

# Gene Selection using Causal Ranking

Daniel Stekhoven

Seminar for Statistics

CC-SPMD retreat - 28.05.10



**CC-SPMD**

Competence Center for  
Systems Physiology  
and Metabolic Diseases

# Outline

- 1 Introduction
- 2 Method
- 3 Results
- 4 Conclusion & Outlook
- 5 References

# Design of experiments

## Given

- a **response** (e.g. phenotype of interest - in our case the number of leaves of an Arabidopsis plant);
- and a set of **observational** data (wild-type gene expressions);

## we want

- a **stable ranking** for the causal effect of the genes on the phenotype (response);

## with the intention to

- offer powerful design for selecting important **candidate genes** for future interventional experiments.

# The data

## *Arabidopsis thaliana* gene expression data

- **observational** microarray data with  $n = 47$  and  $p = 21'326$ ;
- samples are from 35 *A. thaliana* **oecotypes** (D. Weigel, Tübingen);
- samples were hand-picked according to the response (L. Hennig, Grisse Lab, ETH Zurich).

## The response

- point of time, when first flower occurs;
- a robust measure for time: **number of leaves** grown.

# Correlation does not imply causation

## *Cum hoc ergo propter hoc* - or the **false** cause

- **A** is correlated with **B**;
- therefore **A** is caused by **B**.

## Possible conclusions

- 1 A causes B;
- 2 B causes A;
- 3 an unobserved C causes A and B;
- 4 A causes B *and* B causes A;
- 5 the above correlation is a coincidence.

# Correlation does not imply causation

## Example

- Sleeping with your shoes on (**A**) is strongly correlated with waking up having a headache (**B**).
- Therefore, sleeping with your shoes on (**A**) causes headache (**B**).

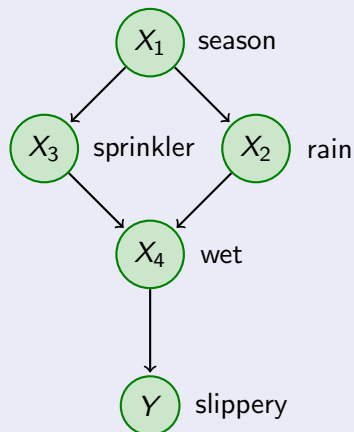
## ...rather

- an unobserved **C** causes **A** and **B**;

...in this case alcohol intoxication (**C**), which thereby gives rise to a correlation.

# Bayesian networks

directed acyclic graphs



directed acyclic graph (DAG)

- all edges are **directed** (one-way);
- and there are no **cycles** in the graph.

causal graph

- arrows encode causal influence;
- graph was directed by causal intuition.

# Predicting causal effects

## method outline

- 1 **Subsampling** for stability reasons;
- 2 estimating **causal graph** using PC algorithm;
- 3 estimating **causal effects** of genes using intervention calculus;
- 4 **repeat** step 1. - 3. many times;
- 5 take summary of rankings as final result.

## proof of concept

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

# Causal gene ranking

	summary	median		error	
Gene	rank	effect	expression	(PCER)	name
1 <b>AT2G45660</b>	1	0.60	5.07	0.0017	<b>AGL20 (SOC1)</b>
2 AT4G24010	2	0.61	5.69	0.0021	ATCSLG1
3 AT1G15520	2	0.58	5.42	0.0017	PDR12
4 AT3G02920	5	0.58	7.44	0.0024	replication protein-related
5 AT5G43610	5	0.41	4.98	0.0101	ATSUC6
6 <b>AT4G00650</b>	7	0.48	5.56	0.0020	<b>FRI</b>
7 AT1G24070	8	0.57	6.13	0.0026	ATCSLA10
8 AT1G19940	9	0.53	5.13	0.0019	AtGH9B5
9 AT3G61170	9	0.51	5.12	0.0034	protein coding
10 AT1G32375	10	0.54	5.21	0.0031	protein coding
11 AT2G15320	10	0.50	5.57	0.0027	protein coding
12 AT2G28120	10	0.49	6.45	0.0026	protein coding
13 AT2G16510	13	0.50	10.7	0.0023	AVAP5
14 AT3G14630	13	0.48	4.87	0.0039	CYP72A9
15 AT1G11800	15	0.51	6.97	0.0028	protein coding
16 AT5G44800	16	0.32	6.55	0.0704	CHR4
17 AT3G50660	17	0.40	7.60	0.0059	DWF4
18 <b>AT5G10140</b>	19	0.30	10.3	0.0064	<b>FLC</b>
19 AT1G24110	20	0.49	4.66	0.0059	peroxidase, putative
20 AT1G27030	20	0.45	10.1	0.0059	unknown protein

- biological validation by gene knockout experiments in progress.

# Conclusion

## caveats

- a DAG is a very strong assumption, e.g. feedback loops;
- estimating a graph with almost 22'000 vertices remains difficult.

## however...

- a rough model, but yet still quite good (cf Maathuis et al., 2010);
- imagine an urn with 21'326 balls of which 100 are white and the rest are black - our method manages to draw 3 white balls in 20 turns. The chance for that to happen at random is  $10^{-6}$ .

# Outlook

in progress

- including **hidden variables** in causal graphs;
- perform real lab experiments with **mutant** plants.

# Acknowledgement

Thank you...








Prof Peter Bühlmann  
*Seminar for Statistics*



PD Dr. Lars Hennig  
*Institute for Plant Sciences*

**...and - thank you for Your attention.**

## References and further reading

-  M. Maathuis, D. Colombo, M. Kalisch & P. Bühlmann.  
*Predicting causal effects in large-scale systems from observational data.*  
Nature Methods 7, 247 - 248, 2010.
-  P. Spirtes, C. Glymour & R. Scheines.  
*Causation, Prediction, and Search.*  
Vol. 1, 2nd edn, The MIT Press, 2000.
-  M. Kalisch & P. Bühlmann.  
*Estimating high-dimensional directed acyclic graphs with the PC-algorithm.*  
Journal of Machine Learning Research 8, 2007.
-  M.H. Maathuis, M. Kalisch & P. Bühlmann.  
*Estimating high-dimensional intervention effects from observational data.*  
The Annals of Statistics, 2009.
-  N. Meinshausen & P. Bühlmann.  
*Stability Selection.*  
Preprint, arXiv: 0809.2932v2, 2009.