Inferring time-varying genetic network using informative priors

Sophie Lèbre

LSIIT, Theoretical Bioinformatics, Strasbourg,

LICSB - 1st April 2009
1 Modelling regulatory networks from gene expression time series
   ⇝ Detecting structural changes
   ⇝ Using available biological knowledge

2 Time-varying network model and inference
   ⇝ temporal variation across Drosophila development

   Joint work with G. Lelandais, F. Devaux, M. Stumpf.

3 Including informative priors
   ⇝ edges, degree distribution

4 Effect on Drosophila development network inference
Recovering genes functions?

- Regulatory relationships:
  - up/down regulation
  - retroaction, feedforwards loops...

⇒ Complex dynamic system

- Objective: identifying this organisation in large scale.

S. Lèbre  slebre@unistra.fr  Time-varying network inference using informative priors
Temporal gene expression data

- Microarrays:
  \( \sim \) **simultaneous** expression of **several thousands** of genes.

- Notations: we consider the stochastic process,

  \[ X = \{ X_{i}^{t}; \ \forall i \in \{1, ..., p\}, \forall t \in \{1, ..., n\} \} \]

  where \( X_{i}^{t} \) is the expression of gene \( i \) at time \( t \),
What information extracting from expression profiles?

- Which genes work together?
- At what instant of the observed process?
DBN modelling of biological motifs

- A biological motif

- **Dynamic** Bayesian Networks (DBNs) allow to model biological cycles

(Friedman et al. 1998, Murphy and Mian 1999, OpgenRhein and Strimmer 2007)
DBN modelling

Assumptions

- \((A_1)\) 1\textsuperscript{st} order Markov process
- \((A_2)\) ‘simultaneous independence’ given the past,
  \[
  \forall t > 1, \forall i, j \in \mathbb{N}, \quad X_t^i \perp \perp X_t^j | X_{t-1}.
  \]
- \((A_3)\) time homogeneity

\[
f(X) = \prod_{1 < t \leq n} f(X_t^1 | X_{t-1}^1, X_{t-1}^2) f(X_t^2 | X_t^1) f(X_t^3 | X_t^2)
\]
DBN for a 1\textsuperscript{st} order auto-regressive process: AR(1).

- AR(1) process: \( \forall t \geq 1, \ X_t = AX_{t-1} + B + \varepsilon_t, \ \varepsilon_t \sim \mathcal{N}(0, \Sigma) \)

\[
\begin{bmatrix}
X_t^1 \\
\vdots \\
X_t^i \\
\vdots \\
X_t^p
\end{bmatrix}
= \begin{bmatrix}
a_{11} & \ldots & a_{1j} & \ldots & a_{1p} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
a_{i1} & \ldots & a_{ij} & \ldots & a_{ip} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
a_{p1} & \ldots & a_{pj} & \ldots & a_{pp}
\end{bmatrix}
\begin{bmatrix}
X_{t-1}^1 \\
\vdots \\
X_{t-1}^j \\
\vdots \\
X_{t-1}^p
\end{bmatrix}
+ \begin{bmatrix}
b_t^1 \\
\vdots \\
\vdots \\
\vdots \\
b_t^p
\end{bmatrix}
+ \begin{bmatrix}
\varepsilon_t^1 \\
\vdots \\
\vdots \\
\vdots \\
\varepsilon_t^p
\end{bmatrix}
\]

- Example:

\[
A = \begin{pmatrix}
a_{11} & a_{12} & 0 \\
a_{21} & 0 & 0 \\
0 & a_{32} & 0
\end{pmatrix}
\]
1. Modelling regulatory networks from gene expression time series
   - Detecting structural changes
   - Using available biological knowledge

2. *Time-varying* network model and inference
   - Temporal variation across Drosophila development
   - Joint work with G. Lelandais, F. Devaux, M. Stumpf.

3. Including informative priors
   - Edges, degree distribution

4. Effect on Drosophila development network inference
Time-varying dynamic Bayesian network model

- Introducing **changepoints** allowing for network **structure change**.

\[
X_t^1 = b_{10} + b_{11} X_{t-1}^1 + b_{12} X_{t-1}^2
\]

\[
X_t^1 = b_{10} + b_{11} X_{t-1}^1
\]

\[
X_t^1 = b_{10}
\]
For each gene $i$, $(1 \leq i \leq p)$,

- a **changepoint vector** $\xi^i = (\xi^i_1, ..., \xi^i_{h-1}, \xi^i_h, ..., \xi^i_{k_i}) \subseteq \{2, ..., n\}$
- in each phase $h$, (for all $\xi^i_h \leq t < \xi^i_{h+1}$),
  - a set of $s^i_h$ **parents** $\tau^i_h = \{j_1, ..., j_{s^i_h}\} \subseteq \{1, ..., p\}$
  - and a set of **parameters** $\theta^i_h = ((b^i_{jh})_{j\in\{0,...,q\}}, \sigma^i_h)$,

define the regression model,

\[
X^i_t = b^i_0 + \sum_{j \in \tau^i_h} b^i_{jh} X^j_{t-1} + \varepsilon^i_t, \quad \varepsilon^i_t \sim \mathcal{N}(0, \sigma^i_h).
\]

\[
\Rightarrow \text{Unknown dimension}
\]

S. Lèbre  slebre@unistra.fr

Time-varying network inference using informative priors
Outline of the tvDBN procedure:

\[ \begin{align*}
\text{Changepoint Birth} & \rightarrow rJ-MH \\
\text{Changepoint Death} & \rightarrow rJ-MH \\
\text{Changepoint Shift} & \rightarrow MH \rightarrow \text{Markov Chain Update}
\end{align*} \]

\[ \begin{align*}
\text{Phase Update} & \rightarrow \chi \rightarrow \text{Edge Birth} \rightarrow rJ-MH \\
& \rightarrow \zeta \rightarrow \text{Edge Death} \rightarrow rJ-MH \\
& \rightarrow \rho \rightarrow \text{Regression Step}
\end{align*} \]

\[ \text{Reversible jump MCMC: Green (1995).} \]
\[ \text{Model selection: Andrieu and Doucet (1999).} \]
Changepoints priors

- number of changepoints \( k \sim \mathcal{P}(\lambda) \)
- changepoints position \( \xi|k \sim \text{Uniform} \)

4 moves: Birth \((b_k)\), Death \((d_k)\), Position shift \((\eta_k)\), Regression model update \((\pi_k)\).

\[
b_k + d_k + \eta_k + \pi_k = 1
\]

\[
b_k = c \min \left\{ 1, \frac{\mathbb{P}_k(k+1)}{\mathbb{P}_k(k)} \right\}, \quad d_k = c \min \left\{ 1, \frac{\mathbb{P}_k(k-1)}{\mathbb{P}_k(k)} \right\}, \quad \eta_k = \frac{1}{2}(b_k + d_k).
\]

Birth: \( \xi^+ = \xi \cup \xi^* \)

- proposal ratio: \( \frac{d_{k+1}}{b_k} \frac{q(\xi^*|\xi^+)}{q(\xi^*|\xi)} = \frac{(n-1)p-k}{\lambda} \)
- Jacobian=1
- posterior distribution ratio: \( \frac{\mathbb{P}(k+1,\xi^+,s^+,\tau^+|y)}{\mathbb{P}(k,\xi,s,\tau|y)} \perp \perp b, \sigma \)

\( \Rightarrow \) Acceptance ratio \( R_{k,k+1}(\xi,\xi^+) = \frac{(n-1)p-k}{\lambda} \frac{\mathbb{P}(k+1,\xi^+,s^+,\tau^+|y)}{\mathbb{P}(k,\xi,s,\tau|y)} \).
2 steps **embedded** *reversible jump MCMC*

⇝ Generation of an ergodic Markov chain.

⇝ Reversible Markov chain: detailed balance satisfied.

⇝ Equilibrium distribution converge to the desired post-distribution,

\[ \mathbb{P}(k, \xi, s, \tau, \theta | x). \]

⇝ Bayes Factor (BF) analysis.
Drosophila melanogaster life cycle data

- Data by Arbeitman et al. (2002)
  - gene expression across the whole life cycle of *D. melanogaster*: embryo, larva, pupa and adult.
  - 4028 genes
  - 66 successive time points

- Results
  - 50 genes have strong evidence for being “regulated” during only a fraction of their life (i.e. BF $> 12$)
  - Implying 52 parents/regulators
  - Changepoints tend to cluster at the development transitions (embryo/larva and larva/pupa) + mid-embryo stage.
Changepoints cluster at phase transition and mid-embryo

B

Timepoint

Embryo Larva Pupa Adult

# changepoints with BF $\geq 3$

10 20 30 40 50 60 70

S. Lèbre slebre@unistra.fr

Time-varying network inference using informative priors
S. Lèbre  slebre@unistra.fr

Time-varying network inference using informative priors
<table>
<thead>
<tr>
<th></th>
<th>Developmental Process</th>
<th>Transcription Activity</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inferred target genes</td>
<td>7</td>
<td>24</td>
<td>50</td>
</tr>
<tr>
<td>Inferred parent genes</td>
<td>7</td>
<td>26</td>
<td>52</td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>50</td>
<td></td>
</tr>
</tbody>
</table>

S. Lèbre  slebre@unistra.fr

Time-varying network inference using informative priors
1. Modelling regulatory networks from gene expression time series
   - Detecting structural changes
   - Using available biological knowledge

2. *Time-varying* network model and inference
   - temporal variation across Drosophila development
   Joint work with G. Lelandais, F. Devaux, M. Stumpf.

3. Including informative priors
   - edges, degree distribution

4. Effect on Drosophila development network inference
Usual priors

Network sparsity

\[ s_h^i \sim \mathcal{P}(\lambda) \] where \( \lambda \sim \text{IG}(\alpha, \beta) \).

Incoming edges: Uniformative priors

\[ \tau_h^i|s_h^i \sim \text{Uniform} \]
Including informative priors

- Network connectivity
  - Some genes share common properties:
    - Genes implicated in drosophila nervous system development are expected to work together (or muscle, heart, ...)
    - Transcription factor activates the transcription of target genes
  - cluster genes sharing common properties
    - 1 cluster with uninformative priors
- $L$ clusters with specific priors
  - Number of incoming edges
  - Inter cluster connectivity
Gene cluster priors

- Number of incoming edges for cluster $l$: hyperparameters $(\alpha_l, \beta_l)$

\[ p(s_h^i) \sim \mathcal{P}(\lambda), \quad \lambda \sim \mathcal{IG}(\alpha_l, \beta_l). \]

S. Lèbre  
slubre@unistra.fr

Time-varying network inference using informative priors
Gene cluster priors

- Weight of the putative parent genes for target gene of each cluster $l$

  $\sim \text{inter cluster connectivity weight matrix } W_{[L \times L]}$

where $W[l, m]$ is weight for genes of cluster $m$ to be a parent of genes of cluster $l$
Informative priors

- Possible changepoints priors
  - Some time point position \textit{may} be considered as more probable than others.

\[ p(\xi|k \text{ Changepoints}) = f(\omega) \]

- Remarks
  - Only commonly accepted knowledge is considered.
  - Bayes factor analysis: results significance taking priors into account.
1. Modelling regulatory networks from gene expression time series
   • Detecting structural changes
   • Using available biological knowledge

2. *Time-varying* network model and inference
   • Temporal variation across Drosophila development
   
   Joint work with G. Lelandais, F. Devaux, M. Stumpf.

3. Including informative priors
   • Edges, degree distribution

4. Effect on Drosophila development network inference
14 genes implicated in drosophila development process:

Minifly - Anarchist - Quo Vadis - Neurexin IV - Bicoid - smi21F - zfh1
Glass - Peter Pan - Rbp9 - Quail - Castor - Mip - CG32486

+ Neighbours ~~~ 36 genes

⇒ tvDBN inference with informative priors specific to 3 clusters:
  - Cluster 1: Developmental process
  - Cluster 2: Transcription factor activity
  - Cluster 3: Unknown
### Inferred target/parent gene vs Gene Ontology

#### Uninformative priors:

<table>
<thead>
<tr>
<th>Developmental process</th>
<th>Transcription Activity</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target genes</strong></td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td><strong>Parent genes</strong></td>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>

#### Informative priors:

<table>
<thead>
<tr>
<th>Developmental process</th>
<th>Transcription Activity</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target genes</strong></td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td><strong>Parent genes</strong></td>
<td>3</td>
<td>8</td>
</tr>
</tbody>
</table>

S. Lèbre  slebre@unistra.fr  
Time-varying network inference using informative priors
CP distribution (with informative priors on gene clusters)

Timepoints

Number of changepoints

EMBRYO

LARVA

PUPA

ADULT

S. Lèbre slebre@unistra.fr

Time-varying network inference using informative priors
Time-varying network inference using informative priors
S. Lèbre  slebre@unistra.fr

Time-varying network inference using informative priors
D. melanogaster: Pupa

CG13601, CG3823, CG3124, CG32486, CG14718, CG1578, CG4407, CG14617, CG17450, CG4199, Fizzy, Quail, Mip, CG2135, Rbp9, CG14303, zfh1, Klp67A, Anarchist, CG1578, Minifly, PeterPan, CG4407, Ipod, CG14718, RfaBp, CG4407, CG14718, CG14617, Klp67A, Bicoid, smi21F

S. Lèbre slebre@unistra.fr
Time-varying network inference using informative priors
D. melanogaster: Adult

Time-varying network inference using informative priors
Results

- A new approach allowing to infer time-dependent networks
- Simultaneous inference of the changepoints position & the models within phases
- Inclusion of some available biological knowledge

Future Work

- Larger scale
- Cluster: muscle / nervous system / heart /...
Results

- A new approach allowing to infer time-dependent networks
- **Simultaneous** inference of the changepoints position & the models within phases
- Inclusion of some available biological knowledge

Future Work

- Larger scale
- Cluster: muscle / nervous system / heart /...

Thank you for your attention!
CP distribution (with uninformative priors)

Timepoints

Number of changepoints

EMBRYO

LARVA

PUPA

ADULT

S. Lèbre  slebre@unistra.fr

Time-varying network inference using informative priors
D. melanogaster: Early Embryo

Time-varying network inference using informative priors
D. melanogaster: End Embryo

Time-varying network inference using informative priors

S. Lêbre  slebre@unistra.fr
D. melanogaster: Larva

Time-varying network inference using informative priors

S. Lèbre  slebre@unistra.fr
D. melanogaster: Pupa

Time-varying network inference using informative priors

S. Lèbre  slebre@unistra.fr
D. melanogaster: Adult

Time-varying network inference using informative priors
The overall parameter space $\Theta$ writes as a finite union of subspaces

$$
\Theta = \bigcup_{k=0}^{\bar{k}} E_k \times \Theta_k
$$

where,

$$
E_k = \left\{ \xi = \{\xi^i\}_{1 \leq i \leq p}; \ \forall i \in P, \xi^i \subseteq \{2, \ldots, n - 1\}, |\xi^i| = k^i, \sum_{i=1}^{p} k^i = k \right\}
$$

$$
\Theta_k = \prod_{i=1}^{p} k^i \prod_{i=1}^{\bar{k}} \left\{ \bigcup_{s^i_{ih}=0}^{s^i_{ih}} \{s^i_h\} \times B_{s^i_{ih}} \right\},
$$

with $B_0 = \mathbb{R} \times \mathbb{R}^+$, $B_{s^i_{ih}} = \mathbb{R}^{s^i_{ih}+1} \times \mathcal{P}_{s^i_{ih}}(Q) \times \mathbb{R}^+$ for $k \geq 1$ and $\mathcal{P}_{s^i_{ih}}(Q)$ contains all subsets of $Q$ of dimension $s^i_h$. 