A Bayesian approach to reconstructing genetic regulatory networks with hidden factors [Beal et al., 2004]

Jean A. Hausser[1]

[1]Institut für Statistik
Ludwig Maximilian Universität München

Friday, July 7th 2006
Modeling, Simulation and Inference
of Complex Biological Systems

Outline

Introduction
- Biological Background
- Experiment
- Properties of the experimental data

Methods
- The linear dynamical system model
- The Bayesian Approach to SSMs
- Variational-Bayes model fitting

Results

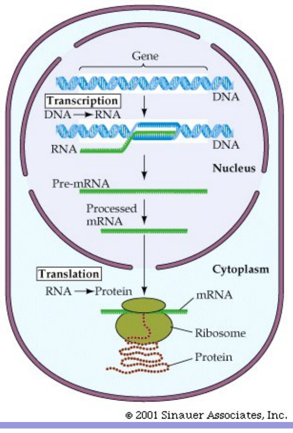
Outline

Introduction
- Biological Background
- Experiment
- Properties of the experimental data

Methods
- The linear dynamical system model
- The Bayesian Approach to SSMs
- Variational-Bayes model fitting

Results

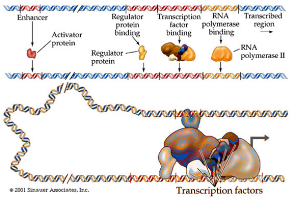
The *Central Dogma* of molecular biology



© 2001 Sinauer Associates, Inc.

- DNA : where the genetic information lies
- RNA : an intermediate product of gene expression
- Proteins : active molecules of life

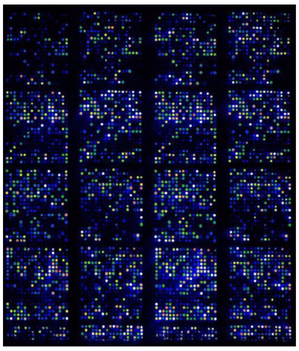
Gene regulation : the big picture



© 2001 Sinauer Associates, Inc.

- coding region : where the gene is
- promoter region : determines gene activation conditions
- RNA polymerase : DNA to RNA transcription enzyme
- Transcription factors complex, repressors
- Regulators, enhancers, and the dynamics

Microarray experiment



For a given gene, what do we call *expression level* ?

- non-proportional
- noise
- reproducibility

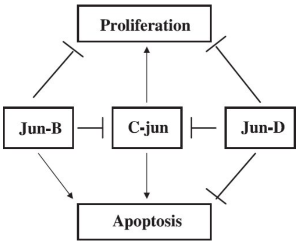
Outline

Introduction
- Biological Background
- Experiment
- Properties of the experimental data

Methods
- The linear dynamical system model
- The Bayesian Approach to SSMs
- Variational-Bayes model fitting

Results

Goal of the study



Proliferation
Jun-B
C-jun
Jun-D
Apoptosis

- genetic regulation cascade occuring during T-cell activation
- what genes are activated / shut down ?

Can statistical modeling help us to better understand T-cell activation ?

Outline

Introduction

Biological Background

Experiment

Properties of the experimental data

Methods

The linear dynamical system model

The Bayesian Approach to SSMs

Variational-Bayes model fitting

Results

Microarray experiment description

- T-cell activation under PMA and iomicin
- time-series : 10 time points
- 58 genes being monitored

What does our model has to take into account ?

- multivariate data (experimental design)
- time-series (experimental design)
- noisy measurements (microarrays)
- missing data (biology is complex)
- causal inference (goal)

Outline

Introduction

Biological Background

Experiment

Properties of the experimental data

Methods

The linear dynamical system model

The Bayesian Approach to SSMs

Variational-Bayes model fitting

Results

Linear State-Space models

aka: Linear Dynamical Systems, Kalman filter models

Assumptions

- hidden state variables
- noisy continuous measurements
- Markovian dynamics

Variables and topology

- observed data : $(\mathbf{y}_1, \dots, \mathbf{y}_T), \mathbf{y}_i \in \mathbb{R}^p$
- $\mathbf{y}_t$ generated from hidden $\mathbf{x}_t$, with $\mathbf{x}_t \in \mathbb{R}^k$
- $\mathbf{x}$ follows 1^{st}-order Markov process

Therefore :

$$p(\mathbf{x}_{1:T}, \mathbf{y}_{1:T}) = p(\mathbf{x}_1)p(\mathbf{y}_1|\mathbf{x}_1)\prod_{t=2}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{y}_t|\mathbf{x}_t)$$

Variables and topology (2)

Assuming

- linear dynamics of hidden variables $p(\mathbf{x}_t|\mathbf{x}_{t-1})$,
- linear dynamics of output function $p(\mathbf{y}_t|\mathbf{x}_t)$,
- model stationarity,
- and state evolution and observation have Gaussian noise

we obtain the linear-Gaussian state-space model (SSM) :

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + \mathbf{w}_t, \qquad \mathbf{w}_t \sim N(\mathbf{0}, Q) \tag{1}$$

$$\mathbf{y}_t = C\mathbf{x}_t + \mathbf{v}_t, \qquad \mathbf{v}_t \sim N(\mathbf{0}, R) \tag{2}$$

where A is the *kxk* state dynamics matrix (HMM: transition) and C the *pxk* observation matrix (HMM: emission)

Variables and topology (3)

Straighforward extension of our model

Let $\mathbf{u}_{1:T}$ be a time-serie of d-dimensional driving inputs. Then,

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + B\mathbf{u}_t + \mathbf{w}_t, \qquad \mathbf{w}_t \sim N(\mathbf{0}, Q) \tag{3}$$

$$\mathbf{y}_t = C\mathbf{x}_t + D\mathbf{u}_t + \mathbf{v}_t, \qquad \mathbf{v}_t \sim N(\mathbf{0}, R) \tag{4}$$

where B is the *dxk* input-to-state matrix and D the *dxp* input-to-observation matrix.
What happens if :

- we provide driving input a constant bias ?
- we want to do control ?
- we define $\mathbf{u}_t := \mathbf{y}_{t-1}$?

Variables and topology (4)

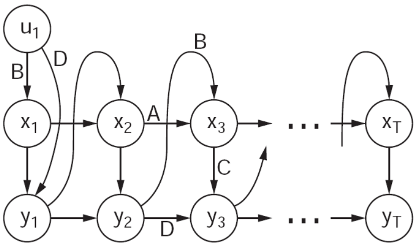
Let $\mathbf{u}_t := \mathbf{y}_{t-1}$. Then,

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + B\mathbf{y}_{t-1} + \mathbf{w}_t, \qquad \mathbf{w}_t \sim N(\mathbf{0}, Q) \qquad (5)$$
$$\mathbf{y}_t = C\mathbf{x}_t + D\mathbf{y}_{t-1} + \mathbf{v}_t, \qquad \mathbf{v}_t \sim N(\mathbf{0}, R) \qquad (6)$$

Consequence

Hidden states can now concentrate on modeling hidden factors while Markovian dependencies between successive outputs are now modeled by output-input feedbacks.

Graphical Model



$$\mathbf{x}_t = A\mathbf{x}_{t-1} + B\mathbf{y}_{t-1} + \mathbf{w}_t, \qquad \mathbf{w}_t \sim N(\mathbf{0}, Q) \qquad (7)$$
$$\mathbf{y}_t = C\mathbf{x}_t + D\mathbf{y}_{t-1} + \mathbf{v}_t, \qquad \mathbf{v}_t \sim N(\mathbf{0}, R) \qquad (8)$$

Genetic network parameters

Which are the parameters of interest for the genetic regulation network inference problem ?
We have :

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + B\mathbf{y}_{t-1} + \mathbf{w}_t, \qquad \mathbf{w}_t \sim N(\mathbf{0}, Q) \qquad (9)$$
$$\mathbf{y}_t = C\mathbf{x}_t + D\mathbf{y}_{t-1} + \mathbf{v}_t, \qquad \mathbf{v}_t \sim N(\mathbf{0}, R) \qquad (10)$$

Plugging in the definition of $\mathbf{x}_t$, $\mathbf{y}_t$ can be written

$$\mathbf{y}_t = (CB + D)\mathbf{y}_{t-1} + \mathbf{r}_t$$

where

$$\mathbf{r}_t = \mathbf{v}_t + C\mathbf{w}_t + CA\mathbf{x}_{t-1}$$

Outline

Introduction
- Biological Background
- Experiment
- Properties of the experimental data

Methods
- The linear dynamical system model
- The Bayesian Approach to SSMs
- Variational-Bayes model fitting

Results

Bayesian ?! Why ?

Bayesian methodology applied to Kalman filtering

- Assessing parameters $(CB + D)$ significativity : a posteriori distribution instead of bootstrap
- Model fitting : ML estimation in low sampling conditions

Model comparison

- Why ? Determine hidden state space dimension
- Cross-validation and the low-sampling issue
- Evidence framework

The Evidence framework

Let's say we want to compare n models $m_1, m_2, \ldots, m_n$ given observed data $\mathbf{y}$.

$$p(m_k|\mathbf{y}) = \frac{p(\mathbf{y}|m_k)p(m_k)}{p(\mathbf{y})} = \frac{\text{likelihood} * \text{prior}}{\text{normalizing-constant}}$$

Assuming θ is the set of all the parameters, the likelihood $p(\mathbf{y}|m)$ can be written

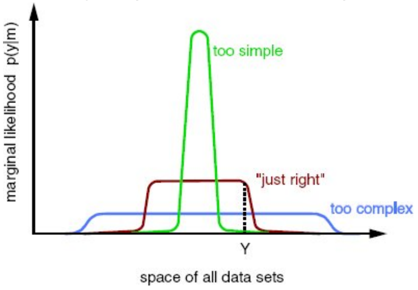
$$p(\mathbf{y}|m) = \int p(\mathbf{y}|\theta, m)p(\theta|m)d\theta$$

Key idea

In the absence of a prior, evidence alone drives model selection [Mackay, 1991].

Occam's razor and the evidence

Parsimony : why does computing the evidence result in choosing simple models over complex ones ?



(picture from [Beal, 2003])

Outline

Introduction

Biological Background

Experiment

Properties of the experimental data

Methods

The linear dynamical system model

The Bayesian Approach to SSMs

Variational-Bayes model fitting

Results

Simplifying assumptions

Remember our model :

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + B\mathbf{y}_{t-1} + \mathbf{w}_t, \qquad \mathbf{w}_t \sim N(\mathbf{0}, Q) \tag{11}$$

$$\mathbf{y}_t = C\mathbf{x}_t + D\mathbf{y}_{t-1} + \mathbf{v}_t, \qquad \mathbf{v}_t \sim N(\mathbf{0}, R) \tag{12}$$

We make simplying assumptions on the noise covariance matrices :

- $Q :=$ Id (no loss of generality, A will adjust)
- $R :=$ diag(σ) (white noise, with sigma one-dimensional σ)

and define the vector of parameters

$$\theta := (A, B, C, D, R)$$

Simplifying assumptions

To fit the model, we want to maximize the log-likelihood (or log-evidence). But this is computationally intractable (too many variables).

$$\begin{aligned}
\ln p(\mathbf{y}|m) &= \ln \int p(\mathbf{y}, \mathbf{x}, \theta|m) d\mathbf{x} d\theta && (13)\\
&= \ln \int q_1(\mathbf{x}) q_2(\theta) \frac{p(\mathbf{y}, \mathbf{x}, \theta|m)}{q_1(\mathbf{x}) q_2(\theta)} d\mathbf{x} d\theta && (14)\\
&\geq \int q_1(\mathbf{x}) q_2(\theta) \ln \left(\frac{p(\mathbf{y}, \mathbf{x}, \theta|m)}{q_1(\mathbf{x}) q_2(\theta)} \right) d\mathbf{x} d\theta && (15)\\
&= \mathcal{F}(q_1(\mathbf{x}), q_2(\theta), \mathbf{y}) && (16)
\end{aligned}$$

where the step with the inequality follows from appealing to Jensen's inequality.

Goal

- Maximize $\mathcal{F}$ with respect to the free distributions $q_1(\mathbf{x})$ and $q_2(\theta)$
- Joint maximization of $q_1(\mathbf{x})$ (hidden process distribution likelihood) and $q_2(\theta)$ (parameter distribution likelihood)

The Variational-Bayes EM Algorithm

At iteration l,

VB-E step

Find $q_1^{(l+1)}(\mathbf{x})$ that maximzes

$$E\{\mathcal{F}(q_1(\mathbf{x}), q_2^{(l)}(\theta), \mathbf{y})\}$$

VB-M step

Find $q_2^{(l+1)}(\theta)$ that maximzes

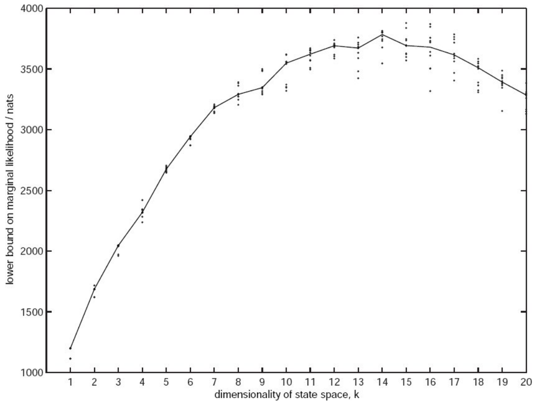
$$E\{\mathcal{F}(q_1^{(l+1)}(\mathbf{x}), q_2(\theta), \mathbf{y})\}$$

and iterate until convergence.

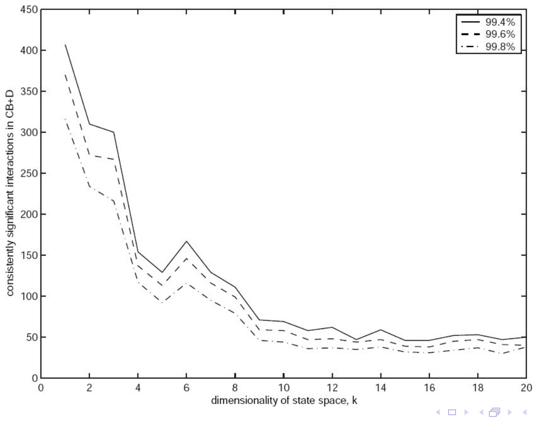
Consequence

- Update hyperparameters and prior parameters to maximize the lower bound on the marginal likelihood
- By applying the VB-EM algorithm, we actually minimize the KL divergence between the approximation $q_1(\mathbf{x})q_2(\theta)$ and the true posterior $p(\mathbf{x}, \theta|\mathbf{y}, m)$ [Beal, 2003]

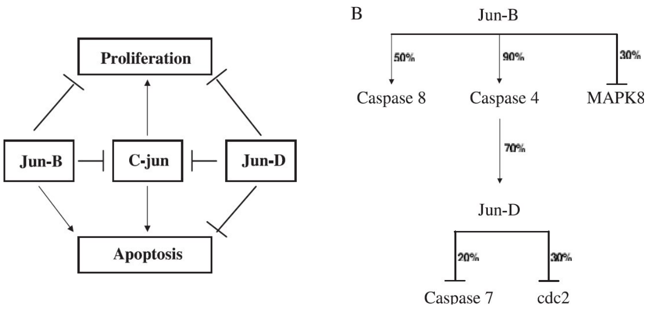
Model comparison



Significant interactions



Biological consequences for the JunB - JunD pathway



Summary

- Kalman filtering : model causal dependencies in auto-correlated, multivariate, noisy data while allowing hidden states
- Bayesian methodology & Kalman filtering : assessing parameter significance, avoiding overfitting when selecting a model (Occam's razor)
- Model fitting with Variational Bayes EM algorithm

- Outlook
  - Test new model biologicaly and investigate what the hidden states account for.
  - Allow for non-linear interactions (expression saturation effects, multiplicative effects).

References

Matthew J. Beal.
Variational Algorithms For Approximate Bayesian Inference.

PhD thesis, University of London, 2003.

Matthew J. Beal, Francesco Falciani, Zoubin Ghahramani, Claudia Rangel, and David L. Wild.
A bayesian approach to reconstructing genetic regulation networks with hidden factors.
Bioinformatics, 2004.

David J. C. Mackay.
Bayesian interpolation.
Neural Computation, 1991.