



# 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*

# *Systems Biology versus Integrative Genomics*

## *Definitions:*

***Systems Biology:*** *Predictive Modelling of Biological Systems based on biochemical, physiological and molecular biological knowledge*

***Integrative Genomics:*** *Statistical Inference based on observations of*

***G*** - *genetic variation*

- *Within species – population genetics*
- *Between species – molecular evolution and comparative genomics*

***T*** - *transcript levels*

***P*** - *protein concentrations*

***M*** - *metabolite concentrations*

***F*** – *phenotype/phenome*

*A few other data types available.*

*Little biological knowledge beyond “gene”*

***Integrative Genomics*** is more top-down and ***Systems Biology*** more bottom-up

***Prediction: Integrative Genomics and Systems Biology will converge!!***

# A repertoire of Dynamic Network Models

To get to networks:

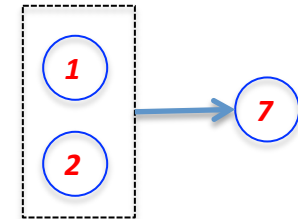
*No space heterogeneity → molecules are represented by numbers/concentrations*

## 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.**

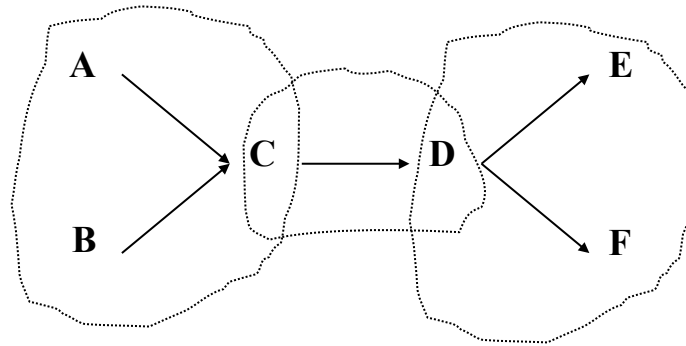
**Boolean – only 0/1 values.**

## **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.*

# 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>1st order ?</i>	<i>Exponential growth/decay as function of some concentration</i>	$2^{36}$	$2^{k*k}$
<i>2nd order ?</i>	<i>Pairwise collision creation: (A, B --&gt; 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*

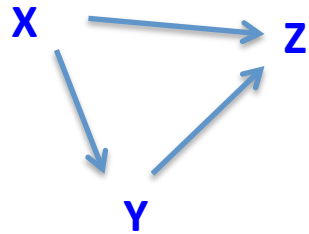
$$a_n = \sum_{k=1}^n (-1)^{k-1} \binom{n}{k} 2^{k(n-k)} a_{n-k}$$

- *Interesting Problems to consider:*

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

# ODEs with Noise

Feed forward loop (FFL) This can be modeled by



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

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

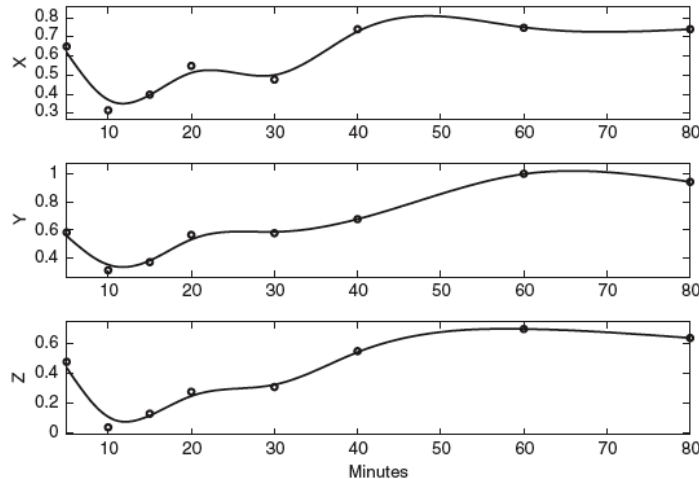
Where

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

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

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

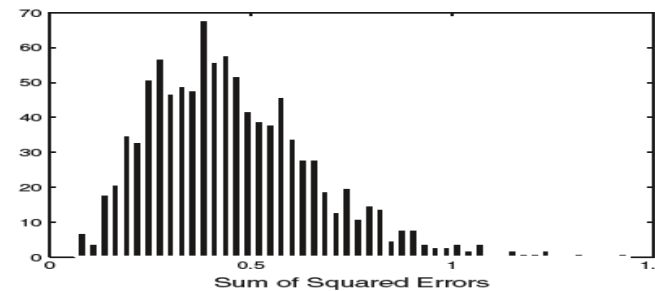
## Data and estimation



Parameters	$\alpha_y$	$\alpha_z$	$K_{xy}$	$K_{xz}$	$K_{yz}$
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

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



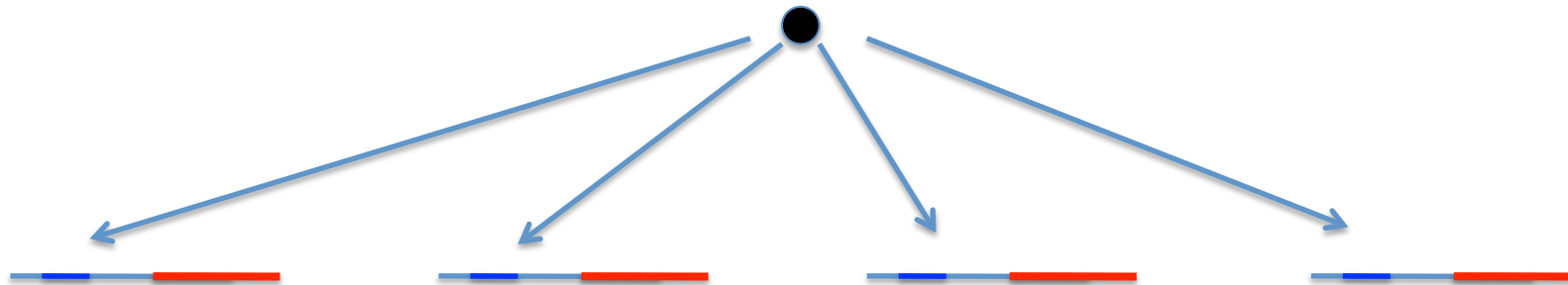
Gene X	Gene Y	Gene Z	SSE	P-values
GCN4	LEU3	ILV5	0.090	0.25
PDR1	PDR3	PDR5	1.17	0.33
GCN4	LEU3	ILV1	0.092	0.34
YLL044W	YER096W	YDR279W	0.84	0.046

# 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  $\mu$  is the  $k$ -dimensional mean and  $\Sigma$  is the  $k \times k$  dimensional covariance matrix.

**Examples:** Brownian Motion: All increments are  $N(0, \Delta t)$  distributed.  $\Delta t$  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.

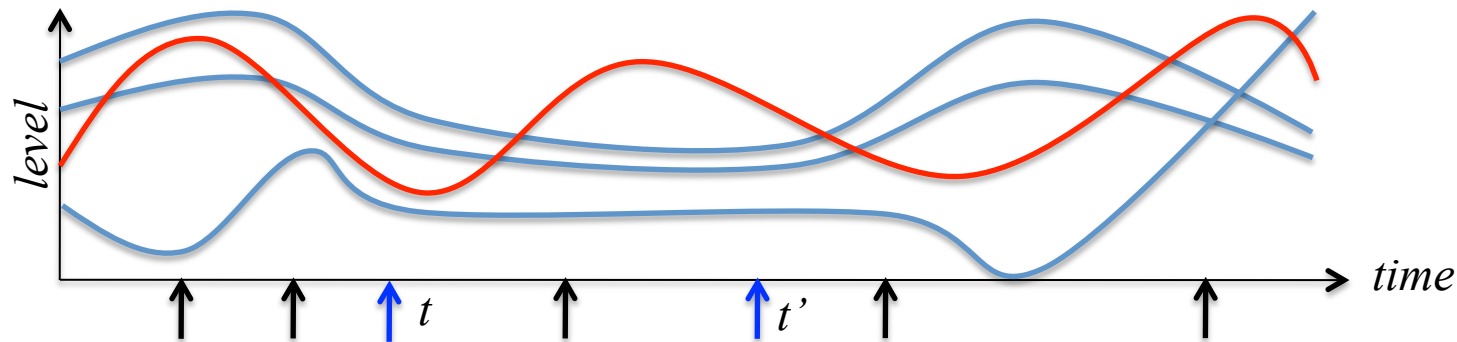


**One TF (transcription factor – black ball) ( $f(t)$ ) whose concentration fluctuates over times influence  $k$  genes ( $x_j$ ) (four in this illustration) through their TFBS (transcription factor binding site - blue). The strength of its influence is described through a gene specific sensitivity,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

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

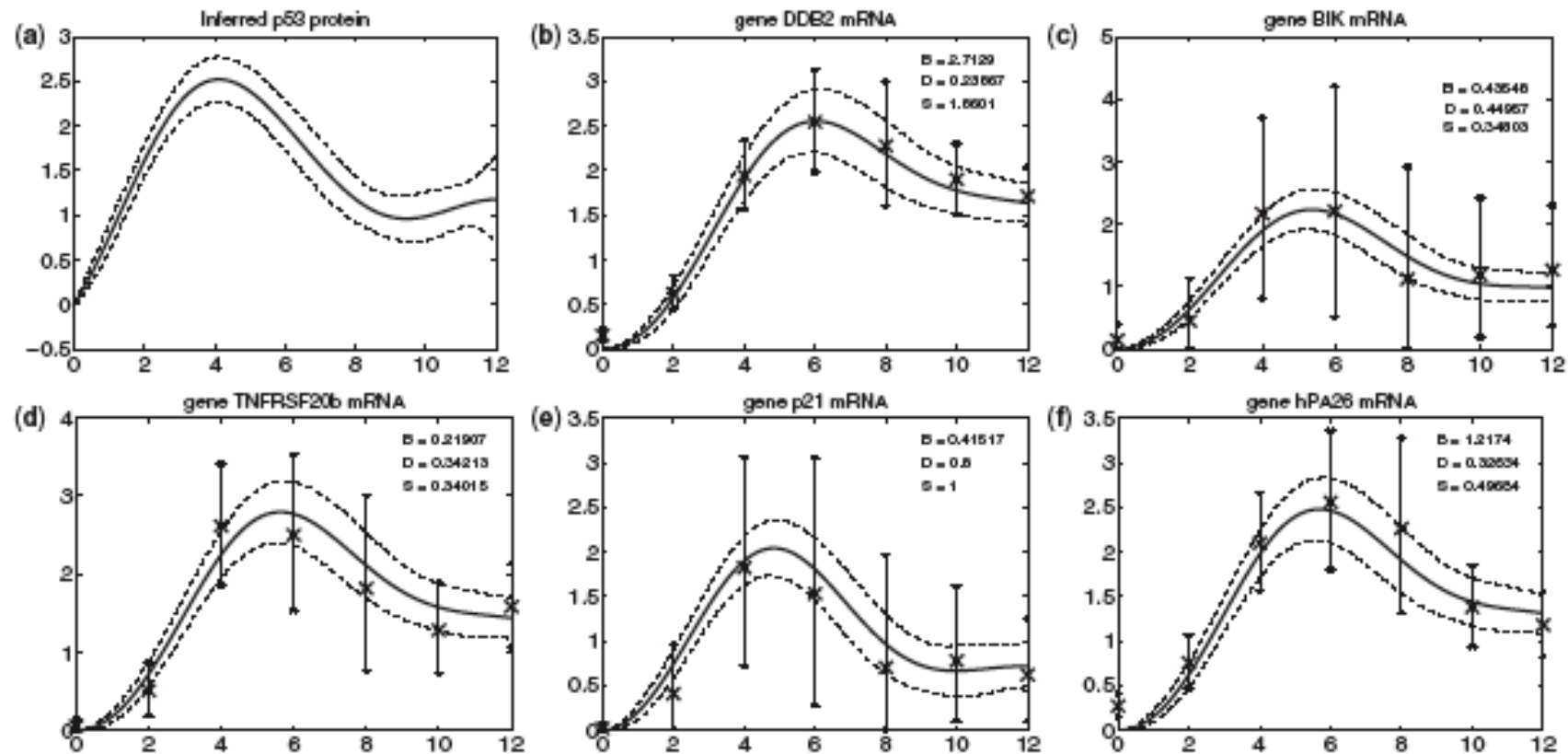
Correlation between two time points of  $x$  and  $f$

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

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*

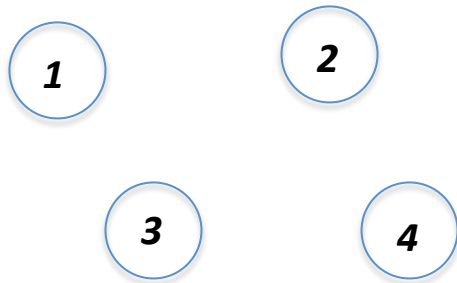
*Network relationship between genes*

*Observations in Multiple Species*

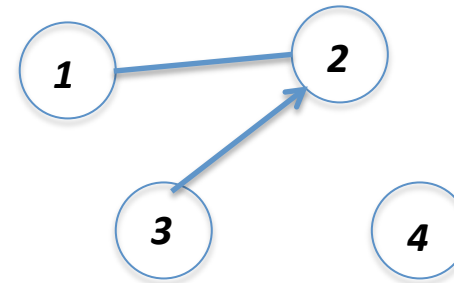
**Comments:** *Inference of Hidden Processes has strong similarity to genome annotation*

# 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**

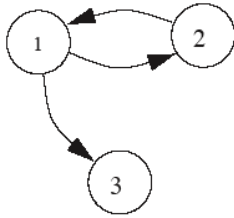
- **Gaussian**

- **Correlation Graphs**

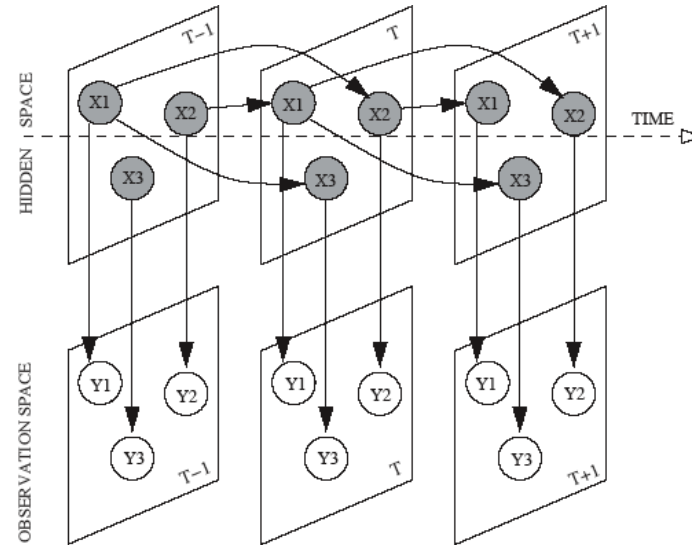
- **Causality Graphs**

# Dynamic Bayesian Networks

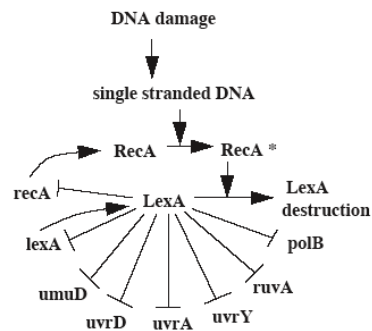
Take a graphical model



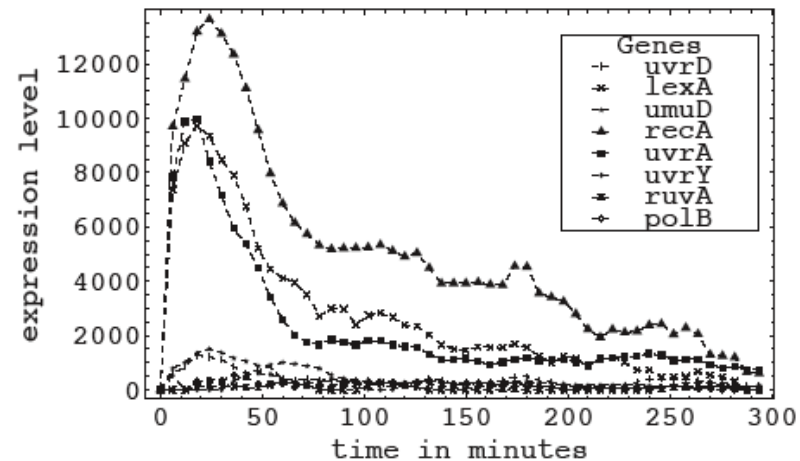
- i. Make a time series of it
- ii. Model the observable as function of present network



Example: DNA repair



Inference about the level of hidden variables can be made



# Equation Discovery I

*Given – Knowledge of System and Data:*

- *Set of labeled quantities*
- *Time Course Data for these quantities*

*Given – Modeling and Inference*

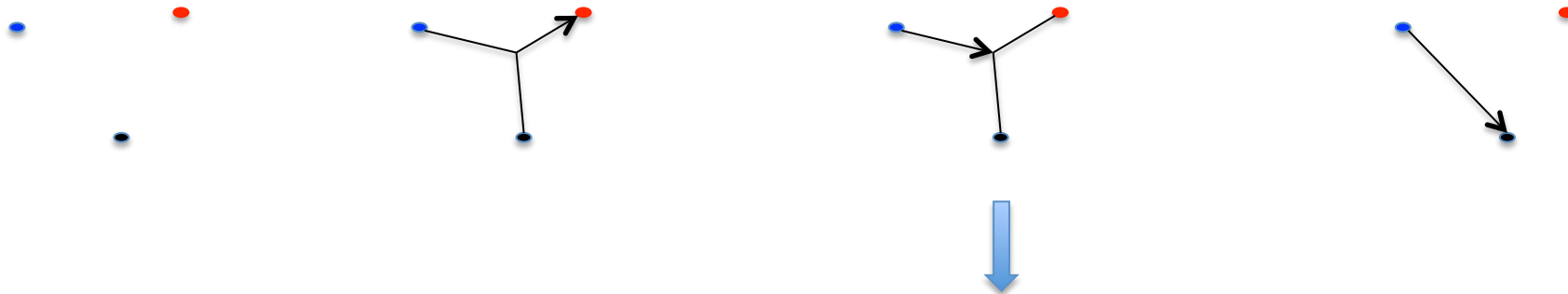
- *Natural Dynamics for the quantities*
- *Optimization Criteria: Root Means Square, Likelihood, Bayesian Integration.*

*Dual Search Problem:*

- *Discrete Search over Equation Structures*
- *Estimating continuous parameters to fit data*

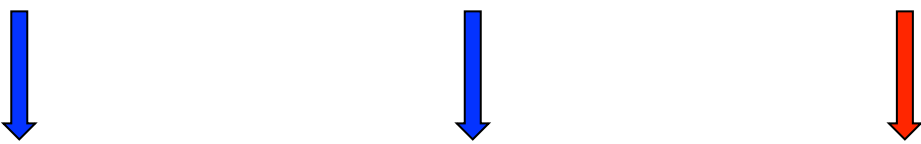
*Dual Search Techniques:*

- *Exhaustive or Heuristic Search*
- *Standard Numerical Optimization*

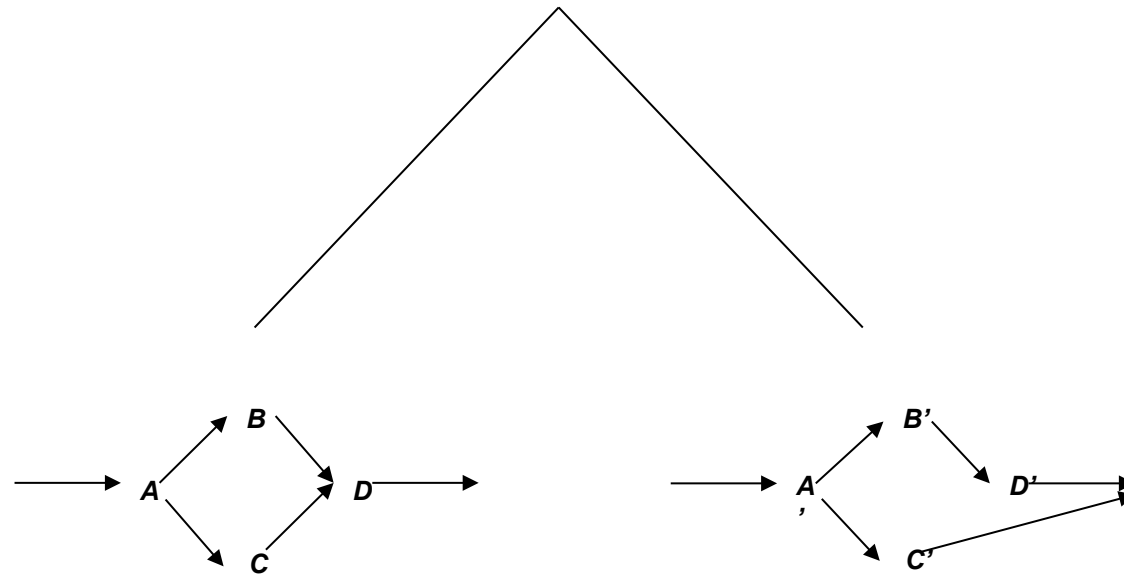


$$a[A], [A_0], [B_0], [C_0]$$

# Inference and Evolution

$$P(D_{mouse}, D_{human}) = \sum_{N_1, N_2} P(D_{human} | N_{human}) P(D_{mouse} | N_{mouse}) P(N_{human}, N_{mouse})$$


Evolve



Infer network



Observe (data)

Human

Mouse

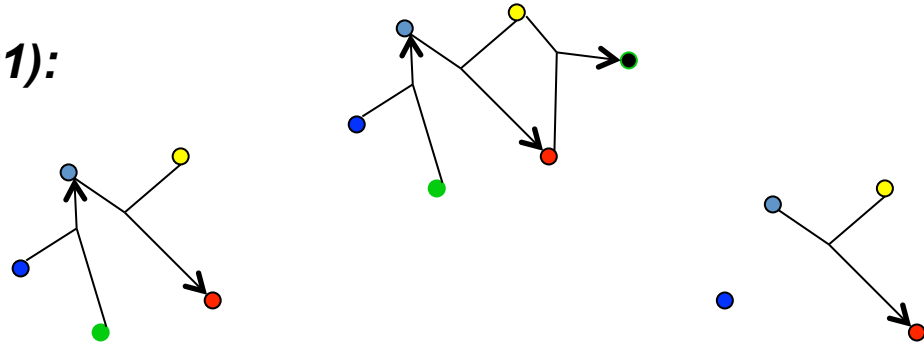
# Suggestion: Evolving Dynamical Systems

- *Goal: a time reversible model with sparse mass action system of order three!!*

**Adding/Deleting components (TKF91):**

**Add rate:  $(k+1)\lambda$**

**Delete rate:  $k\mu$**

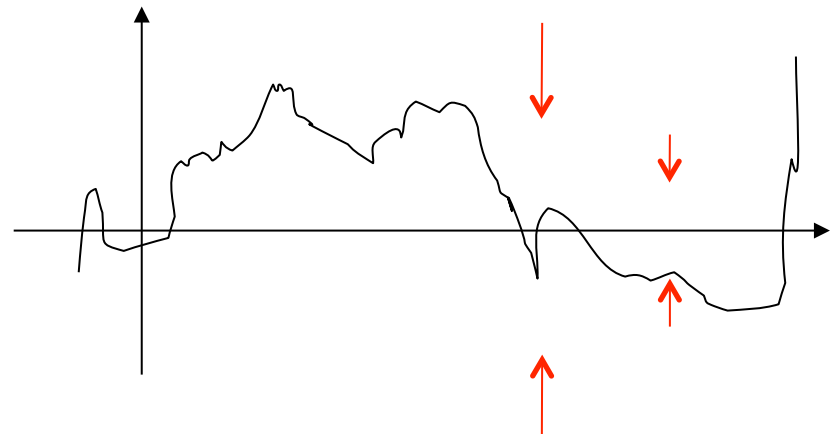


**Adding reactions with birth of component:**

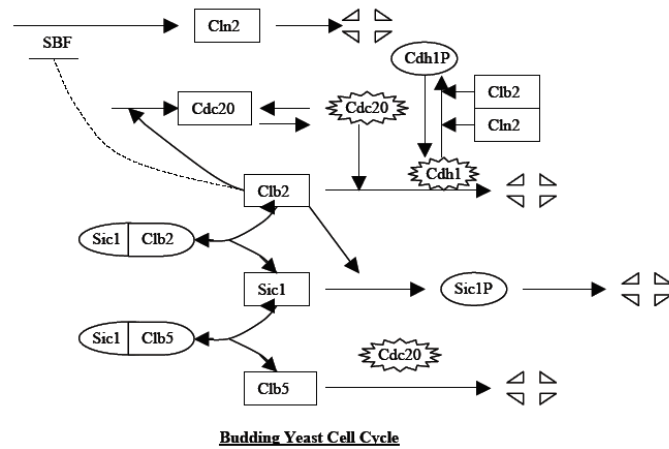
**There are  $3k(k-1)$  possible reactions involving a new-born**

**Reaction Coefficients:**

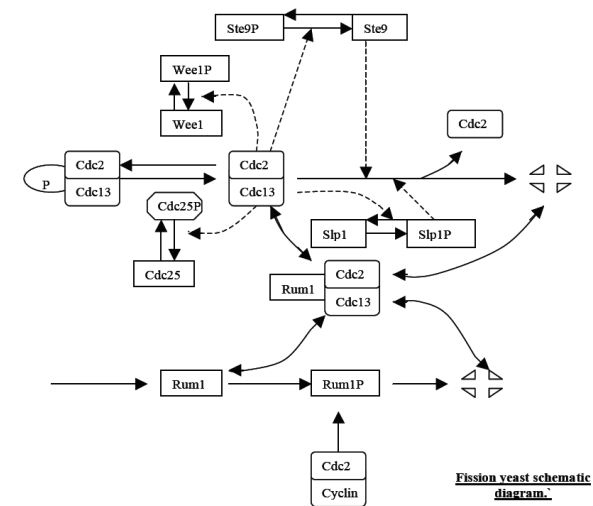
- *Continuous Time Continuous States Markov Process - specifically Diffusion.*
- *For instance Ornstein-Uhlenbeck, which has Gaussian equilibrium distribution*



# Network Example: Cell Cycle



Evolve!



Budding ( $D_1$ )	Clb5	Clb2	Cdh1	Cdc20	Sic1	Cln2	SBF	(N/A)	(N/A)
Fission ( $D_2$ )	Cig2	Cdc13	Ste9	Slp1	Rum1	(N/A)	(N/A)	Cdc25	Wee1

- *What is the edit distance?*
- *Which properties are conserved?*
- *If you only knew Budding Yeast, how much would you know about Fission Yeast?*
- *As  $N_1$  starts to evolve, you can only add reactions. Isn't that strange?*
- *On a path from  $N_1$  to  $N_2$  how close to the minimal has evolution travelled?*
- *What is the number of equation systems possible for  $N_1$ ?*