Causal Stability Ranking

Daniel J. Stekhoven[1,2,3] , Izabel Moraes[4], Gardar Sveinbjörnsson[1],

Lars Hennig[4,5], Marloes H. Maathuis[1] and Peter Bühlmann[1,3]

[1] Seminar for Statistics, Department of Mathematics, ETH Zurich, Switzerland. [2] Life Science Zurich PhD Program on Systems Biology of Complex Diseases. [3] Competence Center for Systems Physiology and Metabolic Diseases, Zurich, Switzerland. [4] Uppsala BioCenter, Department of Plant Biology and Forest Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden. [5] Plant Biotechnology, Department of Biology, ETH Zurich, Switzerland
Correspondence should be addressed to D.J.S. (stekhoven@stat.math.ethz.ch).

Nature Precedings : hdl:10101/npre.2011.6716.1 : Posted 23 Dec 2011

Genotypic causes of a phenotypic trait are typically determined via randomized controlled intervention experiments. Such experiments are often prohibitive with respect to durations and costs. We therefore consider inferring stable rankings of genes, according to their causal effects on a phenotype, from observational data only. Our method allows for efficient design and prioritization of future experiments, and due to its generality it is useable for a broad spectrum of applications.

The growing interest in causal inference[1] has increased not only the need for methods able to handle this task but also for designed experimental validation. It is of general interest to infer the genotypic causes of a complex phenotypic trait[2]. The classical approach relies on randomized controlled intervention experiments, e.g., knocking out a gene and observing the effect on the phenotype relative to the wild-type organism. However, such intervention experiments are time consuming and expensive. We therefore consider the problem of inferring causal effects from data obtained by observing a system without subjecting it to targeted interventions (observational data). This problem is generally ill-posed, but the recently proposed IDA method[3,4] provides estimated lower bounds of causal effects from observational data under some assumptions. However, these bounds come without a measure of uncertainty. We address this issue by introducing a new method combining IDA and a version of stability selection[5], which we call Causal Stability Ranking (CStaR; **Supplementary Section 1 and Figure 1**). The addition of stability selection to IDA provides two advantages. First, CStaR leads to a stable ranking of biomarkers (e.g. genes) according to the size of their causal effects, irrespective of the choice of the tuning parameter in stability selection. Second, under some additional assumptions, CStaR allows controlling an error rate of false positive findings, namely the expected number of false positives and hence also the per-comparison error rate (PCER).

We validated CStaR in two situations. First, we trained CStaR on a publicly available compendium of *Arabidopsis thaliana* gene expression data and performed new biological validation experiments. The compendium contains 47 expression profiles of natural accessions from diverse geographic origins[6] (**Supplementary Sections 2.1 and 2.2**). The phenotypic trait of interest is time to flowering, which is robustly measured by the number of days to bolting or the number of rosette leaves formed before bolting[7]. Timing of flowering according to local climatic conditions is a major determinant of plants' reproductive success and an important agronomical trait that greatly affects yield. Therefore, an improved knowledge about genes controlling flowering time is of great economic value[8].

CStaR scores five known regulators of flowering time (*DWF4, FLC, FRI, RPA2B* and *SOC1*)[7,9,10] in its top 25 (**Table 1**). In particular, *SOC1, FRI* and *FLC* are curated flowering time genes in Arabidopsis Reactome[11] (http://www.arabidopsisreactome.org). This is a highly significant enrichment of known curated regulators when compared to random guessing ($p<10^{-5}$). Interestingly, *FLC* and *FRI* are not only major regulators of flowering time in the model species *A. thaliana* but also in the oil-seed rape crop.

Among the other genes in the top 25, which were not already known to play a role in flowering time, there were 13 genes for which mutant seeds were readily available (**Supplementary Table 1**). These mutants were used for intervention experiments (**Supplementary Section 2.3**) in order to further validate CStaR and to discover new influential genes for flowering time in *A. thaliana*.

The experiments were performed under two photoperiod conditions, short-day (SD) and long-day (LD) with 8h and 16h of light respectively. As phenotypic responses, the number of days to bolting (DTB, for both SD and LD) as well as the rosette leave number (RLN, only for LD) were recorded. Seed viability varied between different genotypes (**Supplementary Tables 2, 3 and 4**) reducing the number of testable mutants to nine

(**Supplementary Table 1**). Differences between the knock-out and control group were tested using a two-sided Welch's t-test. Four new genes were found to have a significant causal effect on the phenotypic responses at level $\alpha=0.05$ in at least one of the three settings (**Table 2, Supplementary Section 2.4**). Among the significant genes is *OTLD1*, a gene involved in chromatin modifications, which may potentially regulate *FLC* expression. Another significant gene is *PDH-E1*, which is involved in carbohydrate metabolism, a known regulation point of flowering time. Future studies of the identified novel genes may increase the biological understanding of flowering time control and provide potential targets for breeding strategies in crops. The entire approach from modeling to biological experiments and findings is schematically described in **Figure 1**.

As a second validation of the CStaR method, we compared it with the plain IDA method, Lasso[12], elastic net[13] and marginal correlation ranking on a publicly available data set of gene expression profiles in *Saccharomyces cerevisiae*[14] (**Supplementary Section 3**). This data set includes both observational and interventional data obtained under similar conditions. Hence, it forms an excellent basis to assess the performance of methods aimed at estimating causal effects from observational data, as the effects estimated from the observational data can be compared to the effects inferred from the interventional data. These data were used to validate IDA[4], and we followed the same approach to validate CStaR. In particular, we used the interventional data to infer the causal effects of the knock-out genes on the remaining genes and defined the top 5% of the effects that were largest in absolute value as the target set. We then trained all methods on the observational data, and compared their receiver operating characteristic (ROC) curves on absolute scale (**Figure 2**) showing a clear improvement of CStaR over plain IDA. Moreover, CStaR and IDA are clearly superior to high-dimensional regression methods and marginal correlation screening, which is in line with the earlier validation of IDA[4].

We propose CStaR as a general method to obtain a stable ranking of genes in terms of the strengths of their causal effects on a phenotype of interest. An added value of our method is that, under some assumptions, this ranking comes with an error measure controlling false positive selections. We showed that CStaR exhibits a large increase in sensitivity when compared to plain IDA and other methods in *S. cerevisiae* (**Figure 2**). Moreover, we demonstrated the success of CStaR for the biologically much more complex multicellular organism *A. thaliana*. This makes it plausible that CStaR is relevant for commercial crops, by providing better target genes for marker-assisted breeding and transgenic approaches. In fact, since CStaR is mathematically justified under clearly stated assumptions[3,5], it has the potential to generalize to many other settings in biology, agriculture and other fields where efficient design and prioritization of new intervention experiments is a core aim.

Note: Supplementary information is available on the Nature Biotechnology website.

Availability

The CStaR method is implemented in the statistical software R. An example script and the full ranking from **Table 1** can be found at http://stat.ethz.ch/~hoven/cstar/.

Acknowledgments

We thank T. Wey for the technical assistance with plant work.

Author Contributions

D.S., L.H., M.M. and P.B. conceived the project; D.S. generated the causal gene ranking for *A. thaliana*; I.M. performed biological validation experiments for *A. thaliana*; L.H. designed biological validation experiments for *A. thaliana* and analyzed the results; D.S. performed tests on biological validation results for *A. thaliana*; G.S. performed validation on *S. cerevisiae*. All authors discussed the results and commented on the manuscript.

Competing Financial Interests

The authors declare no competing financial interests.

References

1. Kruglyak, L. & Storey, J.D. *Nat. Biotechnol.* **27**, 544-545 (2009).
2. Glazier, A.M. & Nadeau, J.H. & Aitman, T.J. *Science* **298**, 2345-2349 (2002).
3. Maathuis, M.H. & Kalisch, M. & Bühlmann, P. *Ann. Stat.* **37**, 3133-3164 (2009).
4. Maathuis, M.H. & Colombo, D. & Kalisch, M. & Bühlmann, P. *Nat. Methods* **7**, 247-248 (2010).
5. Meinshausen, N. & Bühlmann, P. *J. Roy. Stat. Soc. B Met.* **72**, 417-473 (2010).
6. Lempe, J. *et al. PLoS Genetics* **1**, 109-118 (2005).
7. Amasino, R. *Plant J.* **61**, 1001-1013 (2010).
8. Craufurd, P.Q. & Wheeler, T.R. *J. Exp. Bot.* **60**, 2529-2539 (2009).
9. Domagalska, M.A. *et al. Development* **134**, 2841-2850 (2007).
10. Xia, R. *et al. Plant Cell* **18**, 85-103 (2006).
11. Tsesmetzis, N. *et al. Plant Cell* **20**, 1426-1436 (2008).
12. Tibshirani, R. *J. Roy. Stat. Soc. B Met.* **58**, 267-288 (1996).
13. Zou, H. & Hastie, T. *J. Roy. Stat. Soc. B Met.* **67**, 301-320 (2005).
14. Hughes, T.R. *et al. Cell* **102**, 109-126 (2000).

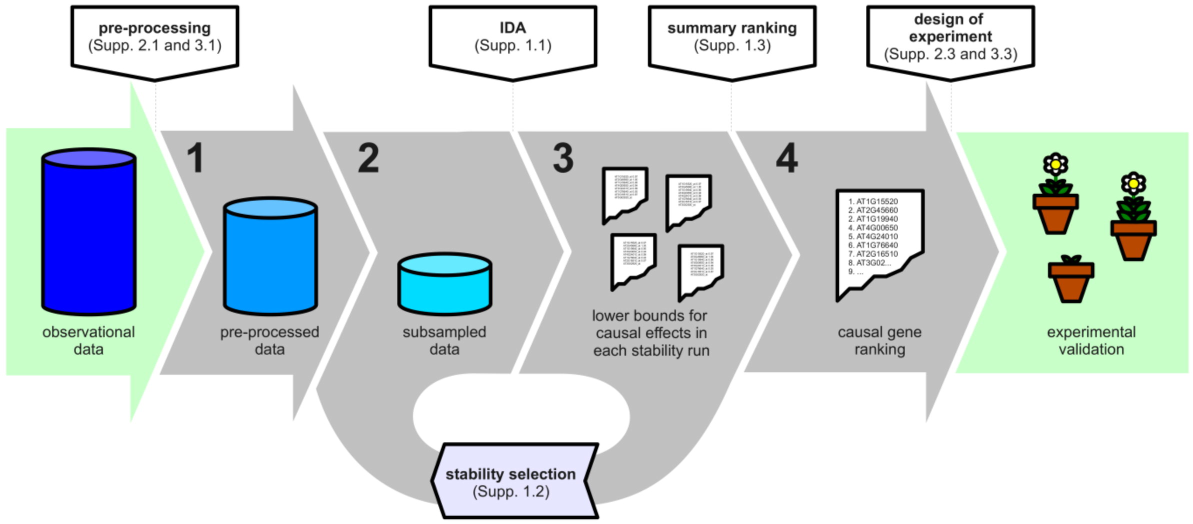

Figure 1 Schematic overview of the methodological framework used in CStaR. After pre-processing the data (Step 1), lower bounds for the causal effects are estimated 100 times using stability selection[5] according to the following procedure: a subsample of size $\left\lfloor \frac{n}{2} \right\rfloor$ is drawn from the total of n pre-processed data points (Step 2). For reasons of comparability, the pre-selected gene expression data are standardized to have gene-wise mean zero and standard deviation one. On this subsample lower bounds for the causal effects are then estimated using IDA[4] and used to rank the genes in each stability run (Step 3, **Supplementary Section 1.1**). Next, for a range of different q-values, we record the relative frequencies over the 100 stability runs that each gene appeared in the top q ranks (**Supplementary Section 1.2**). The median rank over these different q's is used to generate the final ranking of the genes (Step 4). Furthermore, under additional assumptions, an upper bound for the per-comparison error rate (PCER) is estimated for each q-value and its corresponding relative frequency (**Supplementary Section 1.3**). Finally, the gene ranking allows for design of experiment. Thus, a biological validation using intervention experiments can be performed. We tested CStaR in two situations. First, on a publicly available compendium of 31 natural *A. thaliana* accessions consisting of n=47 gene expression measurements, each with 21,326 genes and corresponding flowering time data[6] (**Supplementary Section 2.1**). We performed biological intervention experiments according to the causal gene ranking (**Table 1**) by focusing on candidates that were not already known to control flowering time and for which mutant seeds were readily available (**Supplementary Section 2.3**). The biological experiments were analyzed using a two-sample Welch's t-test (**Supplementary Section 2.4**). The second validation was performed on a publicly available data set in *S. cerevisiae* containing n=63 observational and 234 interventional full-genome expression profiles, with p=5,361 genes[14] (**Supplementary Section 3**). Since this data set includes both observational and interventional data, the validation was analyzed by comparing estimated causal effect on the observational data with inferred effects from the interventional data (**Figure 2**).

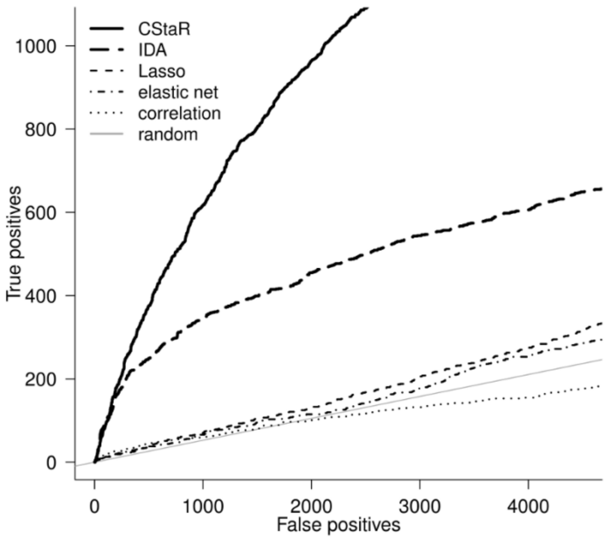

Figure 2 True positive selections (y-axis) versus false positive selections (x-axis) for CStaR (solid) versus plain IDA[4] (long dashed), Lasso[12] (short dashed), elastic net[13] (dash dotted) and marginal correlation ranking (dotted) in the *S. cerevisiae* validation (**Supplementary Section 3**). Random guessing is indicated by the grey line. All methods were trained on the observational data. True positives were defined as the largest 5% of the effects (in absolute value) inferred from the interventional data.

	gene	summary rank	median effect	maximum expression	error (PCER)	name/annotation
1	AT2G45660	1	0.60	5.07	0.0032	SOC1
2	AT4G24010	2	0.61	5.69	0.0033	ATCSLG1
3	**AT1G15520**	**2**	**0.58**	**5.42**	**0.0033**	**PDR12**
4	AT3G02920	5	0.58	7.44	0.0041	RPA2B
5	**AT5G43610**	**5**	**0.41**	**4.98**	**0.0069**	**ATSUC6**
6	AT4G00650	7	0.48	5.56	0.0051	FRI
7	**AT1G24070**	**8**	**0.57**	**6.13**	**0.0040**	**ATCSLA10**
8	**AT1G19940**	**9**	**0.53**	**5.13**	**0.0045**	**ATGH9B5**
9	**AT3G61170**	**9**	**0.51**	**5.12**	**0.0044**	**PPR protein**
10	**AT1G32375**	**10**	**0.54**	**5.21**	**0.0045**	**F-box protein**
11	**AT2G15320**	**10**	**0.50**	**5.57**	**0.0047**	**LRR protein**
12	**AT2G28120**	**10**	**0.49**	**6.45**	**0.0054**	**nodulin protein**
13	AT2G16510	13	0.50	10.7	0.0050	AVAP5
14	AT3G14630	13	0.48	4.87	0.0056	CYP72A9
15	**AT1G11800**	**15**	**0.51**	**6.97**	**0.0053**	**endonuclease**
16	AT5G44800	16	0.32	6.55	0.0079	CHR4
17	AT3G50660	17	0.40	7.60	0.0078	DWF4
18	AT5G10140	19	0.30	10.3	0.0085	FLC
19	**AT1G24110**	**20**	**0.49**	**4.66**	**0.0071**	**peroxidase**
20	**AT2G27350**	**20**	**0.48**	**7.06**	**0.0067**	**OTLD1**
21	AT1G27030	20	0.45	10.0	0.0075	unknown protein
22	**AT2G28680**	**22**	**0.46**	**5.23**	**0.0072**	**cupin protein**
23	AT3G16370	23	0.43	12.4	0.0099	lipase/hydrolase
24	AT5G25640	23	0.33	5.59	0.0091	serine protease
25	AT1G30120	24	0.46	9.97	0.0077	PDH-E1 BETA

Table 1 Top 25 findings by CStaR for the *A. thaliana* data. The genes are ranked by increasing summary rank, where ties are sorted according to the estimated median causal effect taken over 100 stability runs (third column). The maximum expression is taken over the original log$_2$ data. The error is the median PCER over the range of *q* values. *SOC1*, *FRI* and *FLC* are three of 119 curated flowering time genes in Arabidopsis Reactome[11] (http://www.arabidopsisreactome.org). This is a highly significant enrichment of known curated regulators when compared to random guessing ($p<10^{-5}$). Although not curated in Arabidopsis Reactome, also *RPA2B* and *DWF4* are known to affect flowering time[9,10]. Since the ordering of the genes is given by their summary rank, the values of median causal effect and per-comparison error rate (PCER) are not decreasing monotonously. For instance, *ATSUC6* has a smaller median causal effect and a larger PCER than the endonuclease, but since its lower bound for the causal effect is more stable, the former is ranked ten positions higher than the latter. All genes from this list, for which mutant seeds were readily available and which were not already known to control flowering time, were used in the subsequent intervention experiments (indicated in bold). In total, intervention experiments were performed for 13 of the 25 top genes not implemented in flowering yet (**Supplementary Section 2.3**).

	Welch’s t-test		
gene	DTB-SD	DTB-LD	RLN-LD
PDH-E1 BETA	0.04	0.04	*0.91*
ATGH9B5	0.02	*0.15*	0.04
LRR protein (*AT2G15320*)	*0.66*	0.03	*0.47*
OTLD1	*0.43*	0.03	*0.86*

Table 2 P-values from two-sided Welch's t-tests in the *A. thaliana* validation, showing only genes significant in at least one of the following three settings: days to bolting in short days (DTB-SD), days to bolting in long days (DTB-LD), and rosette leave number in long days (RLN-LD). Each mutant was tested versus a control group. P-values larger than 0.05 are written in italics (for complete results see **Supplementary Tables 2, 3 and 4**).