Estimating Causal Effects in Gene Expression from a Mixture of Observational and Intervention Experiments

A. Rau[1], F. Jaffrézic[1], G. Nuel

INSMI, CNRS, Stochastics and Biology Group (PSB),
LPMA, UPMC, Sorbonne Universités, Paris, France

June 2014
Oldenburg, Germany

[1] GABI, INRA, Jouy-en-Josas

Outline

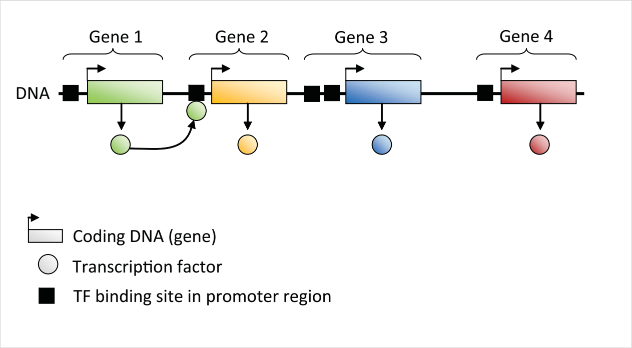
1. Causality in Gene Expression
   - Gene Regulatory Networks
   - Gaussian Bayesian Network
   - Causal Ordering
2. Mixing observation/intervention experiments
   - Maximizing the Likelihood
   - MCMC framework: Mallows
   - Pairwise preferences: Babington-Smith
3. Applications
   - Simulations
   - DREAM 4
   - Rosetta

Outline

1. Causality in Gene Expression
   - Gene Regulatory Networks
   - Gaussian Bayesian Network
   - Causal Ordering
2. Mixing observation/intervention experiments
   - Maximizing the Likelihood
   - MCMC framework: Mallows
   - Pairwise preferences: Babington-Smith
3. Applications
   - Simulations
   - DREAM 4
   - Rosetta

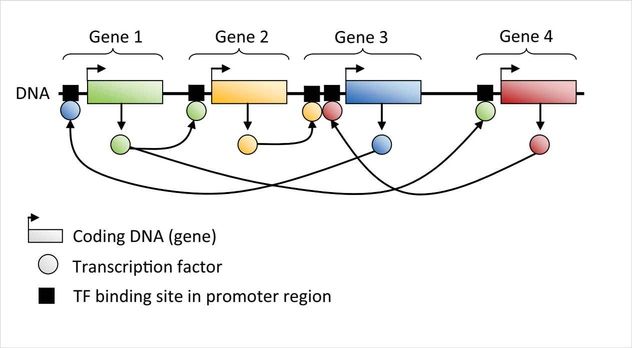
Gene regulatory networks (GRN)

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



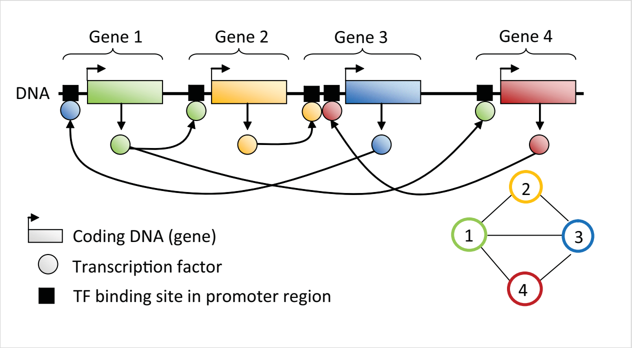
Gene regulatory networks (GRN)

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



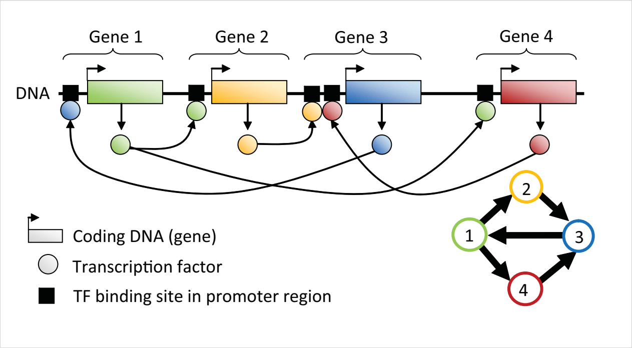
Gene regulatory networks (GRN)

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

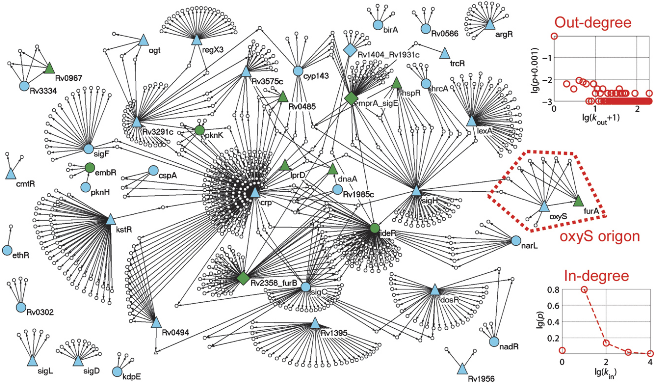


Gene regulatory networks (GRN)

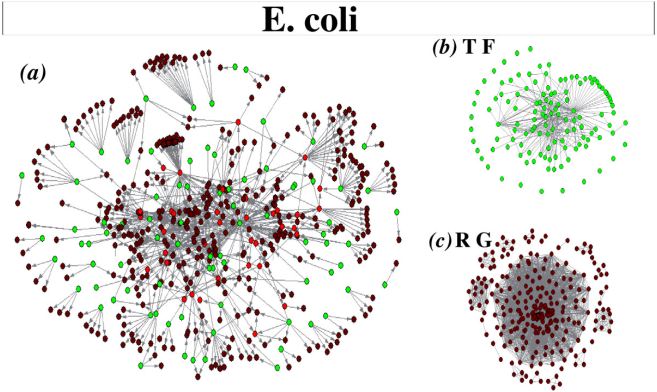
- Groups of coordinated genes that interact indirectly with one another through transcription factors



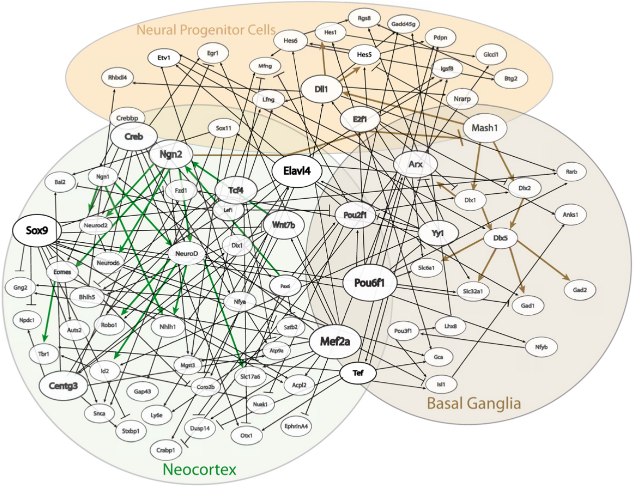
Examples of real-life GRN



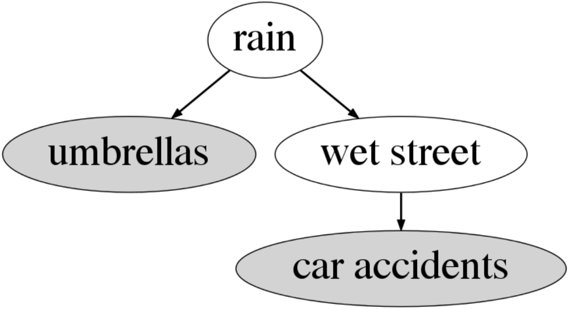
Examples of real-life GRN



Examples of real-life GRN



Correlation *versus* Causality



umbrellas and car accidents are correlated

But:

- provoking car accidents does not make appear umbrellas
- distributing umbrellas in the street does not provoke car accidents

Outline

1. Causality in Gene Expression
   - Gene Regulatory Networks
   - Gaussian Bayesian Network
   - Causal Ordering
2. Mixing observation/intervention experiments
   - Maximizing the Likelihood
   - MCMC framework: Mallows
   - Pairwise preferences: Babington-Smith
3. Applications
   - Simulations
   - DREAM 4
   - Rosetta

Causal Gaussian Bayesian Network

X_j^k is the expression of gene $j \in 1, \ldots, p$ in experiment $k \in 1, \ldots, N$

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

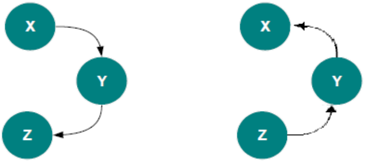
with $W_{i,j} \neq 0$ if and only if $i \in \mathrm{pa}(j)$ and nodes ordered such that that $i \in \mathrm{pa}(j) \Rightarrow i < j$ (i.e., $\mathbf{W} = (W_{i,j})$ is upper triangular). Model parameters are $\theta = (\mathbf{W}, \mathbf{m}, \boldsymbol{\sigma})$.

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

$$W_{i,j} = \frac{d}{dx} \mathbb{E}[X_j | X_{-j}, \mathrm{do}(X_i = x)] \quad L_{i,j} = \frac{d}{dx} \mathbb{E}[X_j | \mathrm{do}(X_i = x)]$$

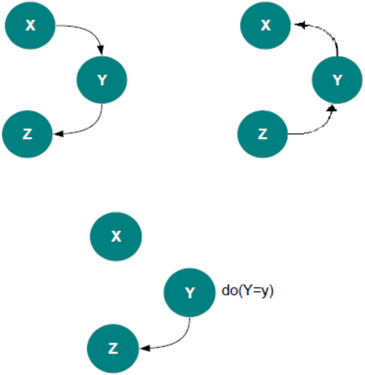
Markov equivalence in DAGs

- 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 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 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

$\Rightarrow$ winner of the DREAM4 100-gene challenge

Outline

1. Causality in Gene Expression
   - Gene Regulatory Networks
   - Gaussian Bayesian Network
   - Causal Ordering
2. Mixing observation/intervention experiments
   - Maximizing the Likelihood
   - MCMC framework: Mallows
   - Pairwise preferences: Babington-Smith
3. Applications
   - Simulations
   - DREAM 4
   - Rosetta

Posterior Causal Ordering

For any given ordering $\boldsymbol{o} = o_1, o_2, \dots, o_p$ we assume the full model: $W_{i,j} \neq \forall i < j$ (not suitable for large p without some kind of regularization).

Posterior Causal Ordering is defined as:

$$\mathbb{P}(\boldsymbol{o}|\text{data}) \propto \mathbb{P}(\text{data}|\hat{\theta}_{\boldsymbol{o}}) \times \mathbb{P}(\boldsymbol{o})$$

where $\hat{\theta}_{\boldsymbol{o}}$ is the MLE of the full model with causal ordering $\boldsymbol{o}$ and $\mathbb{P}(\boldsymbol{o})$ is a prior distribution.

Causal effect estimates:

$$\hat{\boldsymbol{W}} = \sum_{\boldsymbol{o}} \mathbb{P}(\boldsymbol{o}|\text{data}) \times \hat{\boldsymbol{W}}_{\boldsymbol{o}} \quad \text{and} \quad \hat{\boldsymbol{L}} = \sum_{\boldsymbol{o}} \mathbb{P}(\boldsymbol{o}|\text{data}) \times \hat{\boldsymbol{L}}_{\boldsymbol{o}}$$

log-likelihood: observational data only

We can show that the GBN 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:

$$\begin{aligned}\ell(\mathbf{m}, \boldsymbol{\sigma}, \mathbf{W}) = & \quad \text{Cst} - N \sum_j \log(\sigma_j) - \frac{1}{2} \sum_k \sum_j \frac{1}{\sigma_j^2} (x_j^k - \mathbf{x}^k \mathbf{W} \mathbf{e}_j^T - m_j)^2 \\ = & \quad \text{Cst} - N \sum_j \log(\sigma_j) - \frac{1}{2} \sum_k \sum_j \frac{1}{\sigma_j^2} (y_j^k - \mathbf{y}^k \mathbf{W} \mathbf{e}_j^T)^2\end{aligned}$$

with $y_i^k = \left(x_i^k - \frac{1}{N} \sum_{k'} x_i^{k'} \right)$

log-likelihood: observational data only

Simple analytical analysis gives:

$$m_j = \frac{1}{N}\sum_k (x_j^k - \boldsymbol{x}^k \boldsymbol{W}\boldsymbol{e}_j^T) \quad \sigma_j^2 = \frac{1}{N}\sum_k (y_j^k - \boldsymbol{y}^k \boldsymbol{W}\boldsymbol{e}_j^T)^2$$

and $\boldsymbol{W}$ solution of the following linear system, for all (i,j) s.t. $i \in \text{pa}_j$:

$$\sum_{i' \in \text{pa}_j} W_{i',j} \sum_k y_i^k y_{i'}^k = \sum_k y_i^k y_j^k$$

In the full model, $\text{pa}_j = \{i, i < j\}$ we get:

$$\max \ell(\mathbf{m}, \boldsymbol{\sigma}, \mathbf{W}) = \text{Cst} - \frac{N}{2} \log\det\left(\sum_k y_i^k y_j^k\right)$$

$\Rightarrow$ **obs. data are uninformative for the causal ordering**

Outline

1. Causality in Gene Expression
   - Gene Regulatory Networks
   - Gaussian Bayesian Network
   - Causal Ordering
2. Mixing observation/intervention experiments
   - Maximizing the Likelihood
   - MCMC framework: Mallows
   - Pairwise preferences: Babington-Smith
3. Applications
   - Simulations
   - DREAM 4
   - Rosetta

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:

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

Then

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

log-likelihood: Observational + intervention data (2)

The log-likelihood of the model can then be rewritten as:

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

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

Then **W** solution of 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

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

Outline

1. Causality in Gene Expression
   - Gene Regulatory Networks
   - Gaussian Bayesian Network
   - Causal Ordering
2. Mixing observation/intervention experiments
   - Maximizing the Likelihood
   - MCMC framework: Mallows
   - Pairwise preferences: Babington-Smith
3. Applications
   - Simulations
   - DREAM 4
   - Rosetta

Metropolis-Hasting

Objective: draw samples from $\mathbb{P}(\boldsymbol{o}|\text{data})$ (which is only known up to a normalization factor).

Metropolis-Hasting algorithm:

1. start from arbitrary order $\boldsymbol{o}^{(0)}$
2. for $i = 1, \dots, N$:
    - propose $\boldsymbol{o}'$ according to proposal distribution $Q(\boldsymbol{o}'|\boldsymbol{o}^{(i-1)})$
    - compute acceptance rate

$$\min\left(1, \frac{\mathbb{P}(\boldsymbol{o}'|\text{data}) \times Q(\boldsymbol{o}^{(i-1)}|\boldsymbol{o}')}{\mathbb{P}(\boldsymbol{o}^{(i-1)}|\text{data}) \times Q(\boldsymbol{o}'|\boldsymbol{o}^{(i-1)})}\right)$$

    - if move accepted $\boldsymbol{o}^{(i)} = \boldsymbol{o}'$ else $\boldsymbol{o}^{(i)} = \boldsymbol{o}^{(i-1)}$
3. $\boldsymbol{o}^{(0)}, \boldsymbol{o}^{(1)}, \boldsymbol{o}^{(N)}$ is a (dependent) sample of the target distribution.

Mallows' Proposal

Mallows' Ranking Distribution: with parameter $\phi \in]0,1[$ and reference ordering $\boldsymbol{r}$ is defined by

$$\mathbb{P}(\boldsymbol{o};\phi,\boldsymbol{r}) = \phi^{d(\boldsymbol{o},\boldsymbol{r})}$$

where $d(\boldsymbol{o},\boldsymbol{r})$ counts the number of pairwise disagreements.

Properties:

- mode is in $\boldsymbol{r}$
- $\phi \to 0$ corresponds to a dirac distribution
- $\phi \to 1$ corresponds to the uniform distribution
- normalization factor is $1 \times (1+\phi) \times \ldots \times (1+\phi+\ldots+\phi^{p-1})$
- sampling in $O(p)$ with the Repeated Insertion Method

Mallow's distribution in action

$\phi = 0.1$	$\phi = 0.3$	$\phi = 0.6$	$\phi = 0.9$
1 2 4 3 5	1 2 3 4 5	1 3 4 5 2	3 4 2 5 1
1 2 3 4 5	2 1 3 4 5	1 3 4 5 2	1 4 5 3 2
1 3 2 4 5	3 1 2 4 5	1 5 3 2 4	3 2 4 5 1
2 1 3 4 5	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5
1 2 3 4 5	1 2 3 5 4	4 5 3 1 2	2 1 5 3 4
1 2 3 4 5	2 1 4 3 5	1 3 2 4 5	2 4 5 1 3
1 2 3 4 5	1 2 4 3 5	3 1 5 2 4	3 4 2 5 1
1 3 2 5 4	1 2 3 4 5	1 2 3 5 4	4 2 1 3 5
1 2 3 4 5	1 2 3 4 5	1 2 4 3 5	3 4 2 1 5
1 2 4 3 5	1 3 4 5 2	1 3 4 5 2	1 5 3 4 2

Table : Example illustrating ten draws from the Mallows model with a reference ordering of $\boldsymbol{r} = (1\ 2\ 3\ 4\ 5)$ for different temperatures $(\phi = 0.1, 0.3, 0.6, 0.9)$.

Outline

1. Causality in Gene Expression
   - Gene Regulatory Networks
   - Gaussian Bayesian Network
   - Causal Ordering
2. Mixing observation/intervention experiments
   - Maximizing the Likelihood
   - MCMC framework: Mallows
   - Pairwise preferences: Babington-Smith
3. Applications
   - Simulations
   - DREAM 4
   - Rosetta

Babington-Smith ranking distribution

Pairwise preferences: for any pair of distinct genes (i, j) one can easily compute:

$$\pi_{i,j} = \mathbb{P}(i < j|\text{data}_{i,j}) \propto \mathbb{P}(\text{data}_{i,j}|i < j)$$

$$\pi_{j,i} = \mathbb{P}(j < i|\text{data}_{i,j}) \propto \mathbb{P}(\text{data}_{i,j}|j < i)$$

with $\pi_{i,j} + \pi_{j,i} = 1$.

Idea: use pairwise preferences to obtain an approximated support for $\mathbb{P}(\boldsymbol{o}|\text{data})$ using the Babington-Smith distribution.

$$\mathbb{P}(\boldsymbol{o}; \boldsymbol{\pi}) \propto \prod_{i<j} \pi_{o_i,o_j}$$

(ex: if $\boldsymbol{o} = (3\ 1\ 2)$, $\mathbb{P}(\boldsymbol{o}; \boldsymbol{\pi}) \propto \pi_{3,1}\pi_{3,2}\pi_{1,2}$)

Babington-Smith Strategy

Problem: Repeated Insertion Method not applicable for Babington-Smith distribution. MCMC sampling necessary !

Three steps strategy:

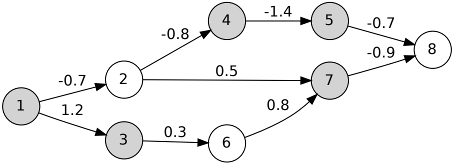
1) compute pairwise preferences π
   $\Rightarrow O(p^2)$ but fast since on restricted datasets
2) sample from Babington-Smith distribution $\mathbb{P}(\boldsymbol{o};\pi)$
   $\Rightarrow$ fast MCMC since likelihood depend only on π
3) compute posterior distribution on approximated support $\mathcal{O}$
   $\Rightarrow$ retain only the most likely orderings, support size arbitrary

Remarks:

- the strategy is fast, only Step 3 is time consuming
- what if Babington-Smith support differs from real support ?

Outline

1. Causality in Gene Expression
   - Gene Regulatory Networks
   - Gaussian Bayesian Network
   - Causal Ordering
2. Mixing observation/intervention experiments
   - Maximizing the Likelihood
   - MCMC framework: Mallows
   - Pairwise preferences: Babington-Smith
3. Applications
   - Simulations
   - DREAM 4
   - Rosetta



$\boldsymbol{N = 30}$: 10 with $\mathcal{J}_k = \{1\}$, 10 with $\mathcal{J}_k = \{3, 4\}$, 10 with $\mathcal{J}_k = \{5, 7\}$

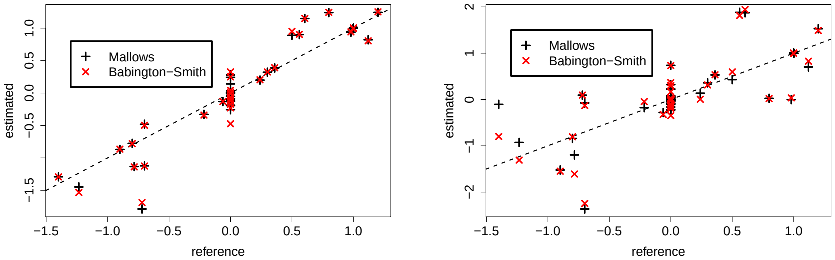
$$\boldsymbol{L}^* = (\boldsymbol{I} - \boldsymbol{W}^*)^{-1} = \begin{pmatrix} 1 & -0.70 & 1.20 & 0.56 & -0.78 & 0.36 & -0.06 & 0.60 \\ 0 & 1 & 0 & -0.80 & 1.12 & 0 & 0.50 & -1.23 \\ 0 & 0 & 1 & 0 & 0 & 0.30 & 0.24 & -0.22 \\ 0 & 0 & 0 & 1 & -1.40 & 0 & 0 & 0.98 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -0.70 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0.80 & -0.72 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -0.90 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

$$\boldsymbol{m}^* = (0.5, 1.2, 0.7, 0.6, 1.4, 0.5, 0.8, 1.2)$$

$$\boldsymbol{\sigma}^* = \eta(0.3, 1.1, 0.6, 0.3, 1.0, 0.5, 0.8, 1.3) \text{ with } \eta = 0.1 \text{ or } \eta = 1.0$$

- **MCMC-Mallows**: $\varphi = 0.2$, iter$= 1000 + 5000$, time$\simeq 100$
- **Babington-Smith**: iter$= 1000 + 5000$, max$= 60$, time$\simeq 1$

$\hat{\boldsymbol{W}}$ versus $\boldsymbol{W}^*$



MSE	MCMC-Mallows	Babington-Smith
$\eta = 0.1$	0.043	0.045
$\eta = 1.0$	0.194	0.174

Top 10 causal orderings

MCMC-Mallows			Babington-Smith sampling		
Gene ordering	log L	DAG err.	Gene ordering	log L	DAG err.
1, 2, 4, 5, 3, 6, 7, 8	-0.8832	0	1, 2, 4, 3, 6, 7, 5, 8	-0.8431	0
1, 3, 2, 4, 6, 7, 5, 8	-1.2104	0	1, 2, 3, 4, 6, 7, 5, 8	-0.8431	0
1, 3, 2, 4, 6, 5, 7, 8	-1.2104	0	1, 2, 4, 3, 6, 5, 7, 8	-0.8431	0
1, 2, 4, 3, 6, 7, 5, 8	-1.2378	0	1, 2, 3, 4, 6, 5, 7, 8	-0.8431	0
1, 2, 3, 4, 6, 7, 5, 8	-1.2378	0	1, 2, 3, 6, 4, 7, 5, 8	-0.9217	0
1, 2, 4, 3, 6, 5, 7, 8	-1.2378	0	1, 2, 3, 6, 4, 5, 7, 8	-0.9217	0
1, 2, 3, 4, 6, 5, 7, 8	-1.2378	0	1, 2, 3, 6, 7, 4, 5, 8	-1.1079	0
1, 3, 2, 6, 4, 7, 5, 8	-1.2890	0	1, 2, 4, 3, 5, 6, 7, 8	-1.3276	0
1, 3, 2, 6, 4, 5, 7, 8	-1.2890	0	1, 2, 3, 4, 5, 6, 7, 8	-1.3276	0
1, 3, 6, 2, 4, 7, 5, 8	-1.2890	0	1, 2, 3, 4, 7, 6, 5, 8	-2.5226	1

$$\eta = \mathbf{0.1}$$

DAG err. = number of ordering inconsistencies with the true DAG.

Top 10 causal orderings

MCMC-Mallows			Babington-Smith sampling		
Gene ordering	log L	DAG err.	Gene ordering	log L	DAG err.
1, 2, 7, 8, 3, 5, 6, 4	-1.3537	3	1, 2, 7, 8, 4, 3, 5, 6	-0.9316	2
1, 2, 7, 3, 5, 6, 8, 4	-1.4674	2	1, 2, 7, 8, 3, 4, 5, 6	-0.9316	2
1, 2, 7, 3, 5, 8, 6, 4	-1.4674	2	1, 2, 7, 3, 8, 4, 5, 6	-1.0712	2
1, 2, 7, 3, 8, 5, 6, 4	-1.4933	3	1, 2, 7, 3, 4, 5, 8, 6	-1.4468	1
1, 7, 8, 3, 2, 5, 6, 4	-1.6368	4	1, 2, 7, 3, 4, 5, 8, 6	-1.4468	1
1, 5, 3, 2, 7, 6, 8, 4	-1.6849	2	1, 2, 7, 3, 4, 5, 6, 8	-1.4468	1
1, 5, 3, 2, 7, 8, 6, 4	-1.6849	2	1, 2, 7, 3, 4, 5, 6, 8	-1.4468	1
1, 7, 3, 2, 5, 6, 8, 4	-1.7490	3	1, 2, 7, 4, 3, 5, 8, 6	-1.4468	1
1, 7, 3, 2, 5, 8, 6, 4	-1.7490	3	1, 2, 7, 4, 3, 5, 8, 6	-1.4468	1
1, 7, 3, 2, 8, 5, 6, 4	-1.7749	4	1, 2, 7, 4, 3, 5, 6, 8	-1.4468	1

$$\eta = \mathbf{1.0}$$

DAG err. = number of ordering inconsistencies with the true DAG.

Outline

1. Causality in Gene Expression
   - Gene Regulatory Networks
   - Gaussian Bayesian Network
   - Causal Ordering
2. Mixing observation/intervention experiments
   - Maximizing the Likelihood
   - MCMC framework: Mallows
   - Pairwise preferences: Babington-Smith
3. Applications
   - Simulations
   - DREAM 4
   - Rosetta

10-genes network challenge

DREAM = Dialogue for Reverse Engineering Assessments and Methods

Data: 5 datasets, each containing 1 wildtype and 10 KO (one for each gene), true network (with feedback loops) known.

Dataset	Pinna	MCMC-Mallows	Babington-Smith
1	0.83 (0.71,0.95)	0.53 (0.35,0.72)	0.60 (0.41,0.79)
2	0.52 (0.35,0.70)	0.52 (0.36,0.68)	0.55 (0.39,0.71)
3	0.82 (0.69,0.94)	0.69 (0.54,0.84)	0.72 (0.56,0.88)
4	0.90 (0.79,1.00)	0.87 (0.76,0.99)	0.90 (0.78,1.00)
5	0.70 (0.53,0.87)	0.81 (0.69,0.93)	0.76 (0.61,0.90)
All	0.73 (0.67,0.80)	0.80 (0.73,0.86)	0.75 (0.68,0.83)

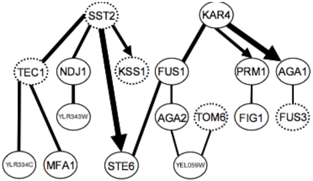
AUC results (with 95% CI) using statistic $|\hat{W}_{i,j}|$

Outline

1. Causality in Gene Expression
   - Gene Regulatory Networks
   - Gaussian Bayesian Network
   - Causal Ordering
2. Mixing observation/intervention experiments
   - Maximizing the Likelihood
   - MCMC framework: Mallows
   - Pairwise preferences: Babington-Smith
3. Applications
   - Simulations
   - DREAM 4
   - Rosetta

Rosetta compendium

300 experiments on yeast, database freely available:
`http://arep.med.harvard.edu/ExpressDB/yeastindex.html`

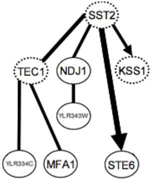


17-genes mating response network (Pe'er *et al*, 2001).

***N* = 300**: 294 wildtypes, 1 KO on TOM6, 4 KD on FUS3, KSS1, SST2, TEC1, 1 MKD on FUS3 and KSS1.

Rosetta compendium

300 experiments on yeast, database freely available:
`http://arep.med.harvard.edu/ExpressDB/yeastindex.html`



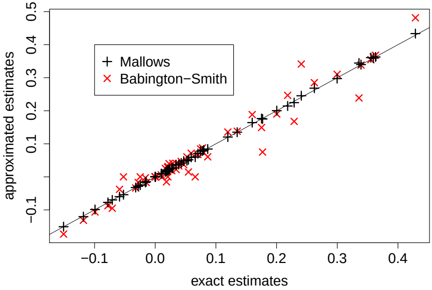
8-genes subnetwork.

***N* = 300**: 294 wildtypes, 1 KO on TOM6, 4 KD on FUS3, KSS1, SST2, TEC1, 1 MKD on FUS3 and KSS1.

Results on the 8-genes subnetwork

$8! = 40,320$ orderings, exhaustive search gives:

$$\hat{W}^{\text{exact}} = \sum_{o} \mathbb{P}(o|\text{data})\hat{W}_o$$



Mallows: $\text{MSE} = 5.7 \times 10^{-6}$
Babington-Smith: $\text{MSE} = 8.6 \times 10^{-4}$

Close-up on Babington-Smith

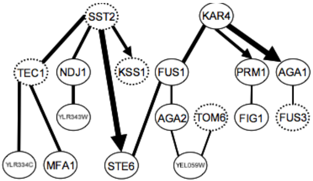
genes	TEC1	MFA1	KSS1	STE6	YLR334C	YLR343W	SST2	NDJ1
TEC1	—	0.50	1.00	0.26	0.50	0.48	0.87	0.51
MFA1	0.50	—	0.66	0.50	0.50	0.50	0.41	0.50
KSS1	0.00	0.34	—	0.01	0.25	0.00	0.04	0.29
STE6	0.74	0.50	0.99	—	0.50	0.50	0.96	0.50
YLR334C	0.50	0.50	0.75	0.50	—	0.50	0.49	0.50
YLR343W	0.52	0.50	1.00	0.50	0.50	—	0.78	0.50
SST2	0.13	0.59	0.96	0.04	0.51	0.22	—	0.34
NDJ1	0.49	0.50	0.71	0.50	0.50	0.50	0.66	—

pairwise preferences

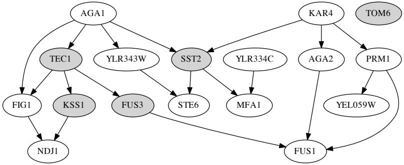
gene order	ties	Δ^{exact}	Δ^{BS}
STE6/YLR334C/YLR343W TEC1 SST2 KSS1 MFA1/NDJ1	12	ref	−0.920
STE6/YLR334C TEC1 SST2 YLR343W KSS1 MFA1/NDJ1	4	−0.003	−2.265
STE6/YLR334C TEC1 YLR343W SST2 KSS1 MFA1/NDJ1	4	−0.009	ref
STE6 TEC1 YLR334C SST2 YLR343W KSS1 MFA1/NDJ1	2	−0.056	−2.265
STE6 TEC1 YLR334C/YLR343W SST2 KSS1 MFA1/NDJ1	4	−0.062	−1.000

most likely causal orderings

Results on the full mating response network



Pe’er *et al* (2001)



MCMC-Mallows

Mating response network inferred from Rosetta dataset. Only the 20 largest direct effects are represented. Grey nodes correspond to genes which have been mutated in some of the samples

Causal ordering:

- DAG condition $\iff$ causal ordering
- observation data only are uninformative for the causal ordering
- we provide likelihood maximization formulas for any given ordering

Statistical inference

- exhaustive search in $O(p!)$ ($p \simeq 10$ max)
- MCMC-Mallows works well
- Babington-Smith fast but unreliable

Further work

- extend Babington-Smith to triplet preferences ?
- large p with regularization (ex: Ridge) and parallel tempering
- using Fisher information to develop adaptive designs