Supplementary: aIDA
Franziska Taruttis 1,* Rainer Spang 1 and Julia C. Engelmann 1*

1 Institute of Functional Genomics, University of Regensburg, 93053 Regensburg, Germany.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

1 GRAPH NOTATION AND DEFINITIONS
A graph $G$ is described by a set of nodes $X = X_1, ..., X_p$ and a set of edges $E = E_1, ..., E_s$. The edges $E$ show relationships between the nodes. The kind of relationship differs depending on the application. Here we are interested in causal relationships between variables, which leads to the use of Causal Bayesian Networks (Pearl (2000)). We distinguish between directed and undirected edges. Directed edges are represented by $\rightarrow$ and describe the fact, that the relationship between two nodes has a defined direction (e.g.: $A$ causes $B$). Undirected edges are shown as $\equiv$ and denote a relationship between two nodes without a certain direction (e.g.: $A$ and $B$ are associated). If all edges in a graph are directed or undirected the graph is called a directed or undirected graph, respectively. If some edges are directed while others remain undirected the graph is called partially directed. If all directed edges in a graph are pointing to the middle node $m$ and denote a relationship between two nodes without a certain direction (e.g.: $A$ causes $B$). Undirected edges are shown as $\equiv$ and denote a relationship between two nodes without a certain direction (e.g.: $A$ and $B$ are associated).

Definition 1.1. (d-Separation by Judea Pearl, definition 1.2.3, p. 16, Pearl 2009)

"A path $p$ is said to be d-separated (or blocked) by a set of nodes $Z$ if and only if

(i) $p$ contains a chain $i \rightarrow m \rightarrow j$ or a fork $i \leftarrow m \rightarrow j$ such that the middle node $m$ is in $Z$, or

(ii) $p$ contains an inverted fork (or collider) $i \rightarrow m \leftarrow j$ such that the middle node $m$ is not in $Z$ and such that no descendant of $m$ is in $Z$.

A set of nodes $Z$ is said to d-separate $X$ from $Y$ if and only if $Z$ blocks every path from a node in $X$ to a node in $Y$.

If $Z$ is d-separating $X$ from $Y$, then $X$ and $Y$ are independent given $Z$.

2 ADDITIONAL PERFORMANCE EVALUATION ON SIMULATED DATA

Fig. S1. Comparison of the partial area under the ROC curve up to 100 false positives (pAUC(FP=100)) for the simulated datasets from dense underlying graphs with $n=50$ samples for different values of $\alpha$ and for using the true underlying network. Black bars show the pAUC(FP=100) for the simulated datasets for aIDA and gray bars show the pAUC(FP=100) of the same datasets for CStaR (Stekhoven et al., 2012). The pAUC was calculated for four different values of the tuning parameter of the PC algorithm and for using the true underlying network. For $\alpha = 0.1, 0.3$ and $0.5$ we observe that larger values of $\alpha$ lead to an improved prediction of causal effects from observational data for the simulated datasets. Both CStaR and aIDA performed better than random guessing for all values of $\alpha$.

* to whom correspondence should be addressed
To achieve the data of a certain node $X_i$, all parents of $X_i$ ($pa_i^1, ..., pa_i^m$) are considered. $X_i$ is calculated by the following equation

$$X_i = w_1 * pa_i^1 + ... + w_k * pa_i^m + Err_i,$$

where $w_1, ..., w_k$ denote the weights of the incoming edges to $X_i$, and $Err_i$ defines the error distribution.

In our scenario all $Err_i$ are sampled from a normal distribution $N(0, 0.001)$.

### 3.2 Hughes et al. (2000) dataset

The Hughes dataset (Hughes et al. (2000)) is an $S.cerevisae$ dataset, used in the IDA (Maathuis et al. (2010)) and in the CStaR (Stekhoven et al. (2012)) articles. It consists of 234 single gene knock-outs, which are treated as interventional data, and 63 wild type gene expression profiles with 5361 genes each (observational data). The preprocessing was analogous to Maathuis et al. (2010) and Stekhoven et al. (2012).

### 3.3 Lenstra et al. (2011) dataset

The second $S.cerevisae$ dataset was taken from Lenstra et al. (2011). Data is available from ArrayExpress under Accession numbers E-TABM-1074 for the interventional dataset and E-TABM-984 for the observational dataset. The interventional dataset is a set of non-essential knock-outs of chromatin modifiers in $S.cerevisae$, while the observational data is a collection of different wild type strains from $S.cerevisae$. Both datasets were measured on A-UMCU-10 - UMC Utrecht S. cerevisae 16K two channel arrays (version 1.3).

As interventional data the file E-TABM-1074-processed-data-2522364894.txt was used. Excluding from the 165 available knockouts those that did not map to two biological replicates in E-TABM-1074.sdrf.txt or knocked out genes that could not be mapped to ensembl IDs via SGD IDs by A-UMCU-10.adf.txt and biomart (Kasprzyk (2011)), we were left with 138 single gene interventions. The log fold changes between the two channels were computed using the R package limma (Smyth (2005)).

As observational data the file E-TABM-984-processed-data-2481698487.txt was used which consists of 47 wild type $S.cerevisae$ strains. The log fold changes between the two channels were computed using the R package limma (Smyth (2005)). In both datasets only genes, which could be mapped to ensembl gene IDs, were considered, leaving a set of 4890 genes. No further gene selection was performed.

This procedure led to an interventional dataset with 138 knockouts and 4890 genes, and an observational dataset with 67 samples and 4890 genes.

Both datasets were standardized to obtain $N(0,1)$ centered and scaled gene expression data for each measured gene.

### 4 THE BACK-DOOR CRITERION IN BIOLOGICAL NETWORKS

We used data from the DREAM3 In-Silico Network Challenge (Marbach et al., 2009; 2010; Prill et al., 2010) to analyze how frequently the Back-door Criterion is fulfilled in biological networks. These datasets are sampled from subgraphs of $E.coli$ and $S.cerevisae$ biological networks. In the DREAM3 challenge, data
Table S1. Overview of the data sets with 100 nodes from the DREAM3 In-Silico Network Challenge

| Network | Nodes | Edges | Regulators |
|---------|-------|-------|------------|
| Ecoli1  | 100   | 125   | 26         |
| Ecoli2  | 100   | 119   | 19         |
| Yeast1  | 100   | 166   | 60         |
| Yeast2  | 100   | 389   | 71         |
| Yeast3  | 100   | 551   | 81         |

Fig. S3. Percentage of causal effects where the estimated parents fulfill the Back-door criterion (black) or are the true parents (gray), for the five DREAM3 data sets over the 100 subsampling runs of 46 time series were sampled from these graphs. Since we do not require time series for aIDA we only use data from time point 0 for network reconstruction. Thus, we have 5 different data sets (2 E. coli, 3 S. cerevisae) with 100 variables and 46 observations each (supplementary Table S1). Supplementary Figure S3 endorses our observation from random networks, that the backdoor criterion is often met, even if true parents are not detected. For these datasets we performed the same analysis as we did for Figure 1 of the manuscript.

5 DEFINITION OF THE TARGET SET OF CAUSAL EFFECTS

5.1 Hughes and Holstege data

The target set of causal effects for both S.cerevisae datasets of gene i on gene j were calculated as described by Maathuis et al. (2010) using the following formula

\[
\beta_{ij} = \frac{|a_{i,j} - \text{mean}(a_{-i,j})|}{|a_{i,c(i)} - \text{mean}(a_{-i,c(i)})|},
\]

where \(a_{i,j}\) are the entries of the interventional data matrix with knock-outs in the rows and genes in the columns and \(\text{mean}(a_{-i,j})\) is a short cut for the mean of the j-th column of the interventional data matrix without considering the ith row.

The top absolute highest 5% causal effects define our target set, which is the set of causal effects we want to predict. Nevertheless, since biochemical experiments that characterize direct interactions are missing, this target set is not the set of the top true causal effects, but we expect an enrichment of causal interactions within that set.

5.2 Simulated datasets

Since the true underlying DAG is known in that case, we apply the second part of the IDA method, the estimation of causal effects given the data and the DAG, to our sampled data and DAG. These causal effects represent our ground truth. The top absolute highest 5% true causal effects define our target set, which is the set of causal effects we want to predict.

6 CSTAR PARAMETERS

For the small simulated datasets with 100 nodes we chose \(q \in \{1\%,\ 1.2\%,\ 1.4\%,\ ...,\ 10\%\}\) of all possible causal effects. For the large simulated dataset with 1000 nodes and the two gene expression datasets from S.cerevisae (Hughes et al., 2000; Lenstra et al., 2011) we used \(q \in \{0.01\%,\ 0.03\%,\ 0.05\%,\ ...,\ 1\%\}\) of all possible causal effects. Stekhoven et al. (2012) showed that the results were insensitive to the choice of the range of qs.
7 EVALUATION OF NETWORK DENSITY IN SIMULATED NETWORKS

Fig. S4. Comparison of the number of parents of the true network and the estimated networks using different values of $\alpha$ for the 5 dense simulated datasets with 1000 nodes, 50 samples and approximately 2500 edges. The distribution of parents estimated using the higher $\alpha$ value is more similar to the true distribution of the number of parents. The error bars show standard deviations over the 5 datasets.

Fig. S5. Comparison of the number of parents of the true network and the estimated networks using different values of $\alpha$ for the 5 sparse simulated datasets with 1000 nodes, 50 samples and approximately 1250 edges. The distribution of parents estimated using the higher $\alpha$ value is more similar to the true distribution of the number of parents. The error bars show standard deviations over the 5 datasets.

REFERENCES

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoungton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepiants, S. B., Shoemaker, D. D., Gachotte, D., Chakraberty, K., Simon, J., Bard, M., and Friend, S. H. (2000). Functional discovery via a compendium of expression profiles. Cell, 102(1), 109–126.

Kalisch, M., Machler, M., Colombo, D., Maathuis, M. H., and Bühlmann, P. (2012). Causal inference using graphical models with the r package pcalg. Journal of Statistical Software, 47, 1.

Kasprzyk, A. (2011). BioMart: driving a paradigm change in biological data management. Database, 2011(0), ba049.

Lenstra, T. L., Benschop, J. J., Kim, T., Schulze, J. M., Brabers, N. A., Margaritis, T., van de Pasch, L. A., van Heesch, S. A., Brok, M. O., Groot Koerkamp, M. J., Ko, C. W., van Leenen, D., Sameith, K., van Hooft, S. R., Lijnzaad, P., Kemmeren, P., Hentrich, T., Kobor, M. S., Buratowski, S., and Holstege, F. C. (2011). The specificity and topology of chromatin interaction pathways in yeast. Molecular cell, 42(4), 536–549.

Maathuis, M. H., Colombo, D., Kalisch, M., and Bühlmann, P. (2010). Predicting causal effects in large-scale systems from observational data. Nature Methods, 7, 247–248.

Marbach, D., Schaffter, T., Mattiussi, C., and Floreano, D. (2009). Generating Realistic In Silico Gene Networks for Performance Assessment of Reverse Engineering Methods. Journal of Computational Biology, 16(2), 229–239. WingX.

Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., and Stolovitzky, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. PNAS, 107(14), 6286–6291. WingX.

Pearl, J. (2000). Causality: Models, reasoning and inference. Cambridge University Press, Cambridge, UK.

Prill, R. J., Marbach, D., Saez-Rodriguez, J., Sorger, P. K., Alexopoulos, L. G., Xue, X., Clarke, N. D., Altan-Bonnet, G., and Stolovitzky, G. (2010). Towards a rigorous assessment of systems biology models: The dream3 challenges. PLoS ONE, 5(2), e9202.

Smyth, G. K. (2005). Limma: linear models for microarray data, pages 397–420. Springer, New York.

Stekhoven, D. J., Moraes, I., Sveinbjörnsson, G., Hemmig, L., Maathuis, M. H., and Bühlmann, P. (2012). Causal stability ranking. Biosinformatics, 28(19), 2819–2823.