

Network Inference & Evolution

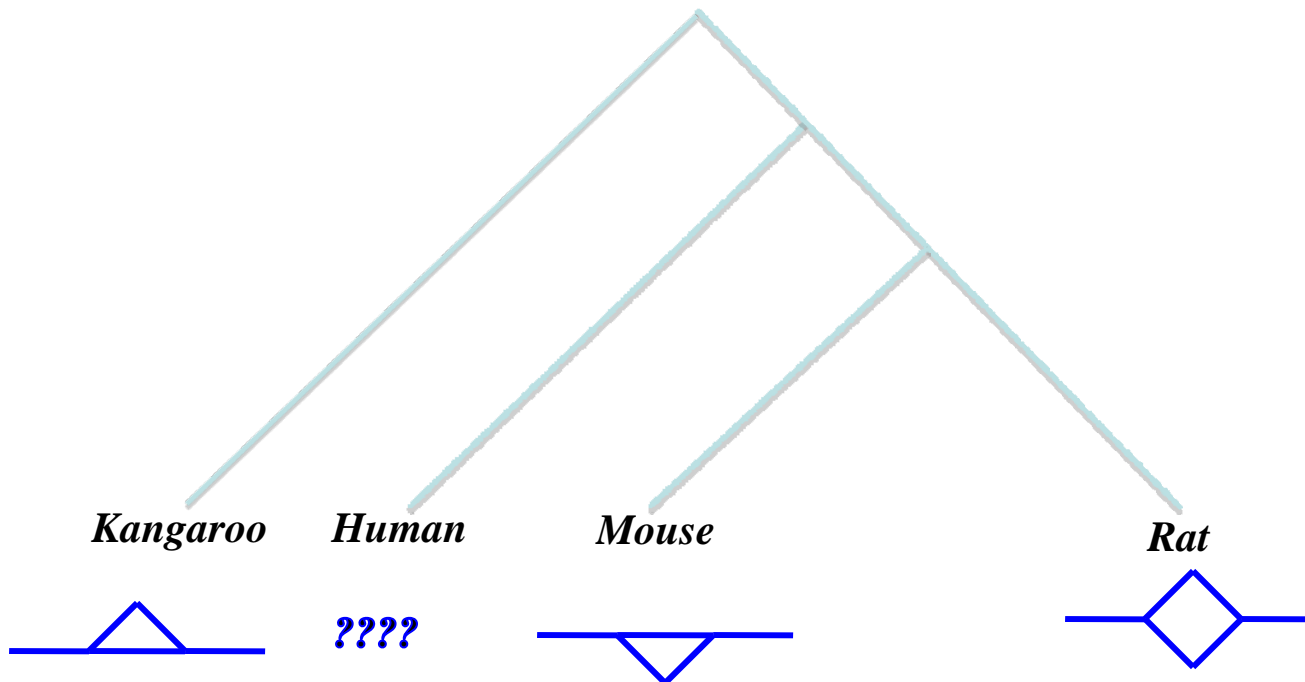
- *Metabolic Pathways*
- *Regulatory Networks*
- *Signaling Pathways*
- *Protein Interaction*

- *Dynamics*
- *Inference*
- *Evolution*

Understanding Evolution

Comparative Annotation

Knowledge and Model Organisms



Networks → A Cell → A Human

- *A cell has $\sim 10^{13}$ atoms.* 10^{13}
- *Describing atomic behavior needs $\sim 10^{15}$ time steps per second* 10^{28}
- *A human has $\sim 10^{13}$ cells.* 10^{41}
- *Large descriptive networks have 10^3 - 10^5 edges, nodes and labels* 10^5
- *What happened to the missing 36 orders of magnitude???*
- *Which approximations have been made?*
 - A** *Spatial homogeneity → 10^3 - 10^7 molecules can be represented by concentration* $\sim 10^4$
 - B** *One molecule (10^4), one action per second (10^{15})* $\sim 10^{19}$
 - C** *Little explicit description beyond the cell* $\sim 10^{13}$
- A** *Compartmentalisation can be added, some models (ie Turing) create spatial heterogeneity*
- B** *Hopefully valid, but hard to test*
- C** *Techniques (ie medical imaging) gather beyond cell data*

A repertoire of Dynamic Network Models

To get to networks:

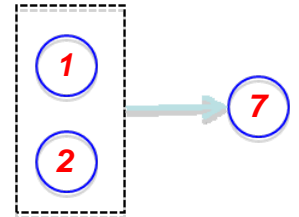
No space heterogeneity → molecules are represented by numbers/concentrations

Definition of Biochemical Network:

- A set of k nodes (chemical species) labelled by kind and possibly concentrations, X_k .



- A set of reactions/conservation laws (edges/hyperedges) is a set of nodes. Nodes can be labelled by numbers in reactions. If directed reactions, then an inset and an outset.



- Description of dynamics for each rule.

ODEs – ordinary differential equations

$$\frac{dX_7}{dt} = f(X_1, X_2)$$

Mass Action $\frac{dX_7}{dt} = cX_1X_2$

Time Delay $\frac{d\bar{X}(t)}{dt} = f(\bar{X}(t - \tau))$

Discrete Deterministic – the reactions are applied.

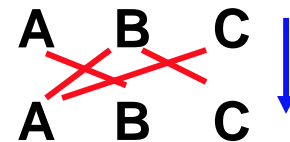
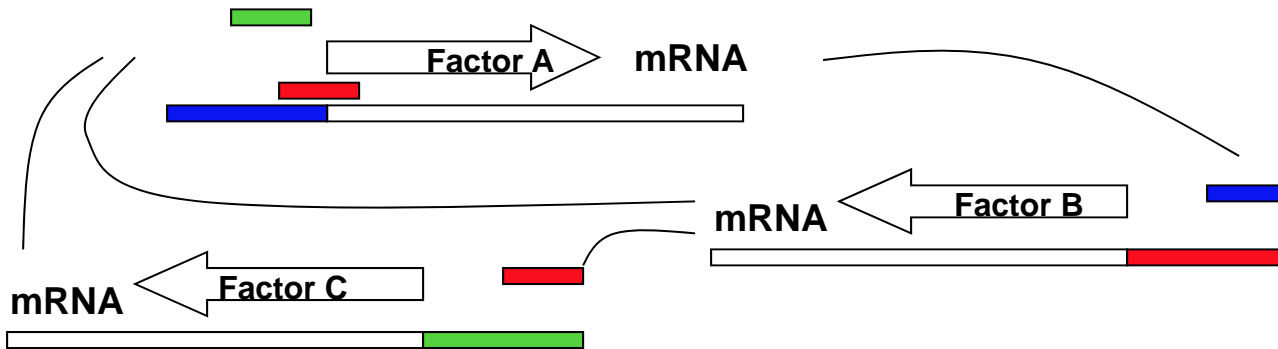
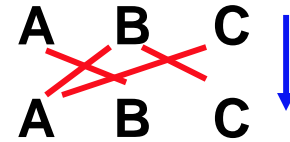
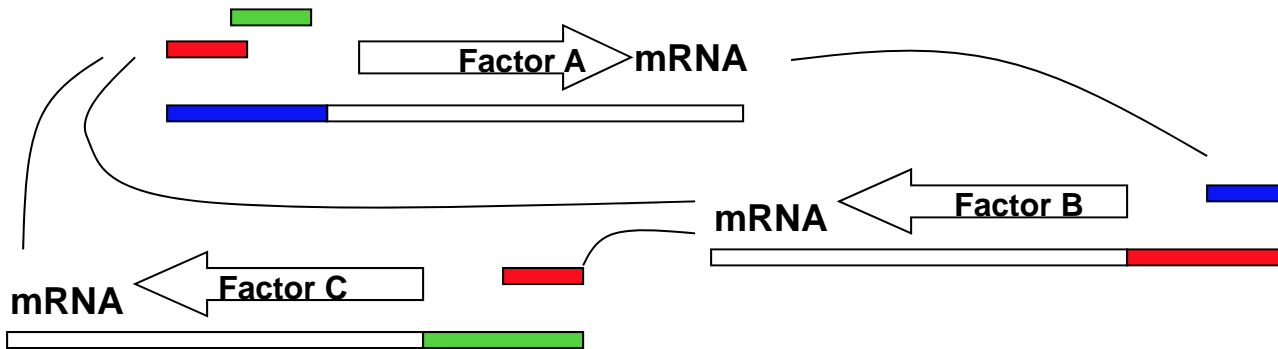
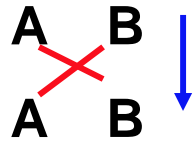
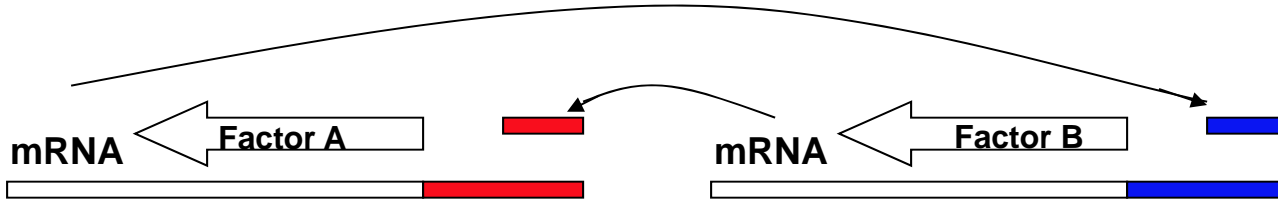
Boolean – only 0/1 values.

Stochastic

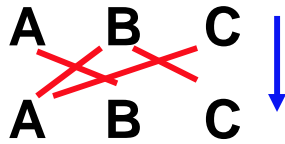
Discrete: the reaction fires after exponential with some intensity $I(X_1, X_2)$ updating the number of molecules

Continuous: the concentrations fluctuate according to a diffusion process.

Boolean Networks



Boolean functions, Wiring Diagrams and Trajectories



A activates B

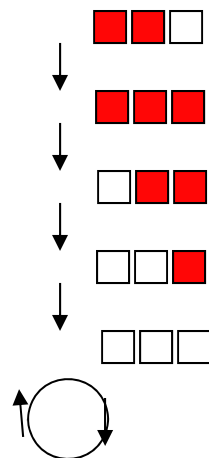
B activates C

A is activated by B, inhibited by (B>C)

Inputs	2	1	1
Rule	4	2	2

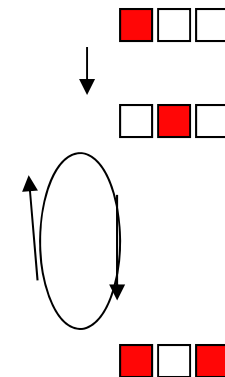
Point Attractor

A	B	C
1	1	0
1	1	1
0	1	1
0	0	1
0	0	0
0	0	0



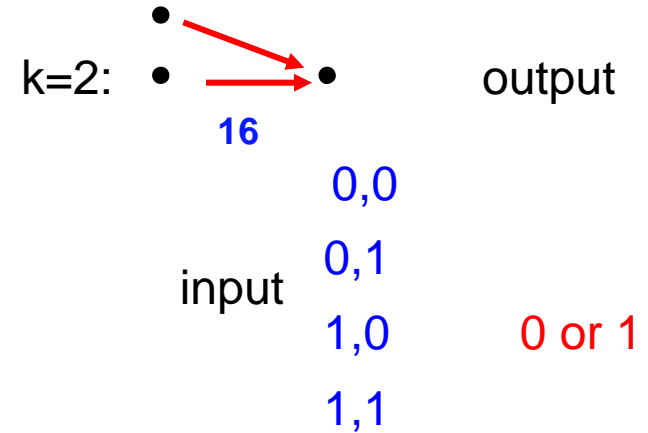
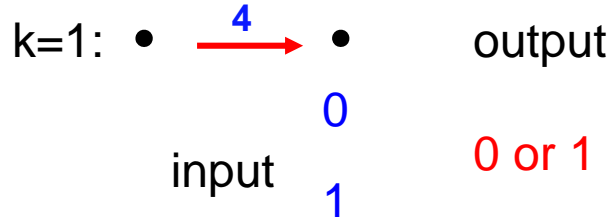
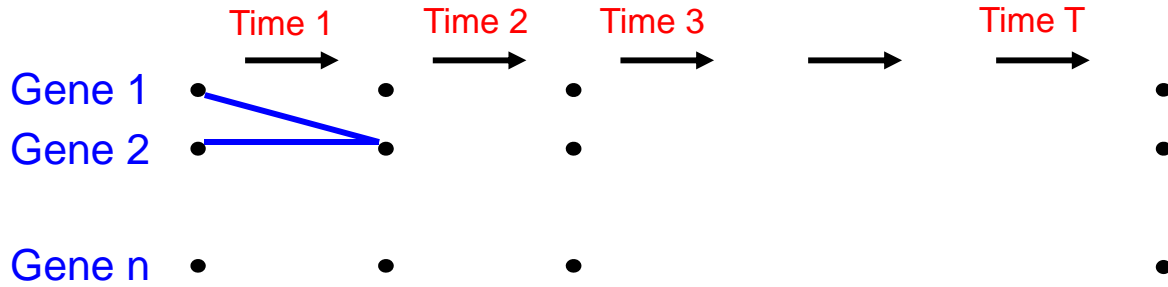
2 State Attractor

A	B	C
1	0	0
0	1	0
1	0	1
0	1	0



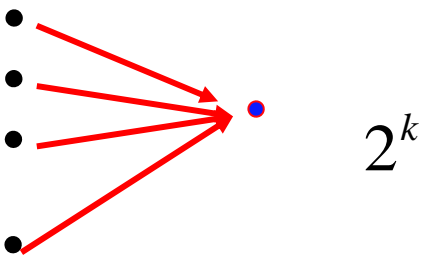
Boolean Networks

R.Somogyi & CA Sniegowski (1996) Modelling the Complexity of Genetic Networks Complexity 1.6.45-64.



A single function:

The whole set:



$$2^{k^k}$$

For each gene dependent on i genes: $\binom{k}{i}$ choices of dependent genes. Number of Boolean Rules $\left(\binom{k}{i} 2^i\right)^k$

Contradiction: Always turned off (biological meaningless) **Tautology:** Always turned on (household genes)

BOOL-1 & 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

Bool-1

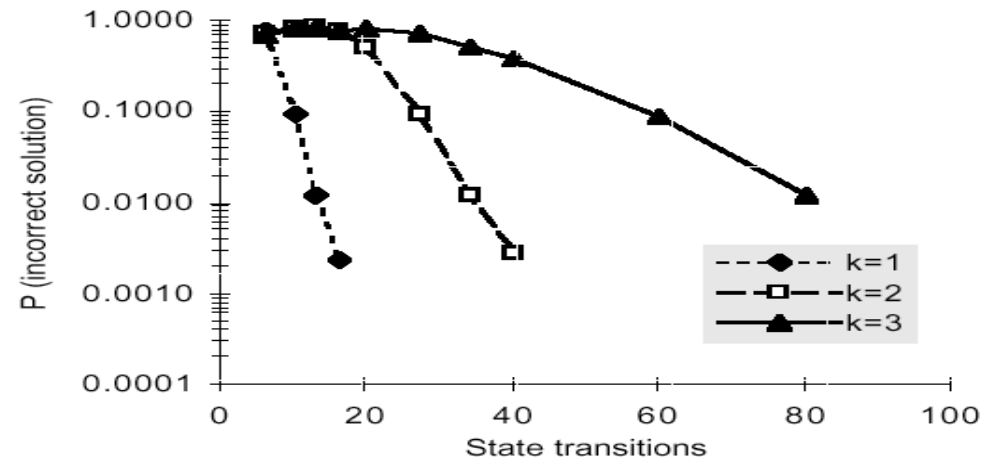
Algorithm

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 .

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

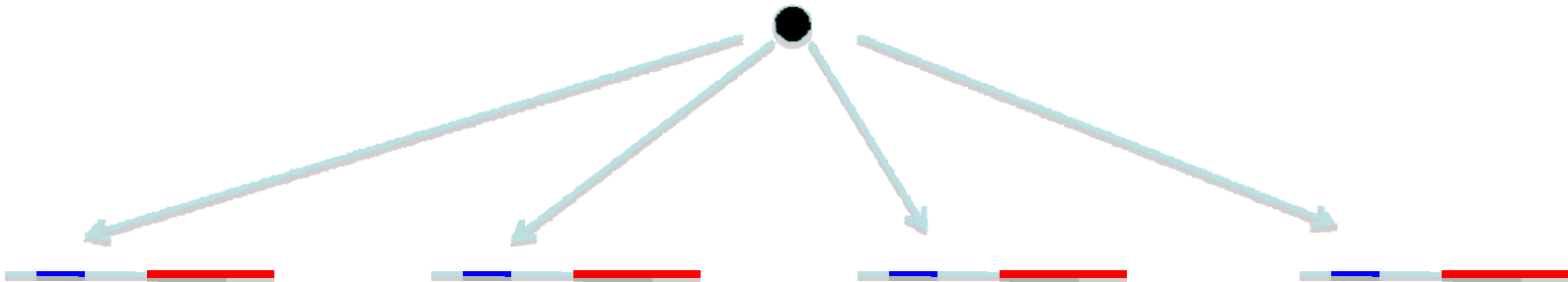


Gaussian Processes

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.

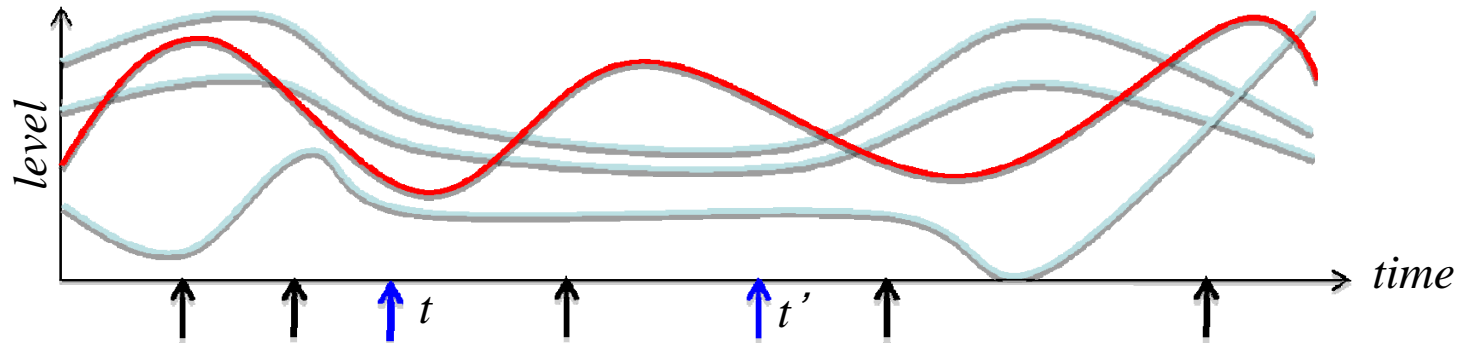


One TF (transcription factor – black ball) ($f(t)$) whose concentration fluctuates over times influence k genes (x_j) (four in this illustration) through their TFBS (transcription factor binding site - blue). The strength of its influence is described through a gene specific sensitivity, S_j . D_j – 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 's

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

Correlation between two time points of different x 's

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

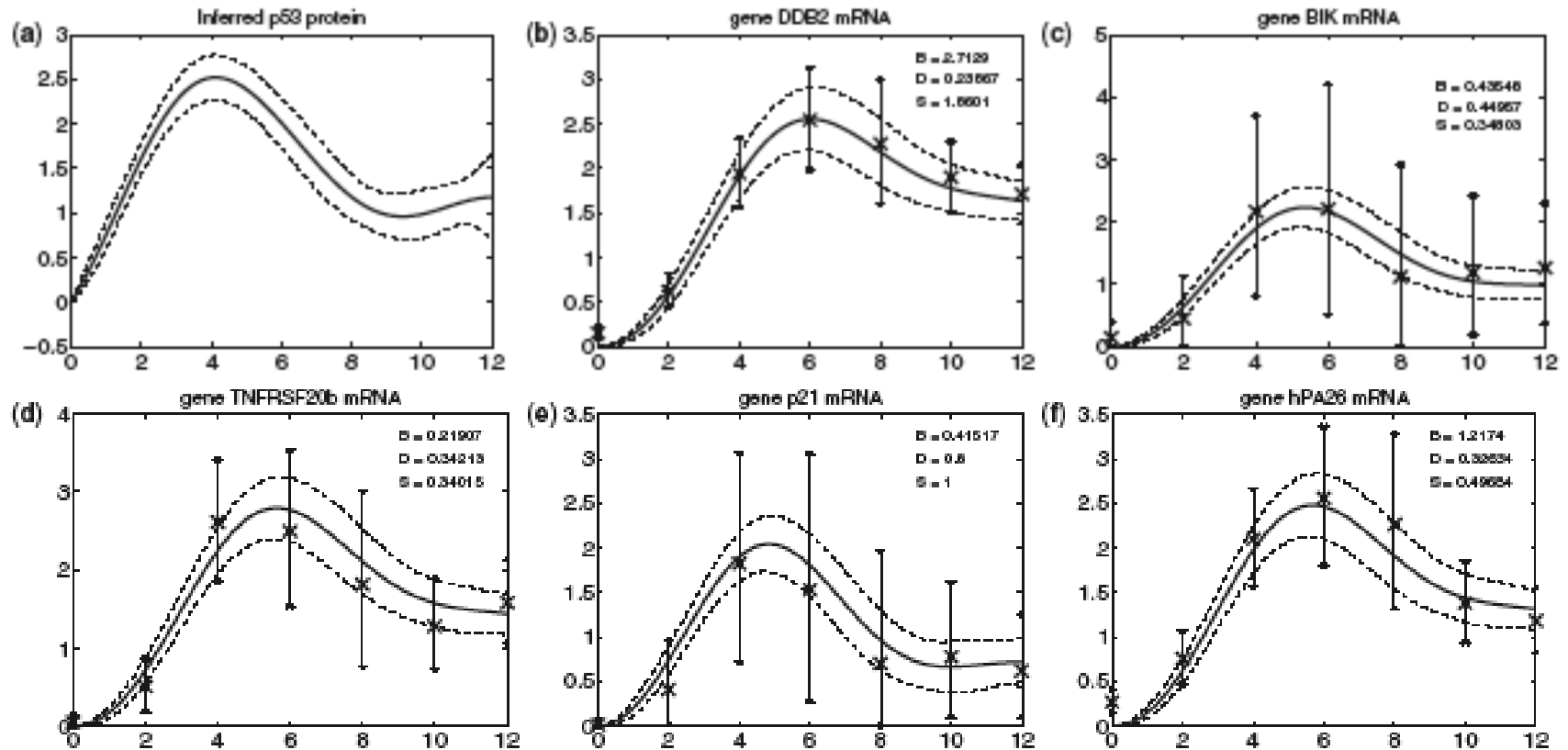
Correlation between two time points of x and f

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

This defines a prior on the observables

Then observe and a posterior distribution is defined

Gaussian Processes



Relevant Generalizations:

Non-linear response function

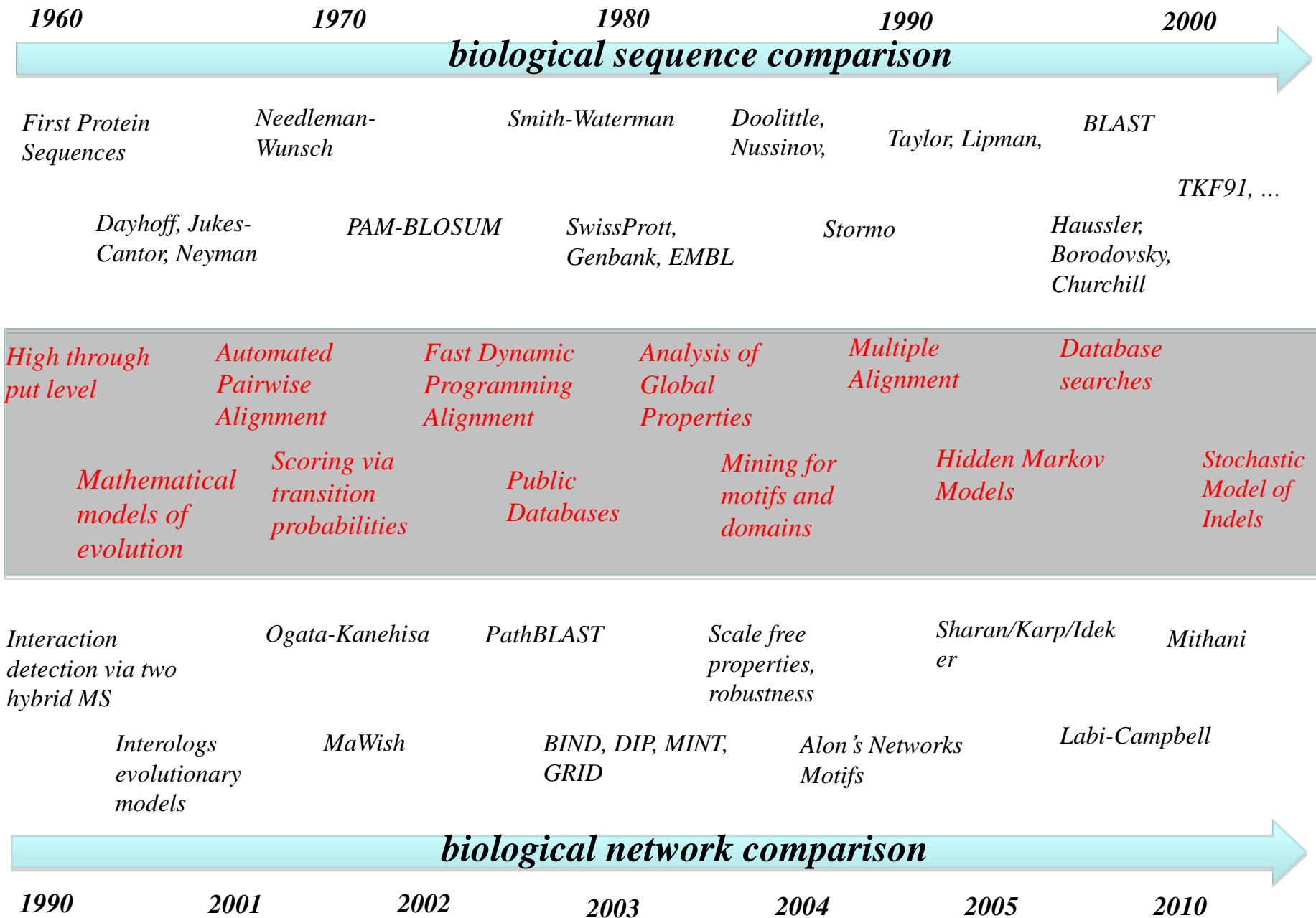
Multiple transcription factors

Network relationship between genes

Observations in Multiple Species

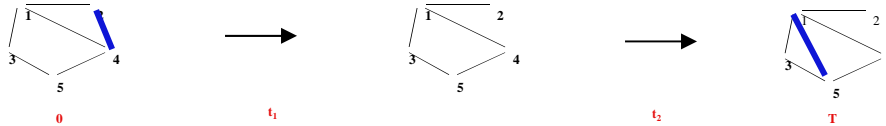
Comments: *Inference of Hidden Processes has strong similarity to genome annotation*

Development of Network/Sequence Analysis



Stochastic Modeling of Network Evolution

Only topology of networks will be considered. I.e. dynamics and continuous parameters often ignored.



What do models of network evolution do?:

Test models

Estimate Parameters in the Evolutionary Process

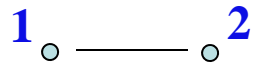
Ancestral Analysis

Framework for Knowledge Transfer

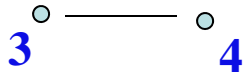
Likelihood of Homologous Pathways

n	Number of all graphs with n nodes	Number of states
1	1	1
2	2	2
3	8	8
4	64	61
5	1024	969
6	32768	31738
7	2097152	2069964
8	268435456	267270033
9	68719476736	68629753641
10	35184372088832	35171000942698

Number of Metabolisms:



+ 2 symmetrical versions



$$P_{\Theta}(\text{graph}_1, \text{graph}_2) = P_{\Theta}(\text{graph}_1) P_{\Theta}(\text{graph}_2 \rightarrow \text{graph}_1)$$



Approaches:

Continuous Time Markov Chains with computational tricks.

MCMC

Importance Sampling



A Model for the Evolution of Metabolisms

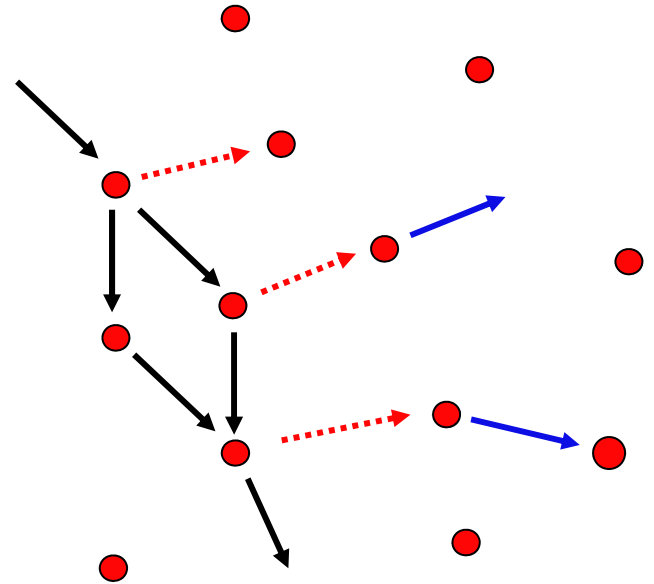
- A given set of metabolites: 
- A given set of possible reactions -
arrows not shown.
- A core metabolism: 
- A set of present reactions - **M**
black and **red** arrows

Restriction R:

A metabolism must define a connected graph

M + **R** defines

1. a set of deletable (dashed) edges **D(M)**: 
2. and a set of addable edges **A(M)**: 



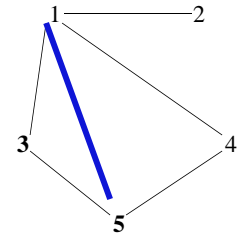
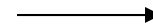
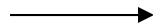
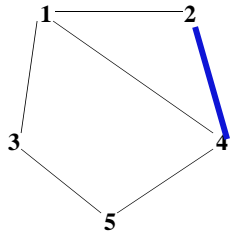
Let μ be the rate of deletion
 λ the rate of insertion

Then

$$\frac{dP(M)}{dt} = \lambda \sum_{M' \in D(M)} P(M') + \mu \sum_{M'' \in A(M)} P(M'') - P(M)[\lambda|D(M)| + \mu|A(M)|]$$

$P(N_1 \rightarrow N_2)$ and Corner Cutting

- How many networks could be visited on "almost shortest" paths?



If $d(N_1, N_2) = k$, then there are 2^k networks are visitable on shortest paths. If 2ε additional steps are allowed, then $2^k (L + L(L-1)/2 + (L(L-1) \dots (L-\varepsilon+1)/\varepsilon!)$ are visitable.

Example. 15 nodes, $L=105$, $\lambda t = \mu t = 0.05$, $\varepsilon = 2$, $d=4$. $P(4) = e^{-.5} \cdot .5^4 / 4! \sim .003$ $P(6) = e^{-.5} \cdot .5^6 / 6! < 10^{-4}$

How can $P(\infty)$ be evaluated?

Can be found in $P(\infty)$ at appropriate rows.

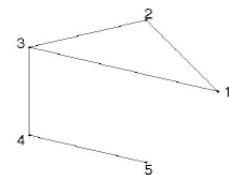
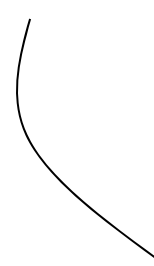
In general not very useful (number of metabolisms).

Simulations

Forward with symmetries could be used in specific cases.

Backward (coupling from the past)

∞

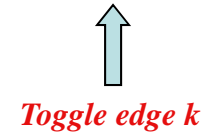


Evolving Networks: MCMC

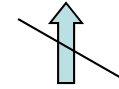
Present pathway:



• *Insertion of an edge pair*



• *Deletion of an edge pair*

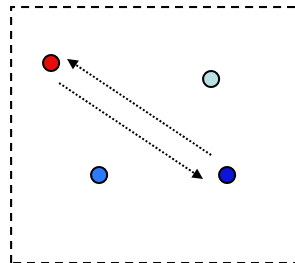


• *Moving of a pair or singles*



• *Metropolis-Hasting integrating of all paths - Green (1995) version:*

Set of paths:



Likelihood - $L(\bullet)$

Probability of going from \bullet to \bullet - $q(\bullet, \bullet)$

J - Jacobian

Acceptance ratio

$$\frac{L(\bullet)q(\bullet, \bullet)}{L(\bullet)q(\bullet, \bullet)} J$$