# Stochastic Network Inference

CRG Summer Course – Modeling for Systems Biology

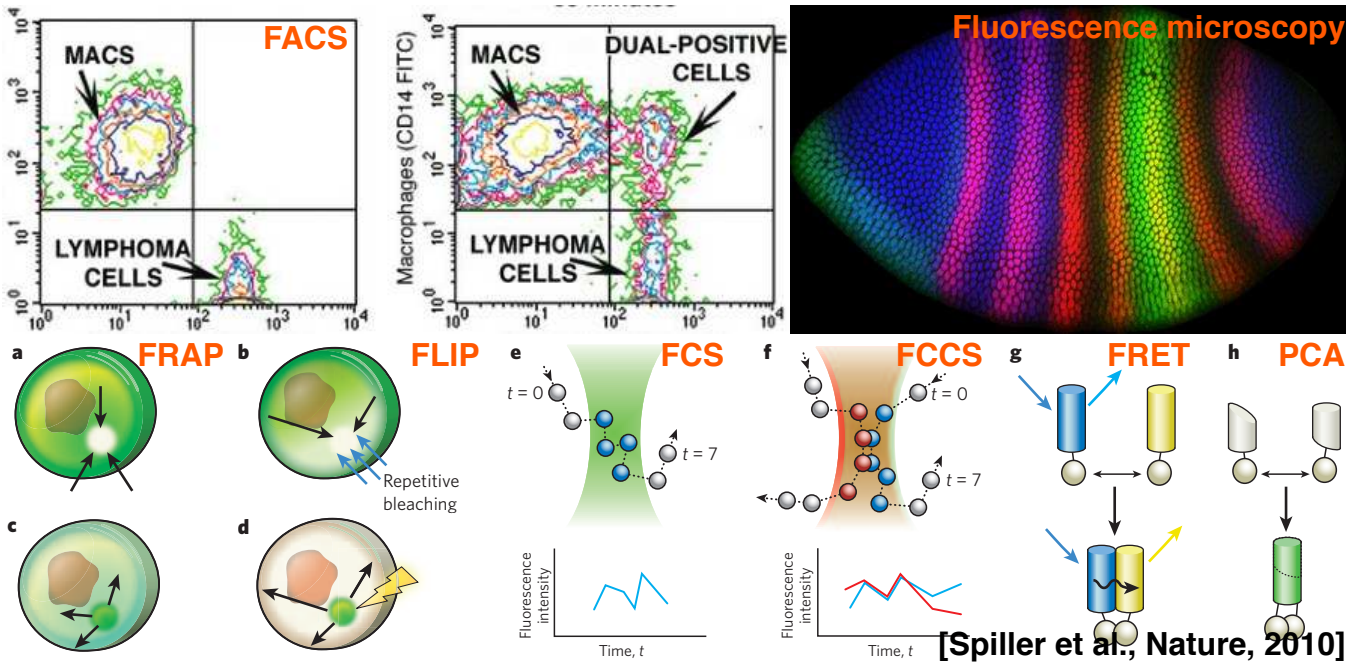
Theodore J. Perkins

www.perkinslab.ca

Ottawa Hospital Research Institute &
University of Ottawa

# Motivation

We've *known* about stochasticity in cellular systems for a long time.




[Spiller et al., Nature, 2010]

Flow cytometry and imaging technologies are allowing us to directly see & measure it!

*Some* of it is functionally significant.

# Outline

- Bayesian networks
  - Definition
  - Parameter learning
  - Structure learning
    - Identifiability
    - Structural limits / penalties
  - Extensions
- Fitting stochastic chemical kinetics models
  - The difficulty with the likelihood function
  - Tian *et al.*'s solution – Gillespie simulation + smoothing
  - Komorowski *et al.*'s solution – Linear noise approximation

# References

- To demonstrate, I'll use a data set and analysis in Madras *et al.* (Stem Cells, 2002) and Nagarajan *et al.* (JTB, 2004).
- An excellent resource for learning more is Kevin Murphy's "A Brief Introduction to Graphical Models and Bayesian Networks" at

`http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html`

# The dataset

- Madras *et al.* (Stem Cells, 2002) measured expression of nine genes during osteoblast differentiation development. What regulates what?

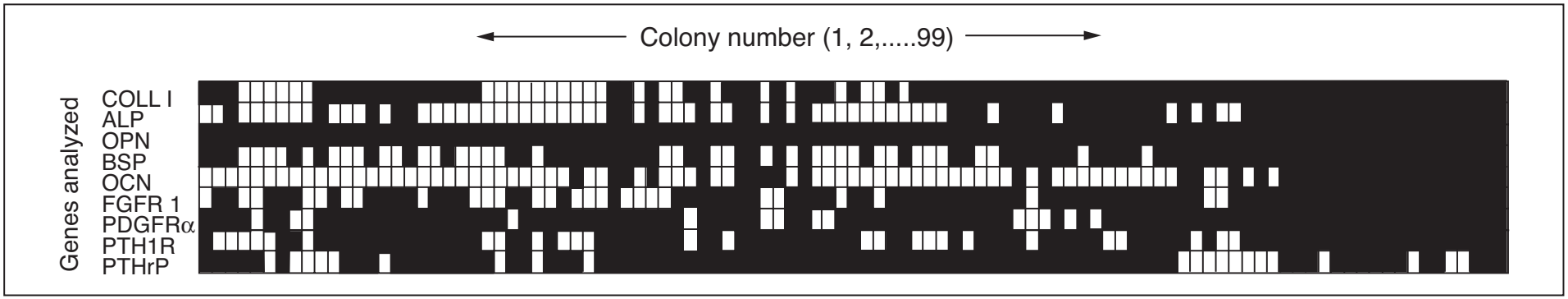



*Figure 2. A schematic of the Southern lineage blot describing expression of nine genes in 99 colonies on a + (expressed; black boxes; value 1) or – (not expressed; white boxes; value 0) scale.*

- The binarized data shows which genes are ON (black) and OFF (white) in 99 (or 103?) replicate colonies.

# Correlation analysis

- Nagarajan *et al.* (JTB, 2004) performed standard correlation analysis, which revealed some apparent relationships.

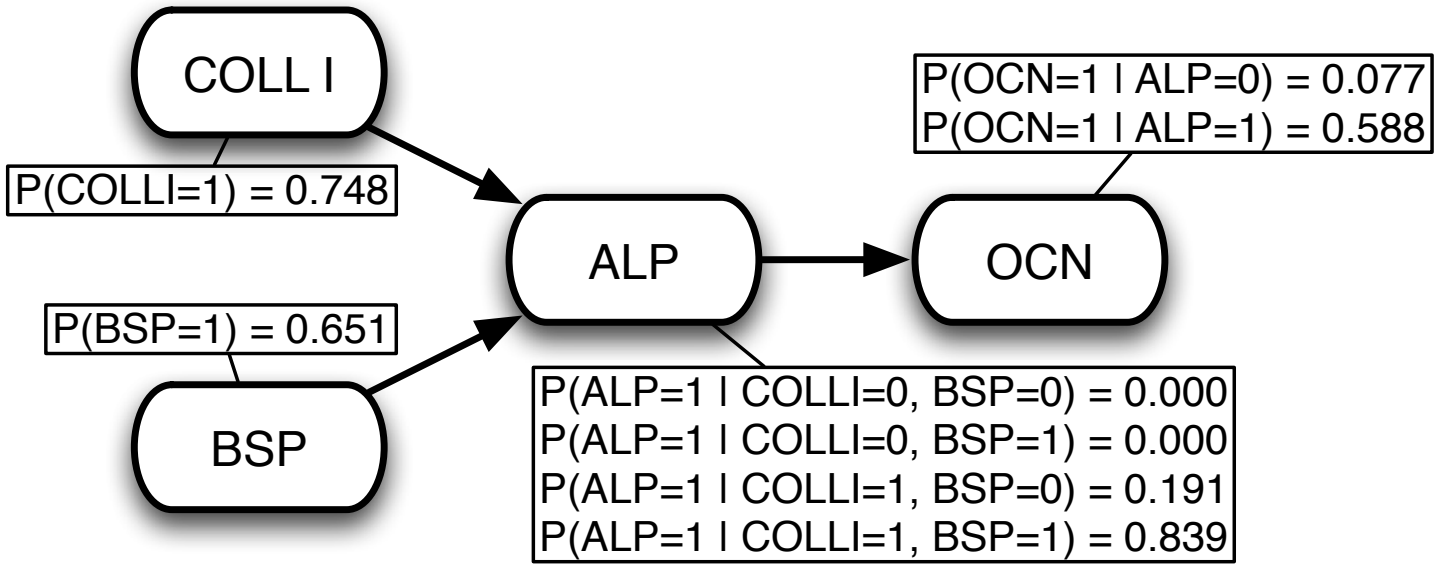
Table 1
Dependencies that had a (power $\sim 0.8$, $\alpha = 0.05$) with ($N_B = 100$) bootstrap realization using the linear correlation, mutual information and Fisher's exact test for the various colony sizes ($n_c = 99$, 66, 33, 22 and 11). The genes *COLL, ALP, BSP, OCN, FGFR1, PTH1R*, *PDGFα* are represented by C, A, B O, F, P and D in the table

| Linear correlation | | | | | Mutual information | | | | | Fisher's exact test | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 99 | 66 | 33 | 22 | 11 | 99 | 66 | 33 | 22 | 11 | 99 | 66 | 33 | 22 | 11 |
| **C–A** | **C–A** | **C–A** | **C–A** | — | **C–A** | **C–A** | **C–A** | **C–A** | — | **C–A** | **C–A** | **C–A** | **C–A** | — |
| **C–B** | — | — | — | — | — | — | — | — | — | **C–B** | — | — | — | — |
| **C–O** | **C–O** | — | — | — | **C–O** | **C–O** | **C–O** | — | — | **C–O** | **C–O** | — | — | — |
| **C–F** | **C–F** | — | — | — | **C–F** | — | — | — | — | **C–F** | **C–F** | — | — | — |
| **C–P** | — | — | — | — | — | — | — | — | — | **C–P** | — | — | — | — |
| **A–B** | **A–B** | **A–B** | **A–B** | — | **A–B** | **A–B** | **A–B** | — | — | **A–B** | **A–B** | **A–B** | **A–B** | — |
| **A–O** | **A–O** | **A–O** | **A–O** | — | **A–O** | **A–O** | **A–O** | **A–O** | — | **A–O** | **A–O** | **A–O** | **A–O** | — |
| **A–F** | — | — | — | — | — | — | — | — | — | **A–F** | — | — | — | — |
| **A–P** | **A–P** | — | — | — | **A–P** | — | — | — | — | **A–P** | **A–P** | — | — | — |
| **B–O** | **B–O** | **B–O** | — | — | **B–O** | **B–O** | **B–O** | **B–O** | — | **B–O** | **B–O** | **B–O** | — | — |
| **O–P** | **O–P** | — | — | — | **O–P** | **O–P** | **O–P** | — | — | **O–P** | **O–P** | — | — | — |
| **F–P** | — | — | — | — | — | — | — | — | — | — | — | — | — | — |

- But causation was not clear, so they wanted to fit a Bayesian network.

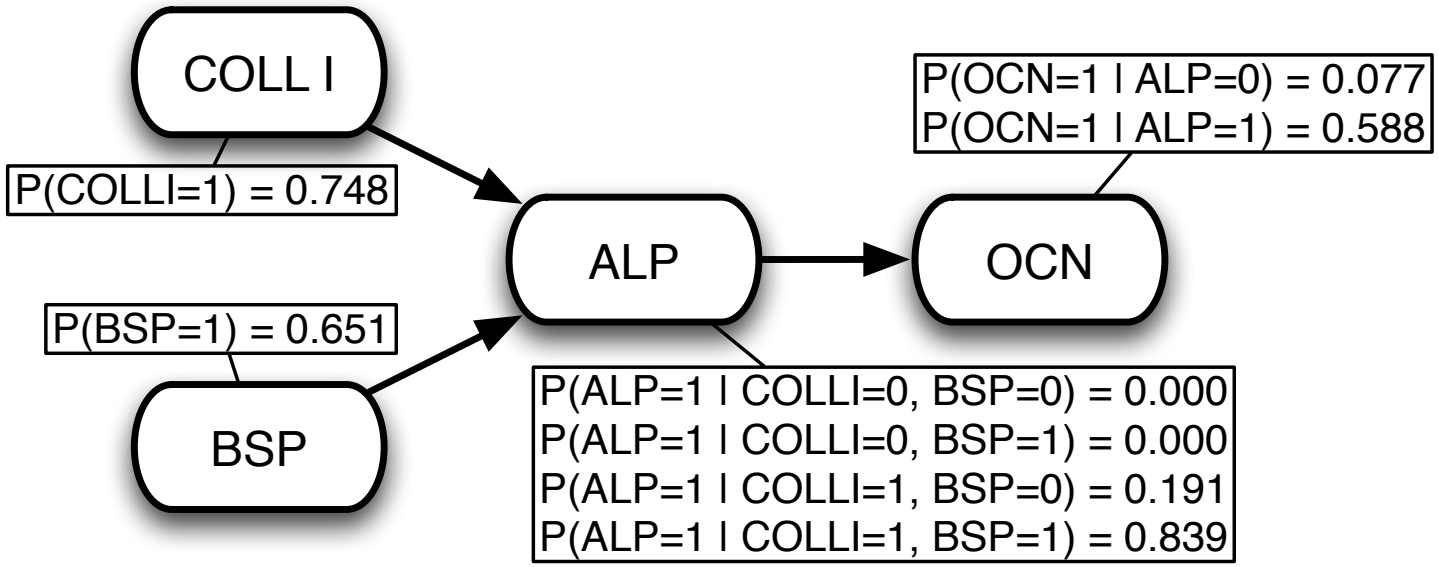
# Bayesian networks

- A Bayesian network describes the behavior of random variables in terms of the values of other random variables (or none).
- The *structure* of the network is the links that say what influences what. It is a directed, acyclic graph.
- The *parameters* of the network specify the nature of those influence (below, given by conditional probability tables, CPTs).



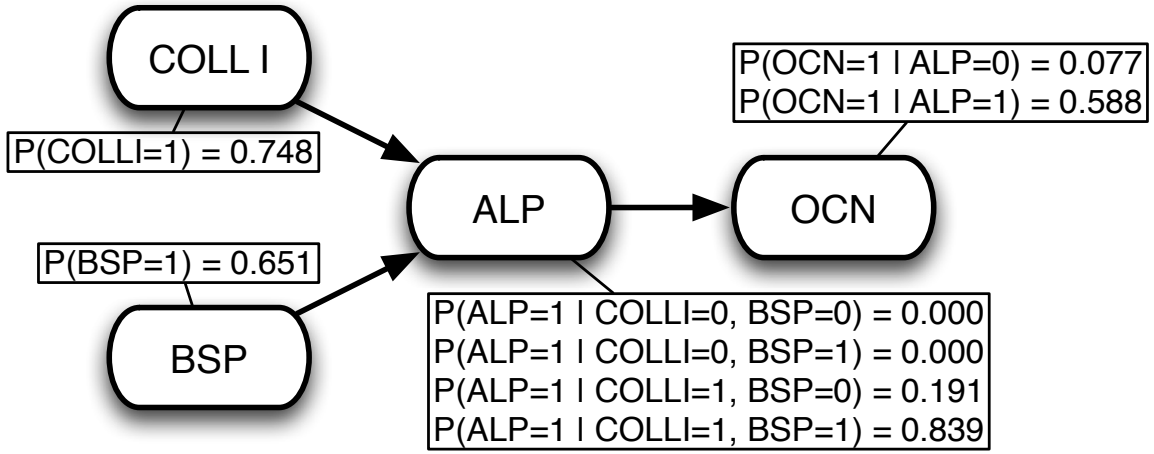
# Computing the probability of a colony's variables' values

The meaning of the network is that it provides a way of computing the probability of particular values for each of the variables.




$$
\begin{aligned}
& P(C=1, B=0, A=1, O=1) \\
= \quad & P(C=1)P(B=0)P(A=1|C=1, B=0)P(O=1|A=1) \\
= \quad & 0.748 \times (1-0.651) \times 0.191 \times 0.588
\end{aligned}
$$

# Aside: Bayes nets can be used to answer other "queries"




Given a network—structure and parameters—we can generally compute the probability of any values for certain variables, given values of other variables. For example:

- What is $P(O = 1)$?
- What is $P(A = 1)$?
- What is $P(A = 0|B = 1)$?
- What is $P(C = 1|O = 1)$?

## Computing the probability of a whole dataset

- Suppose $(C_i, B_i, A_i, O_i)$ for $i = 1 \ldots 99$ are the data for a whole set of colonies.
- We can compute the probability of that whole dataset by multiplying the probabilities of the individual colonies:

$$\begin{aligned} & P(\{(C_i, B_i, A_i, O_i)\}_{i=1}^{N}) \\ = & \prod_{i=1}^{N} P(C_i, B_i, A_i, O_i) \\ = & \prod_{i=1}^{N} P(C_i)P(B_i)P(A_i|C_i, B_i)P(O_i|A_i) \end{aligned}$$

- To estimate a Bayesian network from a data set, we could invoke the Maximum Likelihood "Principle" – choose the network under which the data has highest probability!

# Parameter estimation

- Suppose we knew the structure, where do we get the parameters?
- Max likelihood $\Rightarrow$ just count empirical frequencies!

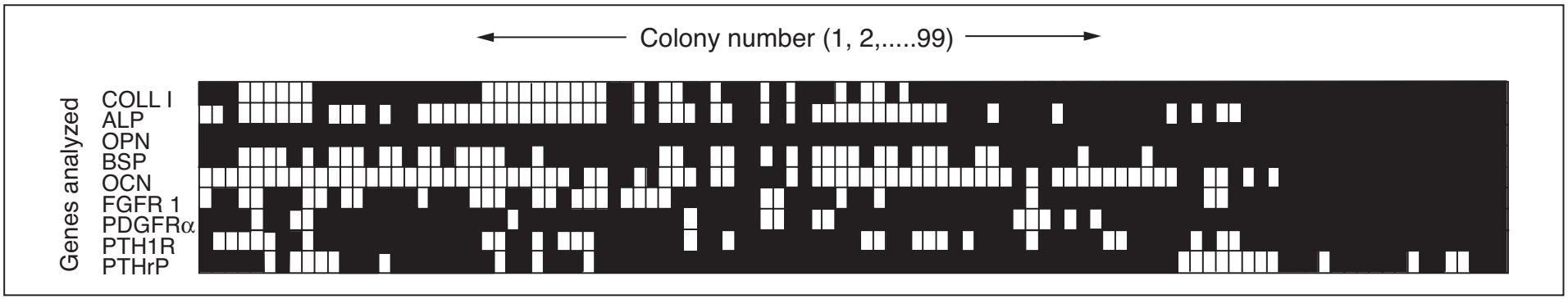
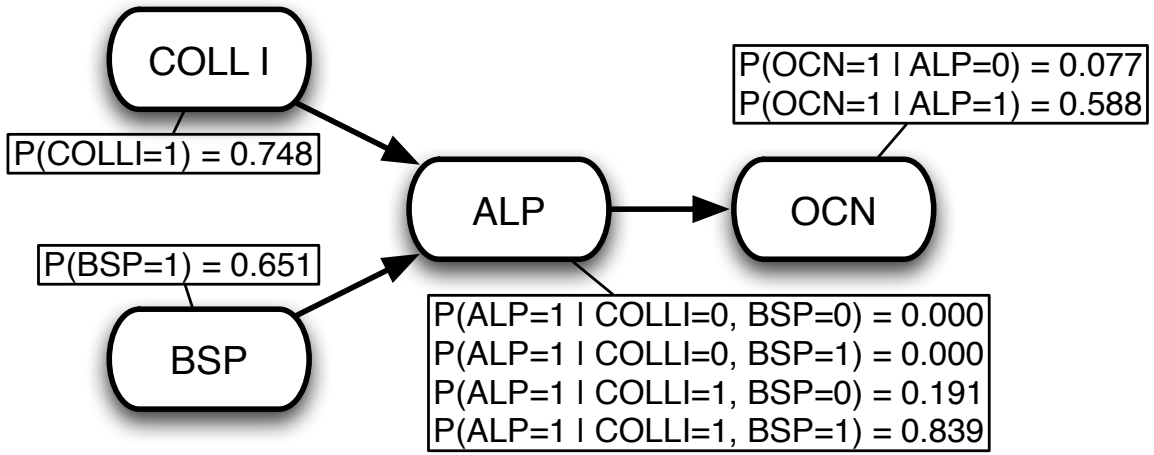
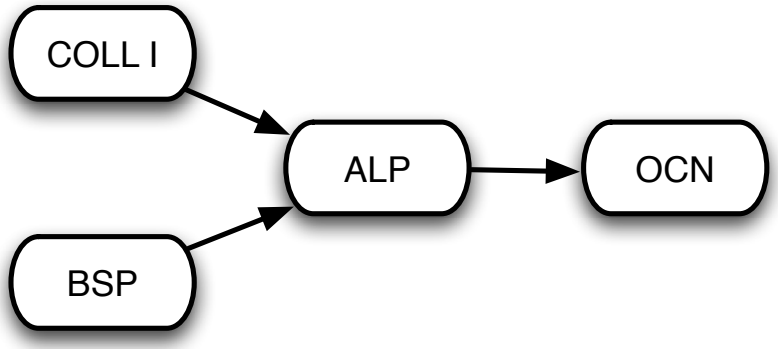


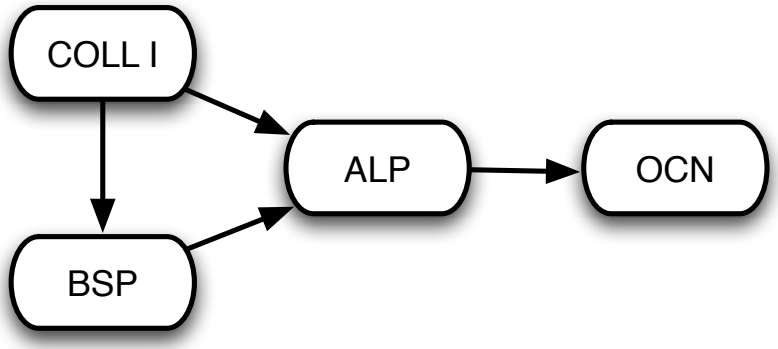

*Figure 2. A schematic of the Southern lineage blot describing expression of nine genes in 99 colonies on a + (expressed; black boxes; value 1) or – (not expressed; white boxes; value 0) scale.*



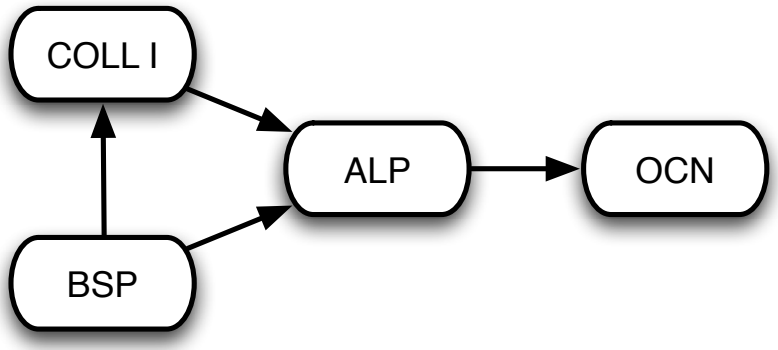
# Structure identification is tricky. . . for several reasons

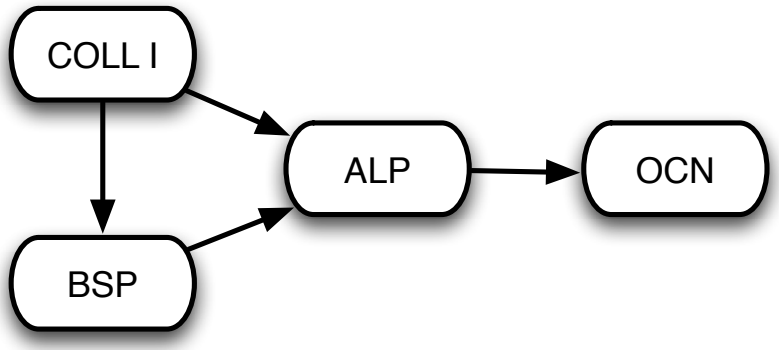
More complex models will always fit the data better (or no worse).




Biologically/causally different models can fit the data equally well.




$$P(C,B)P(A|C,B)P(O|A) = P(B)P(C|B)P(A|C,B)P(O|A) = P(C)P(B|C)P(A|C,B)P(O|A)$$

# How to get a network from a data matrix?

- *Score* a possible network structure by:
  - Compute the parameters from the data.
  - Compute the probability of the data under those parameters. (Intuitively, the data is more probable under a network the has the correct depenencies.)
  - Limit or penalize network complexity. (Because a more complex network will almost always fit the data better. There are a number of options available.)
- *Search* for a network structure that has the best score.
  - If the number of variables is small, all possible network structures can be checked.
  - Otherwise, there's a bajillion heuristic search procedures you can use.

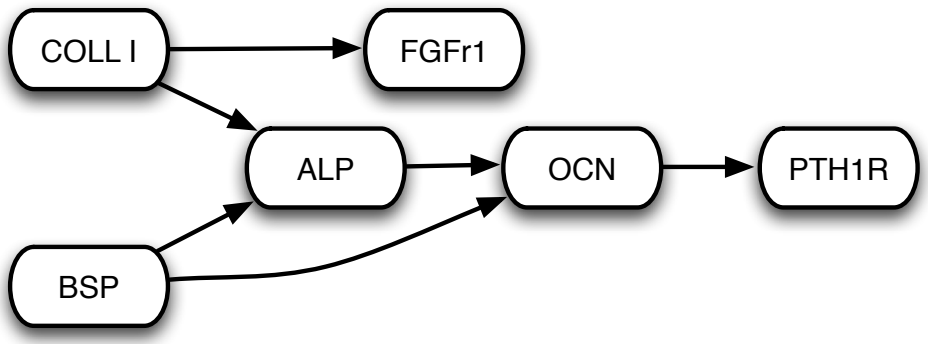
# Back to Nagarajan et al. (JTB, 2004)

They used a Markov chain Monte Carlo approach to search for network structures – basically, a random walk in the space of possible structures, with steps to higher-scoring structures favored.

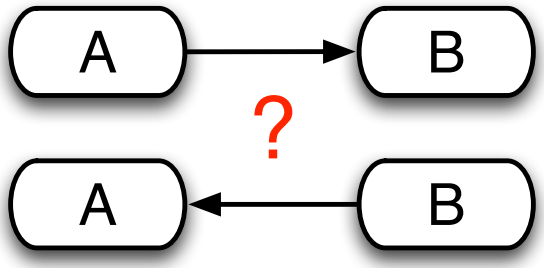
- Limit of three inputs per variable.
- "Bayesian" penalty for model complexity.

The most common links observed in that whole walk were:

| | |
|---|---|
| COLLI $\rightarrow$ ALP | 100% |
| ALP $\rightarrow$ OCN | 100% |
| BSP $\rightarrow$ ALP | 100% |
| OCN $\rightarrow$ PTH1R | 94% |
| COLLI $\rightarrow$ FGFR1 | 85% |
| BSP $\rightarrow$ OCN | 75% |

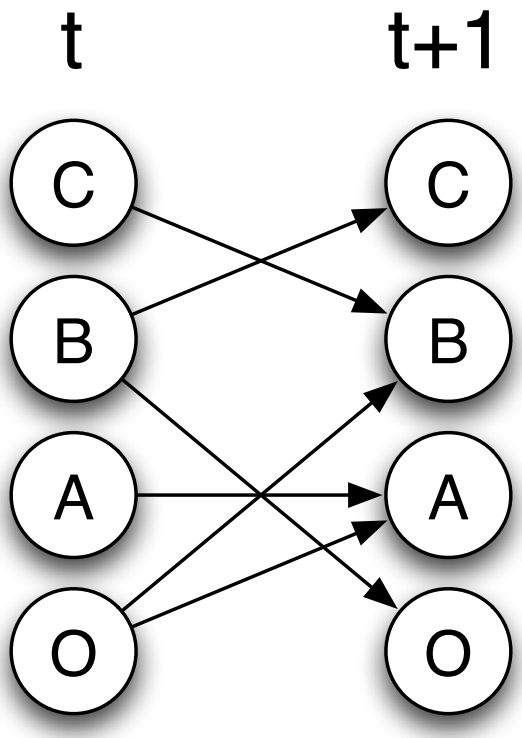


# Comment: Identification via perturbations



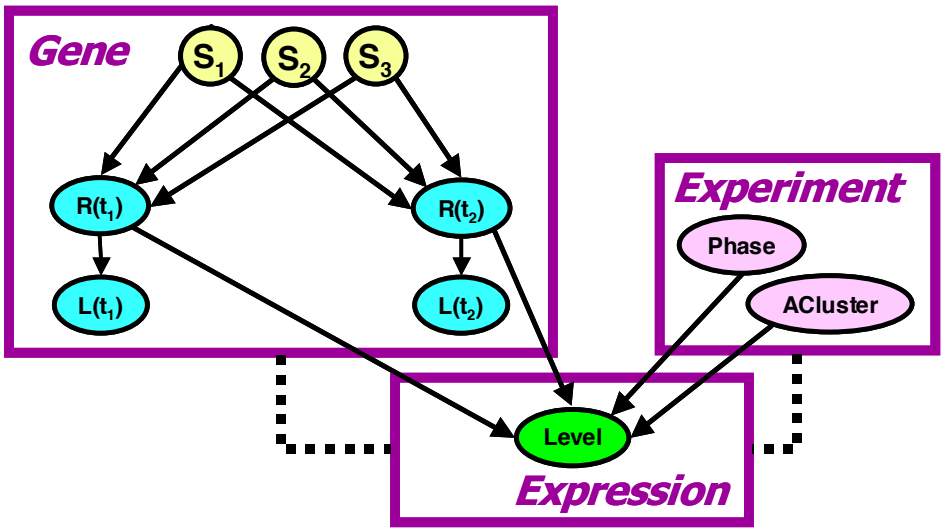

- Suppose we delete gene A. Does gene B change?
- Network structure can *always* be determined fully if we have the ability to fix variables to certain values.

## Comment: Dynamic Bayesian networks




- A dynamic Bayesian network describes the influences of variables at time $t$ on the same variables at time $t + 1$.
- It allows "feedback" / mutual regulation.
- Parameters and structure can be fit readily from time series data.

# Variables don't have to be genes' expression




[Segal *et al.* (RECOMB, 2002)]

- S = promoter sequence of gene
- R = does each TF regulate the gene
- L = chip-chip data on TF-promoter binding
- Level = expression
- Phase / ACluster = experiment information

## Bayesian networks summary

- Bayesian networks are probabilistic models whose structure describes influences between variables, and whose parameters define the nature of those influences.

\+ They are predictive, and directed.

\+ They can be fit to a data matrix, uncovering potential regulatory relationships.

– Finding the right network structure is computationally difficult.

  - Your answer may not be causal, due to equivalent network structures
  - But you can do it with the right interventions.

Questions?

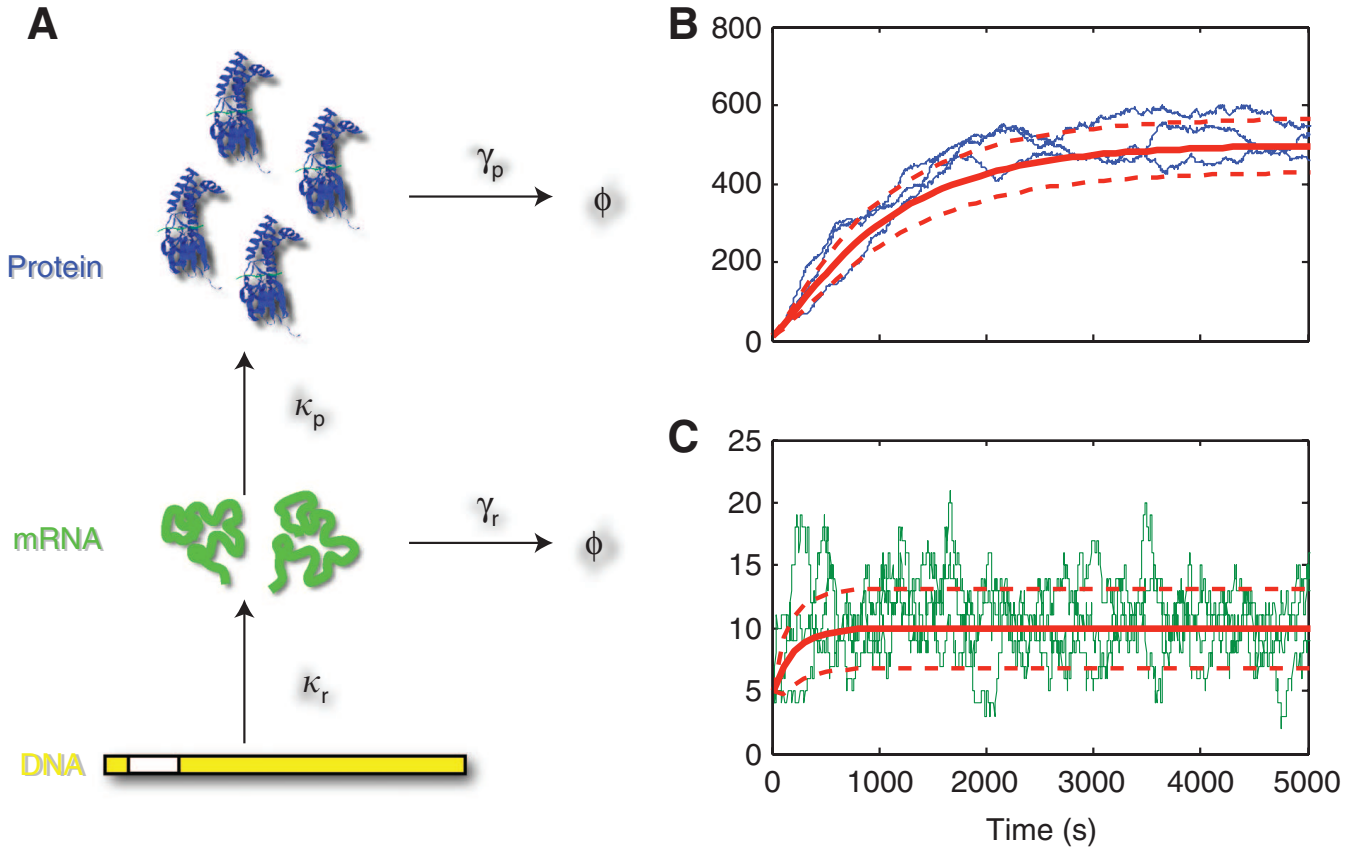
# Outline

- Bayesian networks
  - Definition
  - Parameter learning
  - Structure learning
    - Identifiability
    - Structural limits / penalties
  - Extensions
- Fitting stochastic chemical kinetics models
  - The difficulty with the likelihood function
  - Tian *et al.*'s solution – Gillespie simulation + smoothing
  - Komorowski *et al.*'s solution – Linear noise approximation

# Recall: Stochastic Chemical Kinetic Models (SCKMs)

A stochastic chemical kinetic model describes reactions between chemical species, and defines the rates of those reactions.

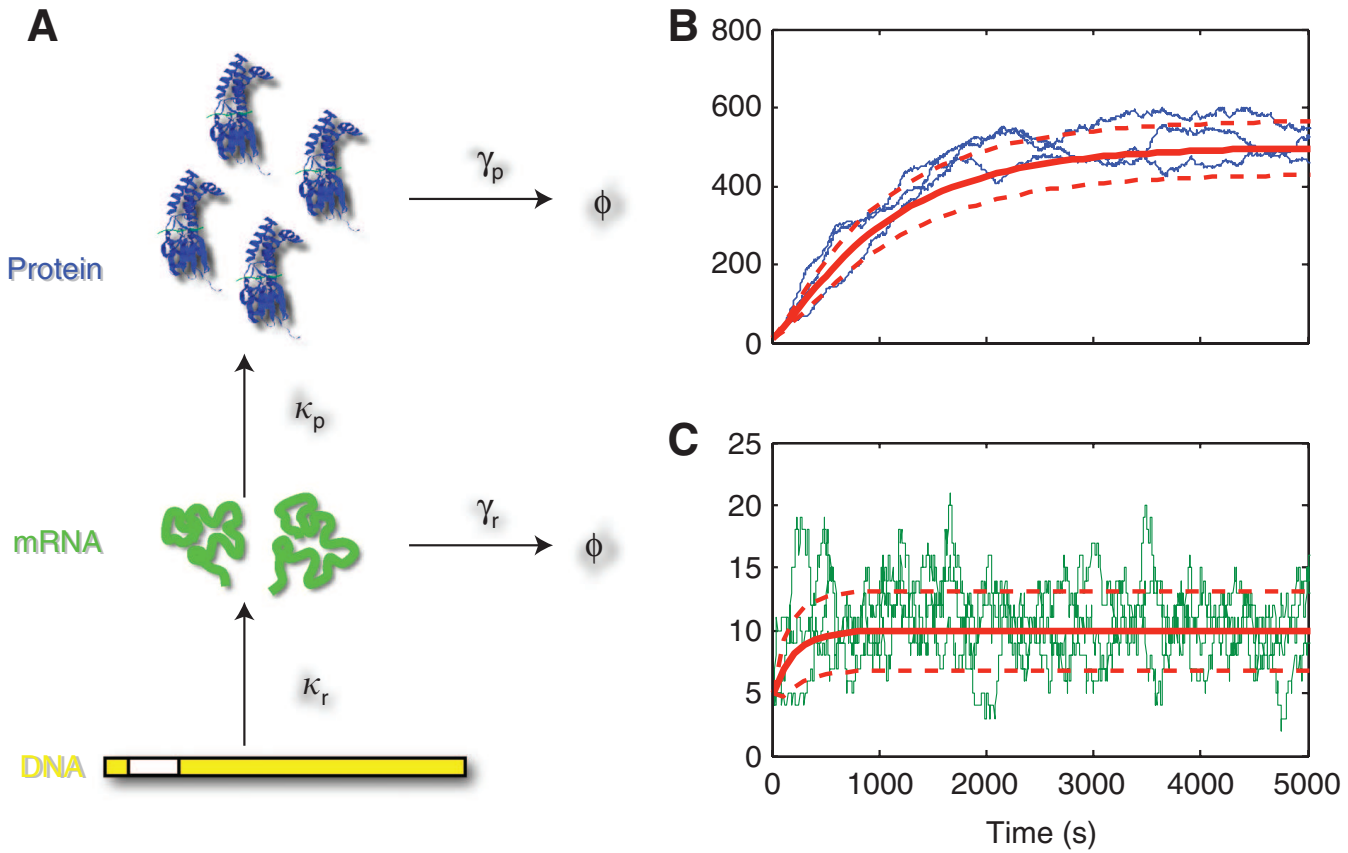
Reactions happen at random, real-valued times, changing the state of the system.




[Munsky *et al.*, 2009]

# Parameter estimation

Suppose we observed trajectories (on the right). How can we estimate the parameters (on the left). (We'll assume model form is known.)




[Munsky *et al.*, 2009]

# The problem of measuring goodness-of-fit

Wilkinson (NRG, 2009) made the comment:

"When tuning the parameters of stochastic models, there is no obvious 'distance' function to optimize, owing to the fact that the likelihood function does not have a simple analytically tractable form."

Why? What does that really mean?

# A first-principles approach to goodness-of-fit

- An SCKM is a statistical model.
- For concreteness, suppose we have a single cell time-series: $X(t_0), \ldots, X(t_M)$.
- Suppose our SCKM has parameters $\theta$ that we want to fit to that data.
- The obvious thing to do is to write down the probability of the data given the parameters:

$$P(X(t_0), X(t_1), \ldots, X(t_M)|\theta)$$

- Which breaks down as:

$$P(X(t_0)|\theta) \times P(X(t_1)|X(t_0), \theta) \times \cdots \times P(X(t_M)|X(t_{M-1}), \theta)$$

- And then we could maximize that to determine $\theta$ – a good choice of $\theta$ is one under which the data is likely to have been observed!

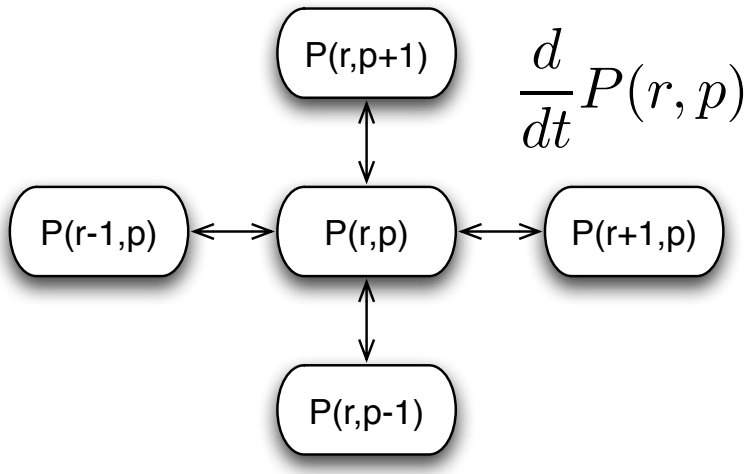
# How to compute the "transition" probability?

How do we compute $P(X(t_{i+1})|X(t_i), \theta)$?

# Recall: Chemical Master Equation

Is a system of linear ordinary differential equations describing how the *probabilities* of different states change over time.

E.g. for our constitutive gene expression model, where $m$ is the number of mRNAs and $p$ is the number of proteins:




$$\frac{d}{dt}P(r,p) = +\kappa_r P(r-1,p) + \gamma_r (r+1) P(r+1,p)$$
$$\kappa_p r P(r,p-1) + \gamma_p (p+1) P(r,p+1)$$
$$-(\kappa_r + \gamma_r r + \kappa_p r - \gamma_p p) P(r,p)$$

Of course, this is an infinite system of ODEs...

# So. . . back to Wilkinson's comment

"When tuning the parameters of stochastic models, there is no obvious 'distance' function to optimize, owing to the fact that the likelihood function does not have a simple analytically tractable form."

$\Rightarrow$ Exact evaluation of the probability of a data set (at least, a single-cell time-series) requires exact solution of the chemical master equation.

(And of course we'd have to do that many times to actually optimize parameters $\theta$.)

(And we're assuming we can observe both mRNA and protein – or more generally, all the chemical species in the system.)

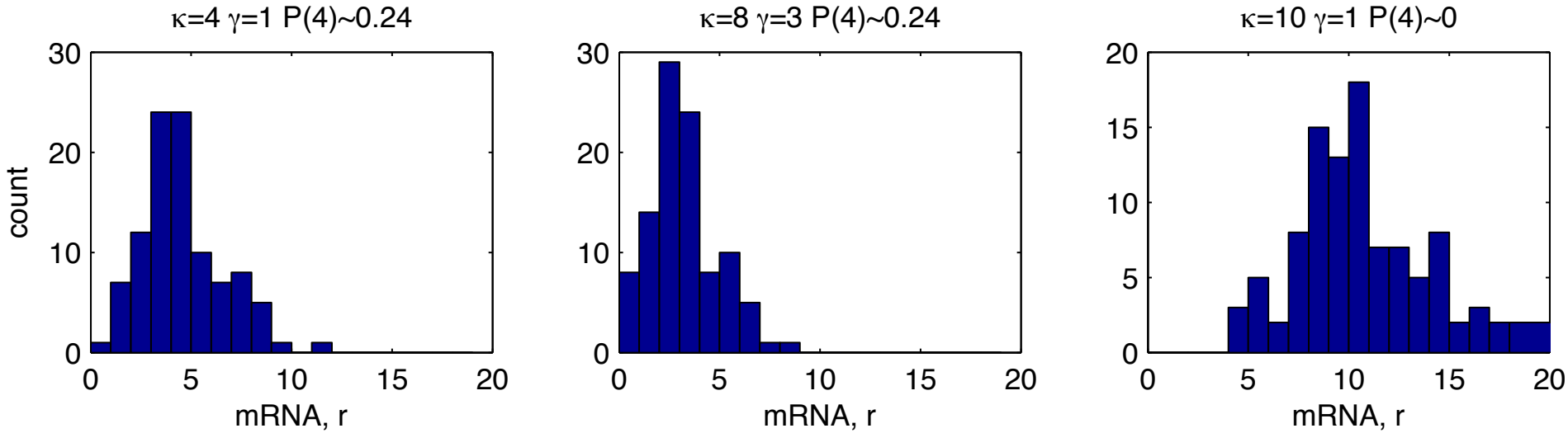
$\Rightarrow$ We're going to give up on exact evaluation, and look at approximations.

# Tian et al., Bioinformatics, 2007

Main idea: Use Gillespie's stochastic simulation algorithm to approximate $P(X(t_{i+1})|X(t_i), \theta)$. For example:

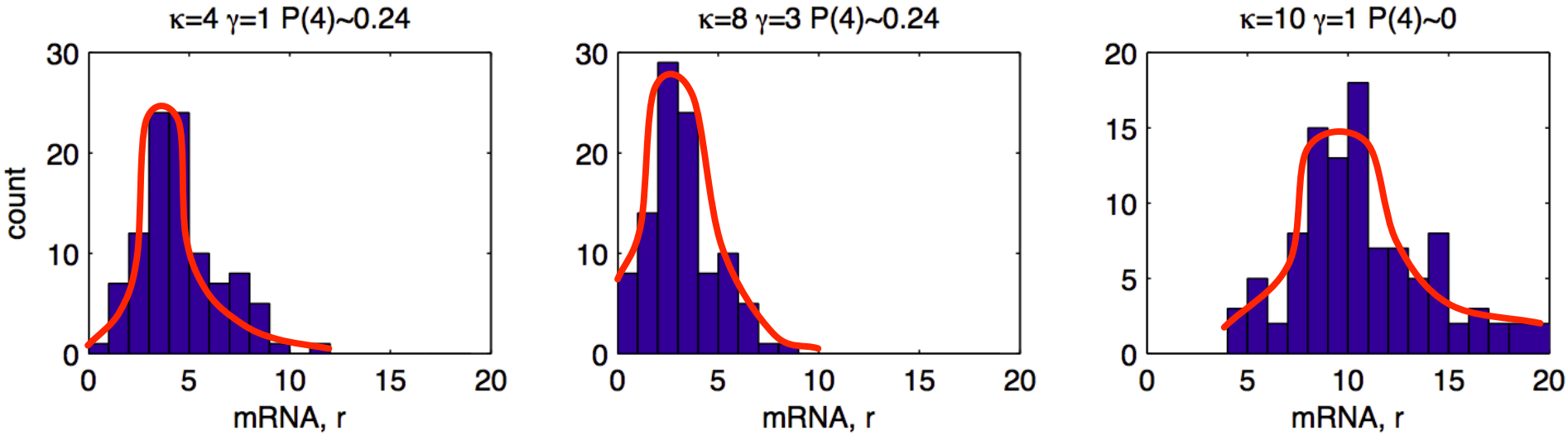
$$\begin{aligned} \emptyset &\rightarrow^{\kappa} r \\ r &\rightarrow^{\gamma} \emptyset \end{aligned}$$

Suppose $r(0) = 0$, $r(10) = 4$, and we take 100 Gillespie simulations to estimate $P(r(10) = 4|r(0) = 0, \kappa, \gamma)$.



# Tian et al., Bioinformatics, 2007

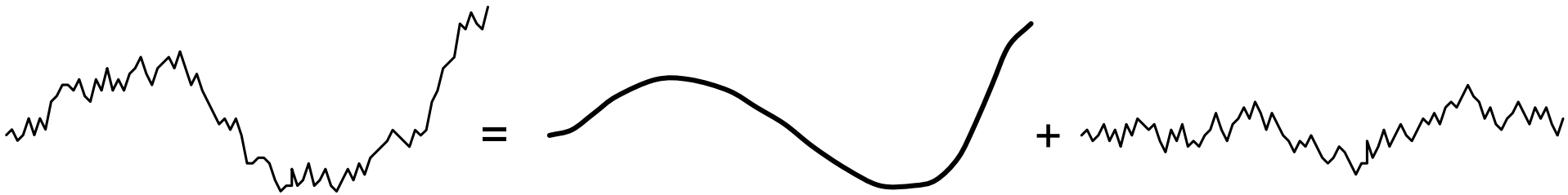
Smoothing is important, to offset effects due to "small" number of simulations.



Main idea: Use the linear noise approximation, LNA, to facilitate the parameter estimation problem.

The LNA separates the system dynamics into a mean behavior and fluctuations around the mean behaviour.

$$X(t) = X_{mean}(t) + \zeta(t)$$



## Mean behaviour

- The mean behavior is assumed to follow the usual deterministic ODE dynamics. This may be linear or nonlinear, depending on the reactions in the system.

$$\frac{d}{dt}X_{mean} = f(X_{mean})$$

- Mean behavior does not depend on the fluctuations.
- We can easily solve for $X_{mean}(t)$ – analytically or numerically, depending on $f(\cdot)$.

## Fluctuation behavior

- The fluctuations are derived by assuming $\sqrt{n}$-size fluctuations in molecular counts (inspired by Poissonian statistics) and expanding the chemical master equation in terms of those fluctuations (sort of). This results in:

$$d\zeta(t) = A(t)\zeta dt + E(t)dW(t)$$

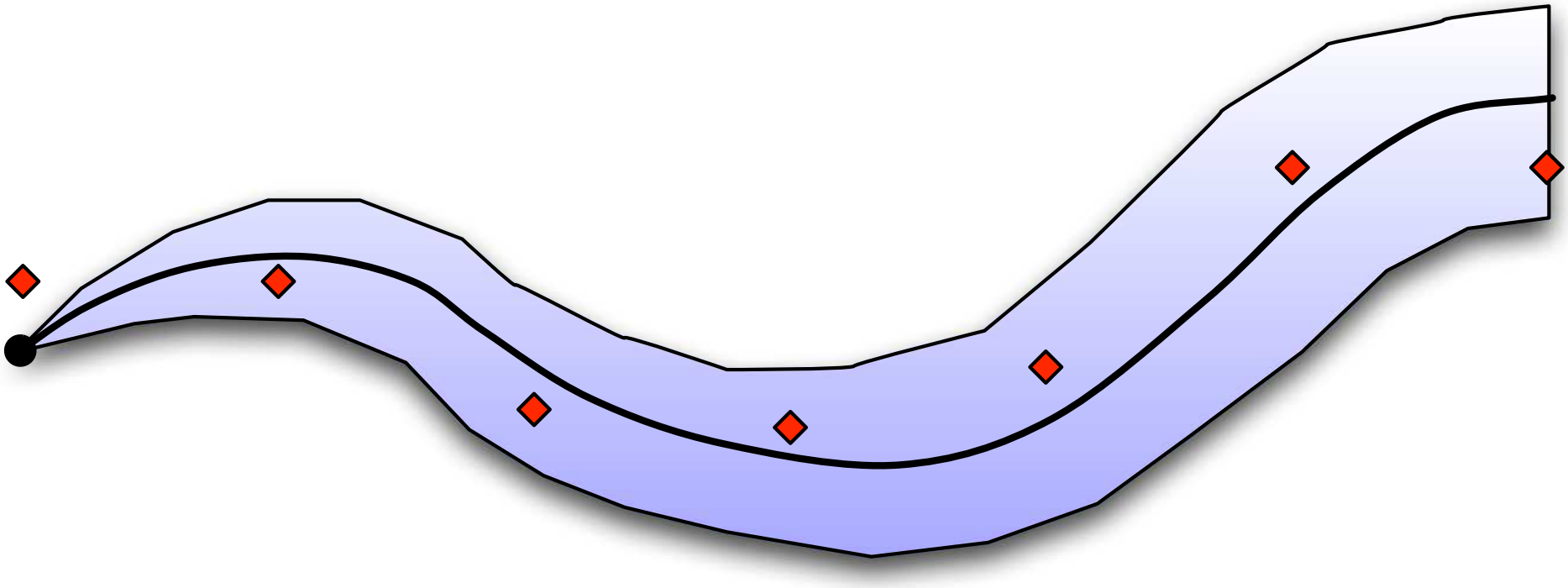
  where $A(t)$ and $E(t)$ are matrices that depend on $X_{mean}(t)$, and $W$ is a Wiener process. (See paper for details; or Van Kampen; or Matt Scott's notes at Waterloo.)

- The fluctuations $\zeta(t)$ obey a linear time-inhomogenous SDE. Therefore, they are normally distributed at all times.

$$L(\zeta(t)) \sim N(\mu(t), \Sigma(t))$$

- The mean $\mu(t)$ and covariance $\Sigma(t)$ can be computed by solving certain ODEs. (See paper.)

# Pictorially. . .



## Summary

In fitting SCKM models:

- Determining the probability of the data for given model parameters is a key computational bottleneck
- Tian *et al.* proposed a scheme that combines Gillespie simulations and state-space smoothing–similar to ideas used in the SDE community
- Komorowski *et al.* proposed a scheme based on the linear noise approximation
- Other approaches, e.g. based on chemical Langevin equation or limits on the number of reactions that may occur appear in the literature
- Identifiability analysis can tell us what parameters can and cannot be extracted from different kinds of data (see Munsky)

# Caveats

- We've not talked about model selection
- We've not talked about extrinsic noise, only intrinsic noise and measurement noise
- Larger models tend to be "sloppy" – some parameters, or combinations of parameters, may be tightly constrained by the data, while other are not

Questions?