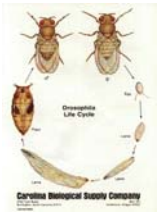
Computational Genomics

10-810/02-710, Spring 2009

Time Series Model for Gene Expression

Eric Xing

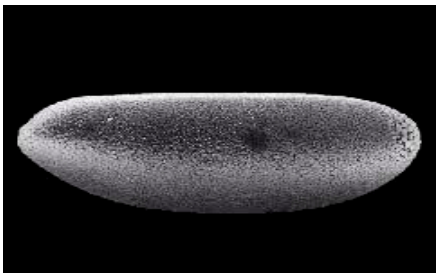


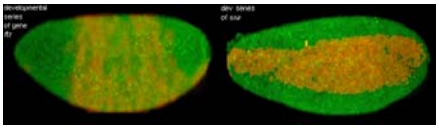
Lecture 18, March 25, 2009

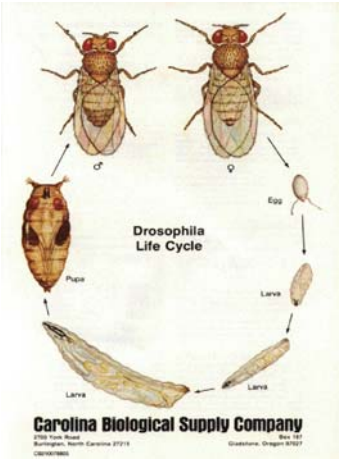
Reading: class assignment

Why Time Series?

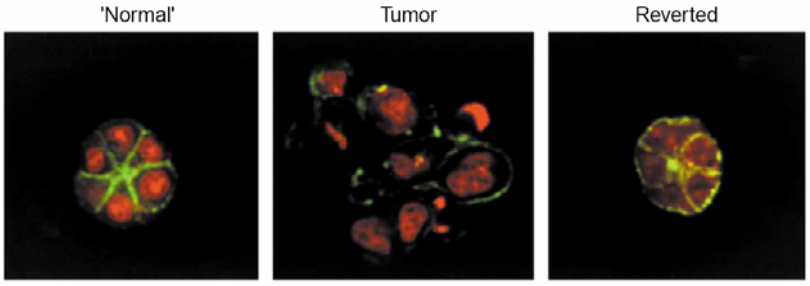
- Biological processes are time evolving!





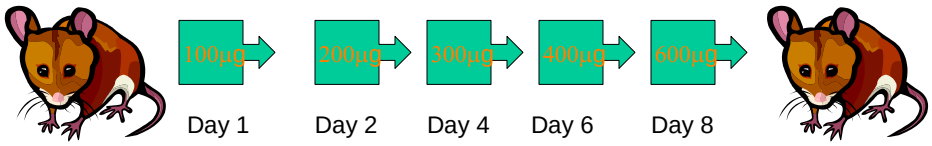


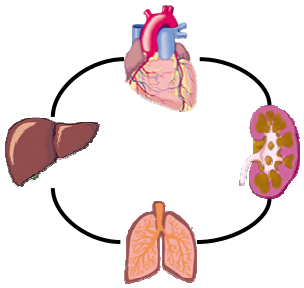
Example II: Breast Cancer Progression and Reversal in Organotypic Culture



Dr. Mina Bissell, Berkeley

Example III: Inflammatory Response in Endotoxinated Mice



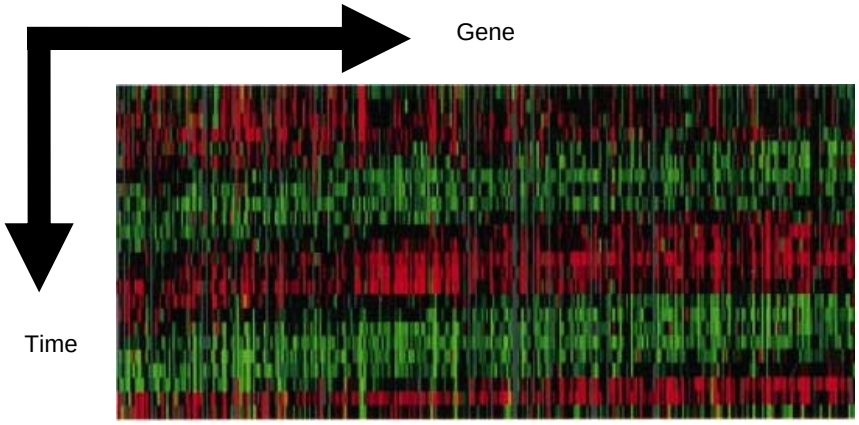


Time Series of Gene Expression

- A sequence of gene expression measured at successive time points at either uniform or uneven time intervals.
- Reveal more information than static data as time series data measure biological systems under different yet related conditions.

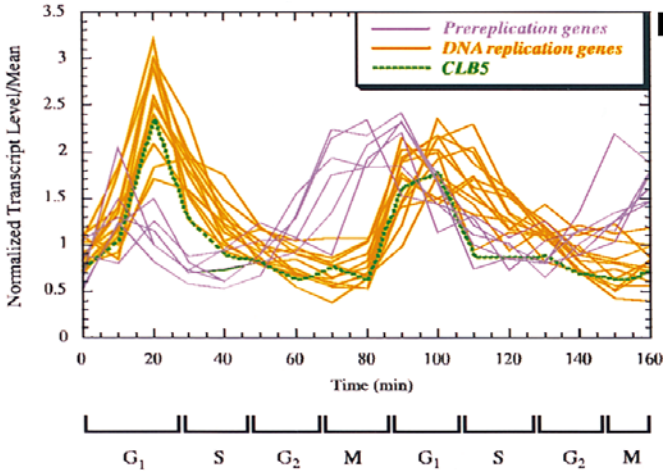
Yeast Cell Cycle

- Spellman et al. Mol. Bio. Cel. 98



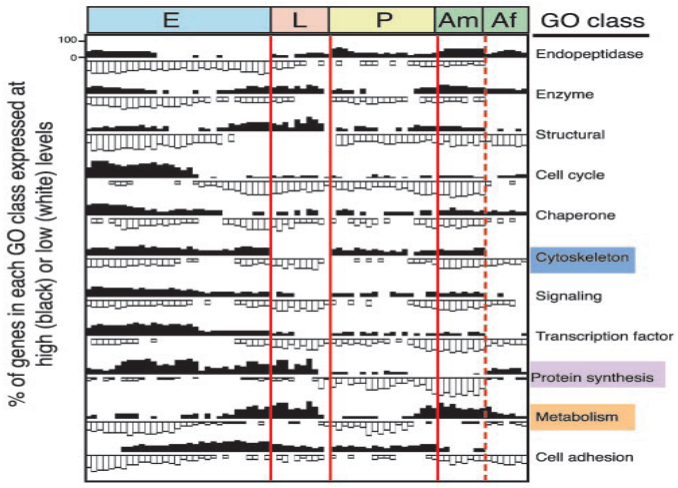
Yeast Cell Cycle (cont'd)

- Period pattern of expression



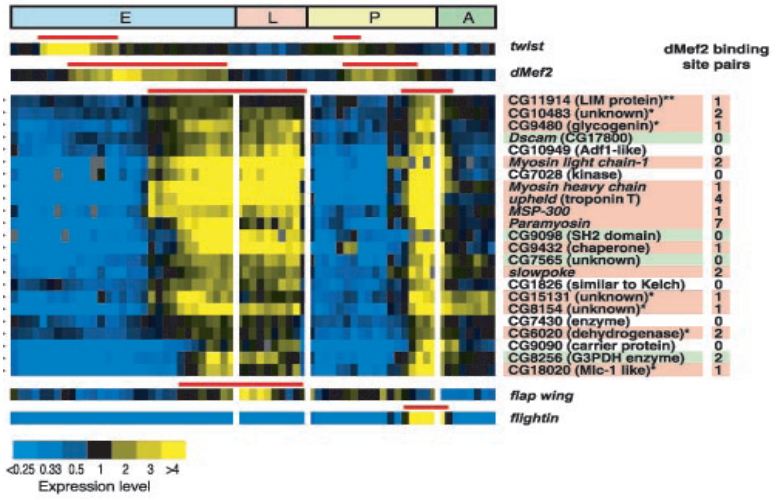
Life Cycle of Drosophila Melanogaster

- Arbeitman et al. Nature 02



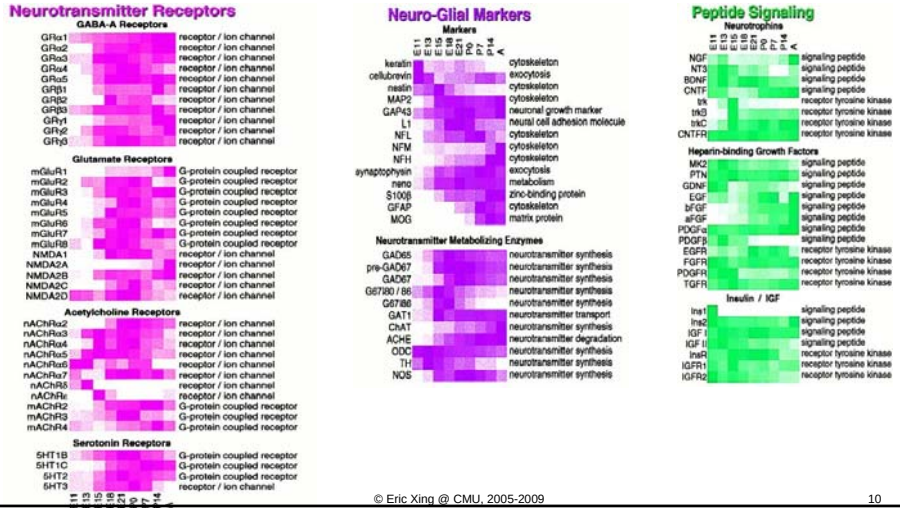
Life Cycle of Drosophila Melanogaster (cont'd)

- Muscle development, timing of transcriptional factors



Spinal Cord Development of Rats

- Wen et al. PNAS 98



The Objectives of Time Series Analysis

- **Interpretation**

  e.g. What are the genes that control the yeast cell cycle?
- **Forecasting**

  e.g. Under stimuli A, what is the growth rate of yeast in 5 hours?
- **Control**

  e.g. How to control the growth of cancerous cells?
- **Hypothesis testing**

  e.g. Is gene A differentially expressed under two different conditions at time point T?
- **Simulation**

  e.g. Can we recreate in-silico model of the organism based on parameters extracted from time series?

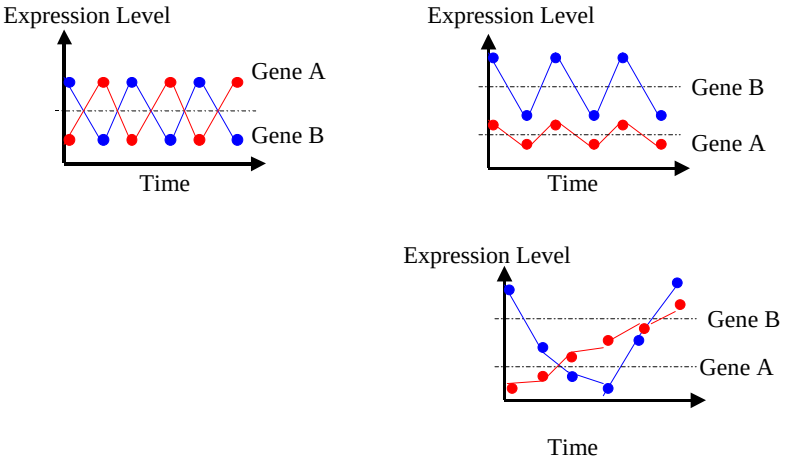
Method of Time Series Analysis

- Cluster Analysis
- Spectrum Analysis
- Smoothing and Trend Analysis
- Dynamic system model
- Learning gene regulatory relations (dynamic networks)

Cluster Analysis

- Treat each gene as a data point
- Treat time series X for a gene as a single vector
- Define similarity score or distance score between two time series X and X'
- Apply any conventional clustering algorithm (hierarchical clustering, k-means, etc.)
- E.g. useful for discovering functional modules

Similarity Measures



Similarity Measures: Correlation Coefficient

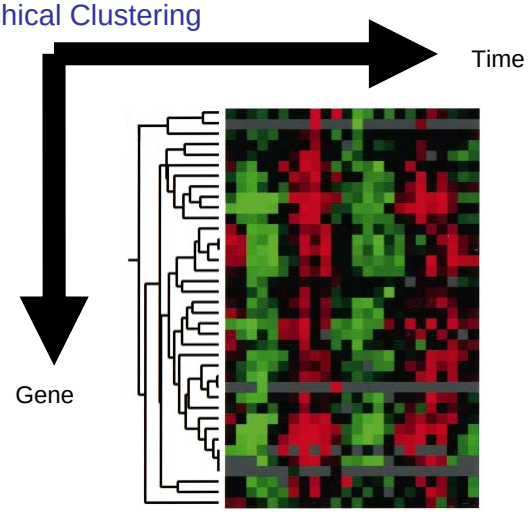
$$s(x,y) = \frac{\sum_{i=1}^{p}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{p}(x_i - \bar{x})^2 \times \sum_{i=1}^{p}(y_i - \bar{y})^2}}$$

where $\bar{x} = \frac{1}{p}\sum_{i=1}^{p} x_i$ and $\bar{y} = \frac{1}{p}\sum_{i=1}^{p} y_i$.

$$|s(x,y)| \leq 1$$

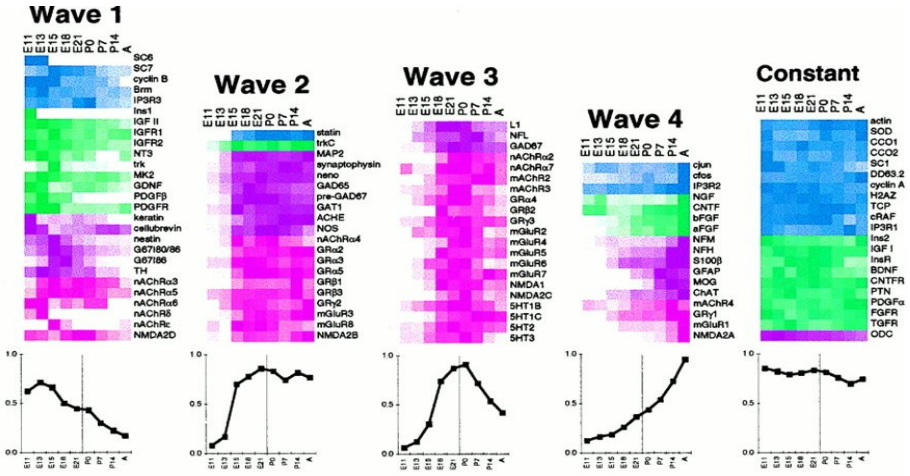
Cluster Analysis

- Hierarchical Clustering



Cluster Analysis

- Clustering genes by their wave patterns



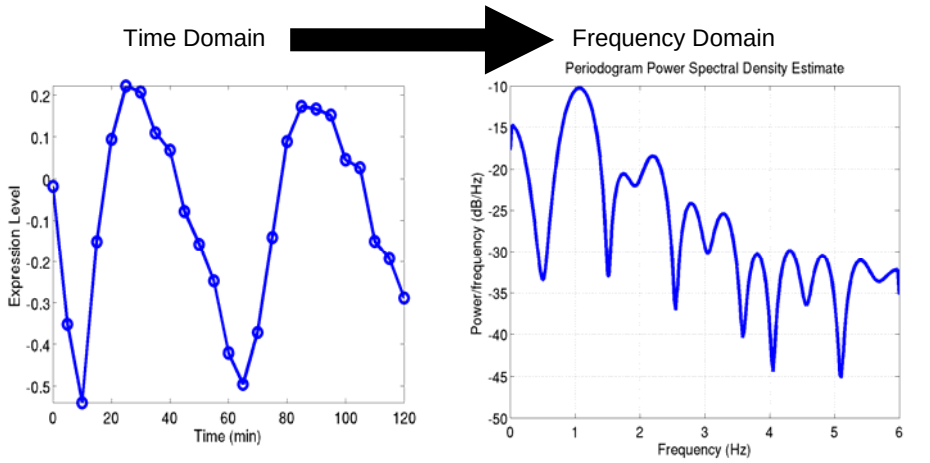
Spectrum Analysis

- Transform gene expression from time domain to frequency domain
- Discrete Fourier Transformation (DFT)

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i 2\pi k \frac{n}{N}} \qquad k = 0, \ldots, N-1.$$

- Significant frequency components were those with large amplitude, ie. $|x_k|$.
- E.g. useful for identifying cell cycle genes

Spectrum Analysis

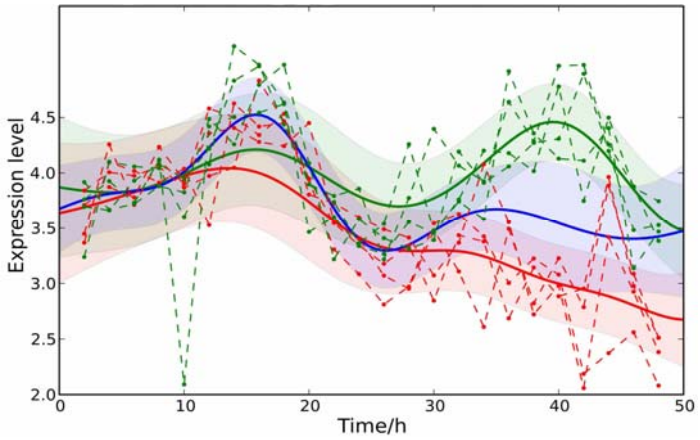


60 min <=> 1 cell cycle

Normalized frequence: 1Hz <=> 1 cell cycle

Smoothing and Trend Analysis

- Eg. how does gene expression change in general?



L_2 and L_1 Regularized Trend Analysis

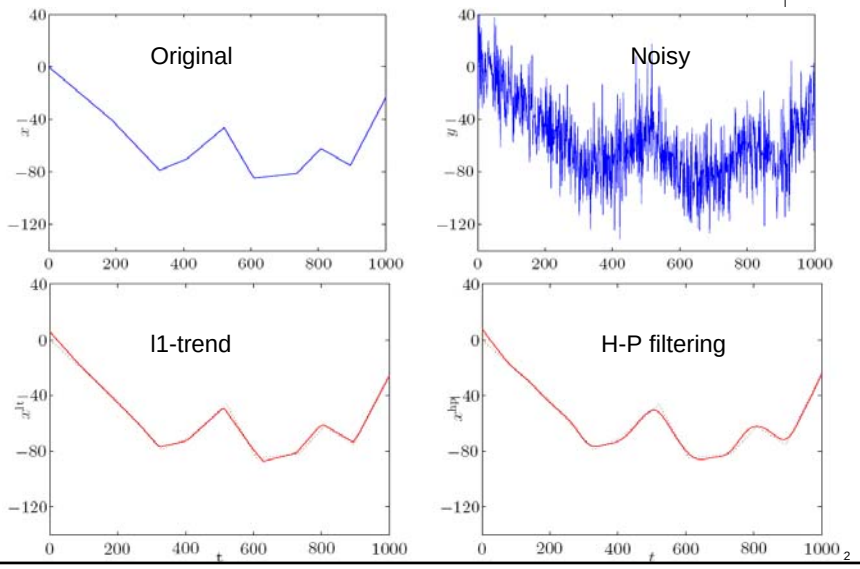
- Hodrick-Prescott filtering: find time series x to smooth time series y s.t. the following objective is minimized (O(N))

$$(1/2)\sum_{t=1}^{n}(y_t - x_t)^2 + \lambda \sum_{t=2}^{n-1}(x_{t-1} - 2x_t + x_{t+1})^2$$

- l_1-trend analysis: slightly different in the regularization (expected O(N), worse case $O(N^{1.5})$)

$$(1/2)\sum_{t=1}^{n}(y_t - x_t)^2 + \lambda \sum_{t=2}^{n-1}|x_{t-1} - 2x_t + x_{t+1}|$$

L_2 vs L_1



Dynamical System Model

- Kalman filter for forecasting
- Estimate the state x of a discrete time controlled process

$$x_k = Ax_{k-1} + Bu_{k-1} + w_{k-1}$$

- With measure process

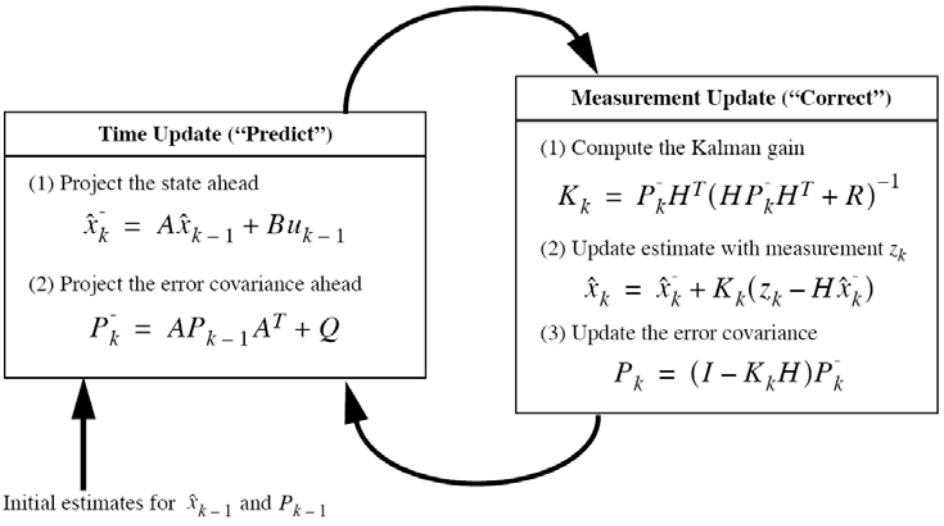
$$z_k = Hx_k + v_k$$

- w_k v_k zero mean Gaussian noise

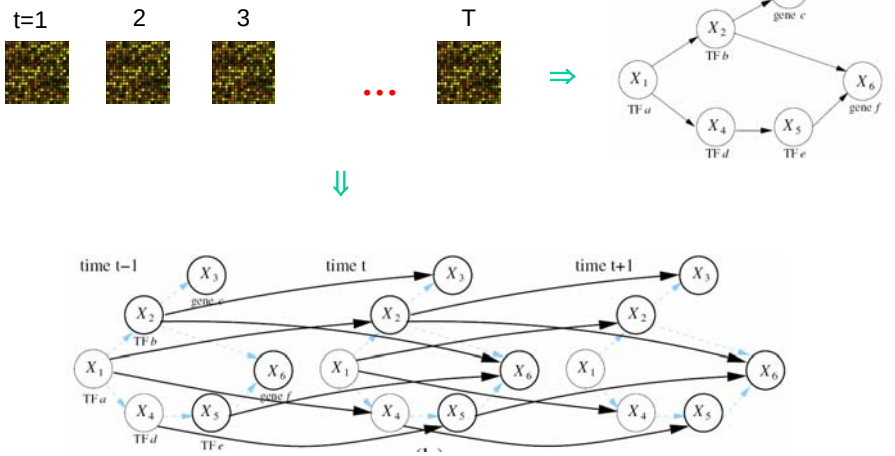
$$p(w) \sim N(0, Q)$$

$$p(v) \sim N(0, R)$$

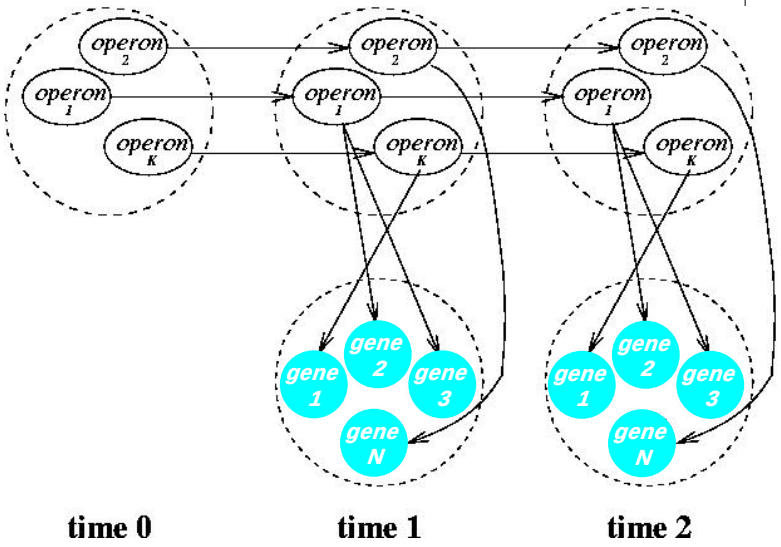
Kalman Filter



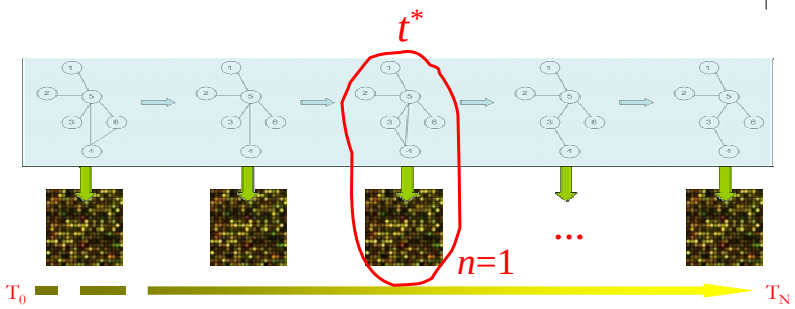
Network Analysis



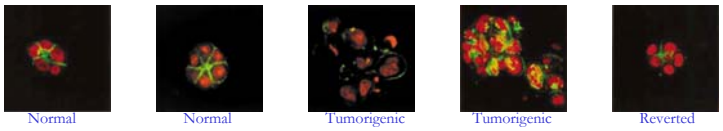
A DBN for E.coli Regulatory Pathways (Ong ISMB 2003)



Temporal/Spatial-Specific "Rewiring" Gene Networks

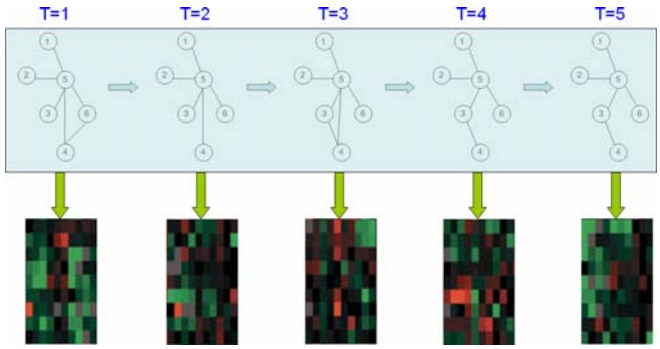


EGFR-induced progression/reversion of breast epithelial cells



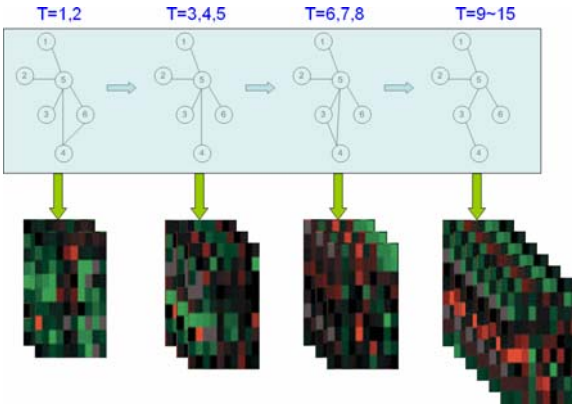
Rewiring Biological Networks

- Networks rewire over discrete timesteps



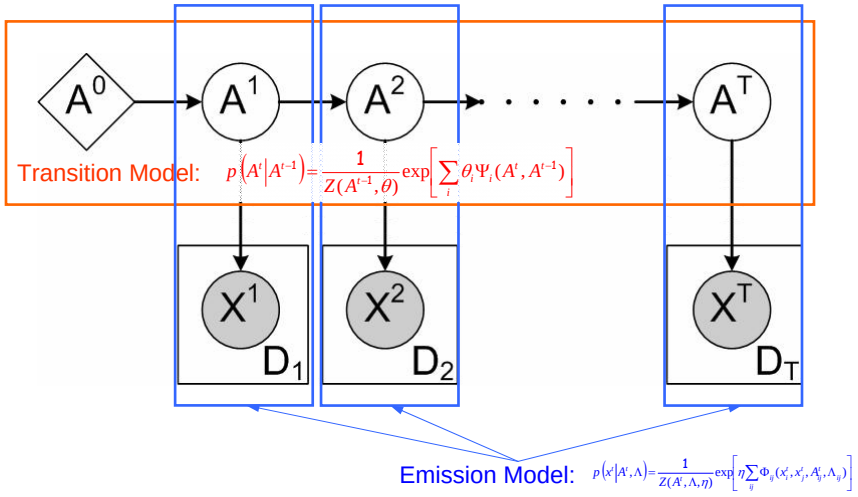
Rewiring Biological Networks (cont.)

- Networks rewire over epochs



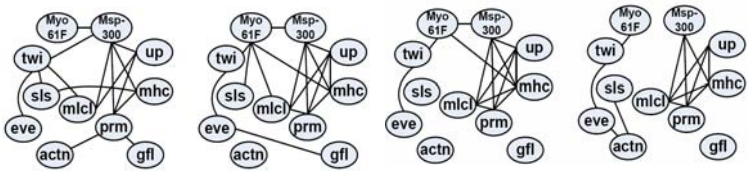
Modeling Time-Varying Graphs

- The temporal exponential graph models *(Fan et al. ICML 2007)*



Results on *Drosophila* data

- The proposed model was applied to infer the muscle development sub-network (Zhao et al., 2006) on *Drosophila* lifecycle gene expression data (Albeitman et al., 2002).
  - 11 genes, 66 timesteps over 4 development stages
  - Further biological experiments are necessary for verification.



Network in (Zhao et al. 2006) | Embryonic | Larval | Pupal & Adult

Evolving Markov Random Fields

(amr and Xing, 2009)

- Assuming the graphs are continuously weighted, then for each time point *t*, we have a MRF model for expression

$$P(\mathbf{x}_\mathbf{d}^\mathbf{t}|\Theta^t) = \exp\left(\sum_{i\in V}\theta_{ii}^t x_{d,i}^t + \sum_{(i,j)\in E^t}\theta_{ij}^t x_{d,i}^t x_{d,j}^t - A(\Theta^t)\right)$$

- **Graphical lasso** has been used to obtain a sparse estimate of *E* with continuous *X*

- Assuming graphs are *smoothly evolving* over time
  - Estimate E^1, E^2, … via temporally smoothed graph lasso

TESLA: Temporally Smoothed L_1-regularized logistic regression

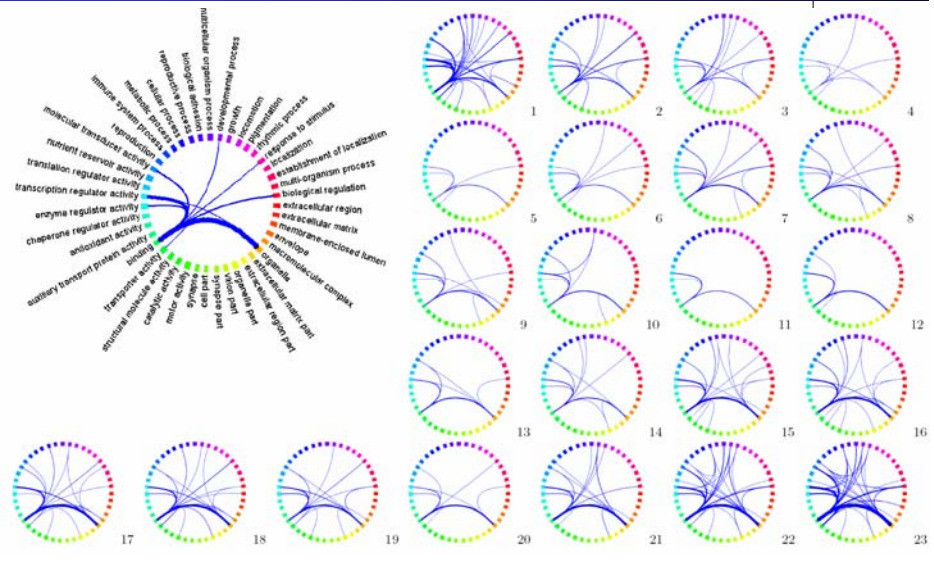
(amr and Xing, 2009)

$$\hat{\theta}_i^1, \ldots, \hat{\theta}_i^T = \arg\min_{\theta_i^1,\ldots,\theta_i^T} \sum_{t=1}^{T} l_{avg}(\theta_i^t) + \lambda_1 \sum_{t=1}^{T} \| \theta_{-i}^t \|_1 + \lambda_2 \sum_{t=2}^{T} \| \theta_i^t - \theta_i^{t-1} \|_q^q,$$

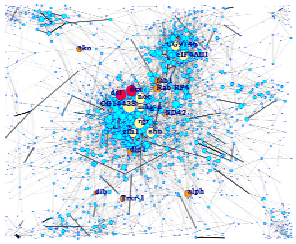
where $l_{avg}(\theta_{\mathbf{i}}^{\mathbf{t}}) = \frac{1}{N^t} \sum_{d=1}^{N^t} \log P\left(x_{d,i}^t | \mathbf{x}_{\mathbf{d},-\mathbf{i}}^{\mathbf{t}}, \theta_{\mathbf{i}}^{\mathbf{t}}\right)$.

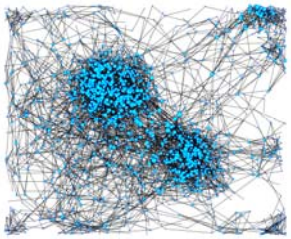
- Convex optimization

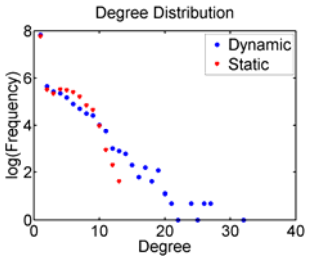
Transient Interaction



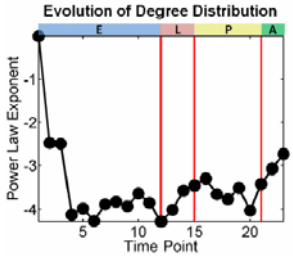
Static Versus Dynamic

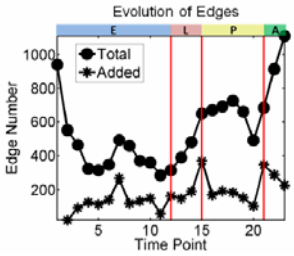


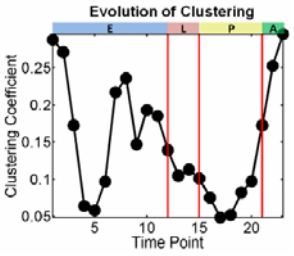




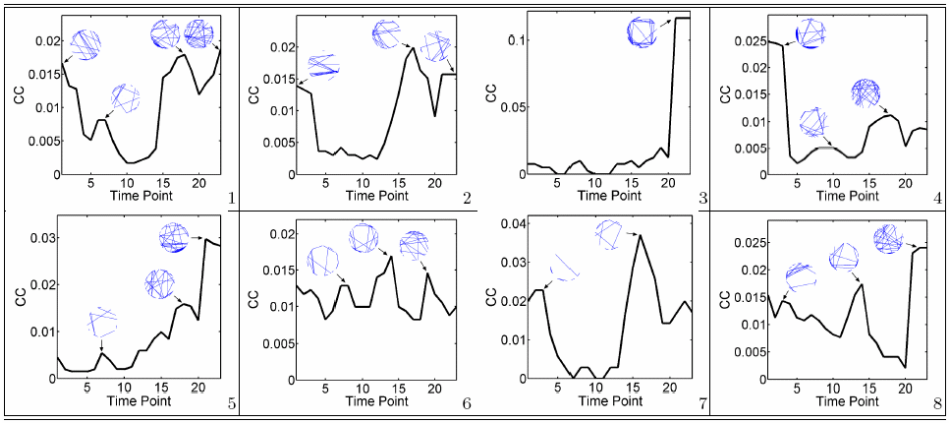
Evolution of Network Signatures







Transient Subgraph



Future Work

- Analyzing time-space data in biological processes
  - Drosophila life cycle
  - Breast cancer progression and reversal
  - Inflammatory response in endotoxinated mice
- Other dynamic behaviors of networks
  - Differentiation: tree of networks
  - Detection of sudden changes
  - Active learning – when to get more samples
- Open theoretical issues
  - Consistence (pattern, value, …)
  - Confidence
  - Stability
  - Sample complexity