Back-door criterion relative to every ordered pair (X, Y) (Pearl, 2000). In Figure 1A the set of nodes {2,3} fulfills the Back-door criterion relative to the ordered pair (6,10). In addition, the set {4,7} fulfills the Back-door criterion relative to the ordered pair (6,10), while for example the set {4} does not fulfill the Back-door criterion, because the path (6,3,7,10) remains unblocked.

Let $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_k)$ be a set of nodes fulfilling the Back-door criterion, and let $\beta = \beta_0, \beta_X, \beta_{Z_1}, \ldots, \beta_{Z_j}$ be the least squares fit of the linear regression

$$Y = \beta_0 + \beta_X X + \sum_i \beta_{Z_i} Z_i + \epsilon, \tag{1}$$

then β_X is a consistent estimator of the causal effect of X on Y (Maathuis *et al.*, 2009). Z adjusts the regression for possible confounders that affect X and Y simultaneously, thus creating spurious correlations between X and Y. An example for an adjustment set Z that satisfies the Back-door criterion of a node X with respect to all other nodes Y is the set of true parents of X.

2.3 IDA and CStaR

For transcriptional networks the underlying causal graph is mostly unknown and needs to be estimated. An obvious idea is to first reconstruct a causal network and then estimate causal effects thereof using Equation (1). Several algorithms have been proposed to estimate the equivalence class of a DAG from observational data (Chickering, 2002, 1996; Dash and Druzdzel, 1999; Madigan *et al.*, 1996; Meek, 1995; Pearl, 2000). IDA uses the PC-algorithm (Spirtes *et al.*, 2000), which estimates the CPDAG in a two-step procedure. In the first step, the skeleton and the v-structures of the CPDAG are estimated by testing for conditional independence. In the second step, orientation rules are applied (Meek, 1995), which exploit the facts that (i) the network cannot be cyclic, and (ii) that no new v-structures are found in an equivalence class. A parameter α is used to calibrate the sparseness of the estimated network. Large values of α increase the number of edges within a network. Importantly, increasing α also increases the runtime of the algorithm significantly, such that one often needs to compromise on the sparsity of the network to keep computations feasible.

The existence of equivalent DAGs complicates the estimation of causal effects. Maathuis *et al.* (2009) have demonstrated that it is still feasible to derive lower bounds of causal effects using their IDA algorithm. In a computationally elegant and efficient way, IDA enumerates all possible parent sets of a node X in a CPDAG, and calculates causal effects of X on all its descendants using any one of the

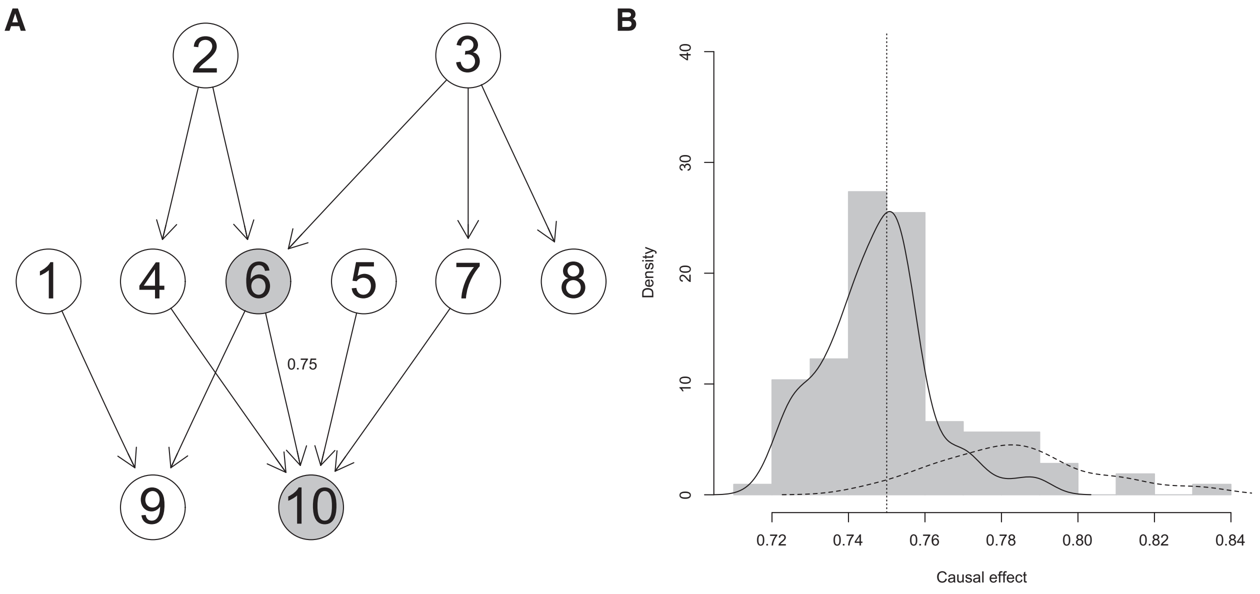

Fig. 1. Example for accumulation of the causal effect for a small simulated dataset. (A) Causal network. The causal effect of interest is the one of node 6 on node 10 (gray nodes). The true causal effect is 0.75. (B) Histogram over all estimated causal effects and the density for the effects that fulfill (solid line) or do not fulfill (dashed line) the Back-door criterion. Causal effects estimated using estimated adjustment sets which do not fulfill the Back-door criterion (dashed line) do not coincide with to the true causal effect, while the causal effects estimated using valid adjustment sets agree with the causal effect (vertical line)

parent sets as adjustment sets Z in Equation (1). This yields a set of estimated effects, and its minimum is used as a lower bound. IDA is asymptotically consistent.

From the systems biology perspective, IDA was ground-breaking, because Maathuis *et al.* (2010) succeeded in predicting expression changes of 5361 genes in 234 yeast deletion strains from 63 expression profiles of wild type yeast, thus constituting the first high throughput analysis of virtual perturbation experiments.

Transcriptional networks consist of tens of thousands of genes. Typically, however, one has not more than at best a few hundred observations of their gene expression values. In situations with many nodes and few observations the accuracy of estimated CPDAGs by the PC algorithm is poor (Kalisch and Bühlmann, 2007). As a remedy Stekhoven *et al.* (2012) suggest applying stability selection (Meinshausen and Bühlmann, 2010) on the IDA algorithm. K subsets of samples are drawn and sets of causal effect bounds are estimated for all of them using IDA. Finally, genes are ranked by how often they appear in the top q genes with the highest effects. This procedure is repeated for several values of q and the median rank is used to calculate the final rank of a gene. This modified IDA algorithm called CStaR was demonstrated to improve on causal discovery (Stekhoven *et al.*, 2012).

2.4 Limitations

A limiting factor in both IDA and CStaR is the poor accuracy of the estimated CPDAGs and the strong dependence of subsequent steps on a valid causal network (Supplementary Figs S1 and S2). Albeit it is difficult to improve on the quality of estimated networks, we argue that there is a more effective way to extract causal information from imperfect networks. More specifically, we have identified two steps in the IDA/CStaR procedure that confine causal discovery.

2.4.1 Selecting a causal effect from a multiset

For a given pair of nodes X and Y and a given CPDAG, IDA calculates multiple causal effects of X on Y, if the CPDAG connects X to other nodes via undirected edges. Both IDA and CStaR are conservative in that they report the minimum of the effect set. While the minimum is a valid lower bound of the causal effect, it represents a biased estimate that can lead to missed causal discoveries.

2.4.2. Summarizing effects across subsampling runs

CStaR applies stability selection to detect stable causal effects. It prefers gene pairs that repeatedly obtain a high score. The focus on gene pairs with the highest scores can lead to missed causal effects with smaller effect sizes.

Whether an estimated effect in the set of causal effects from X on Y associated with a CPDAG, $M(X \rightarrow Y)$, is valid or not depends on the corresponding adjustment set Z used in Equation (1). IDA and CStaR aim at deriving valid effects by identifying the true parents of the node X. If the CPDAG is correct, at least one of the effects in $M(X \rightarrow Y)$ is based on the true parents of X and thus valid, but we do not know which one it is. If the CPDAG is incorrect, it is well possible that all values in $M(X \rightarrow Y)$ are invalid.

3 The aIDA algorithm

3.1 Motivation

We aim at distinguishing between valid and invalid effects both within multisets of effects and across subsampling runs. If the causal network is known, the Back-door criterion and other graph-based criteria (Tian and Pearl, 2002) can be used. If the network is not known, we argue that the distribution of effects across subsampling runs is helpful to estimate the true causal effect. We use estimated causal effects from simulated data using known causal networks to study features of the estimated CPDAGs and the distributions of valid causal effects versus invalid causal effects.

3.1.1 Estimated parent nodes often yield valid estimates even if they are not the true parents

If the network underlying a dataset is estimated correctly, we can estimate valid causal effects from it. But absolutely correct networks

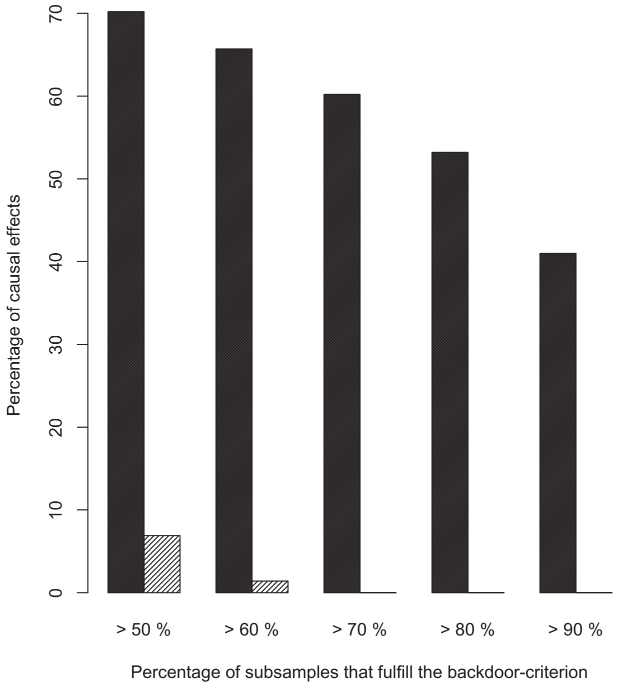

Fig. 2. The Back-door criterion is fulfilled frequently even if the PC algorithm rarely detects the true parents. One thousand cause-effect pairs were sampled randomly from a network with 1000 nodes. For each of these pairs we counted how often the Back-door criterion was fulfilled by the estimated parent sets within the 100 subsamples (black bars) and how often the parent set was the true parent set (gray bars). The *x*-axis shows the percentage of estimated causal effects of this 100 subsampling runs where the Back-door criterion was fulfilled. The *y*-axis displays the proportion of the 1000 sampled cause-effect pairs for which at least x% of the estimated parent sets from the 100 subsampling runs fulfilled the criterion (black) or were the true parent sets (gray). The detection of true parent sets is a very hard task, but meeting the Back-door criterion is not

are no strict requirement to calculate causal effects. Certain benign mistakes do not impede causal effect estimation. If the estimated parents fulfill the Back-door criterion they yield valid estimates even if they are not true parents. For a given DAG with edge weights that represent direct causal effect strengths we can generate artificial gene expression data (Details are given in the Supplementary Materials Section 3.1). We generated 50 samples for a random DAG of 1000 nodes, drew 100 subsamples of size $n = 25$, and ran the PC-algorithm as implemented in (Colombo and Maathuis, 2012) on each subsample, resulting in 100 CPDAGs. From all pairs of nodes $(X \rightarrow Y)$ we calculated the multisets of possible estimated parents of X. Since we know the correct parents of X in the simulation scenario, we can count how often the PC-algorithm correctly identifies a parent set. In these cases IDA can derive valid causal effects. However, finding the correct parent sets is a sufficient but not a necessary criterion to get valid estimates. In fact, every gene set Z that fulfills Pearl's Back-door criterion also yields valid causal effect estimates. In our simulations from known networks (see Supplementary Material for details on data simulation), we can verify whether an estimated parent set fulfills this criterion and count how often this is the case. Figure 2 shows that although estimated parent sets often are not the true parents, they nevertheless frequently fulfill the Back-door criterion and will hence generate valid estimates of causal effects. In other words, finding the true parents is difficult, whereas obtaining a set of estimated parents satisfying the Back-door criterion is not. For example, in more than 40% of the cause-effects pairs the Back-door criterion was fulfilled in more than 90% of the estimated parent sets over the 100 subsampling runs, conversely there were no subsampling runs that contained at least 90% true causal parent sets. Similar observations can be made when reconstructing real biological networks (Supplementary Fig. S3). This explains in part why IDA and CStaR are successful in spite of the poor performance of the PC-algorithm.

3.1.2 Valid effects generate peaks in the effect histograms

We next used Equation (1) to estimate causal effect sets for all pairs of nodes $(X \rightarrow Y)$ and all estimated parent sets that can be derived from the CPDAGs. We pool these numbers both across effect multisets and subsampling runs and visualize their distribution in the light gray histogram of Figure 1B.

The solid curve in Figure 1B is a smoothed density estimate of all effects from node 6 on node 10 across 100 subsampling runs derived from valid adjustment sets. We observe a pronounced peak in this density around the true causal effect (dotted vertical line). This is because estimates derived from valid adjustment sets Z are unbiased estimates of the true causal effect and, hence, scatter around it. The true parents of the cause are always a valid adjustment set, but they are not necessarily the only one. Valid are all sets that fulfill the Back-door criterion relative to the cause and the effect of interest. The dashed curve shows the density of the estimates that are not derived from valid adjustment sets. This distribution is not centered around the true effect. Estimates based on invalid adjustment sets can take any value, and their corresponding distribution is unknown. Since the Back-door criterion is not a necessary condition for Z to be a valid adjustment set, some of these estimates might still be valid but others are not.

If a large fraction of effects is valid, one would expect them to have similar values centered around the true causal effect, and therefore give rise to a peak in the histogram.

This observation is the motivation for the algorithm described below. The basic idea is to pool effects across multisets and subsampling runs and to use the mode of the smoothed histogram as causal effect estimate.

3.2 Algorithm

The input of our algorithm is a set of expression profiles consisting of p genes observed in n samples. This data is observational, i.e. the cells were not perturbed. We assume that all samples were drawn from the same underlying joint distribution of gene expression levels. The output of our algorithm is an ordered list of triples (X, Y, C), were X and Y are genes and C is the estimated causal effect of X on Y. The list is sorted by the absolute value of the effect strength C.

The algorithm performs the following steps:

1. Randomly draw K subsets of samples of size l, resulting in K resampled datasets.
2. For each of these subsets estimate a CPDAG using the PC-algorithm as implemented in Colombo and Maathuis (2012) with sparseness parameter α, resulting in K CPDAGs on the same set of nodes.
3. For every ordered pair of genes $(X \rightarrow Y)$ extract the multisets $M(X \rightarrow Y)$ of possible estimated parents of X and pool them across all CPDAGs.
4. For all elements Z of this multiset estimate a causal effect using the regression model from Equation (1).
5. Generate one histogram of estimated effects per gene pair (Accumulation step). Smooth these histograms using a Gaussian

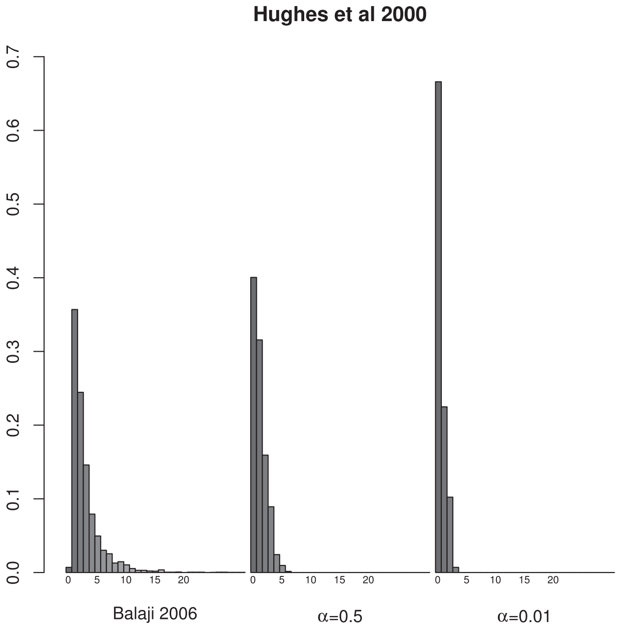

Fig. 3. Comparison of the number of parents of the network from Balaji *et al.* (2006) and the estimated networks using different values of α for the Hughes *et al.* (2000) dataset. The distribution of parents estimated from observational data using the higher value of α is more similar to the true distribution of the number of parents

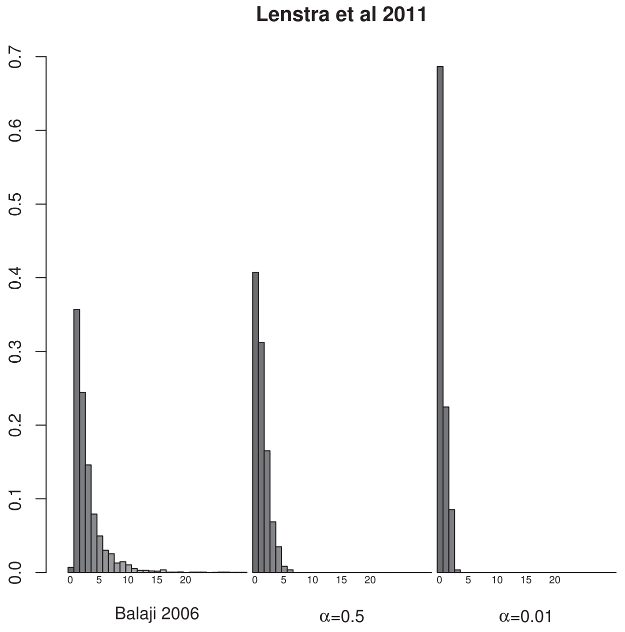

Fig. 4. Comparison of the number of parents of the network from Balaji *et al.* (2006) and the estimated networks using different values of α for the Lenstra *et al.* (2011) dataset. The distribution of parents estimated from observational data using the higher value of α is more similar to the true distribution of the number of parents

kernel, detect the mode in the smoothed histogram and use it as an estimate for the causal effect C of X on Y.

6. Collect all causal effects in a $p \times p$ matrix. Sort the effects by the absolute value of C, and output this sorted list.

The algorithm follows CStaR in steps 1–4 but differs from CStaR in step 5. Step 6 ranks all gene pairs by effect size to achieve results comparable to CStaR. Since the accumulation idea is added to the IDA concept, we call our algorithm aIDA for accumulation based IDA.

4 Results

aIDA and CStaR (Stekhoven *et al.*, 2012) were applied to both simulated datasets and two gene expression microarray datasets from *Saccharomyces cerevisiae* using the order independent implementation of the PC algorithm (Colombo and Maathuis, 2012; Kalisch *et al.*, 2012). For the simulated datasets true causal effects are known from the simulation process. For the microarray datasets the target set of causal effects is calculated as described in Maathuis *et al.* (2010) from interventional experiments (single knock-out strains, see details in Section 5.1 of the Supplementary Materials).

4.1 Parameter calibration

aIDA depends on the calibration of several parameters. The most critical of them is the sparseness parameter α of the PC algorithm (Kalisch and Bühlmann, 2007). Since α is shared by aIDA and CStaR, its calibration does not impede a fair comparison of the methods. The other parameters were set to the recommended values from CStaR (see Supplementary Materials Section 6). For the bandwidth of the Gaussian kernel we used default values of the density() function from basic R, since further calibration did not improve effect estimation. A larger α leads to denser CPDAGs and larger multisets. This in turn leads to more regression parameters that need to be estimated from Equation (1), which might increase the standard error of the effect size β_X. We tested the recommended value for α from the IDA literature but observed for both, the simulated and real data, that resulting CPDAGs can have very different numbers of edges in comparison to previous publications. This is further aggravated by a correction of the implementation of the PC algorithm, which leads to systematically sparser graphs for the same values of α, especially for graphs with many variables and few observations (Colombo and Maathuis, 2012).

For the simulated datasets we have calibrated α such that we obtain CPDAGs with densities similar to the expected density (Supplementary Figs S4 and S5). The histograms show, that $\alpha = 0.5$ is a realistic value for α. Additionally, we show that choosing a higher value of α leads to a better performance (Supplementary Figs S1 and S2). For the datasets from *S.cerevisiae* we compared the densities to a density derived from a transcriptional regulatory network published by Balaji *et al.* (2006) (Figs 3 and 4). We observed that both alpha = 0.01 and alpha = 0.5 underestimate the density of the network. Although the data suggest using an alpha even above 0.5, we did not increase it further due to runtime constraints.

4.2 Performance of aIDA on simulated data

First, we tested the performance of aIDA on simulated datasets generated from known causal Gaussian networks. We simulated random DAGs of various sizes and augmented their edges with weights sampled from a uniform distribution on the interval (0.1,1). The weights represent direct causal effect strengths. For each DAG, we generated artificial datasets of varying size using the R package pcalg (Kalisch *et al.*, 2012). Details on the data generation are given in the supplements, Section 3.1. We use these data to evaluate the performance of aIDA and compare it to that of CStaR. A comparison to plain IDA is not needed, because it has already been shown that it is outcompeted by CStaR with respect to true positive selections versus false positive selections (Stekhoven *et al.*, 2012).

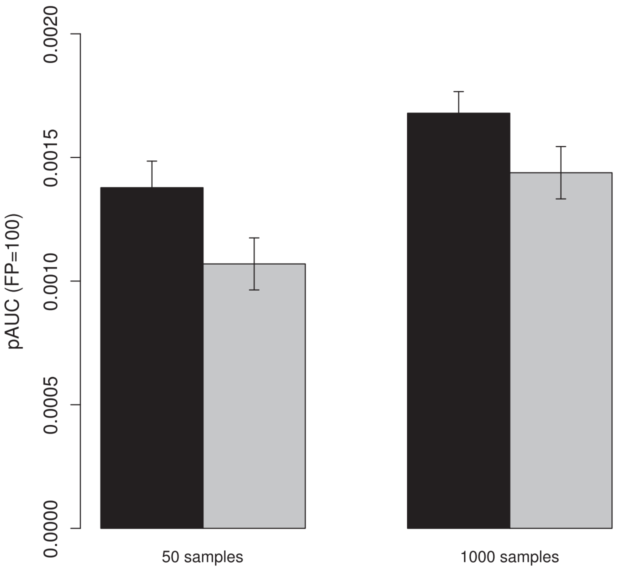

Fig. 5. Comparison of the partial area under the ROC curve up to 100 false positives for 10 simulated datasets with 100 nodes, and $n = 50$ and $n = 1000$ samples. Black bars show values for aIDA, gray bars show values for CStaR. The error bars indicate standard errors across the 10 datasets

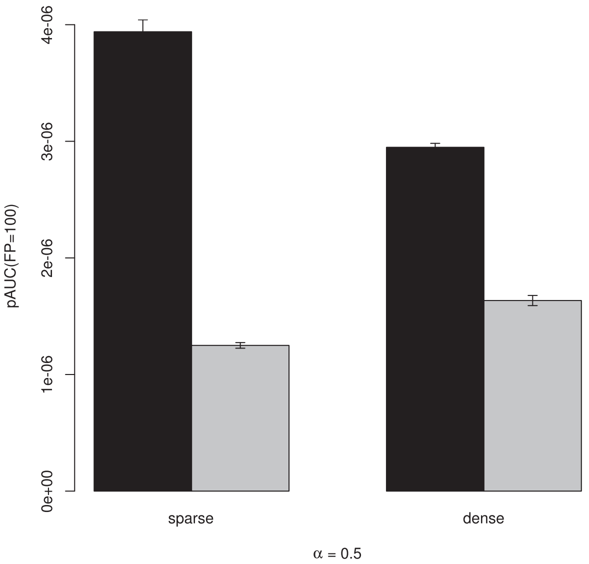

Fig. 6. Comparison of the partial area under the ROC curve up to 100 false positives for the 5 simulated datasets with 1000 nodes, $n = 50$ samples, $\alpha = 0.5$ and a dense and a sparse underlying true graph, respectively. Black bars show values for aIDA, gray bars show values for CStaR. The error bars indicate standard errors across the 5 datasets

In a first analysis we randomly simulated two sets of 10 small DAGs with 100 nodes each. For the first set we generated 10 small datasets with each set containing $n = 50$ samples, and for the second set we simulated 10 large datasets of $n = 1000$ samples. We then run aIDA and CStaR on this data, both using 100 subsamples of size $\frac{n}{2}$ and a sparseness parameter of $\alpha = 0.1$. Figure 5 shows barplots of the partial areas under receiver operating characteristics (ROC) curves (pAUC) up to 100 false positives. The error bars correspond to standard errors across the 10 datasets with 50 and 1000 samples, respectively. aIDA outperforms CStaR for both large and small datasets with respect to pAUC up to 100 false positives.

Both CStaR and aIDA become impractically slow for datasets with more than 1000 nodes and realistic settings of α. In a second analysis we tested aIDA on 5 sparse and 5 dense random networks of size 1000. The sparse and dense random networks contain approximately 1250 and 2500 edges, respectively. We simulated small datasets of 50 samples for each graph. We again used both aIDA and CStaR to rediscover the causal effects in the networks and compared their prediction performances. Figure 6 shows barplots of the pAUC for the 5 datasets generated from dense and sparse graphs, respectively, up to 100 false positives for $\alpha = 0.5$ (for more values of α see Supplementary Figs S1 and S2). Again aIDA outperforms CStaR, and achieves its best results for larger settings of α.

4.3 Application to gene expression data of *S.cerevisiae* deletion strains

aIDA and CStaR were applied to two large scale gene expression studies conducted in yeast. Both datasets analyzed *S.cerevisiae* single-gene deletion mutants (Baudin *et al.*, 1993; Wach, 1996). The datasets contain both observational data of wild type yeast and expression data from many individual deletion strains. We ran aIDA and CStaR on the wild type data to predict causal relationships between genes from observational data and validate these predictions using the data from the corresponding intervention experiments.

4.3.1 Application to data from Hughes *et al.* (2000)

Hughes *et al.* (2000) analyzed 276 deletion mutants and 63 wild type samples on two-color cDNA microarrays with probes for 5361 genes. This reference dataset was previously analyzed using IDA (Maathuis *et al.*, 2010) and CStaR (Stekhoven *et al.*, 2012). Data was preprocessed as described previously by these authors, resulting in expression data of 234 single-gene deletion mutants plus 63 wild type samples.

We ran aIDA and CStaR with the same 100 subsamples of the 63 observational samples. Lists of predicted causal effects were generated and filtered for gene pairs ($X \rightarrow Y$), where X is among the 234 genes, for which deletion strain data was available. The interventional data from deletion strains were used to classify these predictions into true and false positive predictions. Figure 7A shows ROC curves up to 100 false positives for CStaR and aIDA, respectively. Again, aIDA outperforms CStaR.

4.3.2 Application to data from Lenstra *et al.* (2011)

Next, we used a biological dataset from Lenstra *et al.* (2011). This data has not been used previously for causal discoveries. Lenstra *et al.* (2011) analyzed mutants of chromatin machinery components to examine the interactions between chromatin and gene expression in *S.cerevisiae*. Following preprocessing, the data consisted of 138 interventional single-gene deletion profiles and 67 observational wild type samples (for details on data preprocessing see Supplementary Materials Section 3.3). Figure 7B shows that also on this dataset, causal discovery is possible. As can be seen from this Figure, aIDA again performs better in comparison to CStaR.

5 Discussion

We presented aIDA, a method to estimate causal effects from observational data in situations where we have many variables but only few observations, a common situation in biology due to costs of the experiments and availability of biomaterial. aIDA does not require any knowledge about the causal network underlying the data. It uses IDA in a subsampling approach (Maathuis *et al.*, 2009) to account for small sample sizes. While IDA and CStaR take the

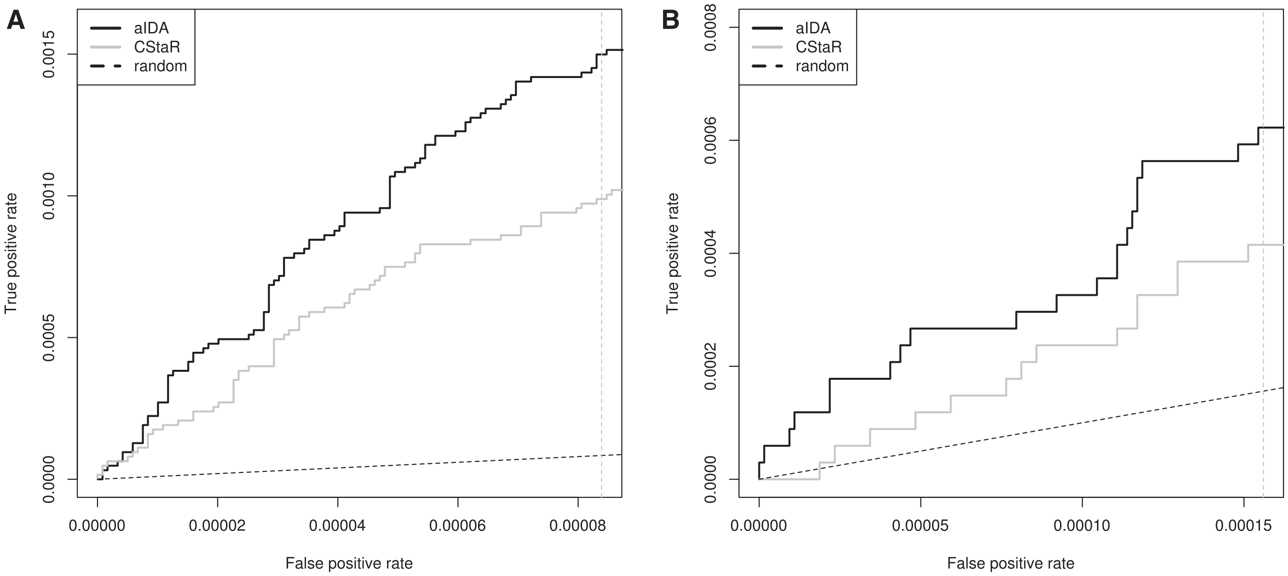

Fig. 7. ROC curves for the Hughes *et al.* (2000) and the Lenstra *et al.* (2011) dataset up to 100 FP using $\alpha = 0.5$. The solid black curves represent the ROC curves for aIDA, the gray curves show the ROC curves for CStaR, and the dotted black lines refer to random guessing. (A) While both methods perform better than random guessing, aIDA shows better performance up to 100 FP than CStaR for the dataset from Hughes *et al.* (2000). (B) aIDA improves over CStaR for the dataset from Lenstra *et al.* (2011). aIDA performs better than random guessing, while CStaR at the beginning of the ranked list does not

minimum absolute value of causal effects from K subsampling runs (Maathuis *et al.*, 2009, 2010; Stekhoven *et al.*, 2012), aIDA uses the whole multisets of causal effects. The estimate for the true causal effect is the mode of the density calculated over all estimated sets of causal effects across K subsampling runs.

IDA and all its extensions require among other assumptions that the multivariate Gaussian distribution is faithful to the DAG (Kalisch and Bühlmann, 2007), i.e. statistical conditional independencies can be inferred from the underlying DAG. Faithfulness in general is not testable (Zhang and Spirtes, 2008) and, of course, there is no evolutionary selection pressure that makes biological networks faithful. An additional limitation of all existing approaches for causal discovery is that feedback mechanisms cannot be captured by a DAG. Therefore, IDA, CStaR and aIDA cannot replace wet lab experiments, but are a good starting point for experimental design.

We show that aIDA outperforms CStaR on both simulated and real datasets with respect to the partial area under the receiver operating characteristics (ROC) curve up to 100 false positives. Besides yielding improved estimates of the top candidates for intervention experiments, the results on simulated data suggest that aIDA performs better than existing algorithms on small sample sizes. In summary aIDA improves the estimation of causal effects from observational data. These findings suggest aIDA as a useful tool for experimental design of functional studies.

Acknowledgement

We thank Peter Oefner and Maurits Evers for valuable discussions.

Funding

This work was funded by the Bavarian Research Network for Molecular Biosystems (BioSysNet) and by the e:bio initiative of the German Federal Ministry of Education and Research (BMBF), grant number 0316166G.

Conflict of Interest: none declared.

References

Balaji,S. *et al.* (2006) Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J. Mol. Biol.*, **360**, 213–227.

Baudin,A. *et al.* (1993) A simple and efficient method for direct gene deletion in *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **21**, 3329–3330.

Chickering,D. (1996) Learning Bayesian networks is NP-complete. In: Fisher,D. and Lenz,H.-J. (eds.) *Learning from Data: Artificial Intelligence and Statistics V*. Springer, New York.

Chickering,D.M. (2002) Learning equivalence classes of Bayesian network structures. *J. Mach. Learn. Res.*, **2**, 445–498.

Colombo,D. and Maathuis,M.H. (2012) A modification of the pc algorithm yielding order-independent skeletons. *CoRR*, **abs/1211.3295**.

Dash,D. and Druzdzel,M.J. (1999) A hybrid anytime algorithm for the construction of causal models from sparse data. In: Laskey,K.B. and Prade,H. (eds.) *UAI*. Morgan Kaufmann, Publishers, San Francisco, CA, USA, pp. 142–149.

Fire,A. *et al.* (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, **391**, 806–811.

Goldszmidt,M. and Pearl,J. (1992) Rank-based systems: a simple approach to belief revision, belief update, and reasoning about evidence and actions. In: Nebel,B. *et al* (eds.) Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning, pp. 661–672.

Hughes,T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

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

Kalisch,M. and Bühlmann,P. (2007) Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *J. Mach. Learn. Res.*, **8**, 613–636.

Kalisch,M. *et al.* (2012) Causal inference using graphical models with the R package pcalg. *J. Stat. Softw.*, **47**, 1.

Larson,M.H. *et al.* (2013) CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat. Protocols*, **8**, 2180–2196.

Lenstra,T.L. *et al.* (2011) The specificity and topology of chromatin interaction pathways in yeast. *Mol. Cell*, **42**, 536549.

Maathuis,M.H. *et al.* (2009) Estimating high-dimensional intervention effects from observational data. *Ann. Stat*, **37**, 3133–3164.

Maathuis,M.H. *et al*. (2010) Predicting causal effects in large-scale systems from observational data. *Nat. Methods*, **7**, 247–248.

Madigan,D. *et al*. (1996) Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs. *Communications in Statistics: Theory and Methods*, **25**, 2493–2519.

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.

Meek,C. (1995) Causal inference and causal explanation with background knowledge. In: Besnard,P. and Hanks,S. (eds.) *UAI*, Morgan Kaufmann, Publishers, San Francisco, CA, USA, pp. 403–410.

Meinshausen,N. and Bühlmann,P. (2010) Stability selection. *J. R. Stat. Soc. Ser. B (Statistical Methodology)*, **72**, 417–473.

Pearl,J. (1988) *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, Publishers, San Francisco, CA, USA.

Pearl,J. (1995) Causal diagrams for empirical research. *Biometrika*, **82**, 669.

Pearl,J. (2000) *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, UK.

Spirtes,P. *et al*. (2000) *Causation, Prediction, and Search*. MIT press, Cambridge, MA, USA, 2nd edn.

Stekhoven,D.J. *et al*. (2012) Causal stability ranking. *Bioinformatics*, **28**, 2819–2823.

Tian,J. and Pearl,J. (2002) A general identification condition for causal effects. In: Dechter,R. and Sutton,R.S. (eds.) *AAAI/IAAI*, AAAI Press/The MIT Press, Cambridge, MA, USA, pp. 567–573.

Verma,T. and Pearl,J. (1991) Equivalence and synthesis of causal models. In: Bonissone, P.P. *et al*. (eds.) *UAI*, Elsevier, Science Publishers B.V., Atlanta, GA, USA, pp. 255–270.

Wach,A. (1996) PCR-synthesis of marker cassettes with long flanking homology regions for gene disruptions in *S. cerevisiae*. *Yeast*, **12**, 259–265.

Zhang,J. and Spirtes,P. (2008) Detection of unfaithfulness and robust causal inference. *Minds Mach*., **18**, 239–271.