Validation of gene knock-out predictions in large-scale gene perturbation experiments

Philip Versteeg
Informatics Institute
University of Amsterdam
Amsterdam, the Netherlands
p.j.j.p.versteeg@uva.nl

Sach Mukherjee
German Center for
Neurodegenerative Diseases
Bonn, Germany
sach.mukherjee@dzne.de

Joris M. Mooij
Informatics Institute
University of Amsterdam
Amsterdam, the Netherlands
j.m.mooij@uva.nl

1 Introduction

To understand the workings of a high-dimensional complex system such as a gene regulatory network, it is important to determine how its single components effects the other components. In other words: can we predict what happens to the thousands of other genes in a genome if we knock out a single one, and without actually performing the experiment?

In a recent large-scale micro-array experiment, both observational and perturbation data has been measured [1], offering unique opportunities for training algorithms and validating those predictions with the interventional data. The data consists of genome-wide expressions for all 6170 genes in the model species *Saccharomyces cerevisiae*, under 262 wild-type, observational settings and 1479 single-gene knockout experiments. Here we investigate several simple methods for scoring a causal effect, i.e. the level change in gene expression when disabling another gene. We validate these predictions with several different scores calculated on the remaining test set.

In earlier work [2], a conservative set of causal effects is estimated with the ICP method applied on the same micro-array dataset. Validation was performed by comparing predictions to a small set of true positives with strong intervention effects and against several external sources from an on-line curated compendium of true edges and a sparse transcription-factor network as "ground-truth" networks. Furthermore, other work by Maathuis et al. [3] compared results of the IDA algorithm trained on the observational part of the data to knock-out data in an experimentally similar setting, but the validation was found to depend sensitively on the definition of the true causal effect [2].

2 Ground-truth scores

Let X_{ij} be the j-th sample of the observed expression level[1] for gene i and let $A_{c(i)j}$ be the expression of gene j when gene i has been knocked-out (intervened), where c is is the gene that is intervened on in the i'th knockout experiment. The sample observational mean and sample standard deviation are notated as μ_j and σ_j respectively.

We define several data-driven measures S_{ij} to score the causal effect, i.e., the change in expression level of gene j under an intervention of gene i.

gt.abs Absolute difference of the interventional expression of a gene and its observational mean, i.e. $S^a_{ij} = |A_{ij} - \mu_j|$.

gt.abs.norm Normalized version of S^a_{ij}, scaled by observational standard deviation, i.e. $S^{an}_{ij} = \frac{|A_{ij}-\mu_j|}{\sigma_j}$.

[1] Log ratio of the mRNA fluorescence intensity of wild-type measurements versus a control.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

gt.rel Similar to S^a_{ij} but relative to the absolute expression that a gene when it is intervened, i.e. $S^r_{ij} = \frac{|A_{ij} - \mu_j|}{|A_{i,c(i)} - \mu_{c(i)}|}$.

gt.rel.norm Normalized version of S^r_{ij}, scaled by observational standard deviation, i.e. $S^{rn}_{ij} = S^r_{ij} \frac{\sigma_{c(i)}}{\sigma_j}$.

3 Predictions and validation

Using observational data X_{ij}, we compute several different scores to T_{ij} to estimate the strength of a causal effect, i.e., the change in expression level in gene j when disabling gene i. For each pair $X_{i.}, X_{j.}$, gene i causing j is scored using Pearson, Spearman and Kendall correlation coefficients, and the regression coefficient obtained from a simple ridge regression with cross-validation. We also add the covariance as estimated by the graphical LASSO. Finally, we have several predictions based on the sample variance:

pred.var.ef Score all i causes j according to the sample variation of j: $T^{\text{ef}}_{ij} = \sigma^2_j$

pred.var.ca.inv Similar to **pred.var.ef** but for the reciprocal of the cause: $T^{\text{ca}}_{ij} = \frac{1}{\sigma_c(i)^2}$

pred.var.ratio The combination of the earlier two scores: $T^{\text{ratio}}_{ij} = \frac{\sigma^2_j}{\sigma_c(i)^2}$

The dataset is randomly split in a training set containing 100 expression levels under observational conditions and 100 expressions under single-gene interventions for all 1479 genes that have knock-out data available. The remainder 162 observations and 1379 interventions are grouped in a test set for validation purposes. The predictions methods above are applied once on the observational training data X^{train}_{ij} and once on the interventional data A^{train}_{ij}, where expression levels of a gene under 100 different interventions are pooled together and interpreted as observations.

The area-under-the-curve (AUC) statistic is used to compare the performance of each training method against a set of true positives consisting of the top m percentage of strongest effects in each ground-truth score. Fig. 1 shows an overview for all methods trained on observational and interventional data separately, compared against the same validation set.[2]

Observational coefficient methods, such as Pearson, Spearman, Kendall coefficients, Ridge regression and GLASSO are showing similar to random performance comparing to the above ground-truth, for both $m = \{1, 5\}$, using either observations or interventions for training.

The variance measures **pred.var.ef**, **pred.var.ca.inv** and **pred.var.ratio** perform very well compared to the other methods for the pooled interventional data. In particular, **pred.var.ef** versus the absolute score **gt.abs** as true positives has an AUC of 0.81, while a near random AUC was found when comparing against **gt.rel.norm**.

4 Conclusions

We defined a set of measures for the strength of causal effect using the data and compared several methods for predicting the strength of an unknown intervention to these measures. Genome-wide inference of causal effects using out-of-the-box methods with observational data is shown to be a difficult task, but pooling interventional data as observational shows promising results. Remarkably, simple prediction scores using the sample variance perform very well against several of the data-driven ground-truth.

References

[1] Kemmeren, P., Sameith, K., van de Pasch, L.A., Benschop, J.J., Lenstra, T.L., Margaritis, T. *et al.* (2014) Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors *Cell* **157**:740-752

[2] Results using the algorithm of [3] and the different ground-truth notion defined there should be available in time for the workshop. An application to the comparable micro-array dataset used in there should be included as well.

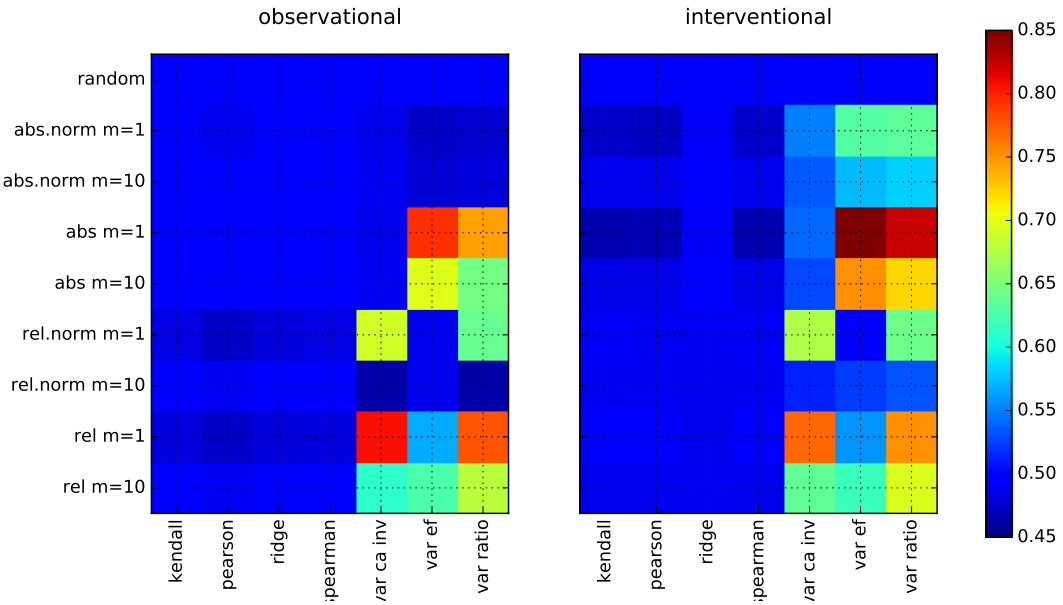

Figure 1: Overview of AUC when comparing prediction methods(horizontally) to different ground-truth measures (vertically). Left: prediction on training set of 100 random observational data-points; Right: training set of 100 random interventions. The true positives are defined as scores different cutoff percentile $m = \{1, 5\}$ and calculated with the remaining 1379 interventions and 162 observations.

[2] Meinshausen, N., Hauser, A., Mooij, J.M., Peters, J., Versteeg, P. & Bühlmann, P. (2016) Methods for causal inference from gene perturbation experiments and validation *Proc. Natn. Acad. Sci.* **113**(27):7361-7368

[3] Maathuis, M.H., Colombo, D. & Bühlmann, P. (2010) Predicting causal effects in large-scale systems from observational data. *Nat. Methods* **7**(4):247-248