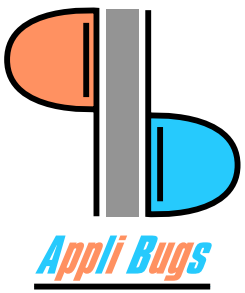
College of Science - Department of Statistics
Statistical Bioinformatics Center

Exploring the identifiability of gene regulatory networks with approximate Bayesian computation

AppliBUGS meeting
AgroParisTech

Andrea Rau

December 9, 2011

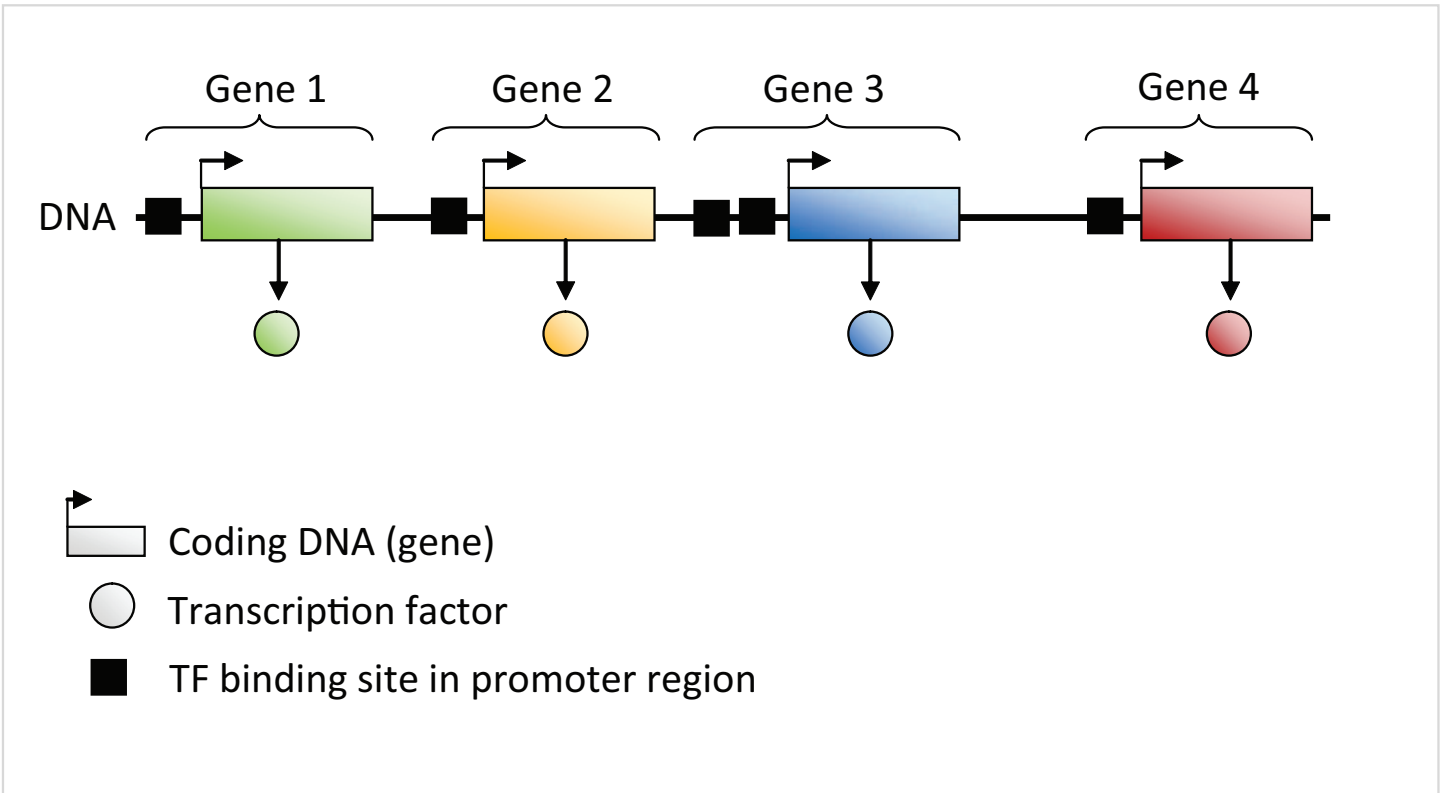


Gene Expression

- Genes: Functional regions of DNA that encode proteins and RNA molecules
- Expression levels of thousands of genes can be measured using "high-throughput" technologies (e.g., microarrays, serial analysis of gene expression, next-generation sequencing)

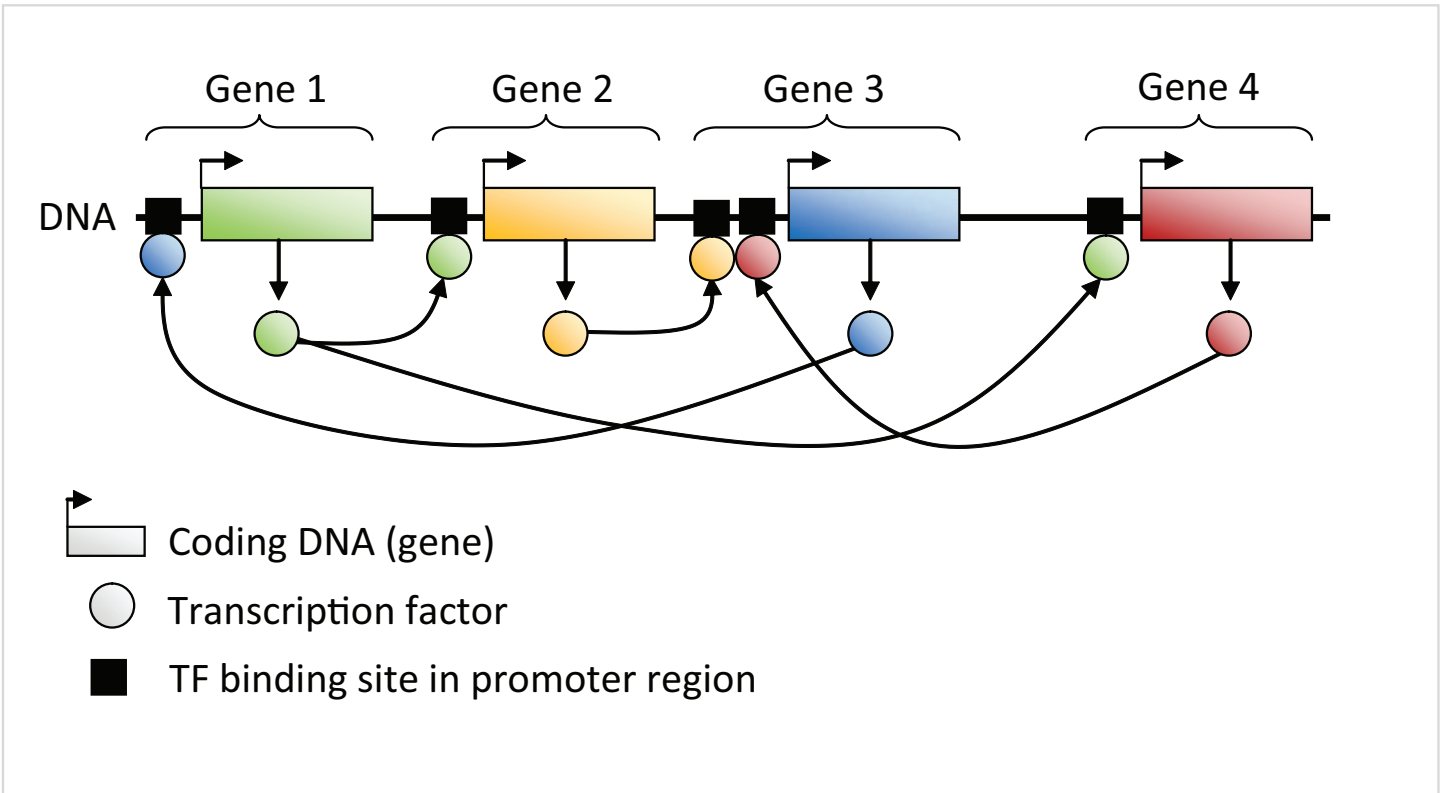
Gene Regulatory Networks

- *Gene regulatory networks*: set of genes that interact indirectly with one another through proteins called transcription factors (TF)
- Abundance of TF is difficult to measure $\Rightarrow$ expression levels of corresponding genes usually used as proxy



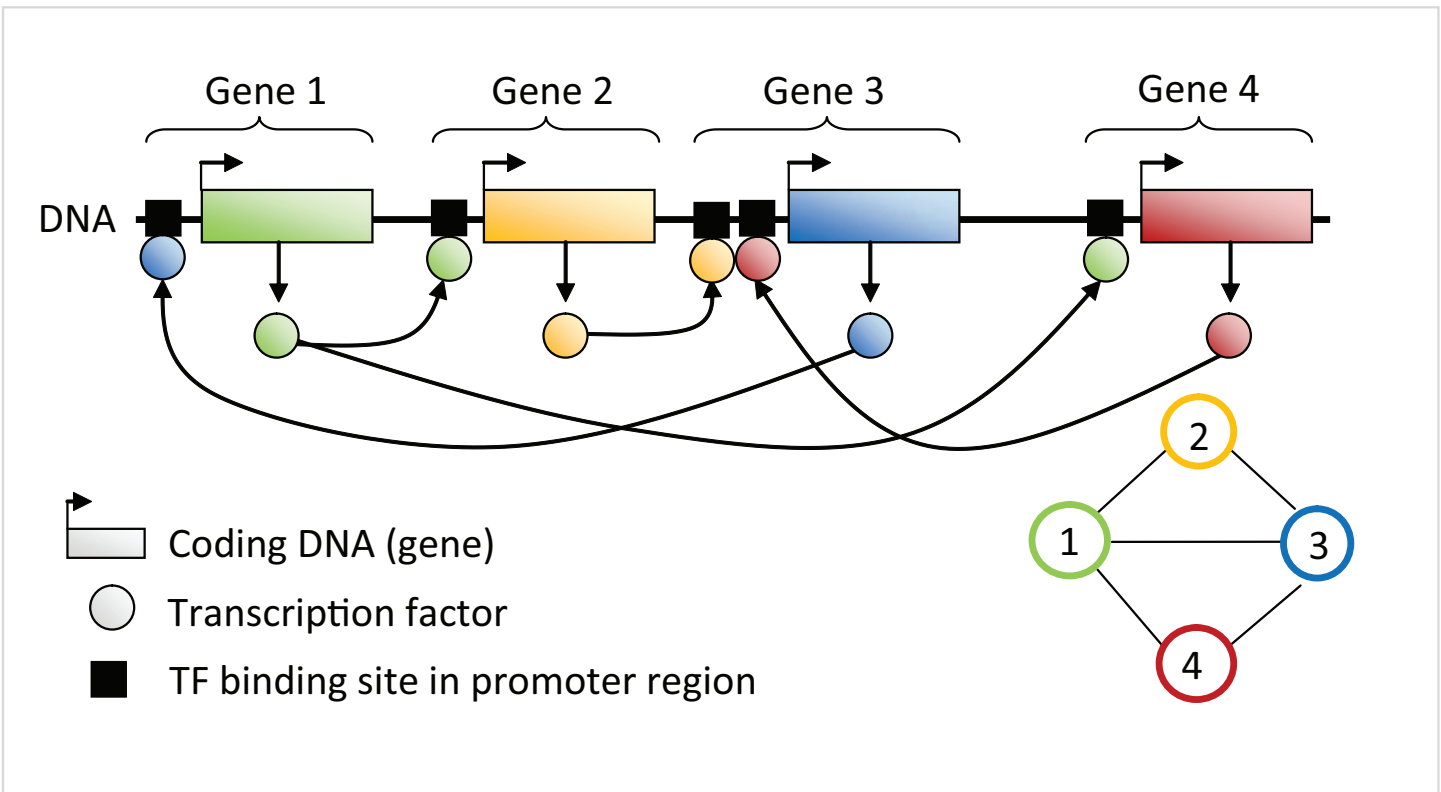
Gene Regulatory Networks

- *Gene regulatory networks*: set of genes that interact indirectly with one another through proteins called transcription factors (TF)
- Abundance of TF is difficult to measure $\Rightarrow$ expression levels of corresponding genes usually used as proxy



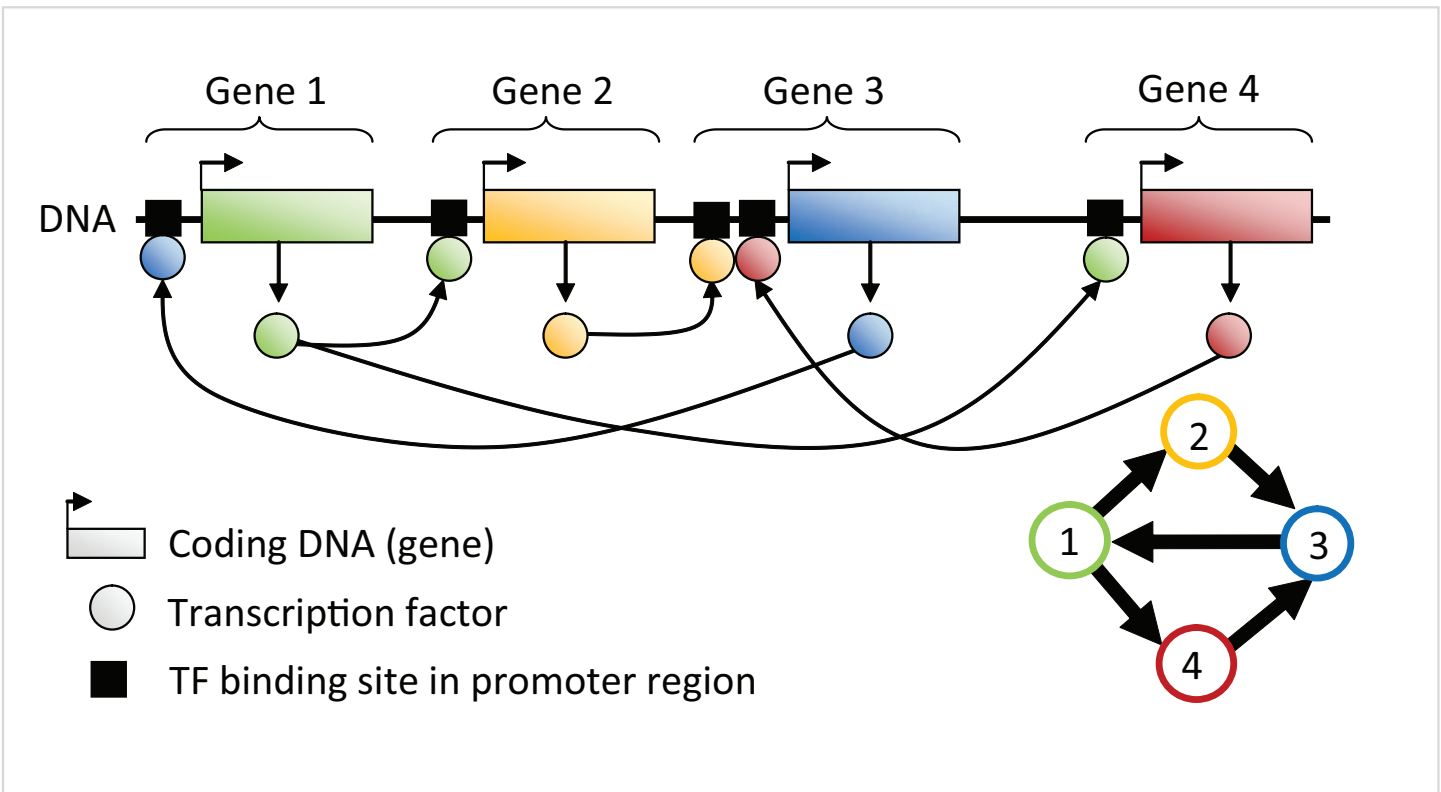
Gene Regulatory Networks

- *Gene regulatory networks*: set of genes that interact indirectly with one another through proteins called transcription factors (TF)
- Abundance of TF is difficult to measure $\Rightarrow$ expression levels of corresponding genes usually used as proxy



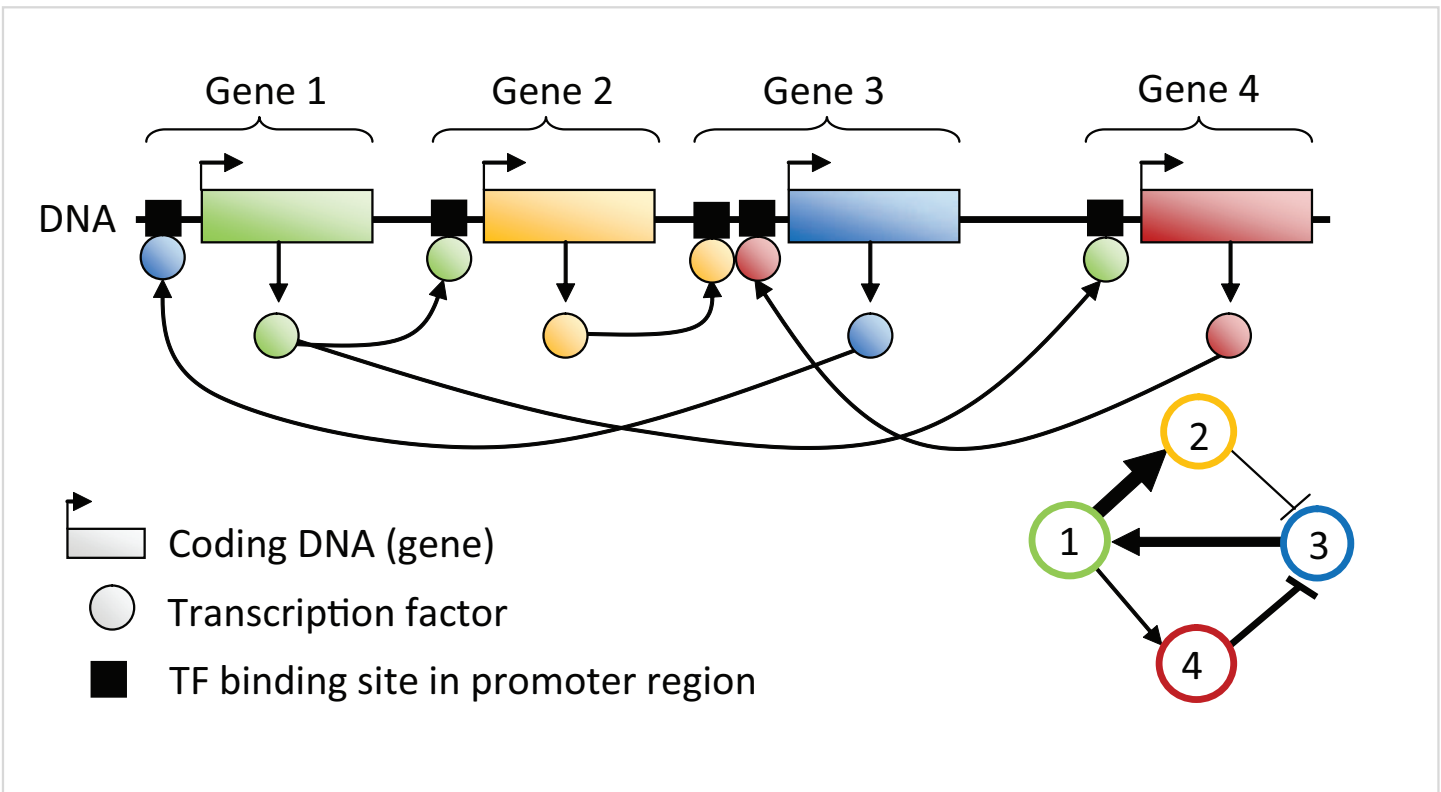
Gene Regulatory Networks

- *Gene regulatory networks*: set of genes that interact indirectly with one another through proteins called transcription factors (TF)
- Abundance of TF is difficult to measure $\Rightarrow$ expression levels of corresponding genes usually used as proxy



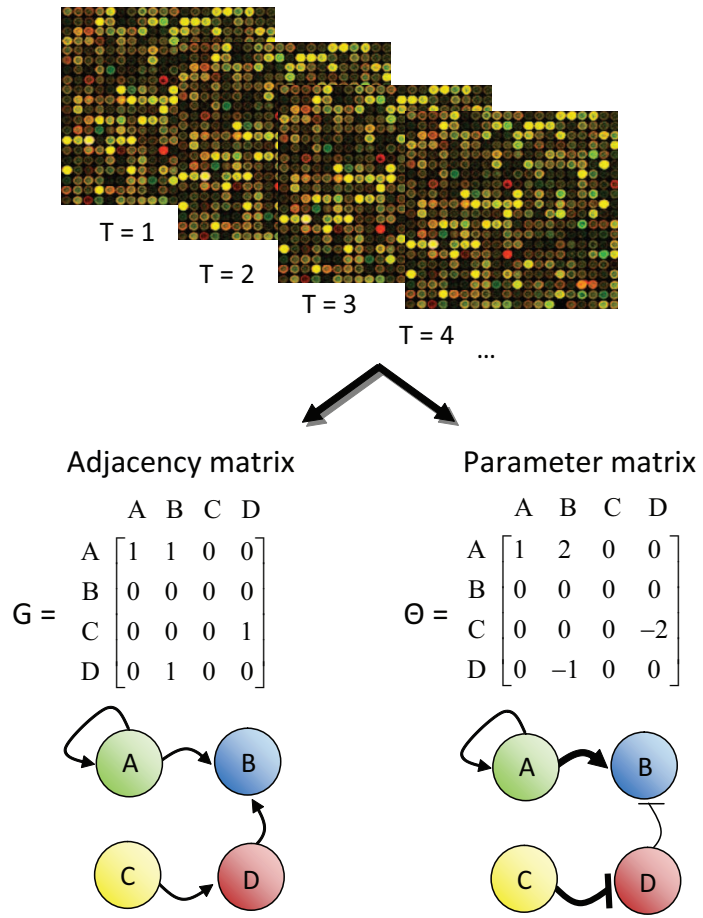
Gene Regulatory Networks

- *Gene regulatory networks*: set of genes that interact indirectly with one another through proteins called transcription factors (TF)
- Abundance of TF is difficult to measure $\Rightarrow$ expression levels of corresponding genes usually used as proxy



Reverse Engineering Gene Regulatory Networks

- Expression levels of thousands of genes can be measured using "high-throughput" technologies (few replicates or time points)
- **Objective**: Use time-course gene expression data to elucidate information about *patterns* of relationships of gene expression



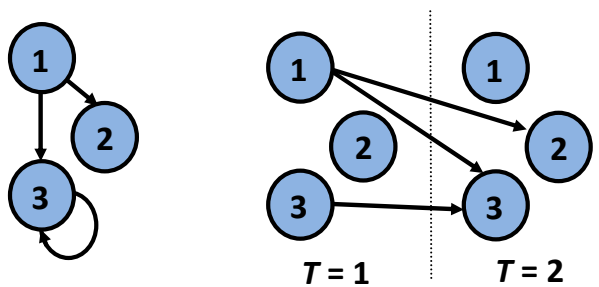
Bayesian Network Framework

Bayesian Network (BN):

- Graphical model to represent *conditional* probabilistic relationships among random variables
- Graphical structure, $\mathcal{G} = (V, E)$ defined by a set of vertices V and edges E and a family of conditional distributions $\mathcal{F}$

Dynamic Bayesian Network (DBN):

- BN limitations: no feedback loops, discrete data, equivalence classes
- Unfold BN over time



Identifiability of gene regulatory (sub-)networks?

- Often, similar inference approaches yield very different network structures on a common dataset
- In addition, complicated network motifs may be difficult or impossible to infer from the available data

- **Question**: Is it possible to determine whether parts of a given network are identifiable, given the available data?

Outline for the rest of the talk

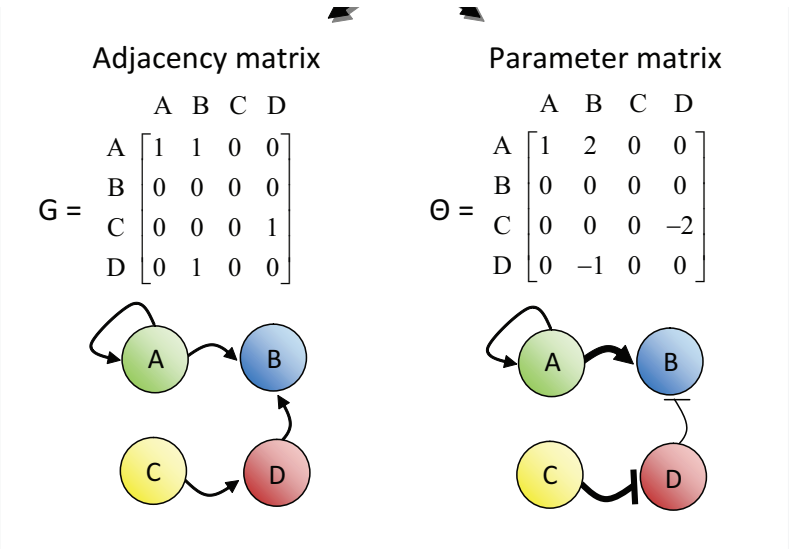
- Approximate Bayesian Computation
  - Background and motivation
  - Monte Carlo approaches
- ABC-MCMC for Networks
  - Simulation studies
  - Real data analysis: SOS DNA repair system in *E. coli*
- Discussion

Some notation

Let observed time-course gene expression data be $Y = \{\mathbf{y}_t : t = 1, \ldots, T\}$, where $\mathbf{y}_t = (y_{t1}, ..., y_{tP})'$ for P genes at T equally spaced time points.

Two related characterizations of a gene regulatory network:

- Adjacency matrix G ($G_{jk} = 1$ if gene k regulates gene j, 0 otherwise)
- Parameter matrix Θ (θ_{jk} represents the relationship between gene k at time $t-1$ and gene j at time t)



Bayesian Framework

- High dimensional problem: many possible gene-to-gene interactions ($\mathcal{O}(P^2)$), usually few time points ($T < 10$)
- Number of possible network structures increases exponentially as the number of genes increases, and many network structures may yield similarly high likelihoods
- Examining the shape of posterior distributions may give additional information about the structure and inferability of specific gene-to-gene interactions
- *A priori* biological information may be encoded into the prior distributions

Likelihood specification

- For a given matrix of gene regulatory network parameters Θ:

$$Y \sim \prod_t f(\mathbf{y}_t; \mathbf{y}_{t-1}, \Theta)$$

  where $\mathbf{y}_0 = 0$.

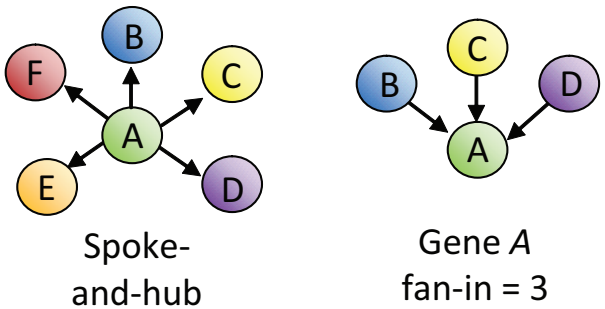
- Simple, linear models (e.g., the first-order vector autoregressive (VAR(1)) model) have been found to be good approximations in some cases to the dynamics of time-course expression data:

$$f(\mathbf{y}_t; \mathbf{y}_{t-1}, \Theta) = \Theta \mathbf{y}_{t-1} + \mathbf{e}_t$$

  where $E(\mathbf{e}_t) = 0$, $E(\mathbf{e}_t \mathbf{e}_t') = \Sigma$ (a positive definite covariance matrix), and $E(\mathbf{e}_t \mathbf{e}_{t'}') = 0$.

Network prior distributions (G and Θ)

- Gene regulatory networks typically sparse with spoke-and-hub structure and few regulators per gene (fan-in)



Prior distributions:

- $\pi(G)$ is uniform over all structures, with maximum fan-in of 5 or less
- $\pi(\theta_{jk}|G_{jk} = 1) \sim \mathcal{U}(-2, 2)$

Approximate Bayesan Computation (ABC)

- **Objective**: infer network from observed expression data Y via the posterior

$$\pi(\Theta, G|Y) \propto f(Y|\Theta)\pi(\Theta|G)\pi(G)$$

- Without restrictive distributional assumptions on model parameters ($\mathbf{e}_t$), likelihood may be difficult to calculate

- **Approximate Bayesian Computation**: Sampling-based Bayesian approach to infer approximate posterior distribution $\pi(\Theta|\rho(Y^\star, Y) \leq \epsilon)$ using simulated data $Y^\star$, a distance function ρ, and tolerance ϵ
  - Approximate when $\epsilon > 0$ and equivalent to simulating from the prior when $\epsilon \to \infty$

ABC rejection method

1. Generate G and Θ from $\pi(G)$ and $\pi(\Theta|G)$, respectively
2. Generate one-step-ahead predictors $\mathbf{y}_t^\star$ from the VAR(1) model, given $\mathbf{y}_{t-1}$ and $\Theta^\star$.
3. Calculate the distance $\rho(Y^\star, Y)$ between Y and $Y^\star$.
4. Accept $(\Theta^\star, G^\star)$ if $\rho \leq \epsilon$.

Very inefficient $\Rightarrow$ Only 5 proposed networks $(\Theta^\star, G^\star)$ are accepted out of a total of 1×10^7 proposals!

ABC rejection method

1. Generate G and Θ from $\pi(G)$ and $\pi(\Theta|G)$, respectively
   $\Rightarrow$ Sequential methods, **Markov chain Monte Carlo**
2. Generate one-step-ahead predictors $\mathbf{y}_t^\star$ from the VAR(1) model, given $\mathbf{y}_{t-1}$ and $\Theta^\star$.

3. Calculate the distance $\rho(Y^\star, Y)$ between Y and $Y^\star$.
   $\Rightarrow$ **Distance criterion**, summary statistics
4. Accept $(\Theta^\star, G^\star)$ if $\rho \leq \epsilon$.
   $\Rightarrow$ Post-sampling regression, nonparametric estimation

Very inefficient $\Rightarrow$ Only 5 proposed networks $(\Theta^\star, G^\star)$ are accepted out of a total of 1×10^7 proposals!

ABC-MCMC (Marjoram et al., 2003)

- ABC-Markov chain Monte Carlo (MCMC): Construct a Markov chain (e.g., using Metropolis-Hastings algorithm) with approximate posterior distribution $\pi(\Theta|\rho(Y^\star, Y) \leq \epsilon)$ as equilibrium distribution
- Given previous $\{\Theta^i, G^i\}$, a proposal $\{\Theta^\star, G^\star\}$ is accepted at iteration $(i+1)$ with probability

$$\alpha = \min\left\{1, \frac{\pi(\Theta^\star, G^\star)q(\Theta^i, G^i|\Theta^\star, G^\star)}{\pi(\Theta^i, G^i)q(\Theta^\star, G^\star|\Theta^i, G^i)}\mathbf{1}(\rho(Y^\star, Y) < \epsilon)\right\}$$

where $q(\cdot|\cdot)$ is the proposal distribution and $\pi(\Theta, G) = \pi(\Theta|G)\pi(G)$

Adapting ABC-MCMC to Networks: ABC-Net

Adaptations must be made to the ABC-MCMC method of Marjoram et al. (2003) for the context of gene regulatory networks:

1. Computationally efficient way to simulate expression data $Y^\star$ from a known regulatory network $\{\Theta^\star, G^\star\}$
2. Appropriate distance function ρ and tolerance ϵ to compare simulated ($Y^\star$) and observed (Y) data
3. Proposal distributions for network structure and parameters

1. Simulating $Y^\star$ for Network $\{\Theta^\star, G^\star\}$

Generally, we simulate gene expression at time t as a function of the gene expression at the previous time point:

$$\mathbf{y}_t^\star = f_t(\mathbf{y}_{t-1}, \Theta^\star)$$

In practice, for continuous data (e.g., microarrays):

- Set $\mathbf{y}_1^\star = \mathbf{y}_1$.
- Generate one-step-ahead predictors based on VAR(1) model on gene expression for $t = 2, \ldots, T$:

$$\mathbf{y}_t^\star = \Theta^\star \mathbf{y}_{t-1}$$

- Note: this is a deterministic simulation procedure...

2. Distance Function and Tolerance

Distance functions (ρ):

- Canberra: $\rho(Y^\star, Y) = \sum_{t=1}^{T}\sum_{i=1}^{P} \frac{|y_{it}^\star - y_{it}|}{|y_{it}^\star + y_{it}|}$
- Euclidean: $\rho(Y^\star, Y) = \sqrt{\sum_{t=1}^{T}\sum_{i=1}^{P}(y_{it}^\star - y_{it})^2}$
- Manhattan: $\rho(Y^\star, Y) = \sum_{t=1}^{T}\sum_{i=1}^{P} |y_{it}^\star - y_{it}|$
- Multivariate Time Series (MVT):

$$\rho(Y^\star, Y) = \frac{1}{T}\sum_{t=1}^{T} [(\mathbf{y}_t - \mathbf{y}_t^\star) - (\hat{\mathbf{y}}_t - \hat{\mathbf{y}}_t^\star)]' \, \hat{\Sigma}^{-1} \, [(\mathbf{y}_t - \mathbf{y}_t^\star) - (\hat{\mathbf{y}}_t - \hat{\mathbf{y}}_t^\star)]$$

2. Distance Function and Tolerance

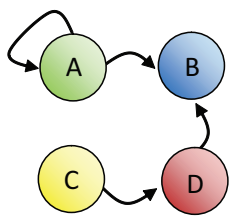
Tolerance (ϵ):

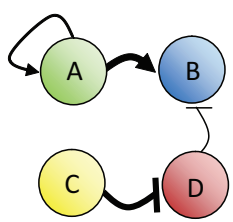
- "Cooling" procedure: decreasing sequence of thresholds, until minimum pre-set threshold ϵ is reached
- $\epsilon = 1\%$ quantile of distances $\rho(Y^\star, Y)$ estimated from 5000 random networks

3. Network Proposals

- With networks, we must propose both a new structure and a new set of parameters
- Recall that we use two representations of a given network: the adjacency matrix G and the parameter matrix Θ

$$G = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} \end{array} \qquad \Theta = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -2 \\ 0 & -1 & 0 & 0 \end{bmatrix} \end{array}$$





- Joint distribution of G and Θ may be seen as a completion to the marginal density of Θ

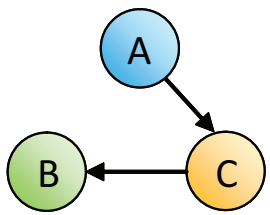
Two-Step Proposal Distribution

- Two-step proposal distribution: $q(G^\star|G^i)$ and $q(\Theta^\star|\Theta^i, G^\star)$:

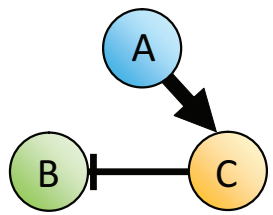
Two-step proposal distribution

Adjacency matrix

$$G^i = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \end{array}$$



Parameter matrix



$$\Theta^i = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -2 \\ 3 & 0 & 0 \end{pmatrix} \end{array}$$

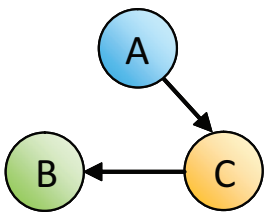
Two-Step Proposal Distribution

- Two-step proposal distribution: $q(G^\star|G^i)$ and $q(\Theta^\star|\Theta^i, G^\star)$:

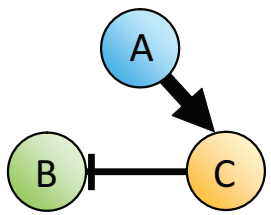
Two-step proposal distribution

Adjacency matrix

$$G^i = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \end{array}$$



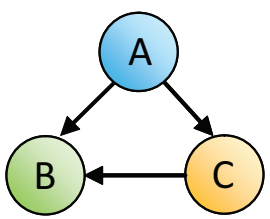
Parameter matrix



$$\Theta^i = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -2 \\ 3 & 0 & 0 \end{pmatrix} \end{array}$$

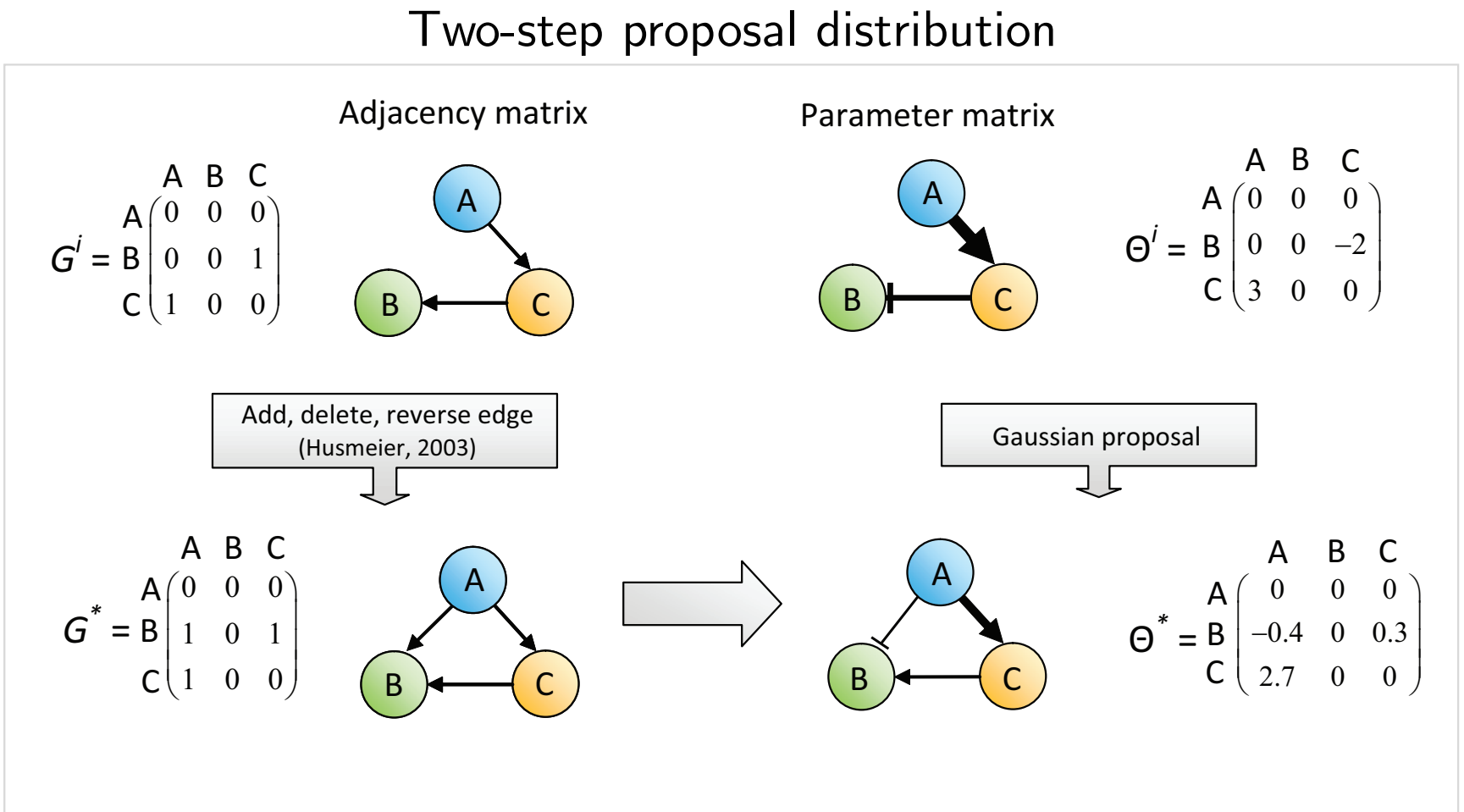
Add, delete, reverse edge
(Husmeier, 2003)

$$G^* = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \end{array}$$



Two-Step Proposal Distribution

- Two-step proposal distribution: $q(G^\star|G^i)$ and $q(\Theta^\star|\Theta^i, G^\star)$:



ABC-MCMC Network Method

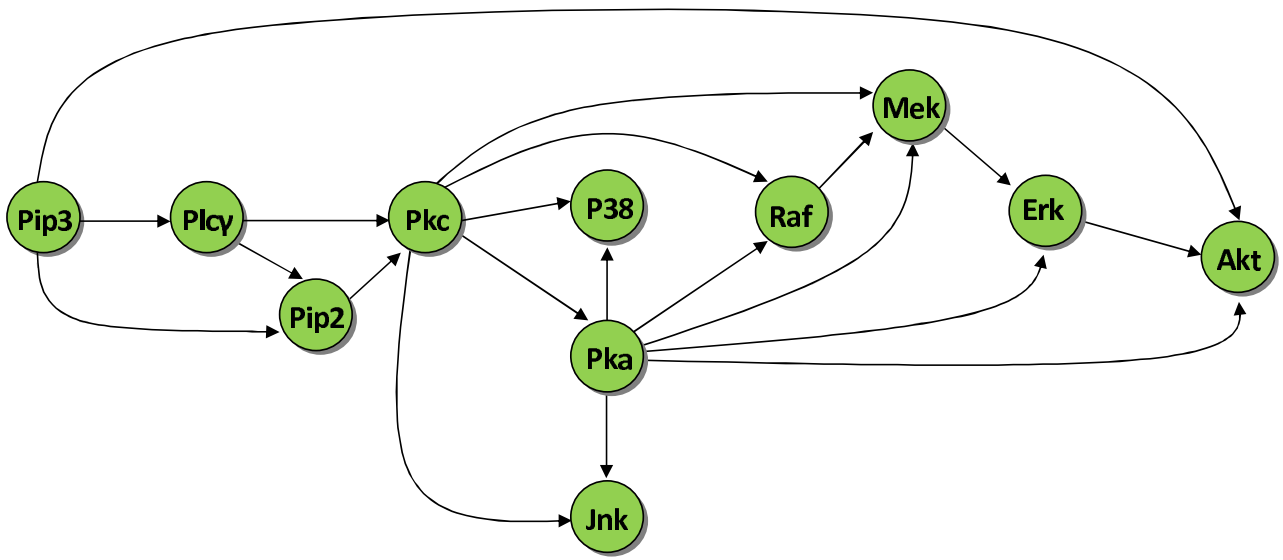
ABC-Net Algorithm:

0. Initialize Θ^i, G^i, $i = 0$.
1. (a) Propose $G^\star$ according to $q(G|G^i)$.
   (b) Propose $\Theta^\star$ according to $q(\Theta|\Theta^i, G^\star)$.
2. Simulate $Y^\star$ from $f(\cdot|\Theta^\star, G^\star)$.
3. Set $\{G^{i+1}, \Theta^{i+1}\} = \{G^\star, \Theta^\star\}$ with probability
   $\alpha = \min\{1, \frac{\pi(G^\star)\pi(\Theta^\star|G^\star)q(G^i|G^\star)q(\Theta^i|\Theta^\star)}{\pi(G^i)\pi(\Theta^i|G^i)q(G^\star|G^i)q(\Theta^\star|\Theta^i)}\mathbf{1}[\rho(\mathbf{y}^\star, \mathbf{y}) \leq \epsilon]\}$
   and $\{G^{i+1}, \Theta^{i+1}\} = \{G^i, \Theta^i\}$ with probability $1 - \alpha$.
4. Set $i = i + 1$. If $i < N$ (a pre-set number of iterations), return to 1.

- Output: dependent samples from the stationary distribution of the chain, $f(\Theta, G|\rho(Y^\star, Y) \leq \epsilon)$
- Burn-in period, number of iterations, chain thinning, ... ▸ Details

Simulations: Raf Signalling Protein Pathway

- Simulations based on currently accepted gold-standard Raf signalling pathway (Sachs et al., 2005) in human immune system cells for 11 genes (20 total edges)



- Simulate $T = 20$ time points, $R = 1$ replicate using VAR model
- Run ABC-Net algorithm for 10 independent chains of length 1×10^6 with thinning interval of 50
- Use Gelman-Rubin statistic to assess convergence across chains

ABC-Net Simulations

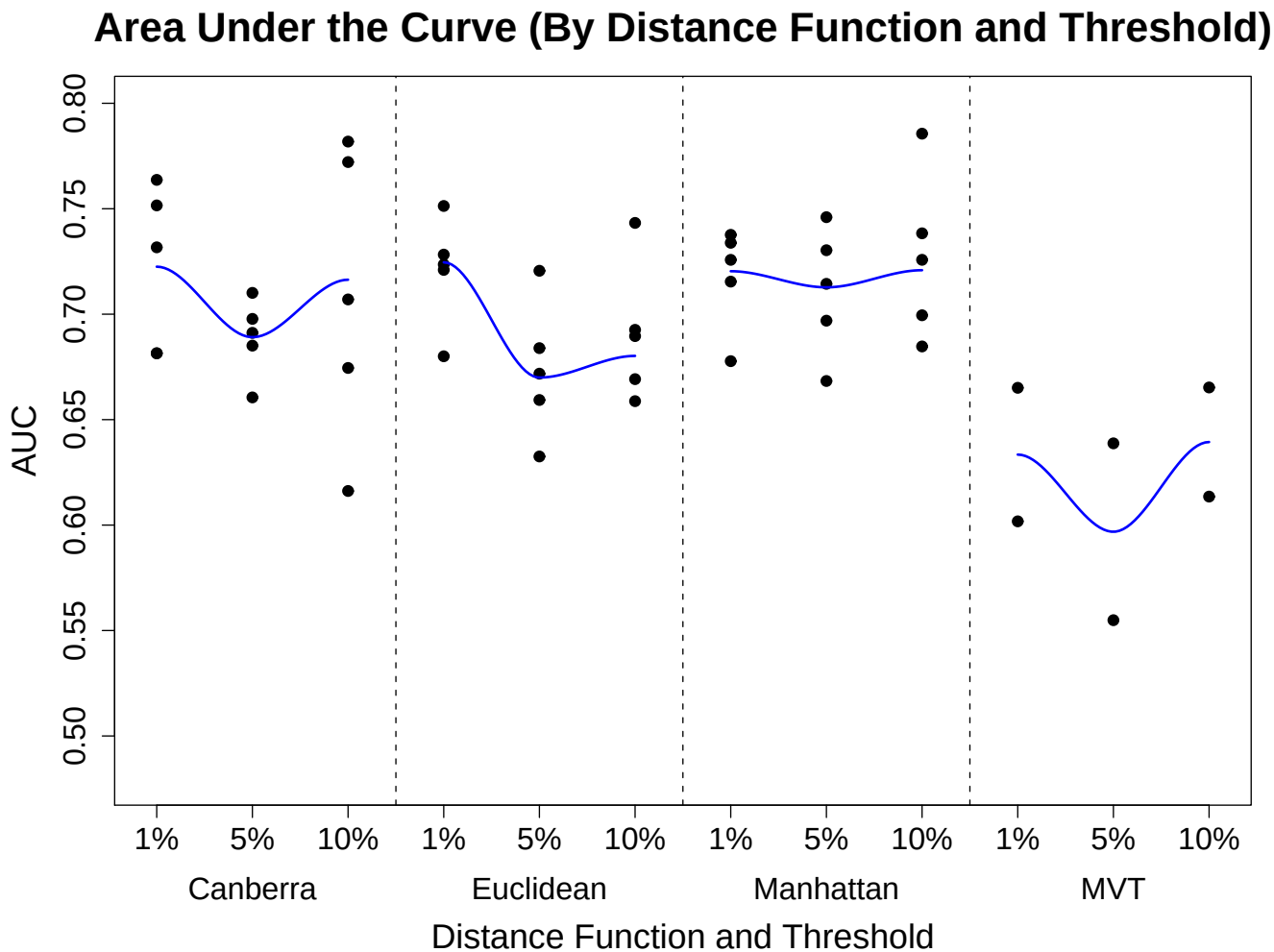
1. Choice of distance function ρ and tolerance ϵ

2. Suitability of VAR simulator for data generated with alternative models (nonlinear models, second-order models, and ordinary differential equations)

3. Qualitative assessment of edge "flexibility"

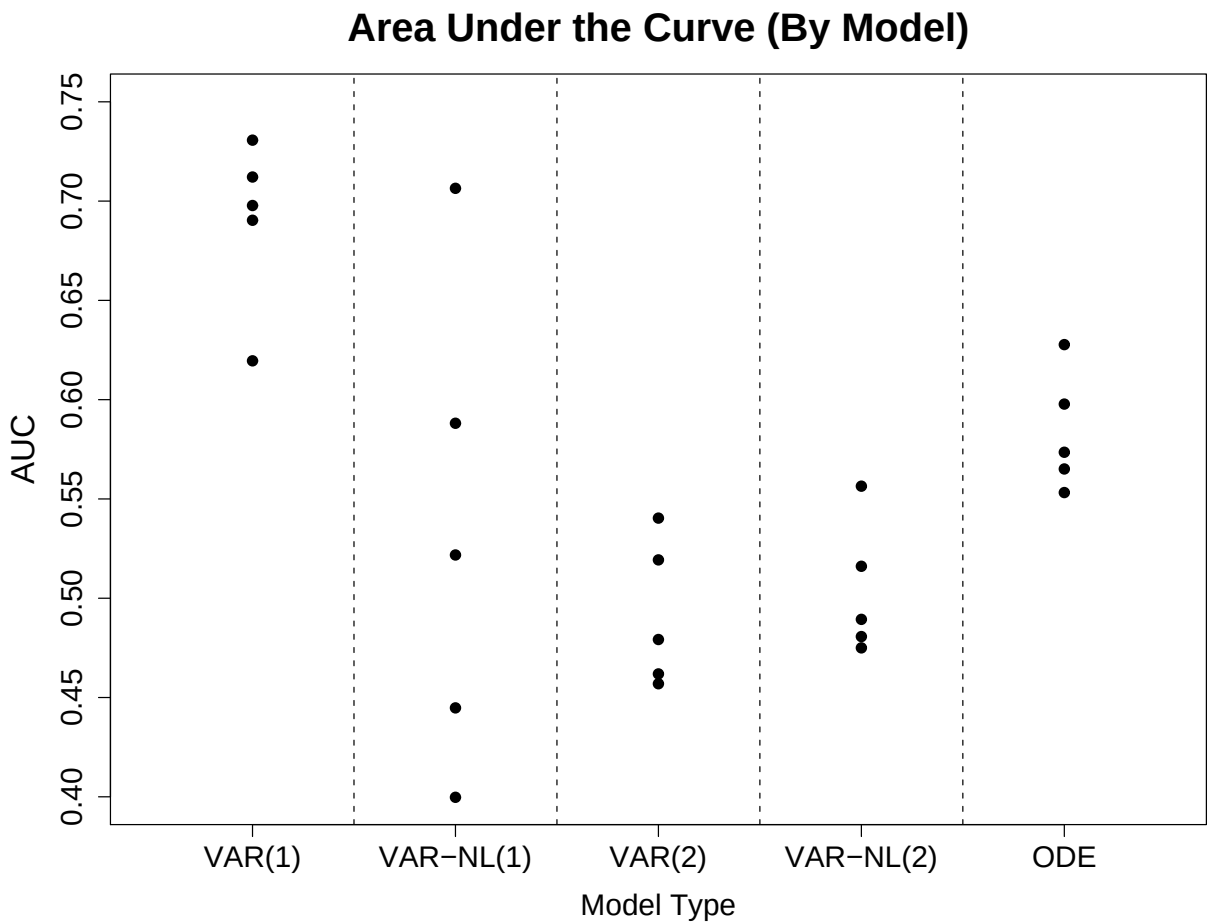
Simulations I: Choice of ρ and ϵ

- Set ϵ to be the 1%, 5%, or 10% quantile of distances ρ from 5000 random networks

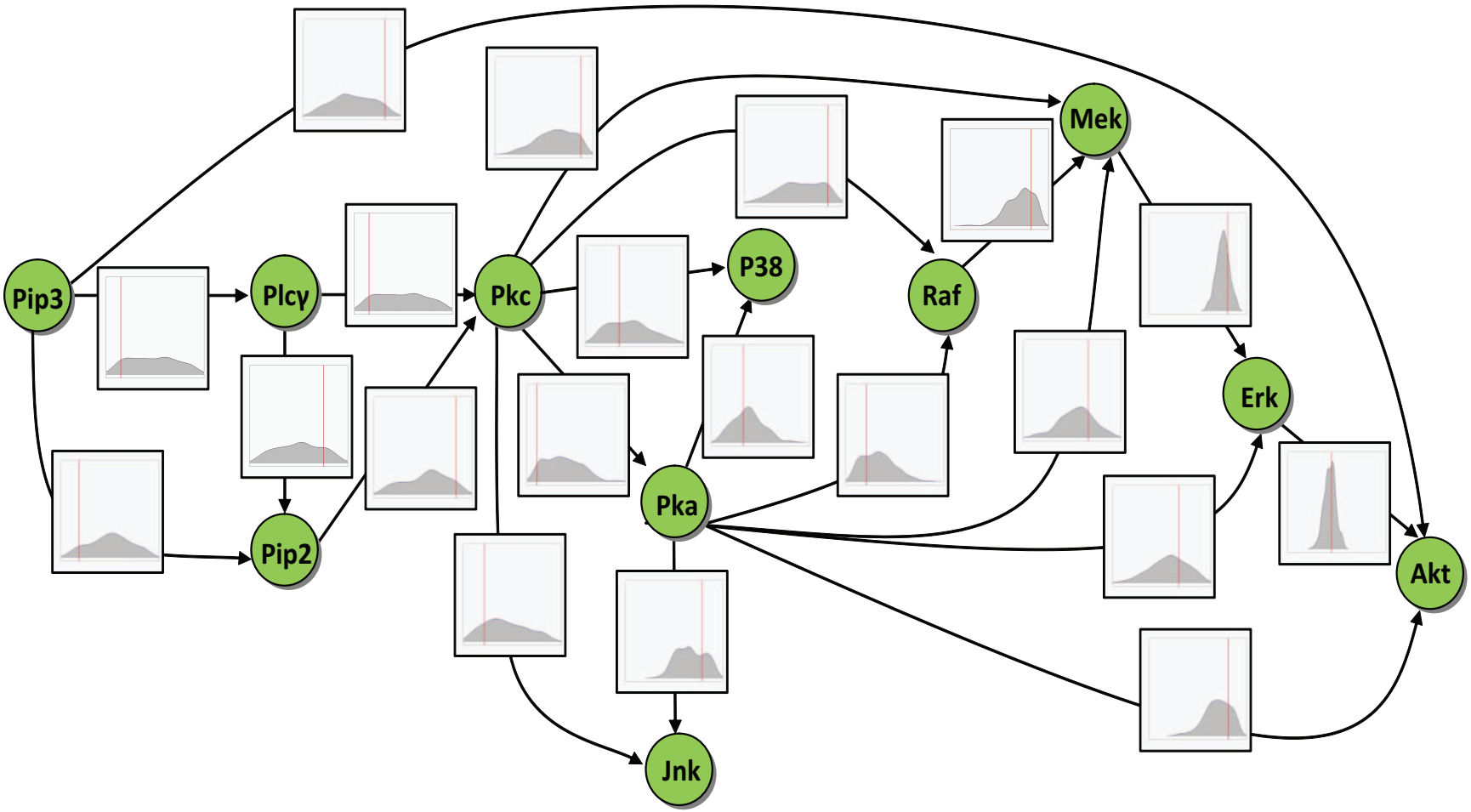


Simulations II: Suitability of VAR Simulator

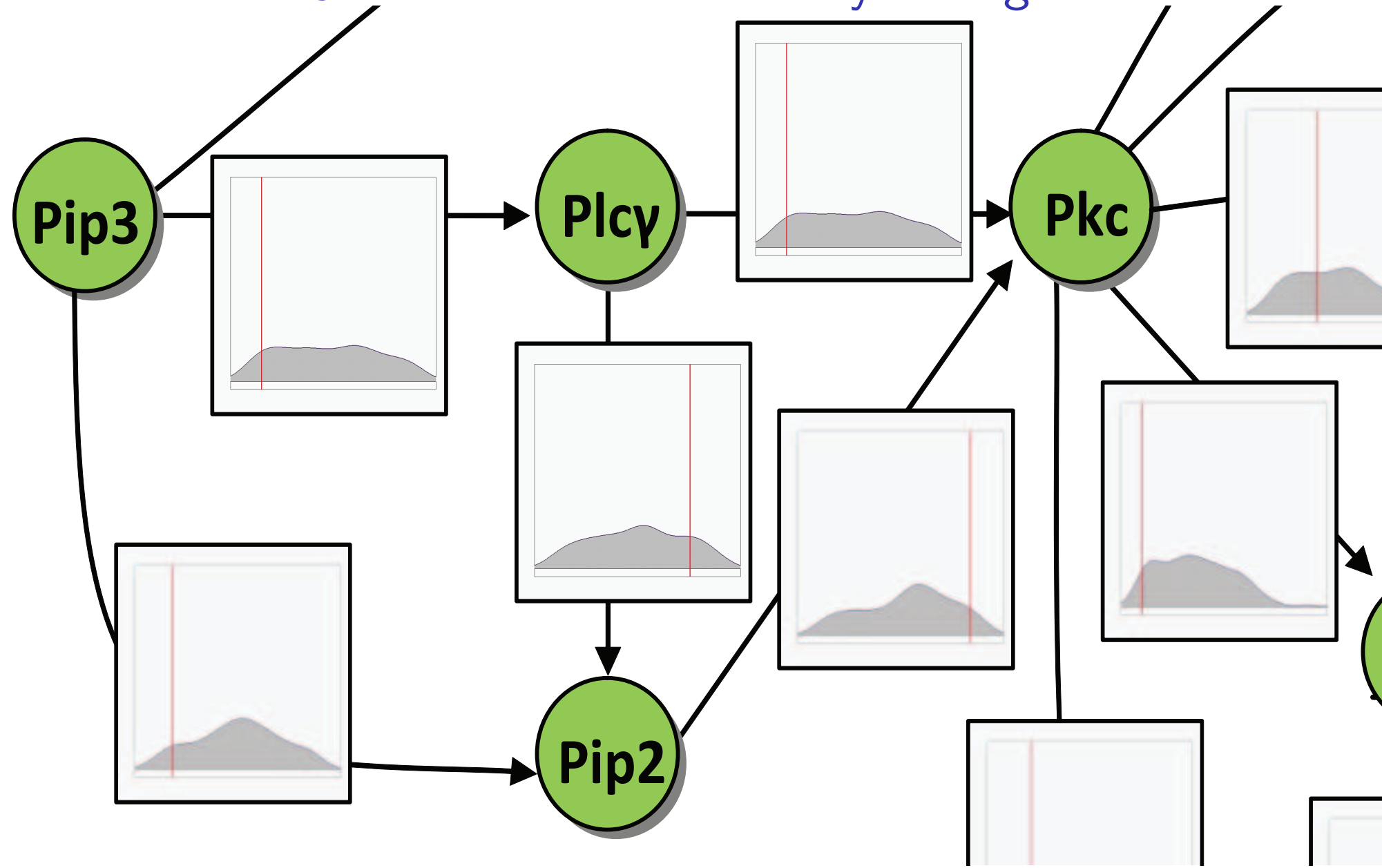
- Alternative models: first-order nonlinear VAR (VAR-NL(1)), second-order VAR (VAR(2)), second-order nonlinear VAR (VAR-NL(2)), and ordinary differential equation (ODE)



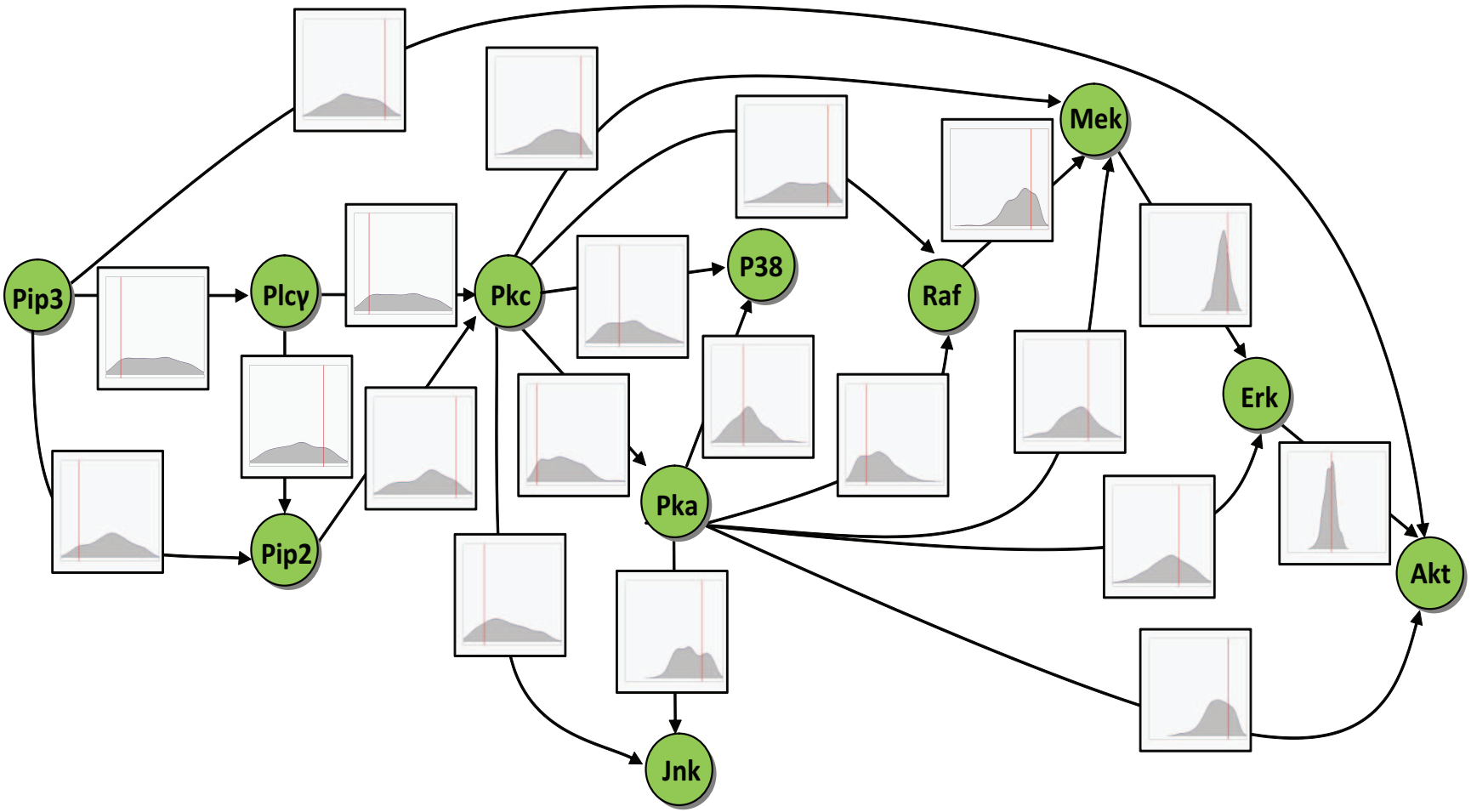
Simulations III: Flexibility of edges



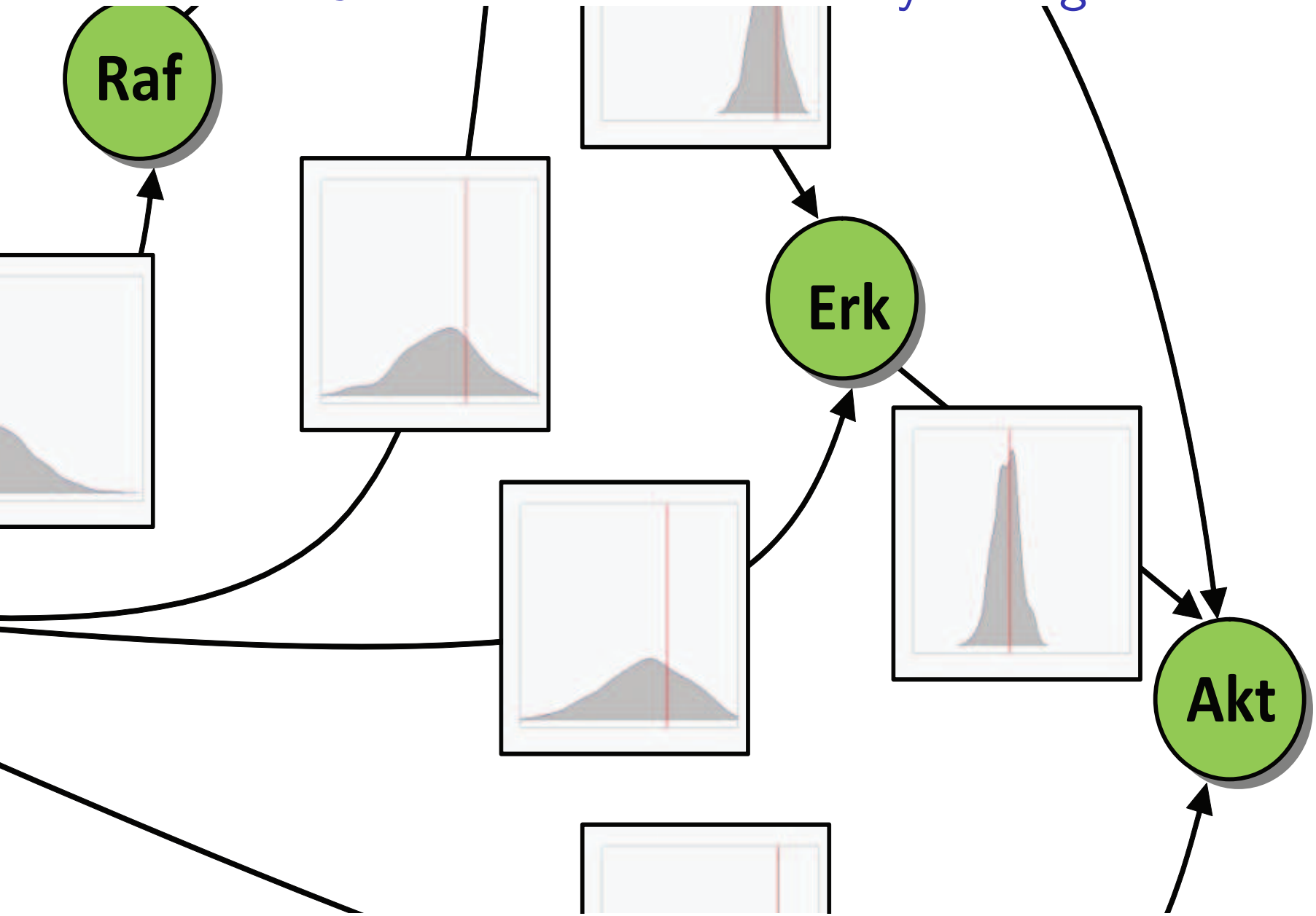
Simulations III: Flexibility of edges



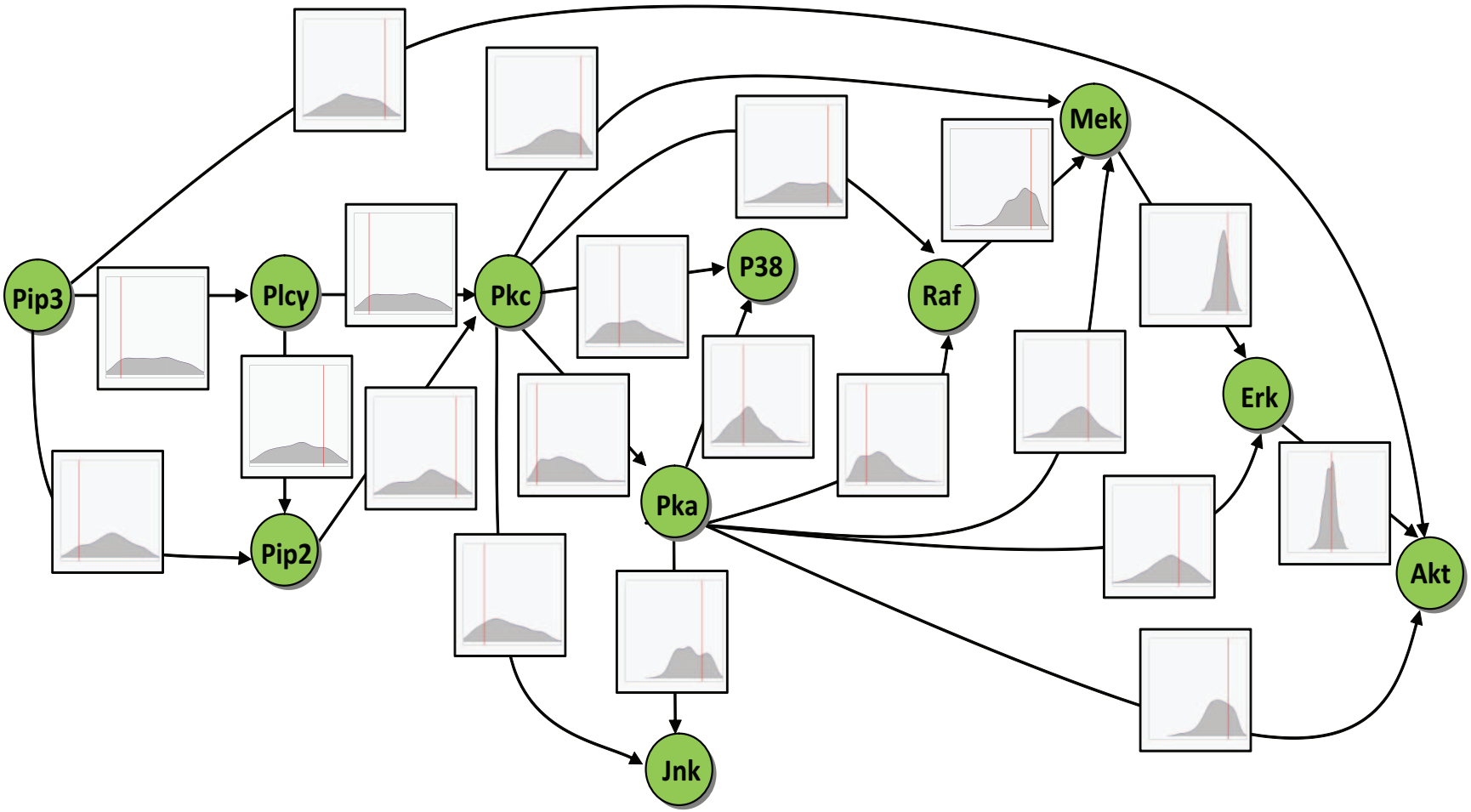
Simulations III: Flexibility of edges

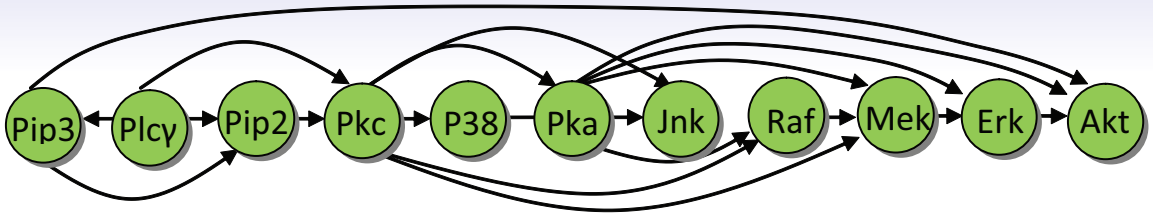


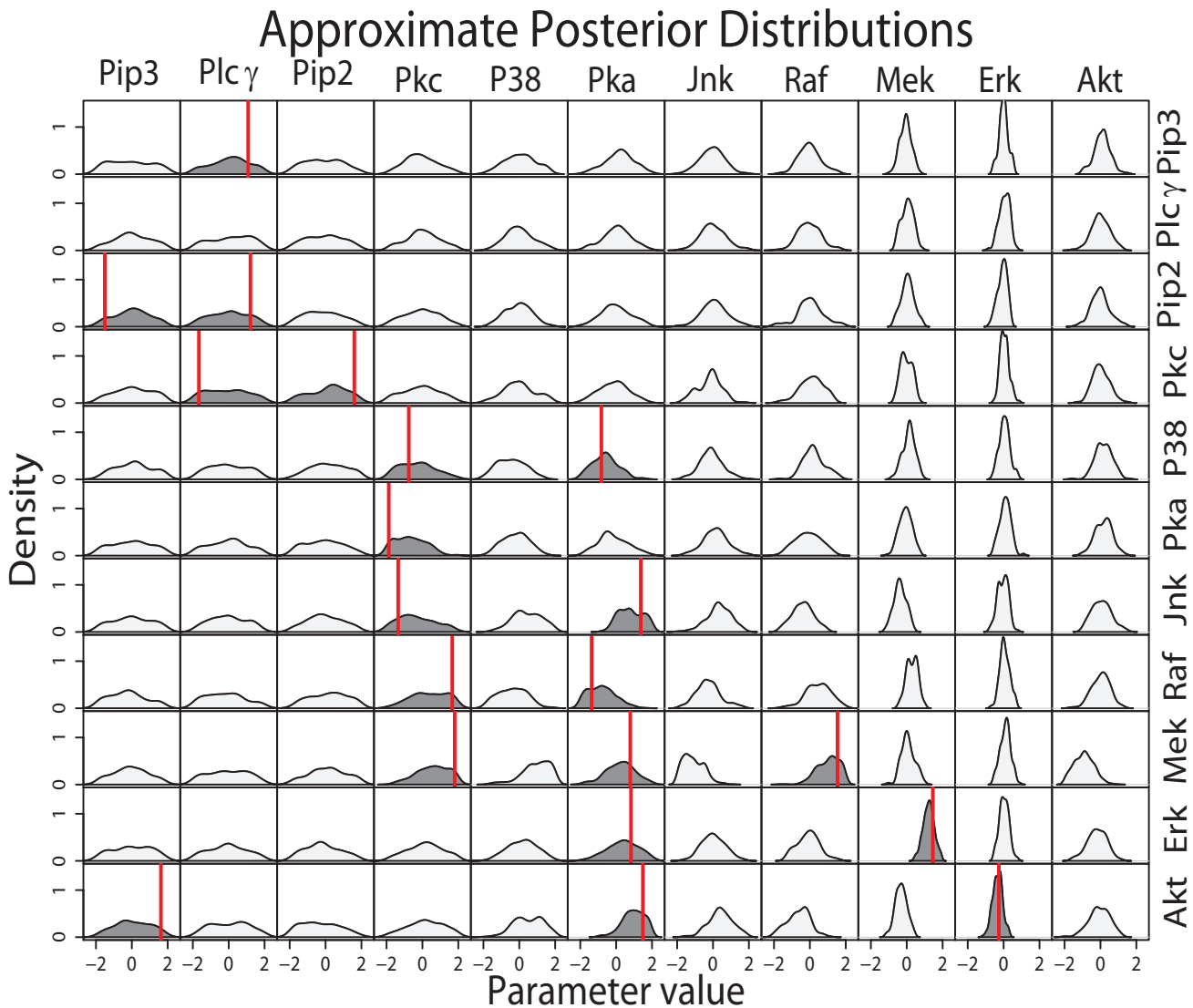
Simulations III: Flexibility of edges



Simulations III: Flexibility of edges




Pip3
Plcγ
Pip2
Pkc
P38
Pka
Jnk
Raf
Mek
Erk
Akt


Approximate Posterior Distributions
Pip3
Plc γ
Pip2
Pkc
P38
Pka
Jnk
Raf
Mek
Erk
Akt
Density
Parameter value

Simulations: Discussion

- Canberra, Euclidean, and Manhattan distances perform similarly in terms of AUC; MVT distance does not perform as well
- Performance of ABC-Net deteriorates for alternative models when a VAR simulator is used
  - Alternative simulators may be used in situations where other models are known to be more appropriate
- "Flexible" and "rigid" edges yield additional information about the dynamics of the network
  - Rigidity and flexibility are closely linked to the network dynamics, robustness, and sensitivity

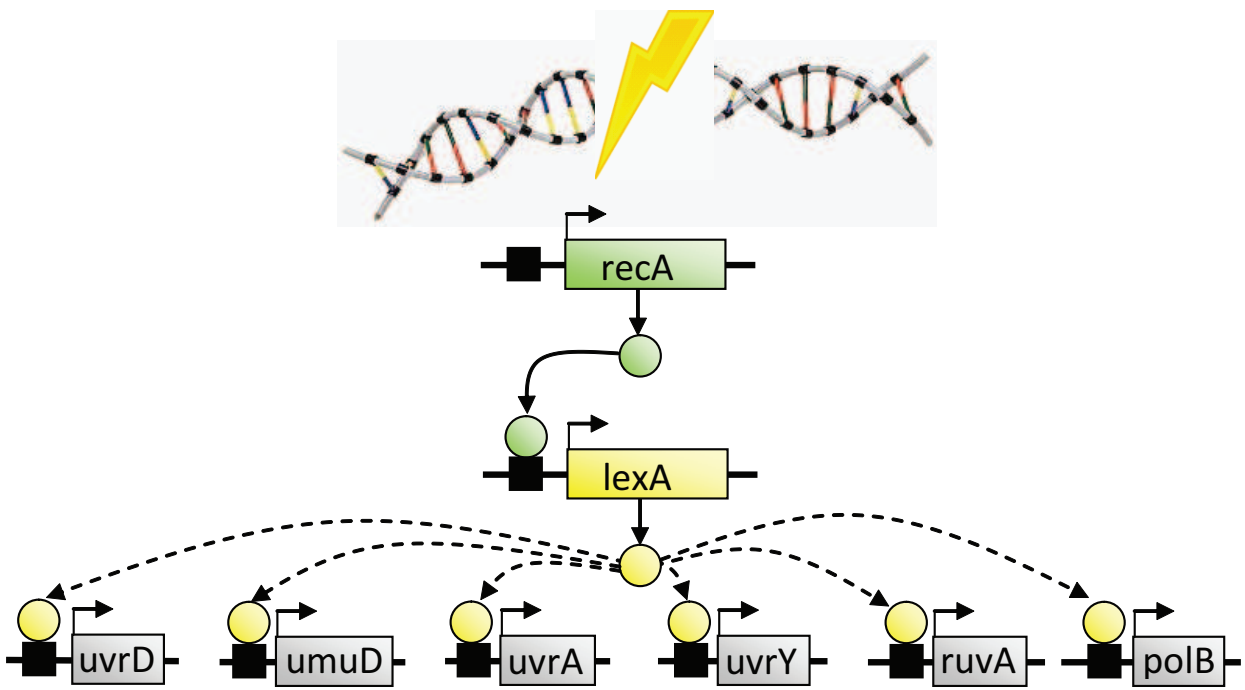
Data Analysis

- Lots of network inference algorithms exist; what additional information can ABC-Net provide?
- Similar methods may yield very different results – **why?**
  - False positives, complicated network structures, small number of time points, ...

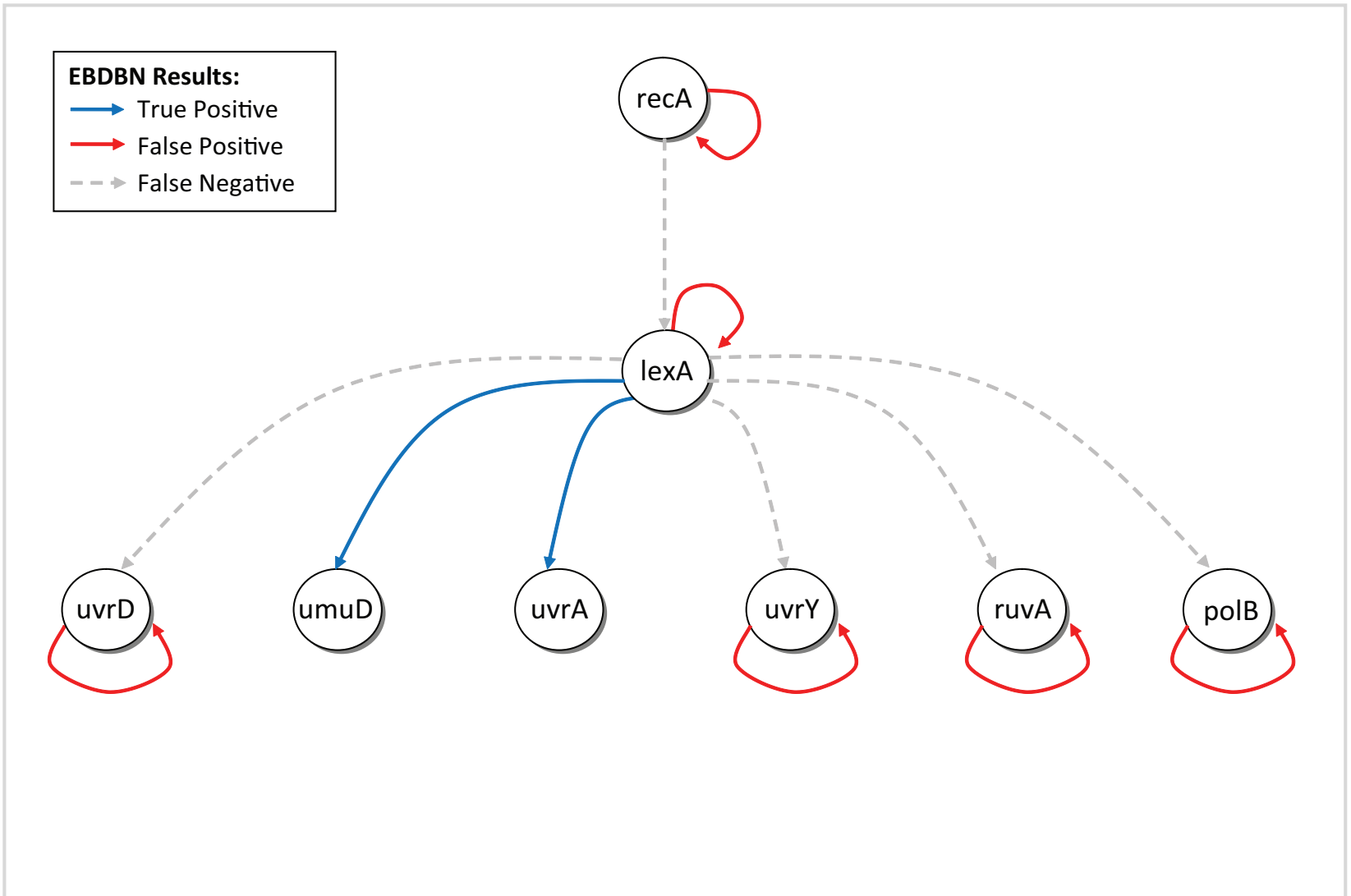
- Real data analysis: S.O.S. DNA repair system in *Escherichia coli*
  - Empirical Bayes Dynamic Bayesian Network (EBDBN) method (Rau et al. (2010)): empirical Bayesian estimation of parameters in a linear state-space model
  - ABC-Net

Data Analysis: S.O.S. DNA Repair System in *E. coli*

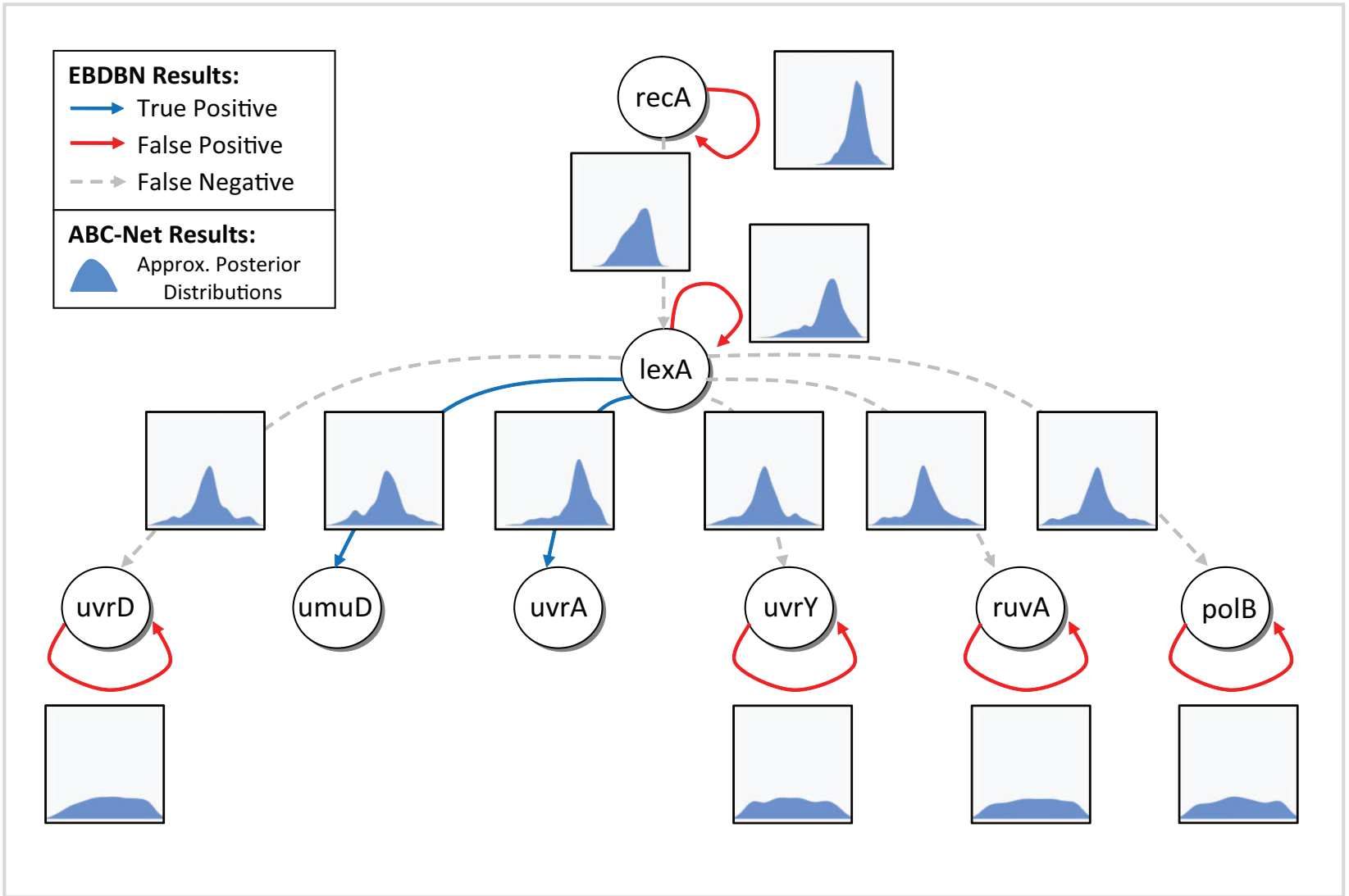
- S.O.S. DNA repair system of *Escherichia coli* (Ronen et al., 2002)
- 8 genes, with lexA as a master regulator that inhibits S.O.S. genes under normal conditions but activates them when DNA damage is sensed by recA ("single-input" module architecture)
- 50 time points, 1 replicate
- Maximum fan-in for ABC-Net method constrained to 2



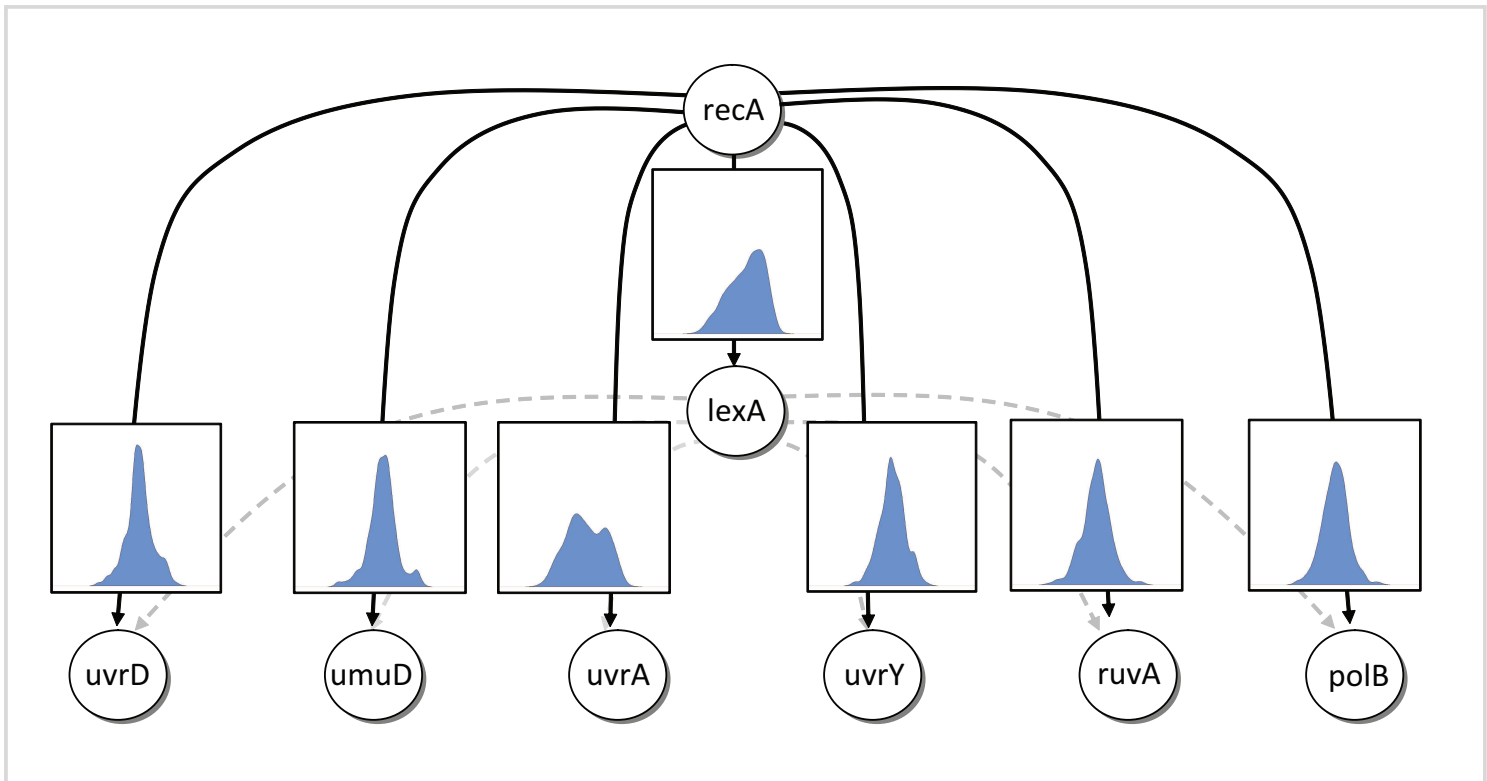
Results: S.O.S. DNA Repair System



Results: S.O.S. DNA Repair System



Results: S.O.S. DNA Repair System



Discussion: S.O.S. DNA Repair System

Recall my original question: **Is it possible to determine whether parts of a given network are identifiable, given the available data?**

- "Rigid" and "flexible" edges identified by the ABC-Net algorithm are a first step to understanding what can be inferred from the given data

- S.O.S. DNA repair is a simple, yet sophisticated network $\Rightarrow$ network is reacting to conditions within the cell
- In S.O.S. system, lexA decreases very rapidly, so S.O.S. genes turn on at about the same time
  - Time-delay models (e.g., autoregressive models) show stronger link between recA and S.O.S. genes

Summary

- Inferring gene regulatory networks is intrinsically difficult: complex network topology, small number of replicates and time points, noise in expression measurements
- Approximate Bayesian Computation methods can reveal information about the dynamics of biological systems from time-series gene expression data
- ABC-MCMC Network (ABC-Net) approach uses a simulation-based Bayesian method with few distributional assumptions to infer approximate posterior distributions in small networks
- Results seem to suggest that given the available data, some gene-to-gene interactions are easier to infer than others...

Future Work

- Further examine components of ABC-Net method:
  - More sophisticated data simulators and techniques to identify optimal simulators for real data
  - Alternative and efficient network structure proposal schemes
  - Objective criterion to characterize approximate posterior distributions (e.g., introduce hierarchical prior on latent indicator variable G in ABC-Net method, and use local Bayes factor to quantitatively examine evidence of network edges)

- Examine alternative simulators and distance functions for count-based measures of gene expression (e.g., RNA sequencing data)

Acknowledgements

Rebecca W. Doerge (Purdue)
Florence Jaffrézic (INRA-GABI)
Bruce Craig (Purdue)
Jayanta Ghosh (Purdue)
Alan Qi (Purdue)

Jean-Louis Foulley (INRA-GABI)
RWD research group @ Purdue
My Truong
Doug Crabill

PURDUE UNIVERSITY

College of Science - Department of Statistics
Statistical Bioinformatics Center

INRA

References

- Rau, A. *et al.* (2011) Reverse engineering gene networks using approximate Bayesian computation. *Statistics and Computing* (in press).
- Rau, A. *et al.* (2010) An empirical Bayesian method for estimating biological networks from temporal microarray data. *SAGMB* 9:1, Article 9.
- Husmeier, D. (2003) Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics* 19, 2271-2282.
- Marjoram, P. *et al.* (2003) Markov chain Monte Carlo without likelihoods. *PNAS* 100, 15324-15328.
- Ronen, M. *et al.* (2002) Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics.*PNAS* 99, 10555-10560.
- Sachs, K. *et al.* (2005) Causal protein-signalling networks derived from multiparameter single-cell data. *Science* 308(5721), 523-529.

LFN Implementation Details

- Burn-in period
  - Cooling procedure: Temper acceptance with exponential cooling scheme, starting at some initial temperature ϵ_0 and cooling to $\epsilon_{i+1} = \lambda\epsilon_i$ until the minimal temperature $\epsilon_{\text{min}} = \epsilon$ is reached. We use $\lambda = 0.90$ and set $\epsilon_0 = \epsilon\lambda^{-10}$.
  - Use each ϵ_i for 200 iterations, then cool to next value.
  - If ϵ_{min} is reached and the acceptance rate for the chain $\leq 1\%$, the burn-in period is reinitialized.
- Chain length:
  - 10 chains for 1×10^6 iterations each (1×10^7 iterations total)
  - Thinning interval of 50 (2×10^5 remaining iterations)
  - Inference made on samples corresponding to smallest 1% of $\rho(\mathbf{y}^\star, \mathbf{y})$ (2000 iterations)

Back