

Dynamic Network Inference

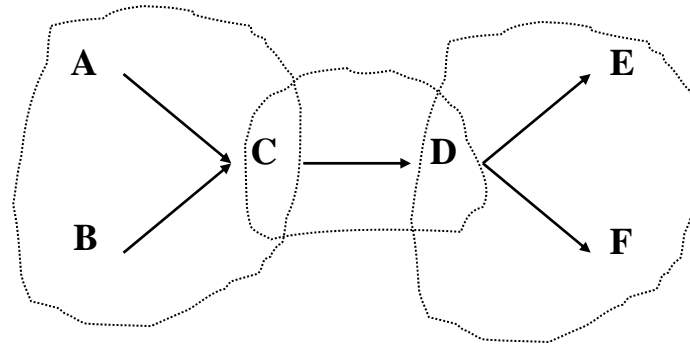
Most statistical work is done on gene regulatory networks, while inference of metabolic pathways and signaling networks are done by other means.

Like in phylogenetics, network inference has two components – graph structure (topology) and continuous aspects, such as parameters of the distributions relating neighboring nodes.

Like genome annotation, networks are often hidden structures that influences something that can be observed.

- *Network combinatorics*
- *Inference of Boolean Networks*
- *ODEs with Noise*
- *Gaussian Processes*
- *Dynamic Bayesian Networks*

Networks & Hypergraphs



$$\frac{dA}{dt} = \frac{dB}{dt} = -k_{A,B}[A][B], \frac{dC}{dt} = k_{A,B}[A][B] - k_C[C], \frac{dD}{dt} = k_C[C] - k_D[D], \frac{dE}{dt} = \frac{dF}{dt} = k_D[D]$$

How many directed hypergraphs are there?

<i>0'th order ?</i>	<i>Constant removal/addition of a component</i>	2^6	2^k
<i>1'st order ?</i>	<i>Exponential growth/decay as function of some concentration</i>	2^{36}	2^{k*k}
<i>2'nd order ?</i>	<i>Pairwise collision creation: (A, B --> C) (no multiplicities)</i>	$2^{6*5*4/3}$	$2^{k*k-1*k-2/3}$
<i>i-in, o-out ?</i>			
<i>Arbitrary ?</i>	<i>Partition components into in-set, out-set, rest-set (no multiplicities)</i>	2^3^6	2^3^k

Number of Networks

- *undirected graphs*

$$\alpha_n = 2^{\frac{n(n-1)}{2}}$$

- *Connected undirected graphs*

$$c_n = \alpha_n - \sum_{k=1}^{n-1} \binom{n-1}{k-1} c_k \alpha_{n-k}$$

- *Directed Acyclic Graphs - DAGs*

$$a_n = \sum_{k=1}^n (-1)^{k-1} \binom{n}{k} 2^{k(n-k)} a_{n-k}$$

- *Interesting Problems to consider:*

- *The size of neighborhood of a graph?*
- *Given a set of subgraphs, how many graphs have them as subgraphs?*

Reverse Engineering Algorithm-Reveal

Discrete known Generations

No Noise

X	0	1	1	1	1	1	1	1	0	0	0
Y	0	0	0	1	1	0	0	0	1	1	1

Shannon Entropies: $H(X) = -\sum p_i \log(p_i)$ $H(X, Y) = -\sum p_{i,j} \log(p_{i,j})$

Mutual Information: $M(X, Y) = H(Y) - H(Y \text{ given } X) = H(X) - H(X \text{ given } Y)$

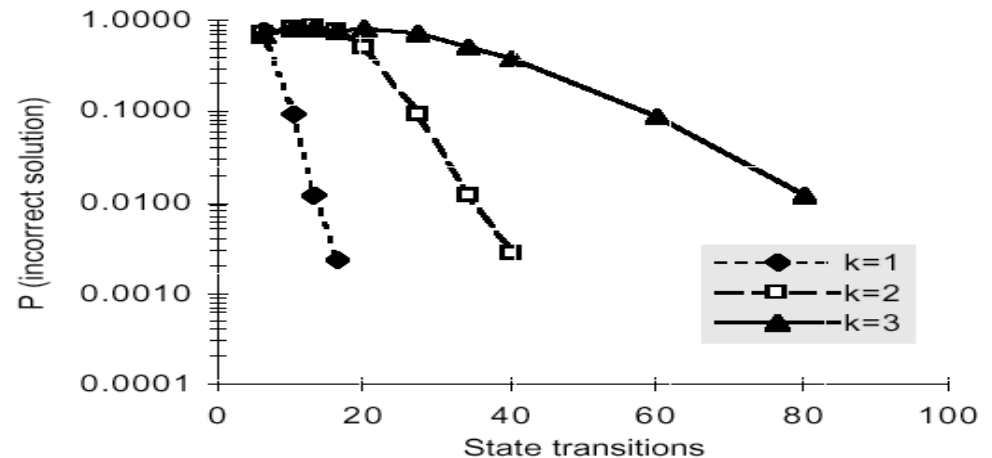
For $j=1$ to k

Find k -sets with significant mutual information.

Assign rule.

	X		
Y	3	2	$H(X) = .97$
	1	4	$H(Y) = 1.00$
			$H(X, Y) = 1.85$

- 50 genes
- Random firing rules
- Thus network inference is easy.
- However, it is not



BOOL-1, BOOL2, QNET1

Akutsu et al. (2000) Inferring qualitative relations in genetic networks and metabolic pathways. Bioinformatics 16.2.727-

Bool-1 Algorithm

For each gene do (n)

For each boolean rule ($\leq k$ inputs) not violated, keep it.

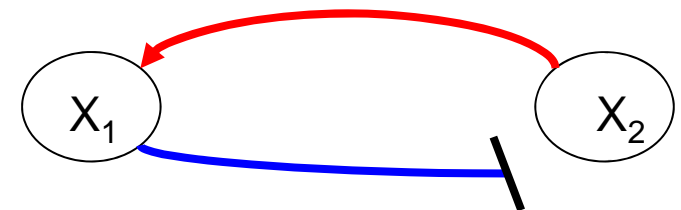
If $O(2^{2k}[2k + \alpha]\log(n))$ INPUT patterns are given **uniformly randomly**, BOOL-1 correctly identifies the underlying network with probability $1-n^{-\alpha}$, where α is any fixed real number > 1 .

Bool-2 p_{noise} is the probability that experiment reports wrong boolean rule uniformly.

Qnet
$$\frac{dX_1}{dt} = a_1 X_{j_1}, \frac{dX_2}{dt} = a_2 X_{j_2}, \dots, \frac{dX_n}{dt} = a_n X_{j_n}.$$

Activation $v_j \rightarrow v_i$ Inhibition $v_j \dashv v_i$

Qualitative Network QNET



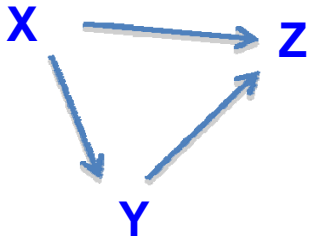
Algorithm

if $(\Delta X_i * X_j < 0)$ delete “ n_1 activates n_2 ” from E

if $(\Delta X_i * X_j > 0)$ delete “ n_1 inhibits n_2 ” from E

ODEs with Noise

Feed forward loop (FFL) This can be modeled by



$$\frac{dY(t)}{dt} = -\alpha_y Y(t) + \beta_y f(X(t), K_{xy}),$$

$$\frac{dZ(t)}{dt} = -\alpha_z Z(t) + \beta_z g(X(t), Y(t), K_{xz}, K_{yz})$$

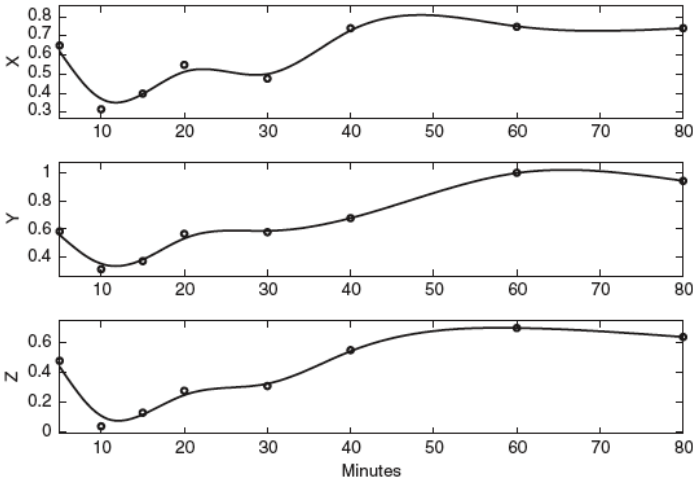
Where

$$f(u, K) = (u/K)^H / (1 + (u/K)^H)$$

$$g(t) = f(X(t), K_{xz}) f(Y(t), K_{yz})$$

Objective is to estimate $\theta = (\beta_y, \beta_z, \alpha_y, \alpha_z, K_{xy}, K_{xz}, K_{yz})$ from noisy measurements of expression levels
 If noise is given a distribution the problem is well defined and statistical estimation can be done

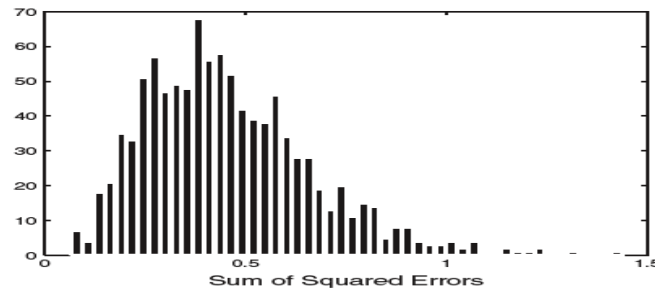
Data and estimation



Parameters	α_y	α_z	K_{xy}	K_{xz}	K_{yz}
FFL 1: X: Gene GCN4; Y: Gene LEU3; Z: Gene ILV5					
Estimates	0.44	0.69	0.90	0.60	0.56
Standard Errors	0.22	0.18	0.33	0.06	0.15

Goodness of Fit and Significance

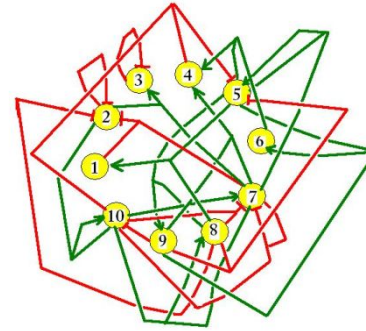
$$SSE(y, s_y, z, s_z) = \sum_{i=1}^{n_y} [y(t_i) - s_y(t_i | \hat{\theta})]^2 + \sum_{i=1}^{n_z} [z(t_i) - s_z(t_i | \hat{\theta})]^2$$



Gene X	Gene Y	Gene Z	SSE	P-values
GCN4	LEU3	ILV5	0.090	0.25
PDR1	PDR3	PDR5	1.17	0.33
GCN4	LEU3	ILV1	0.092	0.34
YLL044W	YER096W	YDR279W	0.84	0.046

Inference in the Presence of Knowledge

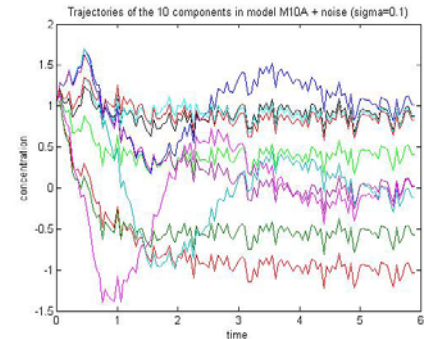
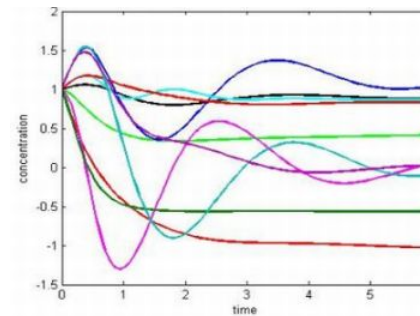
Dynamic mass action systems on 10 components were sampled with a bias towards sparseness



$$\begin{aligned}\frac{dx_1}{dt} &= 0.22 * x(5) * x(8) \\ \frac{dx_2}{dt} &= -1.19 * x(8) - 0.66 * x(2) - 0.59 * x(1) * x(7) \\ \frac{dx_3}{dt} &= -1.03 * x(3) + 0.75 * x(7) * x(2) \\ \frac{dx_4}{dt} &= 1.11 * x(5) * x(6) + 0.95 * x(9) * x(7) \\ \frac{dx_5}{dt} &= -2.01 * x(9) * x(8) + 1.8 * x(5) * x(7) - 0.18 * x(4) * x(10) \\ \frac{dx_6}{dt} &= 0.58 * x(5) * x(9) \\ \frac{dx_7}{dt} &= -1.24 * x(10) * x(8) - 1.41 * x(10) * x(7) + 0.26 * x(10) \\ \frac{dx_8}{dt} &= 0.69 * x(7) * x(10) \\ \frac{dx_9}{dt} &= 2.07 * x(8) * x(5) \\ \frac{dx_{10}}{dt} &= 3.01 * x(2) * x(10) - 0.74 * x(9)\end{aligned}$$

Kinetic parameters were sampled

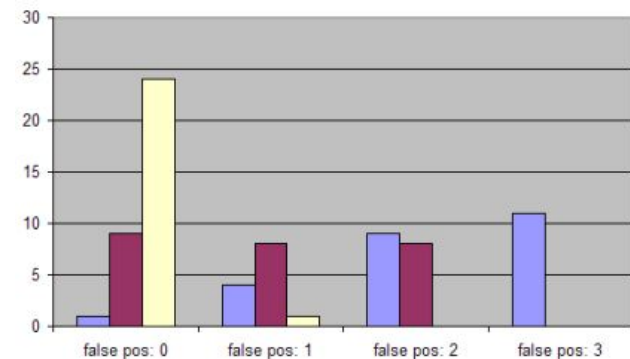
Dynamic trajectories were generated



Normal noise was added

Equation system minimizing SSE was chosen

Adding deterministic knowledge was added

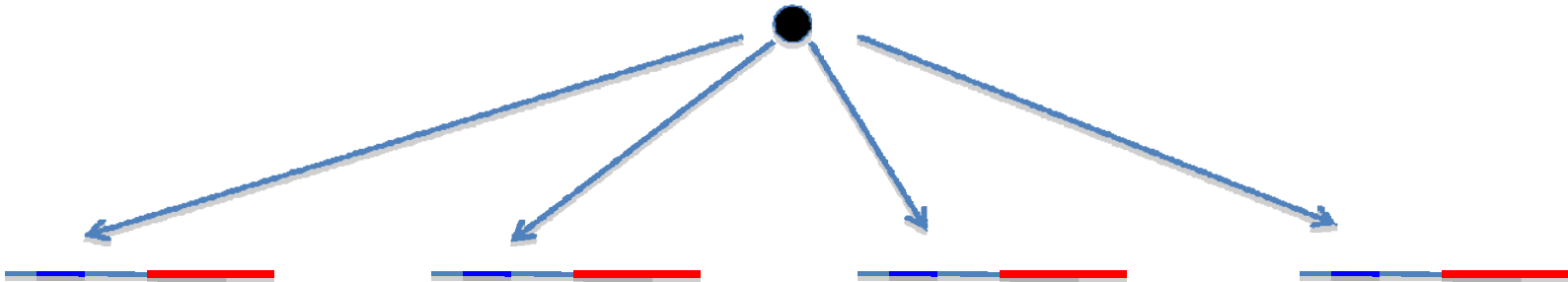


Gaussian Processes

Definition: A Stochastic Process $X(t)$ is a GP if all finite sets of time points, t_1, t_2, \dots, t_k , defines stochastic variable that follows a multivariate Normal distribution, $N(\mu, \Sigma)$, where μ is the k -dimensional mean and Σ is the $k \times k$ dimensional covariance matrix.

Examples: Brownian Motion: All increments are $N(0, \Delta t)$ distributed. Δt is the time period for the increment. No equilibrium distribution.

Ornstein-Uhlenbeck Process – diffusion process with centralizing linear drift. $N(\mu, \sigma^2)$ as equilibrium distribution.

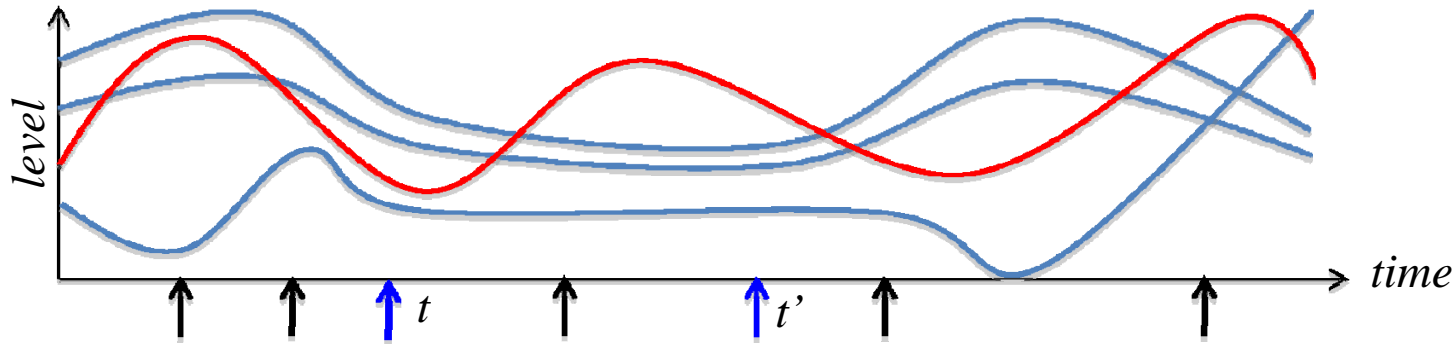


One TF (transcription factor – black ball) ($f(t)$) whose concentration fluctuates over times influence k genes (x_j) (four in this illustration) through their TFBS (transcription factor binding site - blue). The strength of its influence is described through a gene specific sensitivity, S_j . D_j – decay of gene j , B_j – production of gene j in absence of TF

$$\frac{dx_j}{dt} = B_j + S_j f(t) - D_j x_j(t), \quad x_j(0) = \frac{B_j}{D_j} \quad x_j(t) = \frac{B_j}{D_j} + S_j \int_0^t e^{-D_j(t-u)} f(u) du$$

Gaussian Processes

Gaussian Processes are characterized by their mean and variances thus calculating these for x_j and f at pairs of time, t and t' , points is a key objective



Observable

*Hidden and
Gaussian*

Correlation between two time points of f

$$k(t, t') = \exp\left(-\frac{(t-t')^2}{l^2}\right).$$

Correlation between two time points of same x 'es

$$k_{x_j, x_j}(t, t') = S_j^2 \int_0^t \int_0^{t'} e^{-D_j(t-u+t'-u')} k_{f, f}(u, u') du du'.$$

Correlation between two time points of different x 'es

$$k_{x_i, x_j}(t, t') = S_i S_j \int_0^t \int_0^{t'} e^{-D_i(t-u)-D_j(t'-u')} k_{f, f}(u, u') du du'.$$

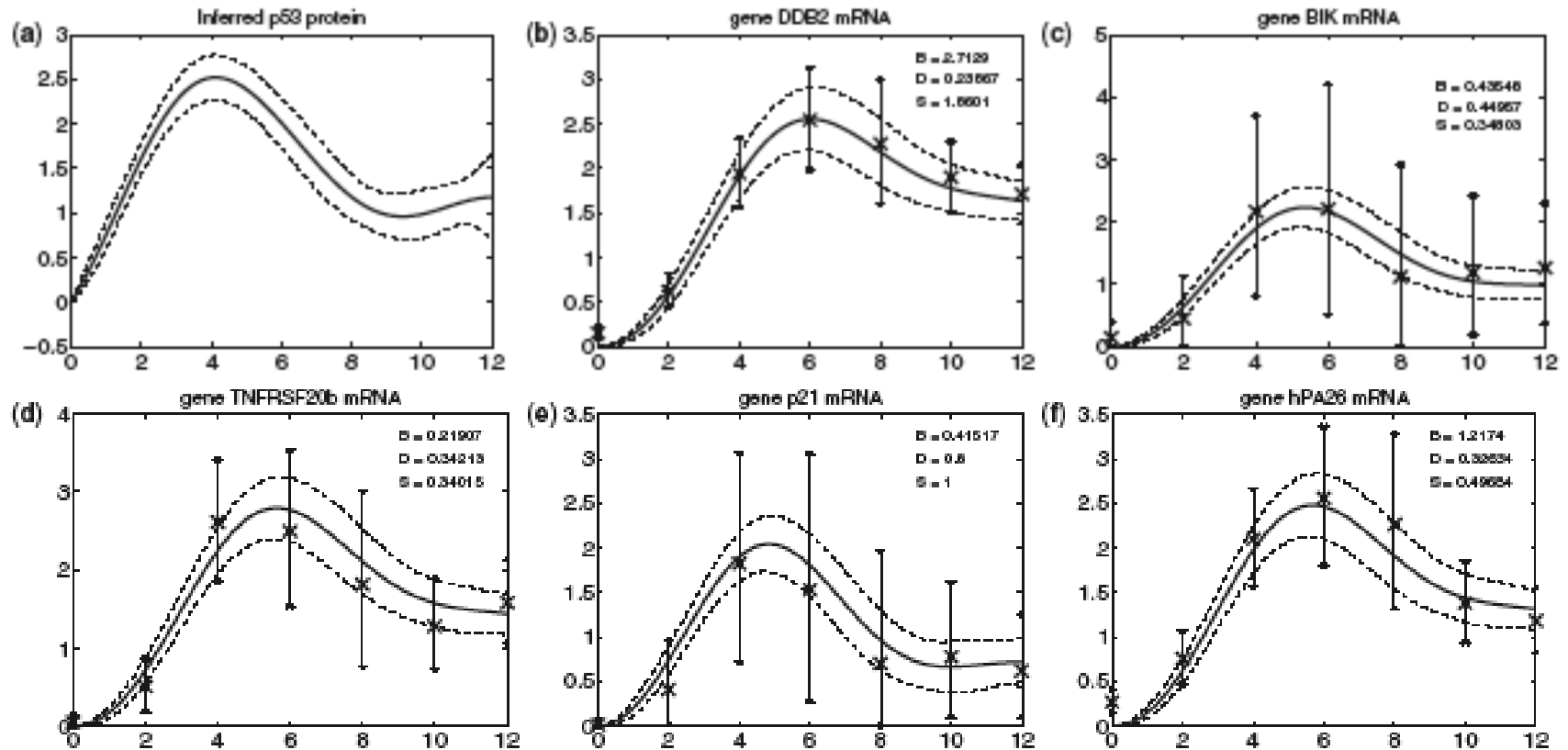
Correlation between two time points of x and f

$$k_{x_j, f}(t, t') = S_j \int_0^t e^{-D_j(t-u)} k_{f, f}(u, t') du.$$

This defines a prior on the observables

Then observe and a posterior distribution is defined

Gaussian Processes



Relevant Generalizations:

Non-linear response function

Multiple transcription factors

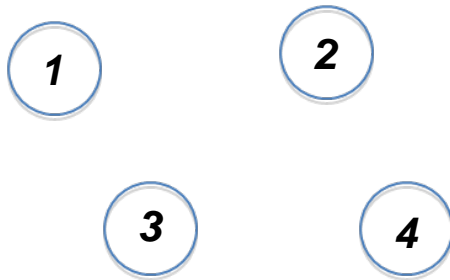
Network relationship between genes

Observations in Multiple Species

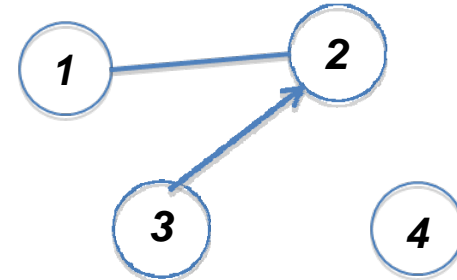
Comments: *Inference of Hidden Processes has strong similarity to genome annotation*

Graphical Models

Labeled Nodes: each associated a stochastic variable that can be observed or not.



Edges/Hyperedges – directed or undirected – determines the combined distribution on all nodes.



- **Conditional Independence**

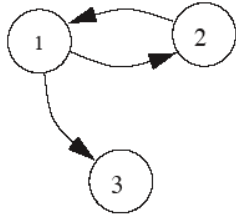
- **Gaussian**

- **Correlation Graphs**

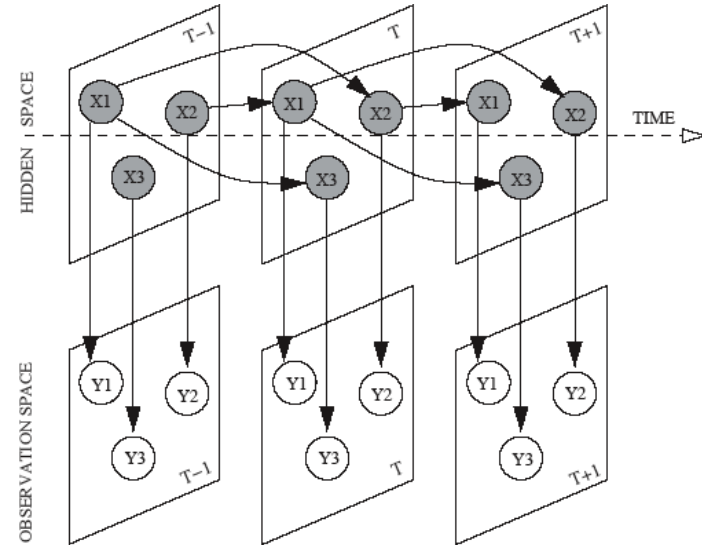
- **Causality Graphs**

Dynamic Bayesian Networks

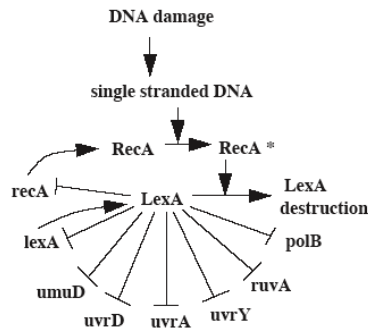
Take a graphical model



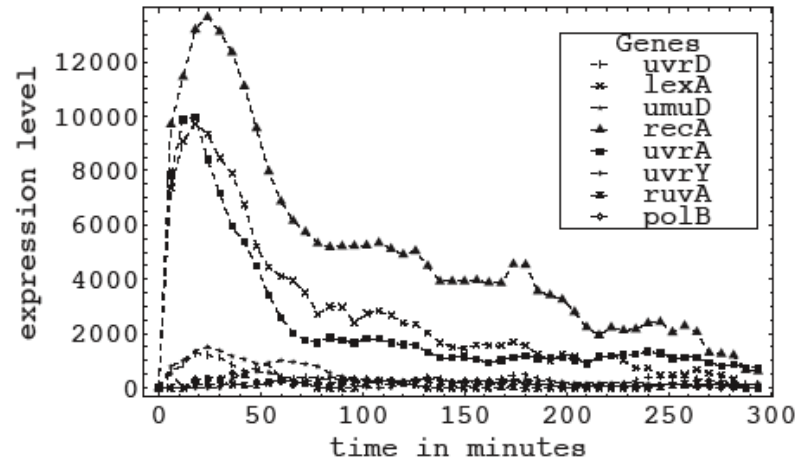
- i. Make a time series of of it
- ii. Model the observable as function of present network



Example: DNA repair



Inference about the level of hidden variables can be made



Feasibility of Network Inference: Very Hard

Why it is hard:

- *Data very noisy*
- *Number of network topologies very large*

What could help:

- *Other sources of knowledge – experiments*
- *Evolution*
- *Declaring biology unknowable would be very radical*

Why poor network inference might be acceptable:

- *A biological conclusion defines a large set of networks*

What statistics can do

- *Conceptual clarification of problem*
- *Optimal analysis of data*
- *Power studies (how much data do you need)*

Statistics can't draw conclusion if the data is insufficient or too noisy (I hope not)

Summary

- *Network Inference – topology and continuous parameters*
- *Network combinatorics*
- *Inference of Boolean Networks*
- *ODEs with Noise*
- *Gaussian Processes*
- *Dynamic Bayesian Networks*
- *Interpretation: From Integrative Genomics to Systems Biology:
Often the topology is assumed identical*