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

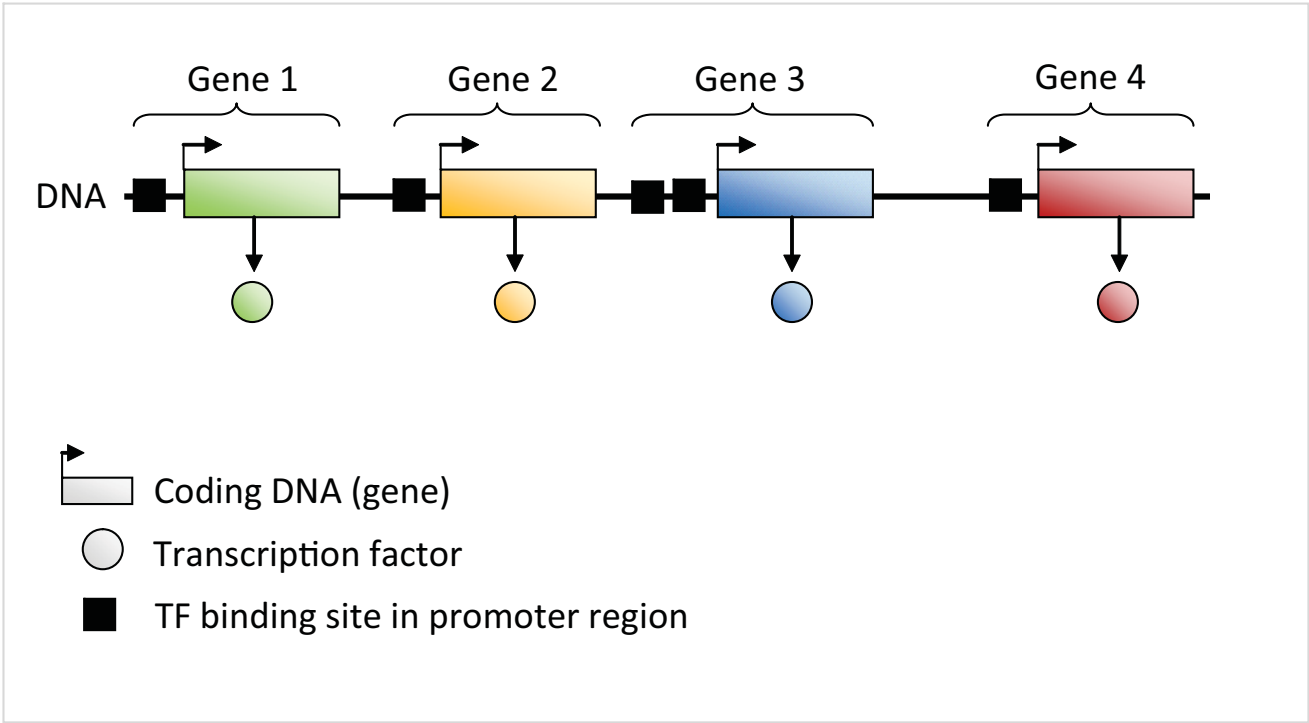
AppliBUGS @ Paris

Andrea Rau, Florence Jaffrézic, Grégory Nuel

June 20, 2013

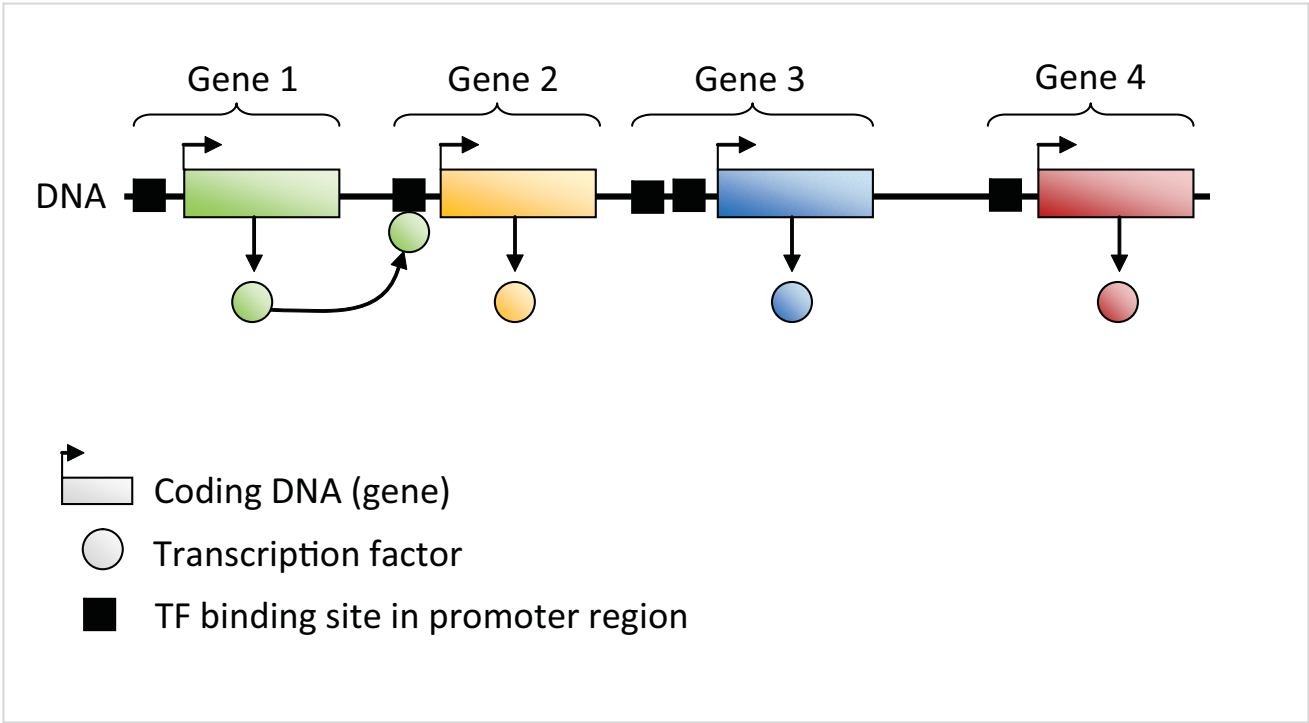
Introduction: Gene regulatory networks (GRN)

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



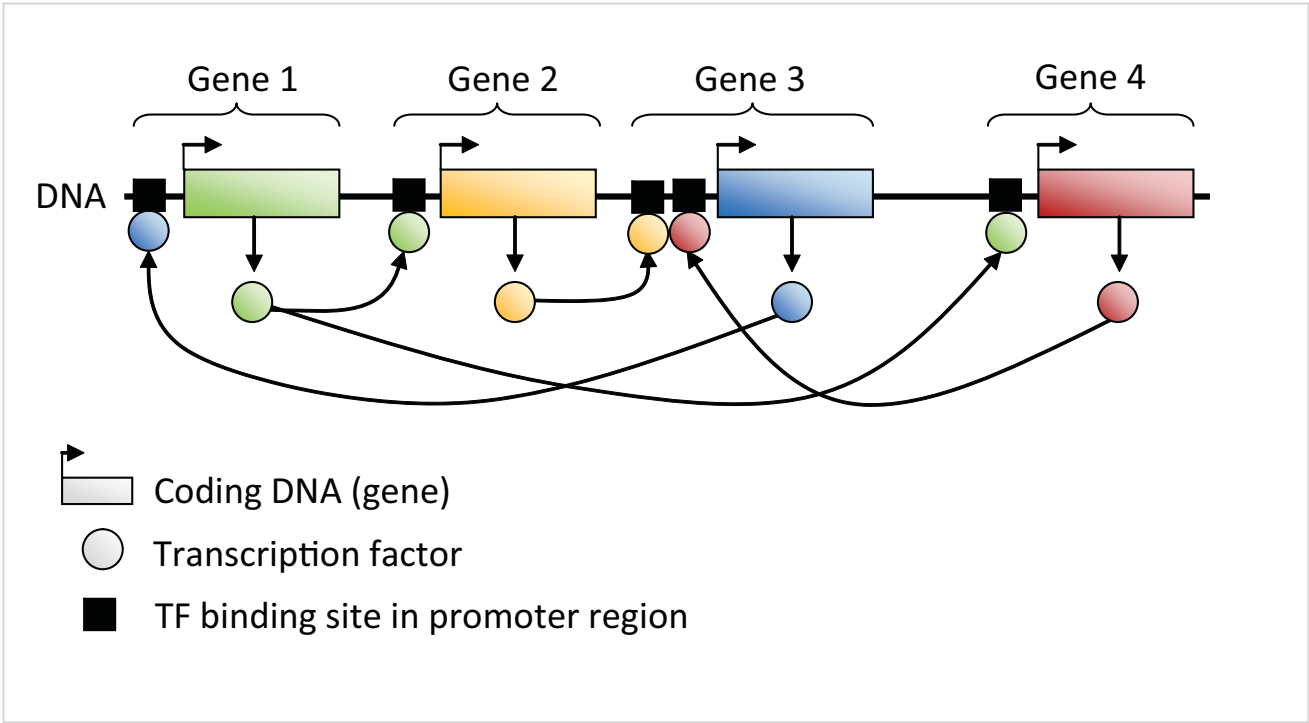
Introduction: Gene regulatory networks (GRN)

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



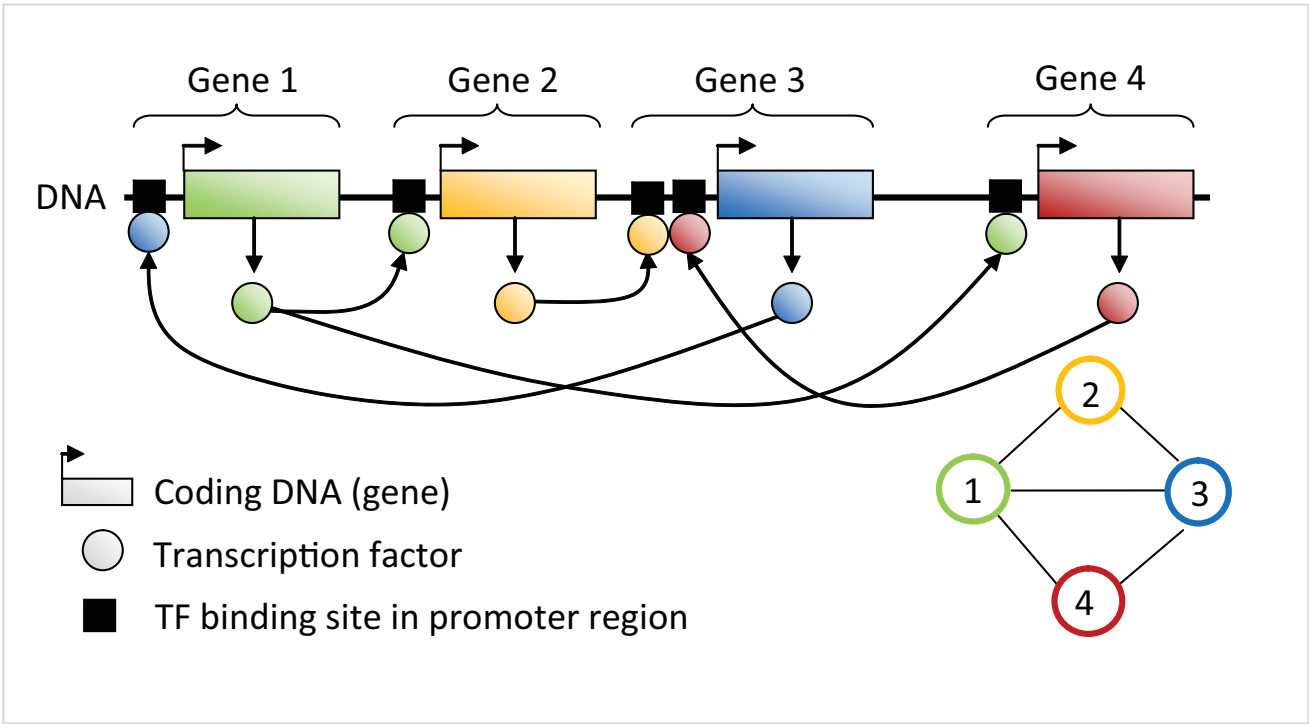
Introduction: Gene regulatory networks (GRN)

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



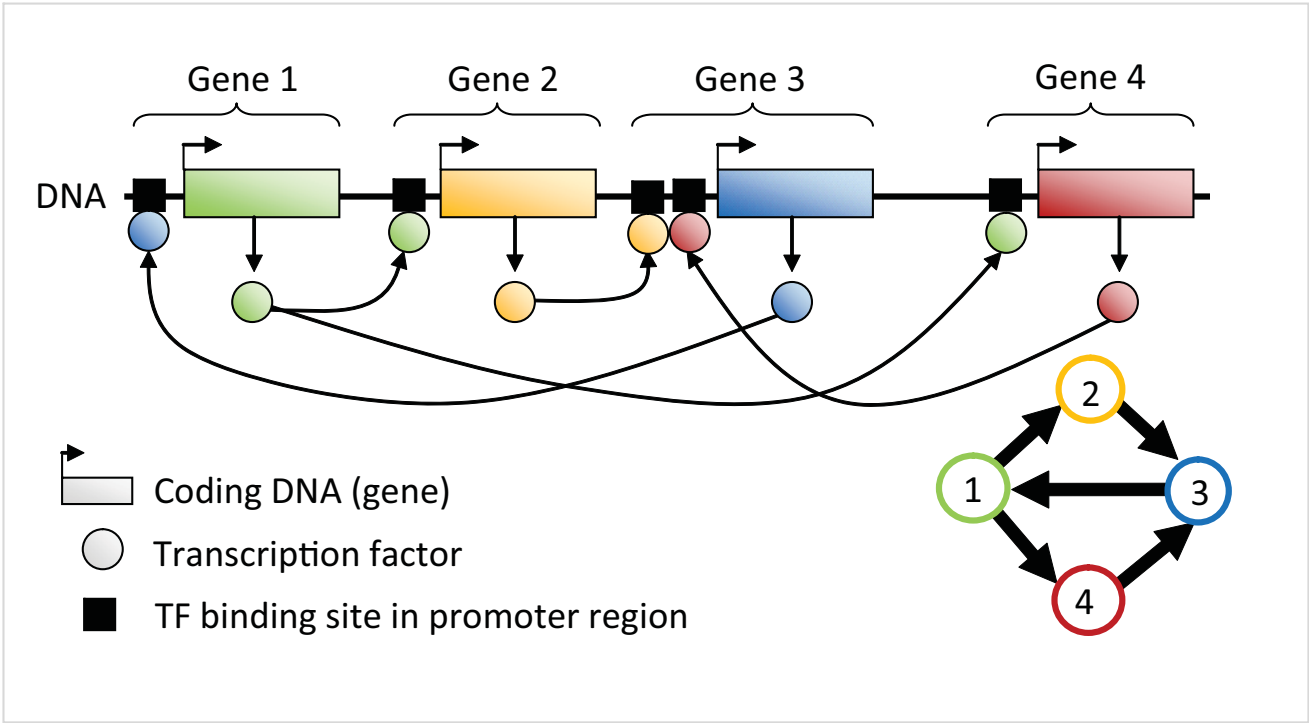
Introduction: Gene regulatory networks (GRN)

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



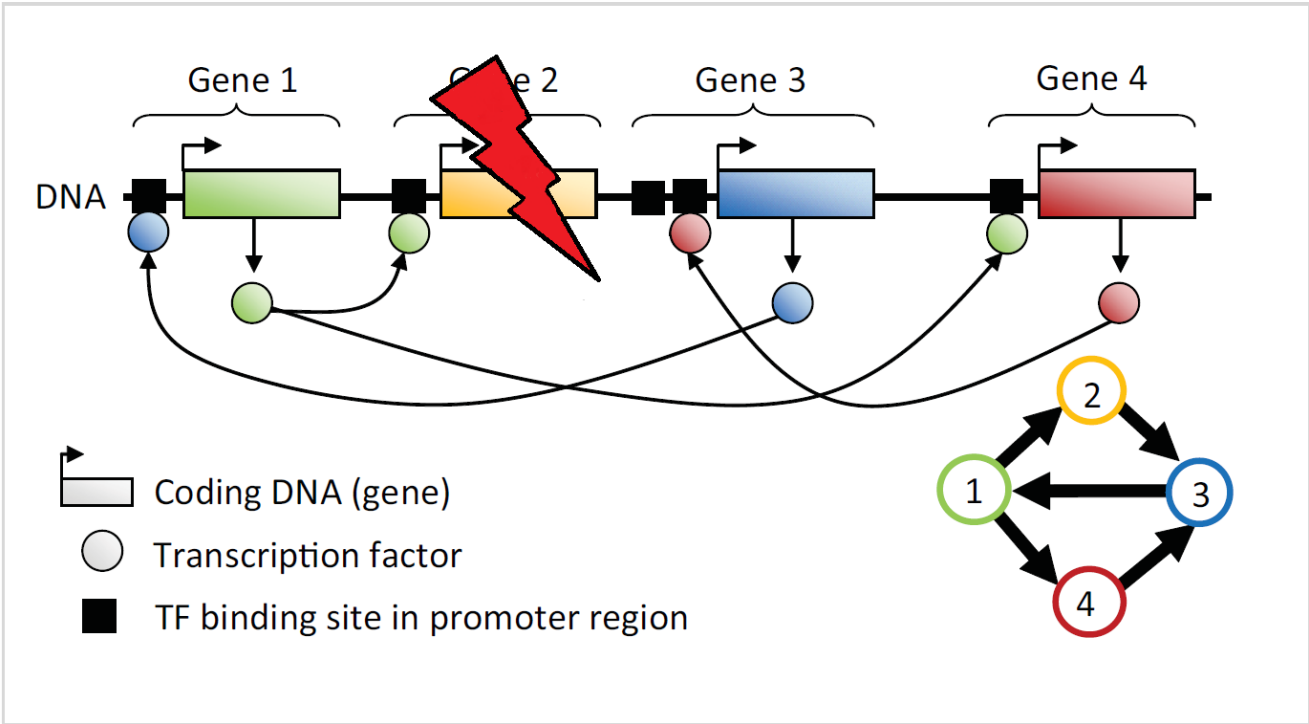
Introduction: Gene regulatory networks (GRN)

- 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 DAG: 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

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

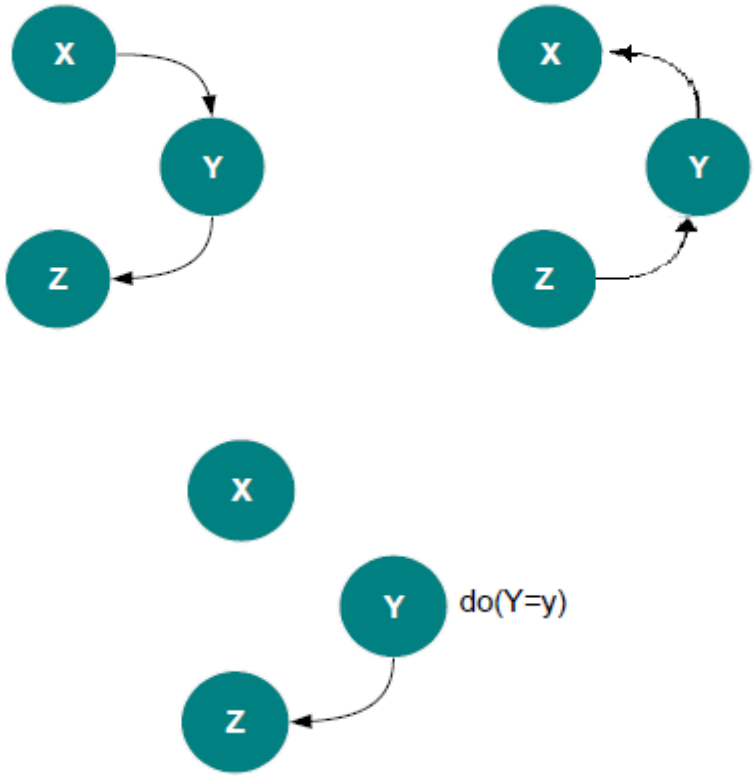
- Equal to 0 if X_i is not an ancestor of X_j

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 node order and estimated causal effects via an empirical MCMC algorithm
   - Node ordering proposal via Mallows model, using node ordering of previous iteration as reference
   - Full estimation of model parameters for a given node ordering using likelihood calculations

Mallows model (Mallows 1957)

- Let R be a modal or reference ordering, $\phi \in (0, 1]$ a temperature parameter, and $r = r_1 r_2 \dots r_m$ be a node ordering:

$$P(r) = P(r|R, \phi) = \frac{1}{Z}\phi^{d(R,r)}$$

where Z is a normalizing constant and

$$d(R, r) = \sum_{i<j} \mathbf{1}\left[r_j \succ r_i\right]$$

is a dissimilarity measure between R and r using the number of pairwise disagreements

- $\phi = 1$ corresponds to a dirac on R, $\phi = 0$ corresponds to a uniform distribution over all node orderings

Sampling performed through repeated insertion model (Doignon *et al.* 2004)

Simulation study: Estimation of causal effects and node ordering

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 $\sigma_j = \{0.01, 0.1, 0.5\}$ for all genes j

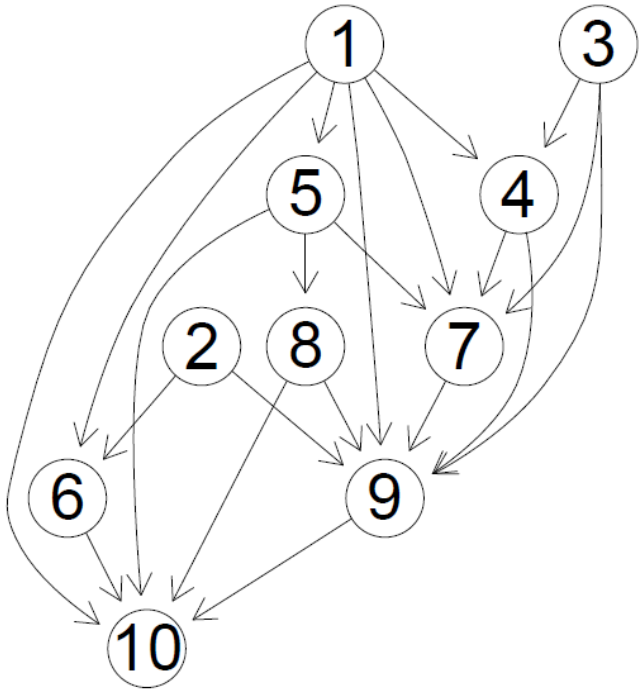
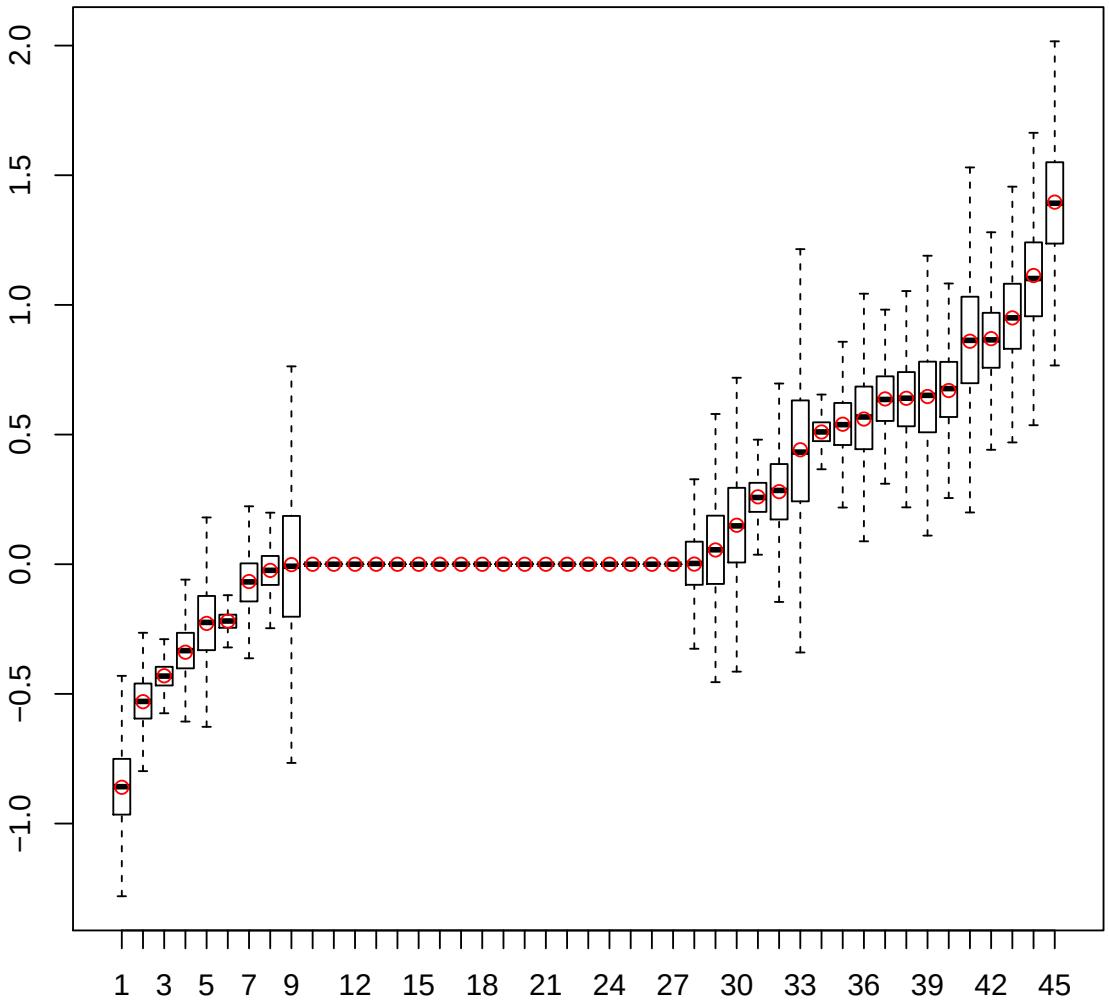


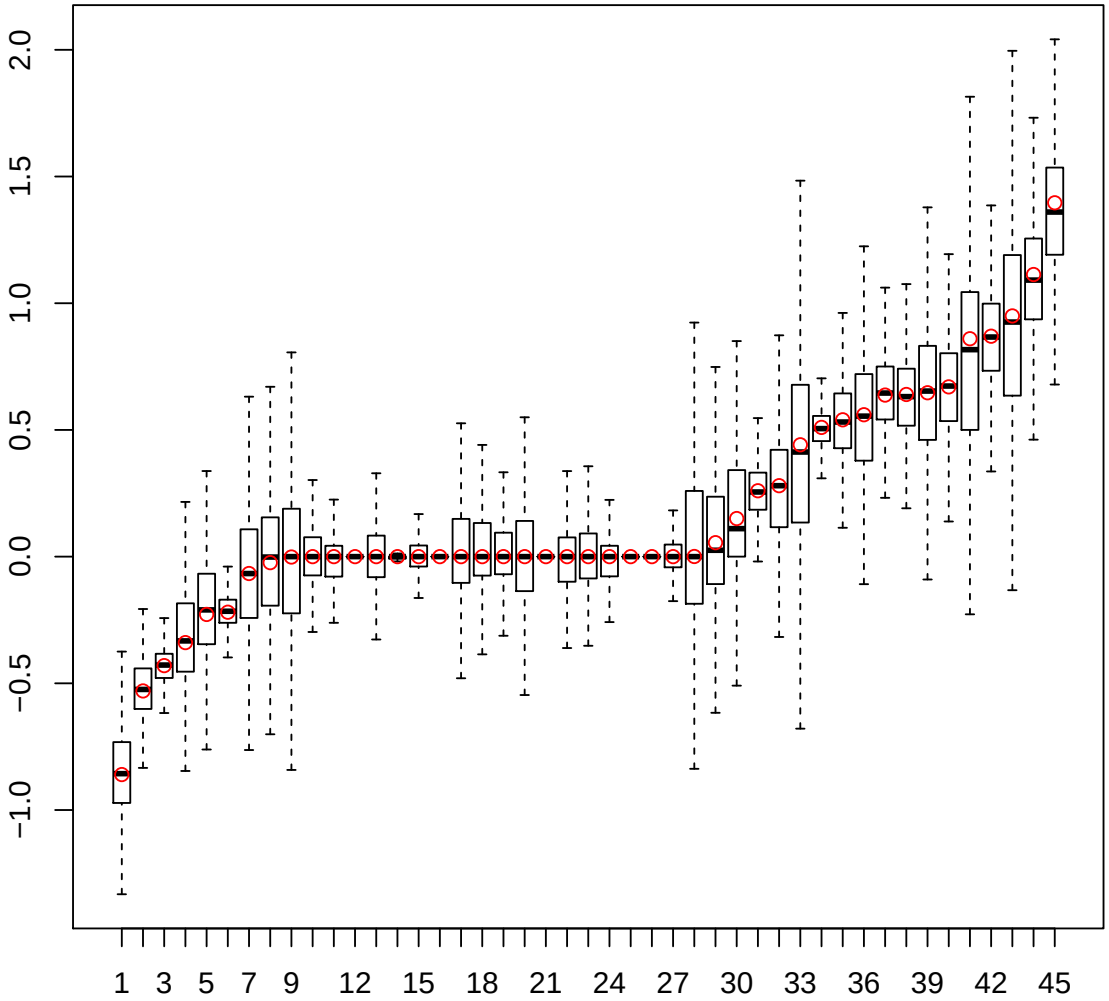
Figure 5 from Kalisch and Bühlmann (2007)

GBN estimation of causal effects: Structure known



(Note: 2000 simulated datasets, $\sigma = 0.1$)

GBN estimation of causal effects: Quick-sort algorithm



(Note: 2000 simulated datasets, $\sigma = 0.1$)

Simulation results: Observational + intervention data

	GBN[1]	Pinna	PC-alg (opt)	PC-alg (pes)
AUROC	**0.790**	0.612	0.718	0.644
AUPRC	**0.654**	0.422	0.551	0.499
Spearman	**0.539**	0.121	0.409	0.27

Table: $\sigma = 0.5$. Results averaged over 100 simulations. AUROC = area under the ROC curve, AUPRC = area under the precision-recall curve, Spearman = Spearman correlation between true and estimated matrices.

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

Simulation results: Observational + intervention data

	GBN[1]	Pinna	PC-alg (opt)	PC-alg (pes)
AUROC	**0.948**	0.821	0.718	0.644
AUPRC	**0.868**	0.732	0.551	0.499
Spearman	**0.815**	0.597	0.409	0.27

Table: $\sigma = 0.1$. Results averaged over 100 simulations. AUROC = area under the ROC curve, AUPRC = area under the precision-recall curve, Spearman = Spearman correlation between true and estimated matrices.

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

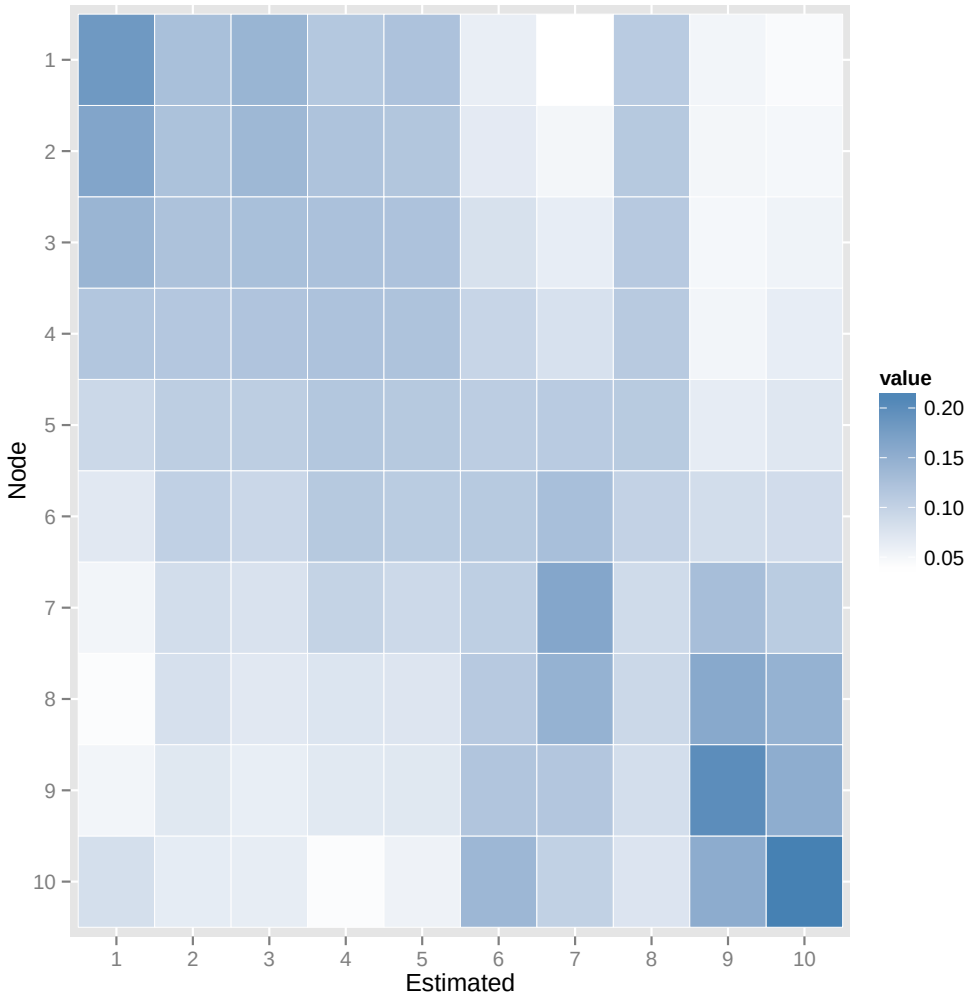
Simulation results: Observational + intervention data

	GBN[1]	Pinna	PC-alg (opt)	PC-alg (pes)
AUROC	**0.984**	0.944	0.714	0.657
AUPRC	**0.934**	0.900	0.546	0.521
Spearman	**0.945**	0.827	0.389	0.265

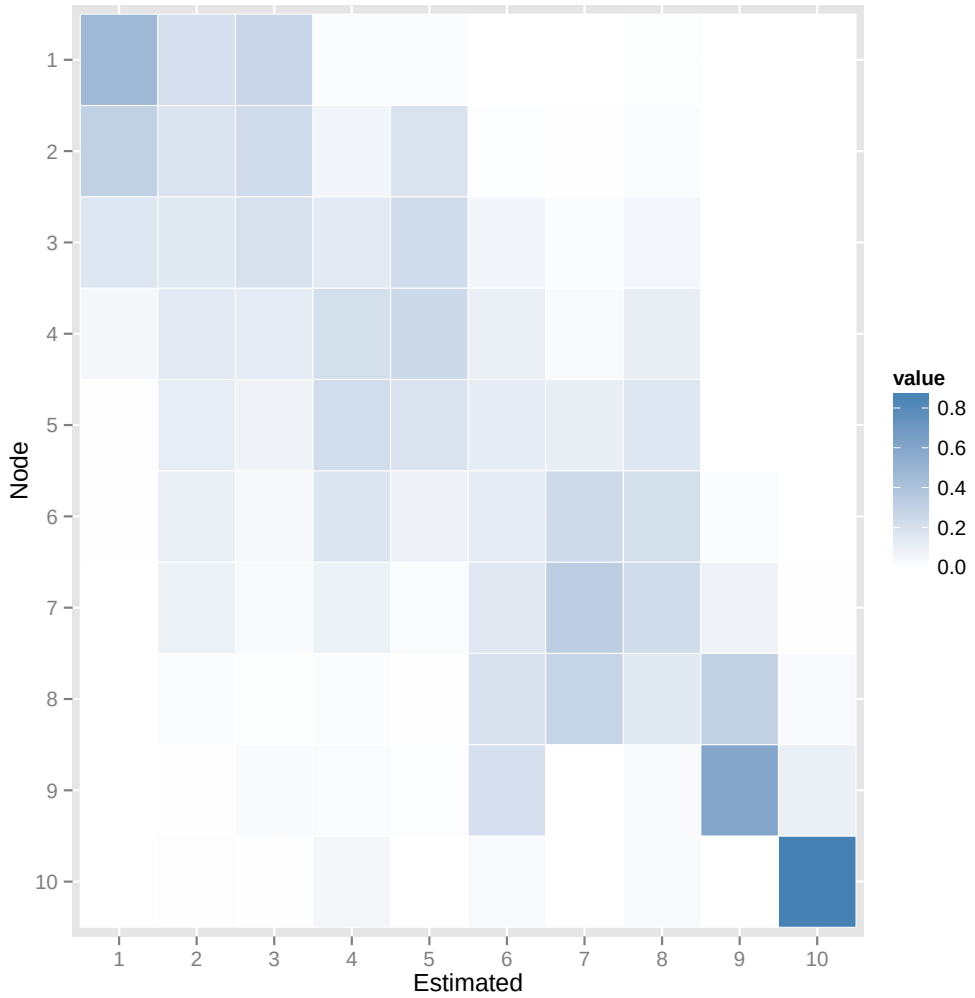
Table: $\sigma = 0.01$. Results averaged over 100 simulations. AUROC = area under the ROC curve, AUPRC = area under the precision-recall curve, Spearman = Spearman correlation between true and estimated matrices.

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

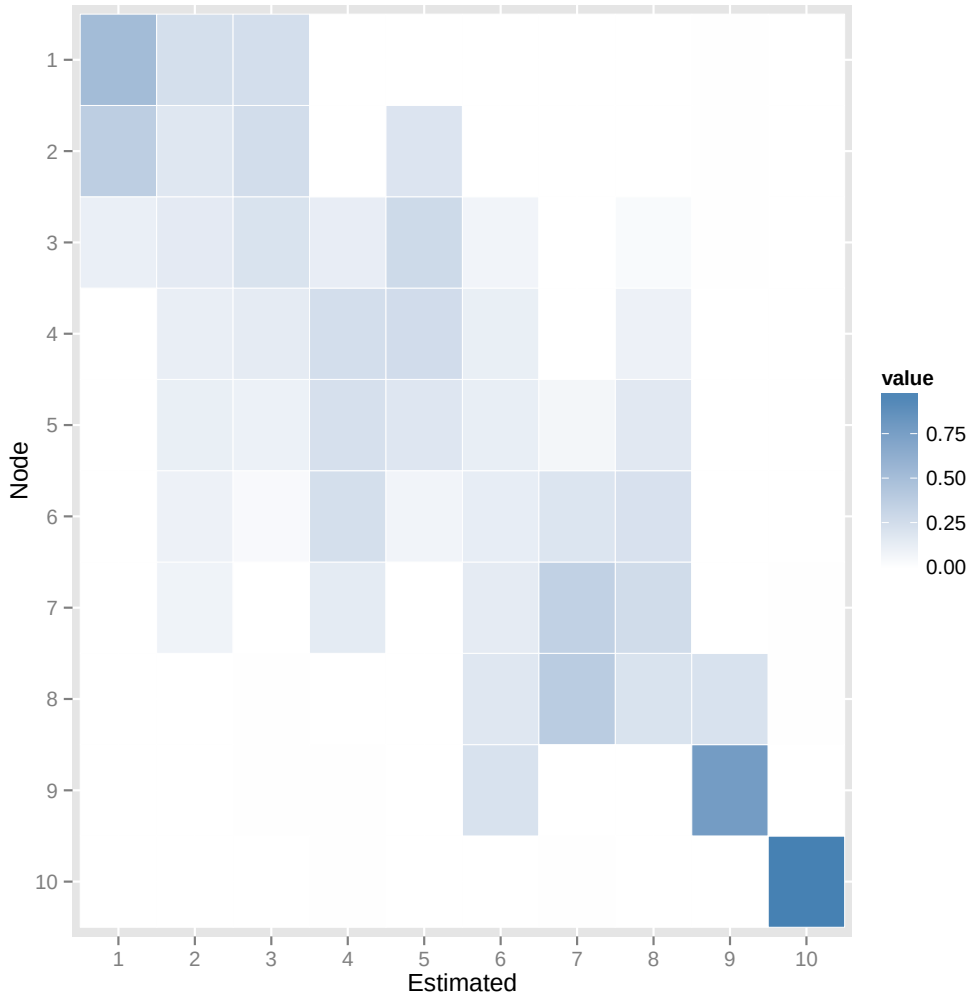
Simulation results: posterior distribution of node ordering ($\sigma = 0.5$)



Simulation results: posterior distribution of node ordering ($\sigma = 0.1$)



Simulation results: posterior distribution of node ordering ($\sigma = 0.01$)



DREAM4 challenge

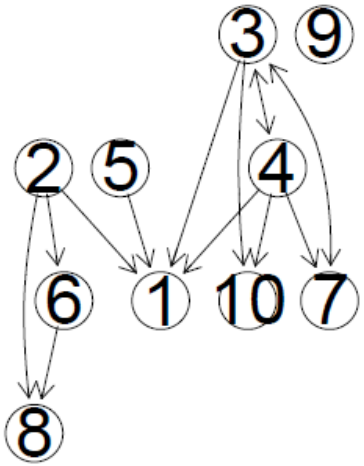
DREAM challenge: international competition held yearly to contribute to the development of powerful inference methods (Stolovitzky *et al.*, 2007)

DREAM4 *in silico* network challenge:

- Goal: Infer directed GRNs from simulated data ($p = 10$, $p = 100$) and provide a level of confidence for the presence of each possible edge
- True network topologies (with feedback loops) extracted from transcriptional regulatory networks of *E. coli* and *S. cerevisiae*
- Data: simulated wild-type, knock-outs, knockdowns, multifactorial perturbations, and time series expression data (stochastic differential equations + measurement noise)
- Pinna *et al.* method was top performer

DREAM4 challenge: Data example

	G_1	G_2	G_3	G_4	G_5
G^{wt}	0.14	0.89	0.01	0.87	0.14
G^1	0.00	0.96	0.00	0.86	0.06
G^2	0.68	0.00	0.04	0.90	0.05
G^3	0.17	0.86	0.00	0.88	0.02
G^4	0.13	0.86	0.08	0.00	0.09
G^5	0.12	0.78	0.09	0.91	0.00



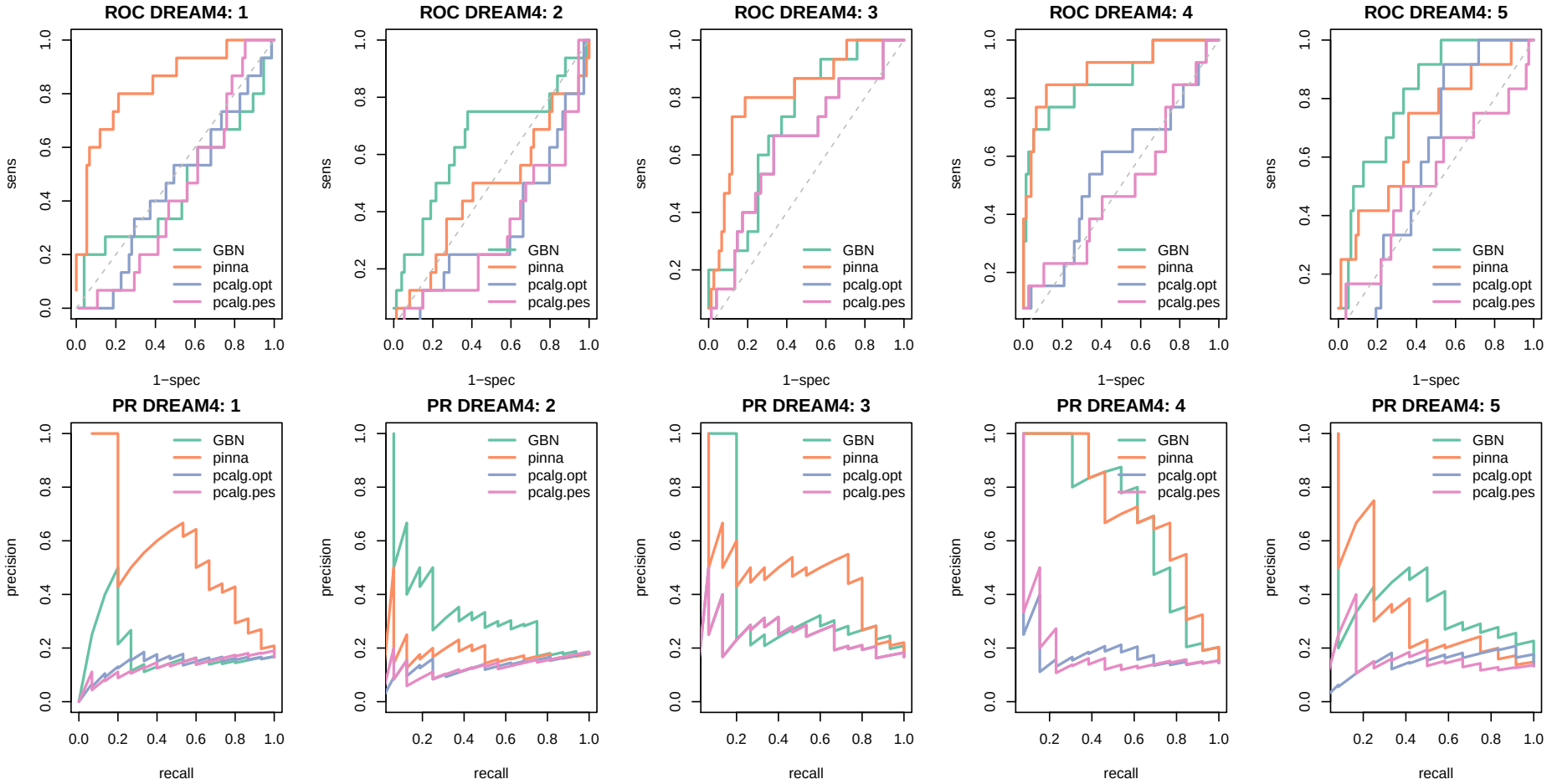
DREAM4 challenge details

- 50k iterations run, with burn-in of 5k and thinning every 50 iterations
- Trial run to select $\phi = \exp(-1/0.8)$ such that acceptance rate is $\approx$ 35%

Compare GBN MCMC total causal effect posterior means compared to Pinna W^D matrix and IDA method

- GBN MCMC: wild-type, knock-out, and multifactorial perturbation data
- IDA: wild-type and multifactorial perturbation data
- Pinna: wild-type and knock-out data

DREAM4 challenge results



Discussion

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

- MCMC algorithm to explore posterior distribution of node ordering
- Initial results very encouraging and suggest the benefit in jointly analyzing steady-state and intervention data, as well as multiple intervention (i.e., double or triple knock-out) data

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

Thanks to Rémi Bancal (M2 intern)

References:

- Doignon, Pekeč, Regenwetter (2004) The repeated insertion model for rankings: Missing link between two subset choice models. *Psychometrika*, 69(1):33-54.
- Kalisch and Bühlmann (2007) Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research* 8, 613-636.
- Lu and Boutilier (2011) Learning Mallows models with pairwise preferences. *Proceedings of the 28th International Conference on Machine Learning (ICML)*, 145-152.
- 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.