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

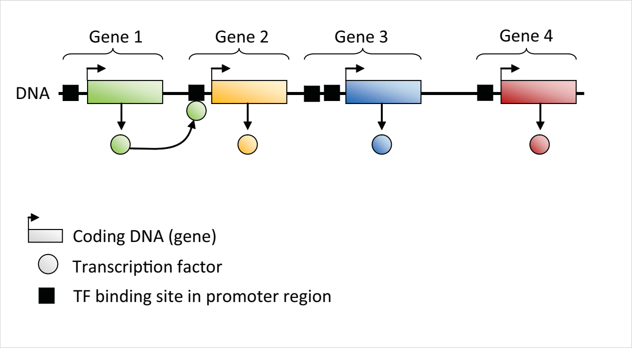
NETBIO @ Paris

Andrea Rau, Florence Jaffrézic, Grégory Nuel

September 12, 2013

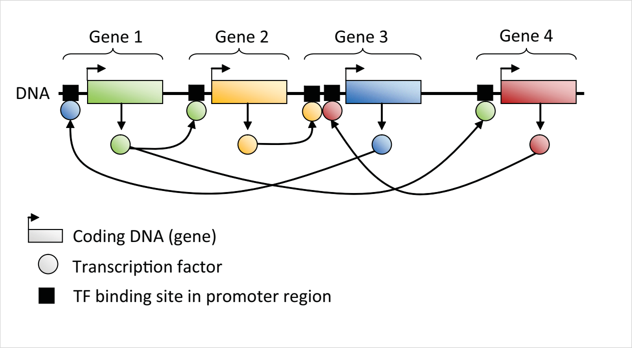
Gene regulatory networks (GRN)

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



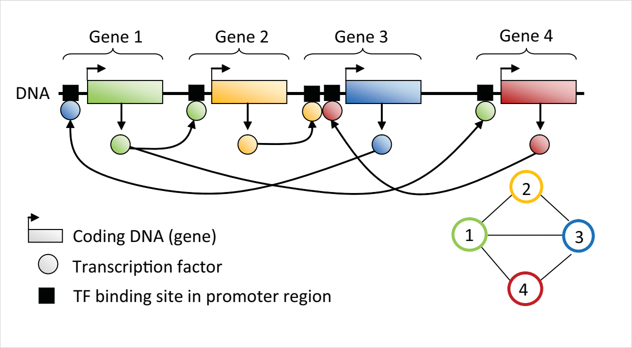
Gene regulatory networks (GRN)

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



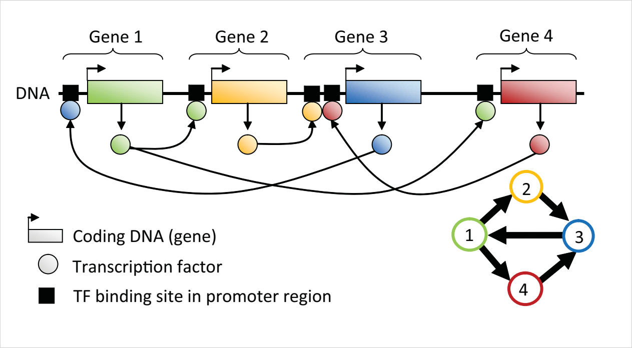
Gene regulatory networks (GRN)

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



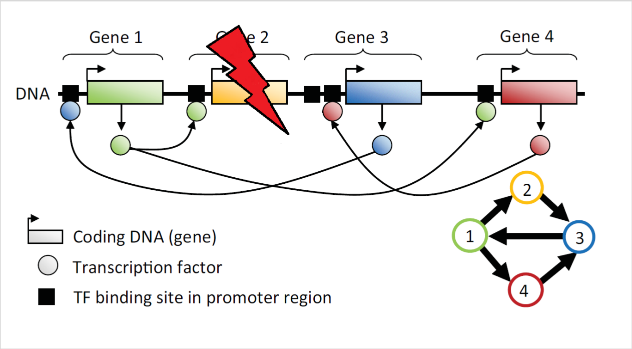
Gene regulatory networks (GRN)

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



Gene regulatory networks (GRN)

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



Observational vs. intervention expression data

Observational data

Wild-type or steady-state expression over multiple biological replicates (or time points), easy and less expensive to obtain

Intervention data

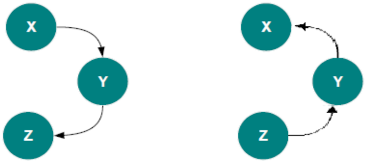
Observe the expression levels of every gene in the network in the presence of one or multiple perturbations:

- Genetic (e.g., knock-out or knock-down experiments)
- Biological (e.g., alter growth media or temperature)

⇒ Generate information about (indirect or direct) causal relationships, ... but can be $$$ and labor-intensive

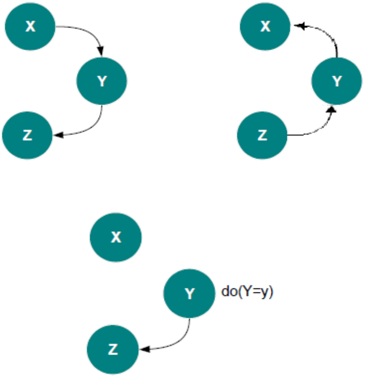
Markov equivalence in DAGs

- 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

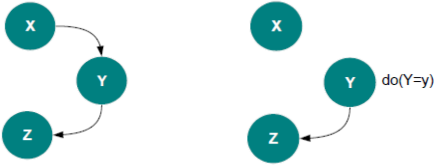


Effect of an intervention on a DAG

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

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

Note: $\mathbb{P}(X_j|\mathrm{do}(X_i = x)) \neq \mathbb{P}(X_j|X_i = x)$



Causal effects

Definition: **Total causal effects**

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

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

Definition: **Direct causal effects** (graph edges)

$$\alpha_{ij} = \frac{\partial}{\partial x}\mathbb{E}(X_j | \mathrm{pa}(X_i), \mathrm{do}(X_i = x))$$

- Equal to 0 if X_i is not a parent of X_j

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

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 100-gene challenge

Some motivating questions...

- Can more complicated intervention designs (partial knock-outs, multiple knock-outs) be jointly modeled with observational data to estimate causal effects?
- Does the inclusion of multiple intervention data improve inference of causal effects?
- Can the information provided by a given gene knock-out experiment be quantified?

Notation

- X_j^k is the expression of gene $j \in 1, \ldots, p$ in experiment $k \in 1, \ldots, N$
- Gaussian Bayesian network (GBN):

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

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

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

Joint log-likelihood: Observational data only

We can show that this model is equivalent to $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = \mathbf{mL} \quad \text{and} \quad \boldsymbol{\Sigma} = \mathbf{L}^T \text{diag}(\boldsymbol{\sigma}^2)\mathbf{L} = \sum_{j \in \mathcal{I}} \sigma_j^2 \mathbf{L}^T \mathbf{e}_j^T \mathbf{e}_j \mathbf{L}$$

where $\mathbf{e}_j$ is a p-dimensional null row-vector except for its j^{th} term

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

$$\ell(\mathbf{m}, \boldsymbol{\sigma}, \mathbf{W}) = \text{Cst} - \sum_j N_j \log(\sigma_j) - \frac{1}{2} \sum_k \sum_j \frac{1}{\sigma_j^2} (x_j^k - \mathbf{x}^k \mathbf{W} \mathbf{e}_j^T - m_j)^2$$

Joint log-likelihood: Observational + intervention data (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 now be written as:

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

Then

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

Joint log-likelihood: Observational + intervention data (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 then be rewritten as:

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

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

Then $\mathbf{W}$ 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

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

Identifying an appropriate causal node ordering in the graph

Some possibilities:

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

2. Explore the posterior distribution of the causal node order and estimated causal effects via an empirical Metropolis-Hastings 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 $r = r_1 r_2 \ldots r_m$ 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 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

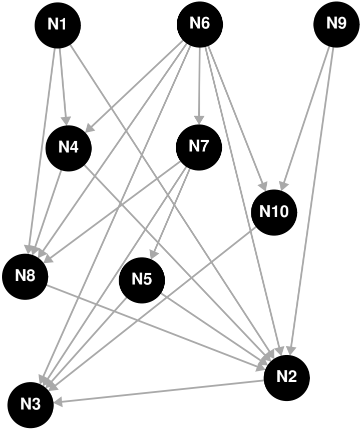
Simulation study: Description

Simulated data following a GBN ($p = 10$ genes):

- Non-zero $w_{ij} \in (-1, -.25) \cup (.25, 1)$
- $m_j = 0.5$ and $\sigma_j = \{0.01, 0.1, 0.5\}\ \forall\ j$

Five settings considered:

1. Observational only
2. Systematic single knock-outs
3. Partial single knock-outs
4. Multiple knock-outs
5. Multiple knock-outs and 3 hidden genes



Trial run to select ϕ such that acceptance rate is ≈ 30-40%.

Simulation setting 1: Observational only

20 observational (wild-type) replicates with no interventions

Table : Area under the ROC curve (AUROC), area under the precision-recall curve (AUPRC), Spearman correlation with true total causal effects, and mean squared error (MSE) of estimated total causal effects, averaged over 100 datasets (sd).

Criterion	MCMC-Mallows[1]	Pinna	IDA (opt)	IDA (pes)
AUROC	0.749 (0.043)	—	0.76 (0.062)	0.643 (0.079)
AUPRC	0.638 (0.053)	—	0.628 (0.078)	0.527 (0.088)
Spearman	0.48 (0.091)	—	0.491 (0.128)	0.254 (0.177)
MSE	0.056 (0.007)	—	0.182 (0.054)	0.126 (0.034)

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

Simulation setting 2: Systematic single KO

10 wild-types and one knock-out per gene

Table : Area under the ROC curve (AUROC), area under the precision-recall curve (AUPRC), Spearman correlation with true total causal effects, and mean squared error (MSE) of estimated total causal effects, averaged over 100 datasets (sd).

Criterion	MCMC-Mallows[1]	Pinna	IDA (opt)	IDA (pes)
AUROC	0.948 (0.03)	0.825 (0.048)	0.733 (0.068)	0.67 (0.073)
AUPRC	0.868 (0.042)	0.737 (0.059)	0.569 (0.087)	0.53 (0.091)
Spearman	0.696 (0.053)	0.553 (0.097)	0.42 (0.14)	0.318 (0.186)
MSE	0.026 (0.012)	0.104 (0.011)	0.334 (0.137)	0.196 (0.067)

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

Simulation setting 3: Partial single KO

15 wild-types and one knock-out for five genes {N1, N4, N6, N7, N9}

Table : Area under the ROC curve (AUROC), area under the precision-recall curve (AUPRC), Spearman correlation with true total causal effects, and mean squared error (MSE) of estimated total causal effects, averaged over 100 datasets (sd).

Criterion	MCMC-Mallows[1]	Pinna	IDA (opt)	IDA (pes)
AUROC	0.845 (0.059)	0.795 (0.017)	0.736 (0.056)	0.646 (0.085)
AUPRC	0.734 (0.078)	0.725 (0.038)	0.588 (0.075)	0.514 (0.092)
Spearman	0.587 (0.104)	0.636 (0.034)	0.449 (0.099)	0.285 (0.187)
MSE	0.035 (0.015)	0.081 (0.008)	0.215 (0.066)	0.146 (0.049)

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

Simulation setting 4: Multiple KO

10 wild types, one knock-out per gene and five double knock-outs: $\{N1, N5\}$, $\{N1, N6\}$, $\{N4, N7\}$, $\{N6, N9\}$, and $\{N7, N10\}$

Table : Area under the ROC curve (AUROC), area under the precision-recall curve (AUPRC), Spearman correlation with true total causal effects, and mean squared error (MSE) of estimated total causal effects, averaged over 100 datasets (sd).

Criterion	MCMC-Mallows[1]	Pinna	IDA (opt)	IDA (pes)
AUROC	0.959 (0.016)	0.83 (0.035)	0.733 (0.068)	0.67 (0.073)
AUPRC	0.886 (0.028)	0.725 (0.039)	0.569 (0.087)	0.53 (0.091)
Spearman	0.712 (0.028)	0.625 (0.058)	0.42 (0.14)	0.318 (0.186)
MSE	0.015 (0.006)	0.107 (0.008)	0.334 (0.137)	0.196 (0.067)

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

Simulation setting 5: Multiple KO and 3 hidden genes

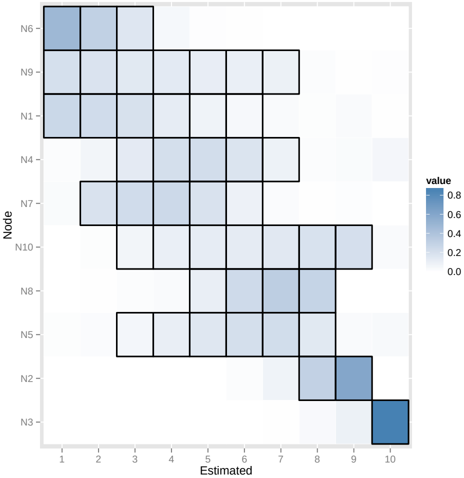
10 wild types, one knock-out per gene, five double knock-outs:
{N1, N5}, {N1, N6}, {N4, N7}, {N6, N9}, and {N7, N10}
and 3 randomly chosen hidden genes

Table : Area under the ROC curve (AUROC), area under the precision-recall curve (AUPRC), Spearman correlation with true total causal effects, and mean squared error (MSE) of estimated total causal effects, averaged over 100 datasets (sd).

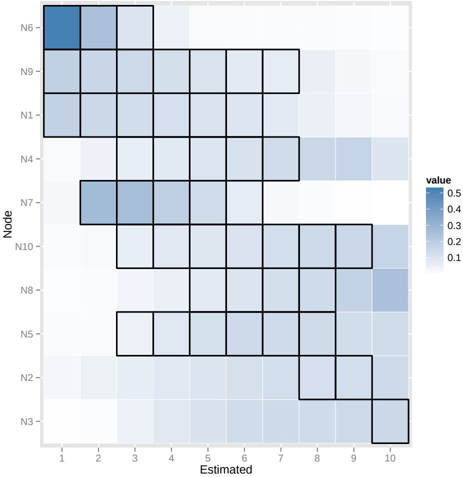
Criterion	MCMC-Mallows[1]	Pinna	IDA (opt)	IDA (pes)
AUROC	0.932 (0.046)	0.574 (0.165)	0.58 (0.145)	0.562 (0.121)
AUPRC	0.539 (0.078)	0.36 (0.105)	0.353 (0.086)	0.35 (0.08)
Spearman	0.67 (0.109)	0.037 (0.372)	0.076 (0.316)	0.076 (0.31)
MSE	0.044 (0.034)	0.15 (0.041)	0.45 (0.225)	0.294 (0.124)

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

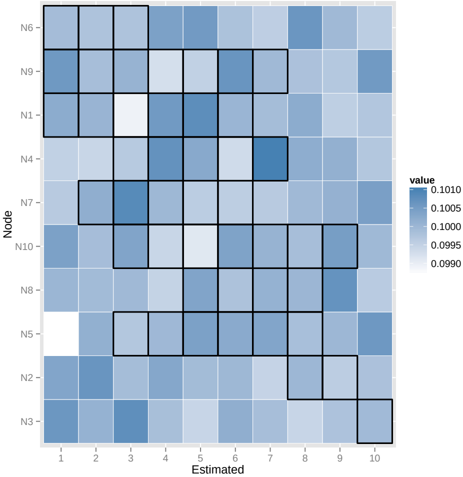
Posterior distribution of node ordering: Systematic single KO



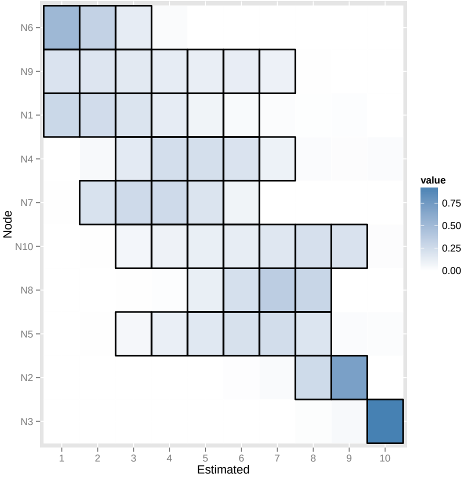
Posterior distribution of node ordering: Partial single KO



Posterior distribution of node ordering: Observational only



Posterior distribution of node ordering: Multiple KO



DREAM4 challenge

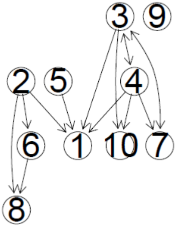
DREAM challenge: international competition held yearly to contribute to the development of powerful inference methods (Stolovitzky *et al.*, 2007)

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
- 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 for 100-gene networks

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 data: Partial knock-out setting

- For each of the five DREAM4 datasets, remove half of the knock-outs (chosen at random)

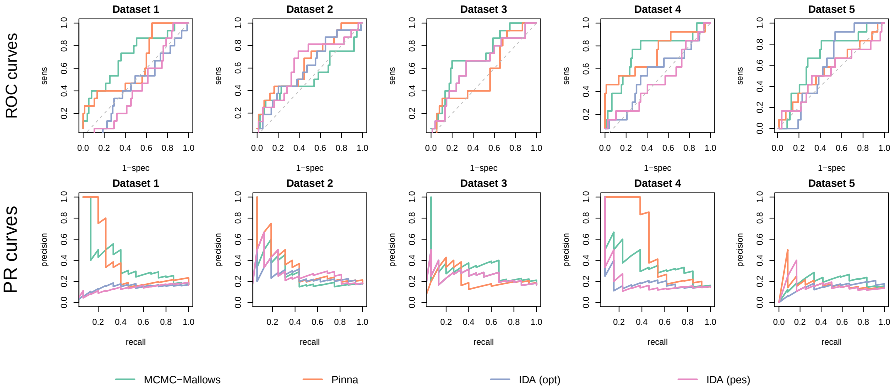
Compare GBN-Mallows total causal effect posterior means to Pinna and IDA

- GBN-Mallows[a]: wild-type, knock-out, & multifactorial perturbation data
- IDA: wild-type and multifactorial perturbation data
- Pinna: wild-type and knock-out data

[a] 50k iterations run, with burn-in of 5k and thinning every 50 iterations.

Trial run to select ϕ such that acceptance rate is $\approx$ 30-40%.

DREAM4 data: Partial knock-out setting



Discussion

GBN model for an arbitrary mixture of observational and knock-out (and multiple or partial knock-out!) data to enable calculation of causal effects:

- MCMC algorithm to explore posterior distribution of node ordering via Mallows proposal model
- Results suggest the benefit in jointly analyzing steady-state and (even incomplete) intervention data, as well as including multiple interventions

Future work

- Extension to larger-scale networks: MCMC with parallel tempering and sparsity constraints (ridge or Lasso) for **W**
- Experimental design to plan future (multiple) knock-out experiments...

Thanks to Rémi Bancal (M2 intern)

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.