Joint estimation of causal effects from observational and intervention gene expression data

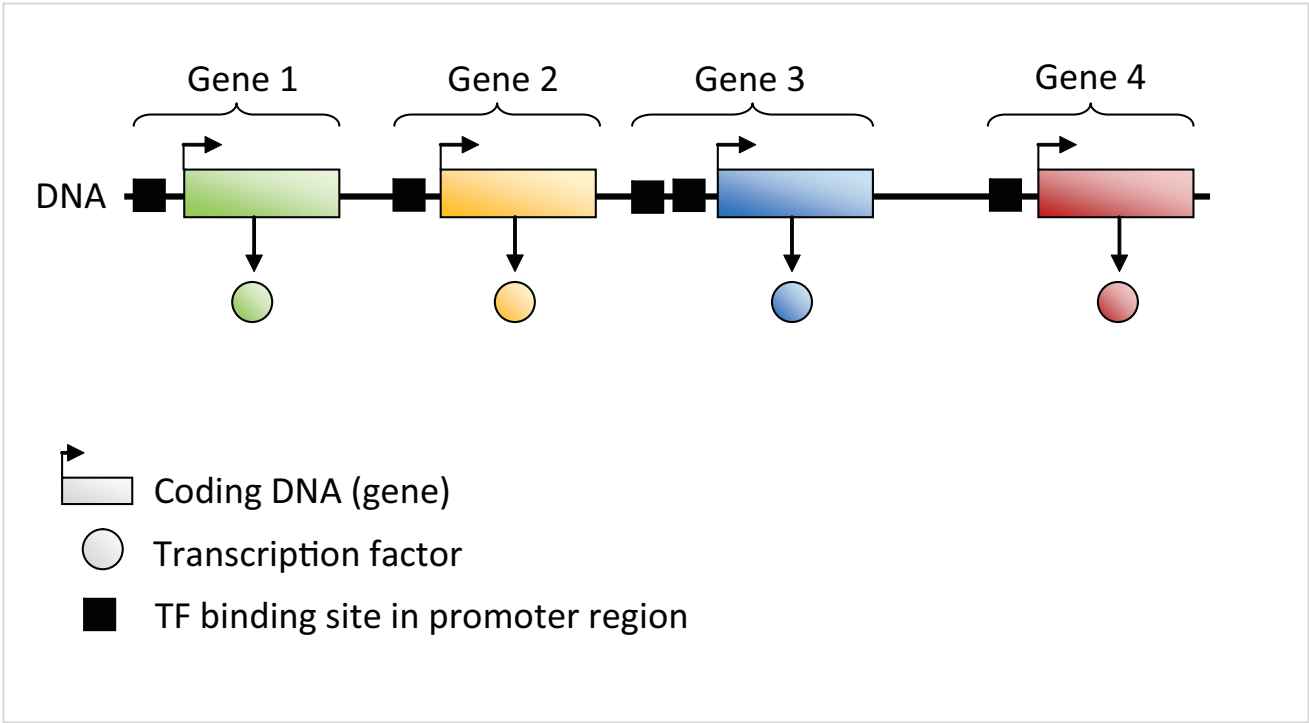
StatSeq @ Paris

Andrea Rau, Florence Jaffrézic, Grégory Nuel

March 29, 2013

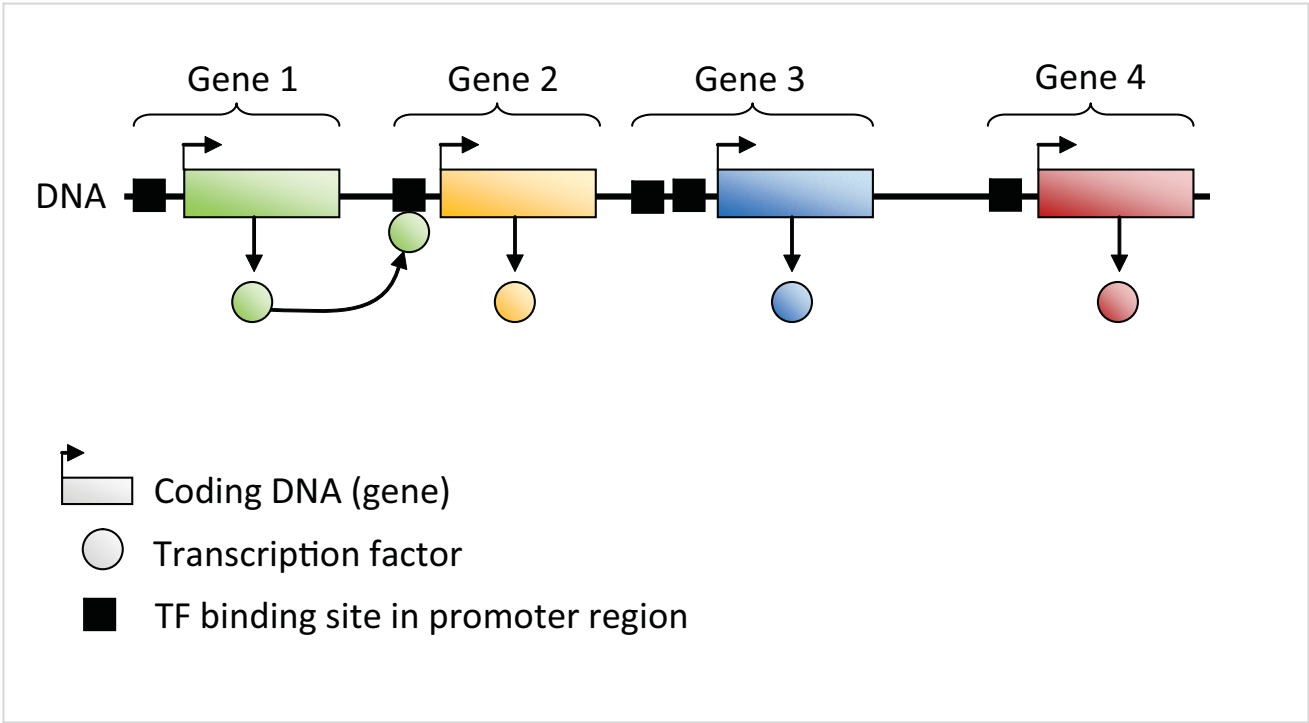
Introduction: Gene regulatory networks (GRN)

- Groups of coordinated genes that interact indirectly with one another through transcription factors



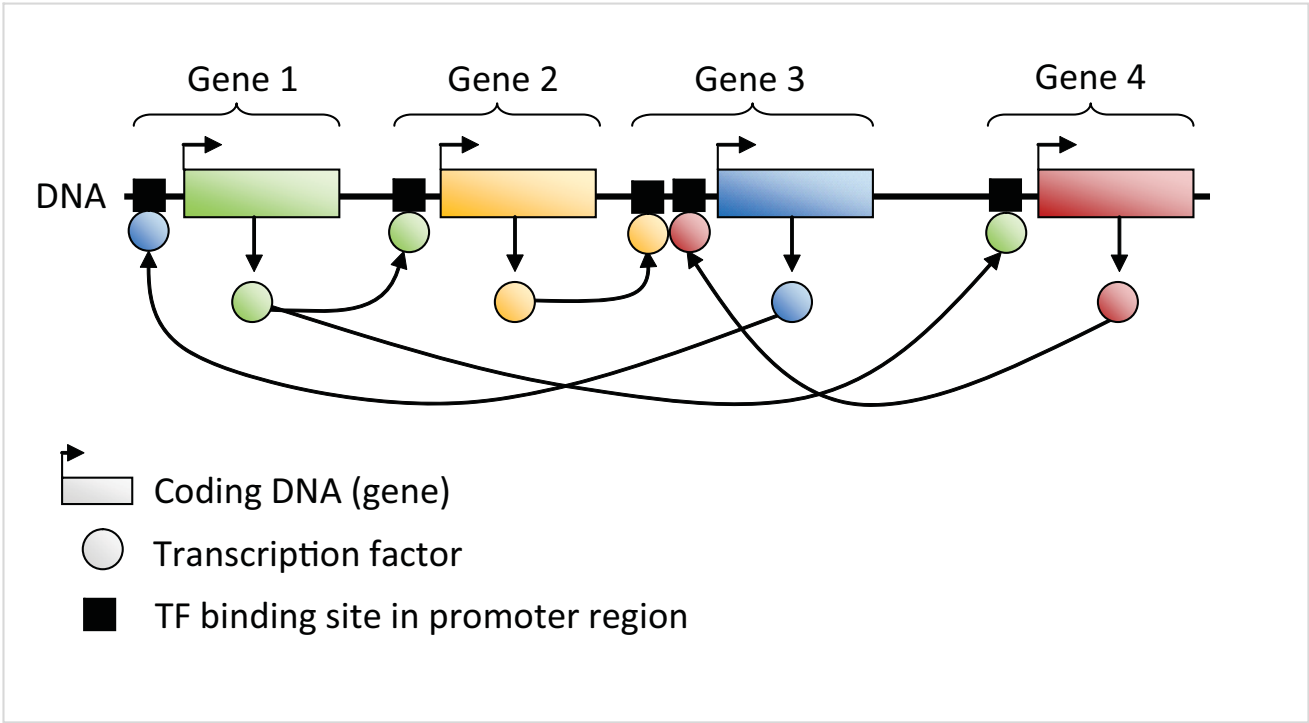
Introduction: Gene regulatory networks (GRN)

- Groups of coordinated genes that interact indirectly with one another through transcription factors



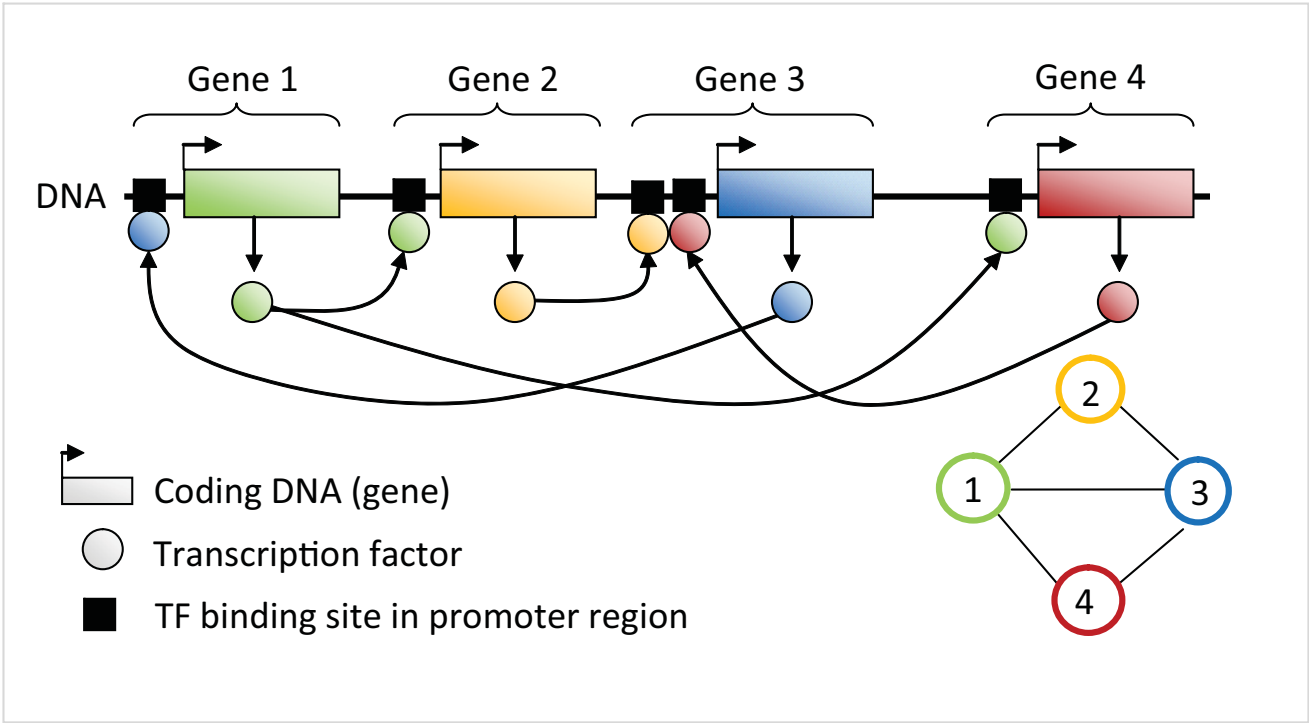
Introduction: Gene regulatory networks (GRN)

- Groups of coordinated genes that interact indirectly with one another through transcription factors



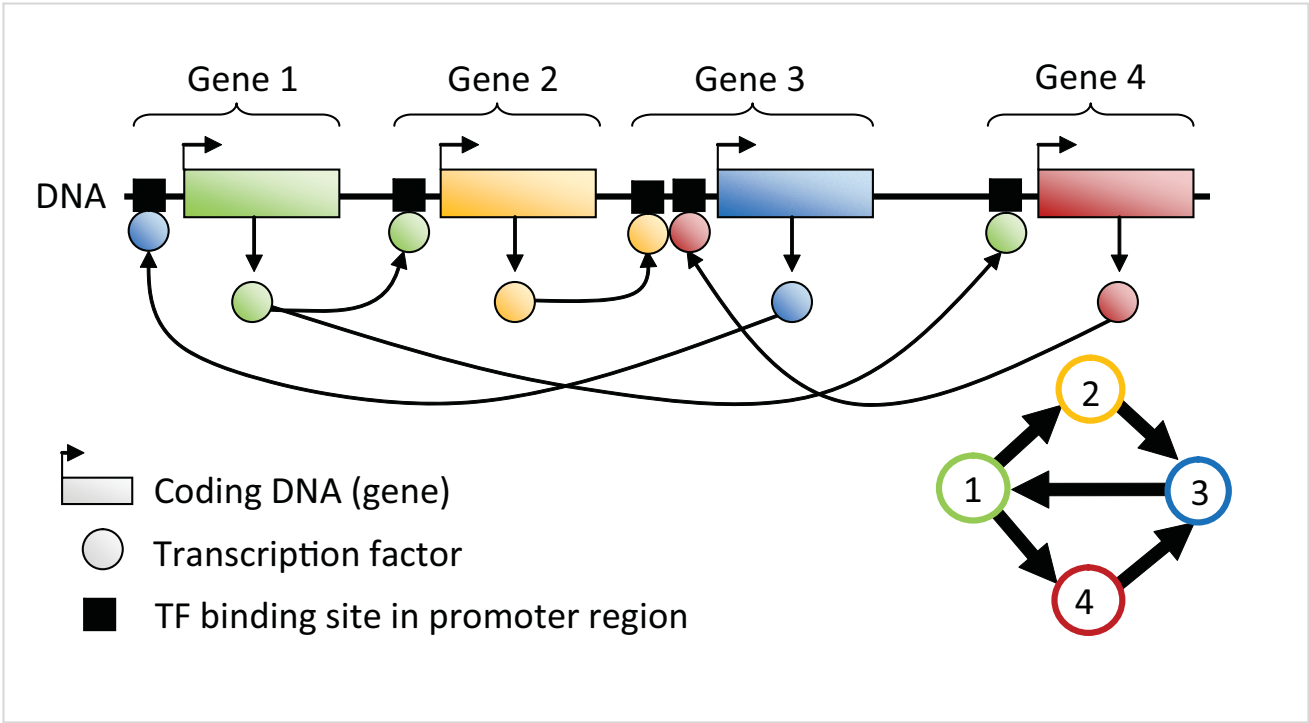
Introduction: Gene regulatory networks (GRN)

- Groups of coordinated genes that interact indirectly with one another through transcription factors



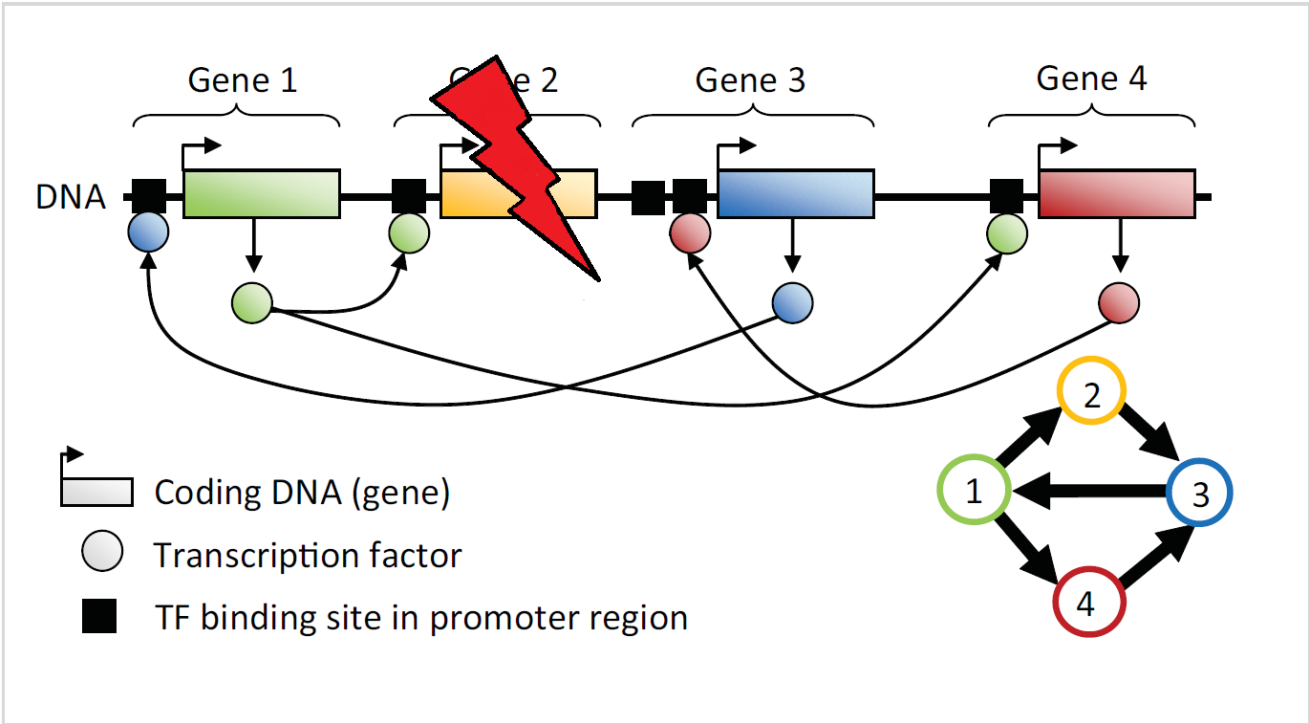
Introduction: Gene regulatory networks (GRN)

- Groups of coordinated genes that interact indirectly with one another through transcription factors



Introduction: Gene regulatory networks (GRN)

- Groups of coordinated genes that interact indirectly with one another through transcription factors



Effect of an intervention on a graph: Total causal effects

Following an intervention $\text{do}(X_i = x_i)$, consider the expected value of each gene via do-calculus (Pearl, 2000):

$$\mathbb{E}(X_j|\text{do}(X_i = x_i)) = \begin{cases} \mathbb{E}(X_j) & \text{if } X_j \in \text{pa}(X_i) \\ \int \mathbb{E}(X_j|x_i, \text{pa}(X_i))\mathbb{P}(\text{pa}(X_i))d\text{pa}(X_i) & \text{if } X_j \notin \text{pa}(X_i) \end{cases}$$

Note: $\mathbb{P}(Y|\text{do}(X = x)) \neq \mathbb{P}(Y|X = x)$

Definition: Total causal effects

$$\beta_{ij} = \frac{\partial}{\partial x}\mathbb{E}(X_j|\text{do}(X_i = x_i))$$

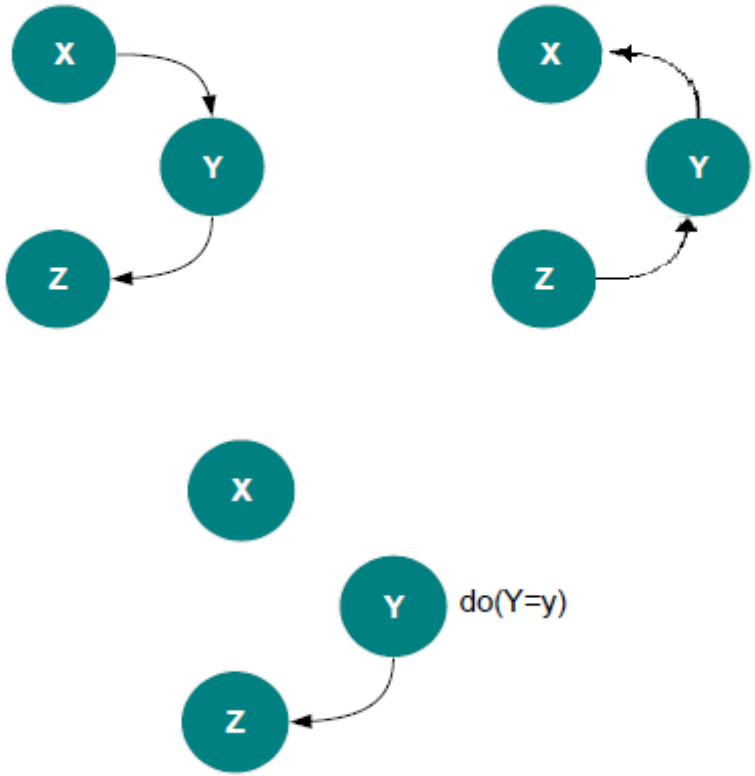
- Equal to 0 if X_i is not an ancestor of X_j

Markov equivalence in DAGs

- Markov equivalence: two different network structures can yield the same joint distribution and observational data alone generally cannot orient edges

Markov equivalence in DAGs

- Markov equivalence: two different network structures can yield the same joint distribution and observational data alone generally cannot orient edges



Estimating causal effects from intervention data

Idea: if gene X_1 is regulated by gene X_2, its expression level after knock-out of X_2 should differ considerably compared to its wild type (steady-state) expression

Pinna *et al.* (2010):

- Data: one wild-type (X_j^{wt} for gene j), and one knock-out experiment for each gene (X_j^i for gene j under knock-out of gene i)
- Four different deviation matrices calculated, feed-forward edges down-ranked, and causal links ranked in order of absolute value

Note: winner of the DREAM4 challenge

Estimating causal effects from observational data

Some causal information can be recovered from observational data alone...

Intervention-calculus when the DAG is Absent (Maathuis *et al.*, 2009)

1. Estimate the equivalence class of the DAG via the PC-algorithm (Kalisch and Bühlmann, 2007)
2. Use intervention calculus to estimate bounds for causal effects across equivalence classes, and rank causal effects

- Shown to be better able to predict strong causal effects using observational data alone (Maathuis *al.*, 2010) than Lasso and elastic-net

Notation

- X_j is the expression of gene j
- Gaussian Bayesian network (GBN):

$$X_j = m_j + \sum_{i \in \text{pa}(j)} w_{ij} X_i + \varepsilon_j \text{ with } \varepsilon_j \sim \mathcal{N}(0, \sigma_j^2)$$

  for $j = 1, \ldots, p$
- $w_{ij} \neq 0$ if and only if $i \in \text{pa}(j)$
- Directed acyclic graph (DAG), and nodes have been ordered so that $i \in \text{pa}(j) \Rightarrow i < j$ (i.e., $\mathbf{W} = (w_{ij})$ is upper triangular)
- Model parameters are $\theta = (\mathbf{W}, m, \sigma)$

- Total causal effects are $\beta = (\mathrm{I} - \mathbf{W})^{-1} = \mathrm{I} + W + \ldots + W^{p-1}$

Joint log-likelihood (1)

Consider experiment k with intervention on $\mathcal{J}_k$ ($\mathcal{J}_k = \emptyset$ means no intervention), where $\mathcal{K}_j = \{k, j \notin \mathcal{J}_k\}$ and $N_j = |\mathcal{K}_j|$.

The log-likelihood of the model can be written as:

$$\ell(m, \sigma, w) = \text{Cst} - \sum_j N_j \log(\sigma_j) - \frac{1}{2} \sum_k \sum_{j \notin \mathcal{J}_k} \frac{1}{\sigma_j^2} (x_j^k - x^k \mathbf{W} e_j^T - m_j)^2$$

Then

$$m_j = \frac{1}{N_j} \sum_{k \in \mathcal{K}_j} (x_j^k - x^k \mathbf{W} e_j^T)$$

Joint log-likelihood (2)

Consider experiment k with intervention on $\mathcal{J}_k$ ($\mathcal{J}_k = \emptyset$ means no intervention), where $\mathcal{K}_j = \{k, j \notin \mathcal{J}_k\}$ and $N_j = |\mathcal{K}_j|$.

The log-likelihood of the model can now be written as:

$$\ell(\sigma, w) = \text{Cst} - \sum_j N_j \log(\sigma_j) - \frac{1}{2} \sum_k \sum_{j \notin \mathcal{J}_k} \frac{1}{\sigma_j^2} (y_j^{k,j} - y^{k,j} \mathbf{W} e_j^T)^2$$

where for (k, j) such that $j \notin \mathcal{J}_k$: $y^{k,j} = x^k - 1/N_j \sum_{k' \in \mathcal{K}_j} x^{k'}$

Then w can be estimated by solving the following linear system:

$$\sum_{i', (i', j) \in \mathcal{E}} w_{i',j} \sum_{k \in \mathcal{K}_j} y_i^{k,j} y_{i'}^{k,j} = \sum_{k \in \mathcal{K}_j} y_i^{k,j} y_j^{k,j} \quad \text{for all } (i, j) \in \mathcal{E}$$

and

$$\sigma_j^2 = \frac{1}{N_j} \sum_{k \in \mathcal{K}_j} (y_j^{k,j} - y^{k,j} \mathbf{W} e_j^T)^2$$

Identifying the best ordering of nodes

Some possibilities:

1. Deterministic quick-sort algorithm to determine optimal node ordering

2. Explore the posterior distribution of the DAG structure space and estimated causal effects via an MCMC algorithm
   - Fixed number of edges, graph structure proposal via edge deletion/addition

Simulation study: DAG structure

Simulated data following a GBN ($p = 10$ genes), with 10 wt and 1 KO for each gene:

- Non-zero $w_{ij} \in (-1, -.25) \cup (.25, 1)$
- $m_j = 0.5$ and $s_j = 0.1$ for all genes j

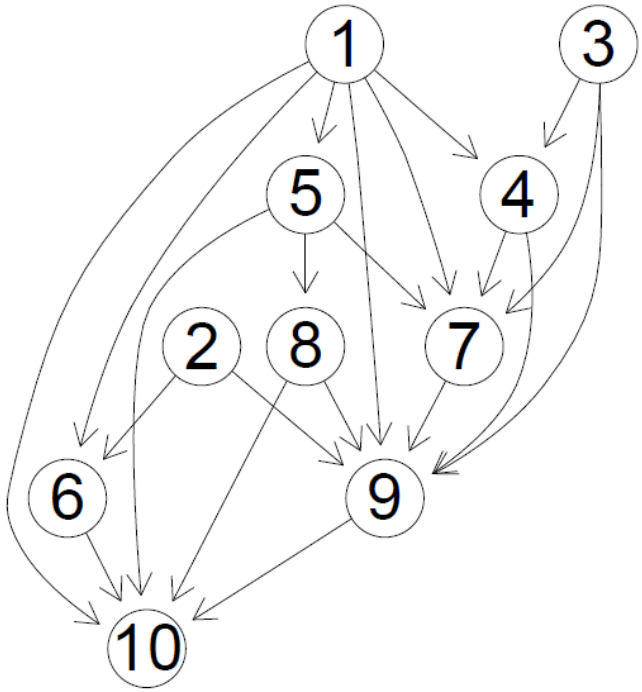


Figure 5 from Kalisch and Bühlmann (2007)

Simulation study: DAG structure

Simulated data following a GBN ($p = 10$ genes), with 10 wt and 1 KO for each gene:

- Non-zero $w_{ij} \in (-1, -.25) \cup (.25, 1)$
- $m_j = 0.5$ and $s_j = 0.1$ for all genes j

- Also consider multiple KO: $\{1, 2\}$, $\{1, 3\}$, $\{3, 8\}$, $\{4, 5\}$, and $\{5, 6\}$

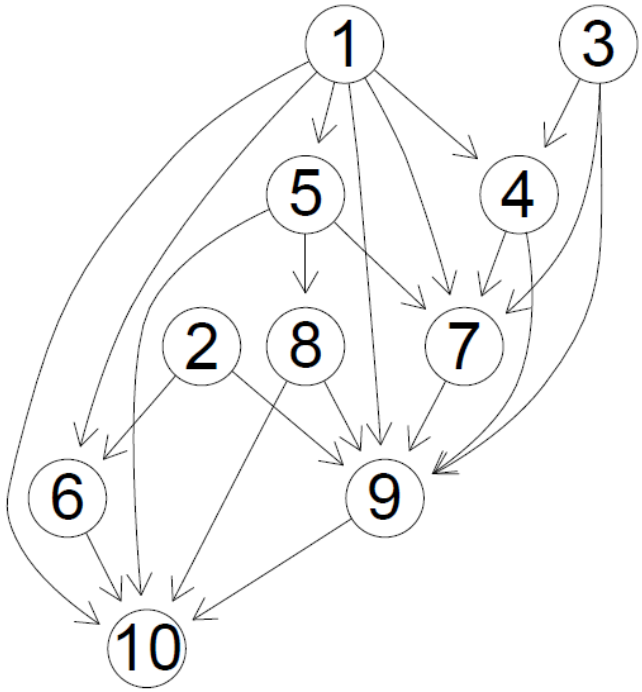
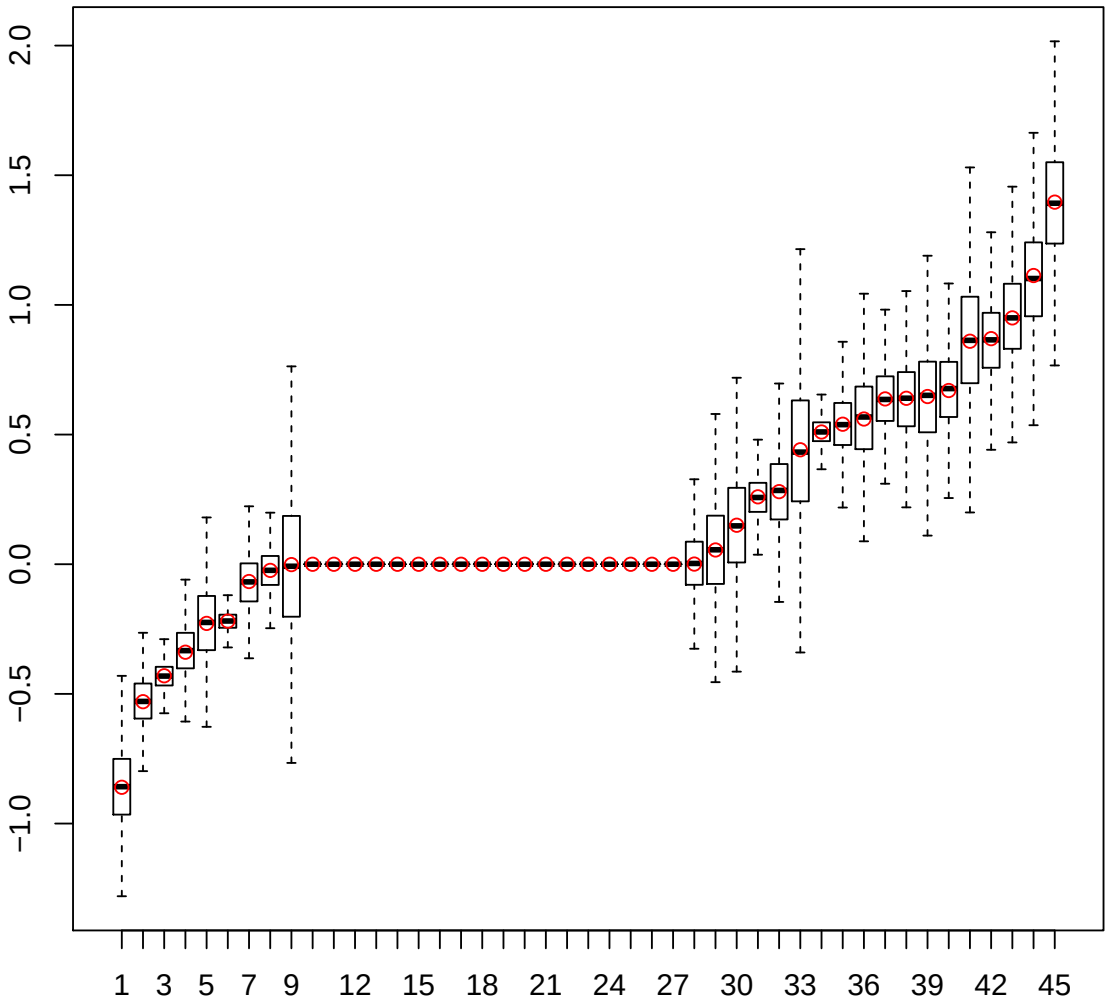


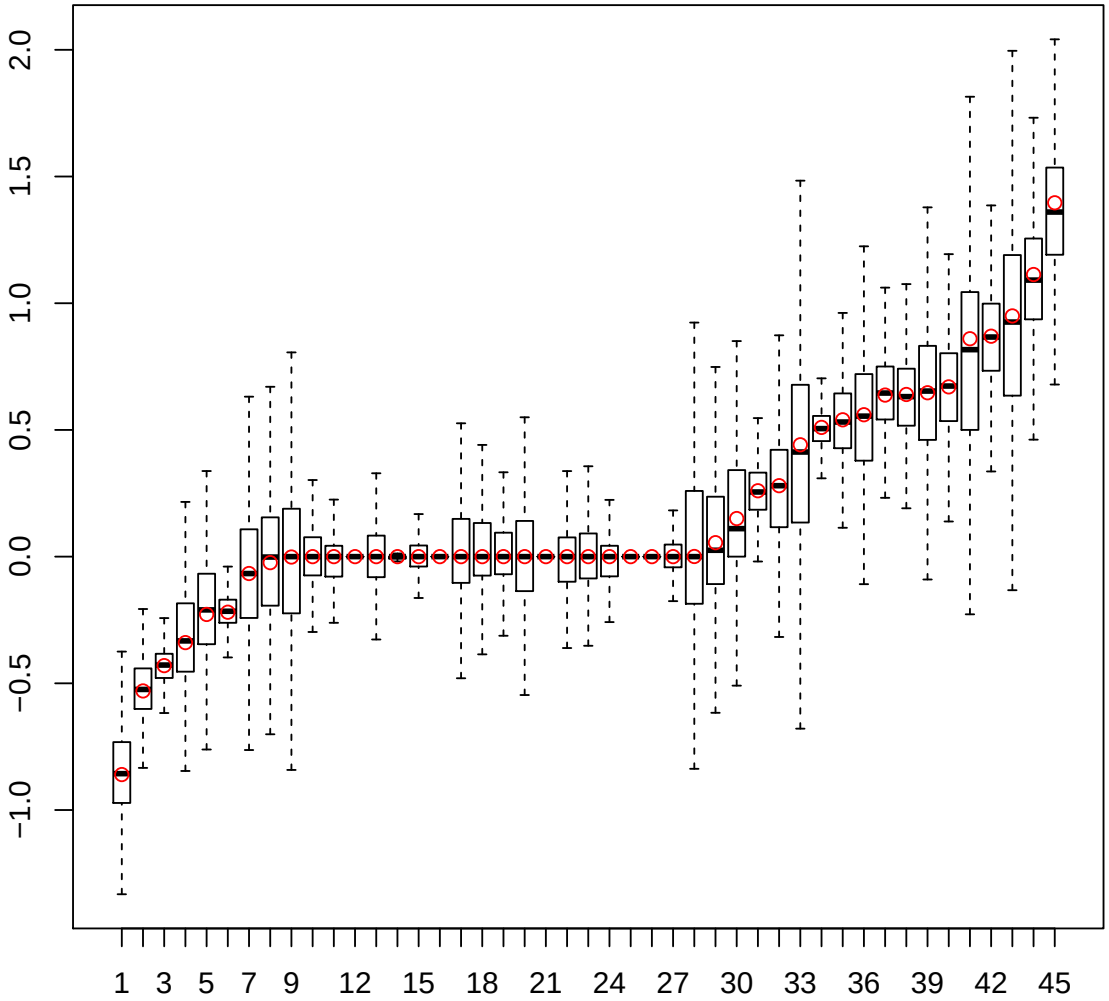
Figure 5 from Kalisch and Bühlmann (2007)

GBN estimation of causal effects: Structure known



(Note: 2000 simulated datasets)

GBN estimation of causal effects: Quick-sort algorithm



(Note: 2000 simulated datasets)

Simulation results: Only observational data

Table: TP and FP out of top 21, results averaged over 100 datasets (sd).

	GBN[1]	Pinna	PCalg min	PCalg max
TP	11.06 (1.78)	—	10.19 (1.89)	11.89 (1.54)
FP	9.94 (1.78)	—	10.81 (1.89)	9.11 (1.54)
Spearman	0.37 (0.09)	—	0.24 (0.12)	0.42 (0.09)

[1] GBN MCMC: 50k iterations, 5k burn-in, thinning every 50 iterations

Simulation results: Observational + intervention data

Table: TP and FP out of top 21, results averaged over 100 datasets (sd).

	GBN[1] (multiple KO)	GBN[1] (single KO)	Pinna	PCalg min	PCalg max
TP	**18.74** (1.3)	**17.8** (1.5)	14.13 (1.55)	10.2 (1.94)	11.05 (1.53)
FP	**2.26** (1.3)	**3.2** (1.5)	6.87 (1.55)	10.8 (1.94)	9.95 (1.53)
Spearman	**0.72** (0.04)	**0.69** (0.05)	0.5 (0.07)	0.28 (0.11)	0.37 (0.09)

[1] GBN MCMC: 50k iterations, 5k burn-in, thinning every 50 iterations

Discussion

GBN for a mixture of steady-state and knock-out (and multiple knock-out!) data to enable calculation of total causal effects:

- MCMC algorithm / Quick-sort node ordering
- Initial results very encouraging and suggest the benefit in jointly analyzing steady-state and intervention data

- Future work: Experimental design to plan future (multiple) knock-out experiments...

Thanks to Rémi Bancal (M2 intern)

References:

- Kalisch and Bühlmann (2007) Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research* 8, 613-636.
- Maathuis *et al.* (2009) Estimating high-dimensional intervention effects from observational data. *Annals of Statistics* 37:6A, 3133-3164.
- Maathuis *et al.* (2010) Predicting causal effects in large-scale systems from observational data. *Nature Methods* 7:4, 247-248.
- Pearl (2000) *Causality: Models, Reasoning, and Inference.* Cambridge University Press, Cambridge.
- Pinna *et al.* (2007) From knockouts to networks: Establishing direct cause-effect relationships through graph analysis. *PLoS One* 5:10.
- Stolovitzky *et al* (2007) Dialogue on reverse-engineering assessment and methods. *Ann NY Acad Sci* 1115, 1-22.