An Introduction to Systems Biology from a Machine Learning Perspective

Neil D. Lawrence

Tampere University of Technology, Finland

22nd June 2009

Roadmap

1. Introduction & Philosophy
2. Biological Systems
3. Modelling the System
4. p53 and SOS Response
5. Signalling Pathway
6. Acknowledgements

Outline

1. Introduction & Philosophy
2. Biological Systems
3. Modelling the System
4. p53 and SOS Response
5. Signalling Pathway
6. Acknowledgements

Roadmap

1. Introduction & Philosophy
2. Biological Systems
3. Modelling the System
4. p53 and SOS Response
5. Signalling Pathway
6. Acknowledgements

Talk Overview

- Feynman on Biology:
  "There's Plenty of Room at the Bottom: An Invitation to Enter a New Field of Physics" Feynman (1959).
- Systems biology: Interaction of Biological Components
- Differentiates from a *Reductionist* approach to biology that previously dominated.
- Course based on edited volume from MIT Press "Learning and Inference in Computational Systems Biology" Lawrence et al. (2010) (in press).
  - Models of Transcriptional Regulation.
  - Gaussian Process Inference for Linear Activation.
  - Non-linear Response Models.
  - Cascaded Differential Equations.

Can a Biologist Fix a Radio? Lazebnik (2002)

"It is difficult to find a black cat in a dark room, especially if there is no cat."

- Biological systems are immensely complicated.
- Lazebnik argues the need for models that are quantitative.
  - Such models should be predictive of biological behaviour.
  - Such models need to be combined with biological data.
- Our objective:
  - Establish the need for such models.
  - Describe some approaches to constructing these models.

Reductionist Thinking

- Reductionism Garfinkel (1991) breaks down a system into its component parts.
  - A human body *is just* a collection of biological cells.
  - A biological cell *is just* a collection of biochemical interactions.
  - Chemical compounds are just *are just* a collection of atoms.
  - Atoms *are just* a collection protons, neutrons and electrons.
  - Conclusion: to understand a human we must just understand protons, neutrons and electrons.

Systems Approach

- There is little point in reducing a system as far as its component parts (quarks and leptons ... strings (M-theory) ...) if the questions are at a higher level: How does a plane fly?
- A disease may be caused by a mutation in one gene, but curing the disease may involve the entire pathway.
- Study the system at the level in which we want to ask questions:
  - *e.g.* **Which proteins interact in the ERK/MAPK signalling pathway?**
    This is a critical pathway in cell proliferation. Of strong interest in cancer.

Machine Learning and Systems Biology

- Where learning and inference come in?
  - Models of interaction are not fully characterised. Use inference and learning to deal with unknowns.
- Is this what we normally do in machine learning?
  - No — models are mechanistic in inspiration not black box.
  - However, perhaps it's what we *will* do in the future!
- My prediction: Machine learning in the future will have two major foci.
  1. V. large data sets, *e.g.* prediction of relevant adverts.
  2. Small data set relative to complexity of the system.
     - ⋆ Not enough information to describe the model. Need to turn to mechanistic models to help.
     - ⋆ Lessons from large data set will still apply as system may be very complex!
- A big focus for our research in Manchester: *inference in mechanistic models*.

Systems Biology

- Systems biology provides an opportunity to analyse complex systems by combining prior knowledge with data.
- Integrate data with knowledge of the chemical kinetics of the system.
- These lectures:
  - Philosophical motivation.
  - Review of transcription.
  - Chemical kinetics in a simple synthetic biology system.
  - Inference of hidden variables in single input motifs.
  - Model selection through Bayes' factors.
  - Gaussian process models and differential equations.

What is Systems Biology?

- We think of systems as in "system identification".
- We want to develop computational techniques to uncover, ideally, the true underlying mechanisms.
- We refer to this field as "Computational Systems Biology".
- Develop algorithms that uncover the structure and parametrization of the underlying mechanistic model.
- More precisely: we wish to answer specific questions about the underlying mechanisms.
- We can think of this process as learning or inference.

Reverse Engineering Biological Networks

- Traditional approaches to system identification cannot be directly applied.
  - observation of the state variables of the system is only possible at relatively low sampling rates.
  - there is a strong stochastic component to biological systems. (either intrinsic or from the measurement process).
  - Biological systems have *evolved*. Organism survival dictates they are robust to environmental changes and mutations. This implies redundancy.
- Taken together these characteristics distinguish biological systems from standard engineering systems.

Indentifiability and Biological Systems

- A major challenge is the *identifiability* of the system.
- Selection pressure for robustness results in highly redundant pathways.
- Biological systems are dependent on kinetic rates that are highly temperature dependent.
- Contrast this with a human designed analogue electrical circuits: we use ceramic capacitors for timing circuits because of high tolerance to temperature changes.
- As we will see from the repressilator example, (Elowitz and Leibler, 2000), a simple biological oscillator has a strong variation in amplitude, phase and period of oscillation.
- If a system's output is insensitive to its parameters, it is difficult to identify the parameters given the system output.

Systems Biology and the Scientific Process

- Available data has increased dramatically for biology, however, data is scarce relative to the complexity of the system.
- Identifiability problems are made worse by data scarcity.
- What are the prospects for serious progress in Computational Systems Biology?
- Despite these issues, prospects are *good*.
- Identifiability problems exist, but do they effect our ability to answer fundamental scientific questions?

A Scientific Framework for Systems Biology

- Popper's philosophy of science: creation of models that are *predictive* and *refutable*.
- An ongoing process of hypothesis generation, followed by observation followed by an update of the hypothesis.
- Quoting from Popper:

> *The problem 'Which comes first, the hypothesis (H) or the observation (O)?' is soluble; as is the problem, 'Which comes first, the hen (H) or the egg (O)?'. The reply to the latter is, 'An earlier kind of egg'; to the former, 'An earlier kind of hypothesis'. Popper (1963)*

Hypothesis Driven Research

- Some modern biological work decried for "Not being hypothesis driven".
- Popper allows for this approach.
- Our observations must be couched in a frame of (even vague) expectations.
- Some expectations are refuted by observation: expectations are updated.
- This is an excellent model for research in systems biology.
  - Expectations are couched in the form of mathematical idealizations of the system (the earlier egg).
  - Refutation through observation arises through biological experiment (the hen).
  - This gives rise to a modified hypothesis (the new egg).
- Systems biology as an iterative *experimental spiral*: incrementally improving our understanding of the biological system.

Formal Expression of Framework

- Can we express such a framework more formally?
- Dawid's *prequential approach* Dawid (1984).
- Prequential: contraction of *probability forecasting with sequential prediction*.
  - Dawid argues we should consider the parameters only as summarizing the acquired knowledge.
  - Predictions of the model are the critical element.
  - The 'egg' is a prequential forecaster.
  - Selected biological experiments are designed to refute those predictions.
  - Dawid (1982) shows how prequential forecasts which fail the 'complete calibration' are refuted.

Prequential Forecasting: Jeffrey's Law

Quoting from Dawid (Jeffrey's Law):

> *... Consequently, all non rejected [prequential forecasting systems] end up making the same forecasts.*
>
> *This is an interesting and unexpected boon: in just those cases where we cannot choose empirically between several forecasting systems, it turns out that we have no need to do so! This property has implications for Philosophy of Science, giving some support to Popper's methodology, wherein a number of alternative hypotheses about Nature may be put forward, each being retained until it is refuted because its forecasts depart from observation. In our context, such refutation follows evidence that the complete calibration criterion is violated. This approach need not pick out, even asymptotically, a single "true model". (Indeed there is no need even to assume the existence of an underlying "true" law generating the data.) Using it, we should, however, eventually be left only with [prequential forecasting systems] that can all be expected to continue to make essentially identical predictions. Dawid (1984)*

Systems Biology in Practice

- Choose a mathematical framework that allows us to sustain multiple hypotheses (*e.g.* a Bayesian approach).
- Our mathematical idealization needs to be combined with a large quantity of data. (system is very complex even scarce data could be gigabytes of data.)
- This practical requirement becomes difficult to fulfil as the complexity of the mathematical idealization increases.
- The idea of starting with a family of hypotheses representing all conceivable variations on the system design is not practicable.

Modelling Compromise

- Simplify the model.
  - Even if the underlying biochemical system exhibits dynamical behavior, we might assume that high decay implies system is at steady state.
  - Leads to practical savings.
  - Experiments at high sample rates will refute hypothesis.
  - Simplification is worth pursuing: the scientific question may be answered before high sample rate experiments are required.
  - This set up is a **modelling compromise**.

Inference Compromise

- Use of approximate inference techniques.
- Bayes' rule,

$$p(\boldsymbol{\theta}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int p(\mathbf{Y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}}$$

  requires integral over parameters.
- Replacing integrals with approximations is an **inference compromise**.
- Examples include the Laplace approximation, variational approximations and sampling.

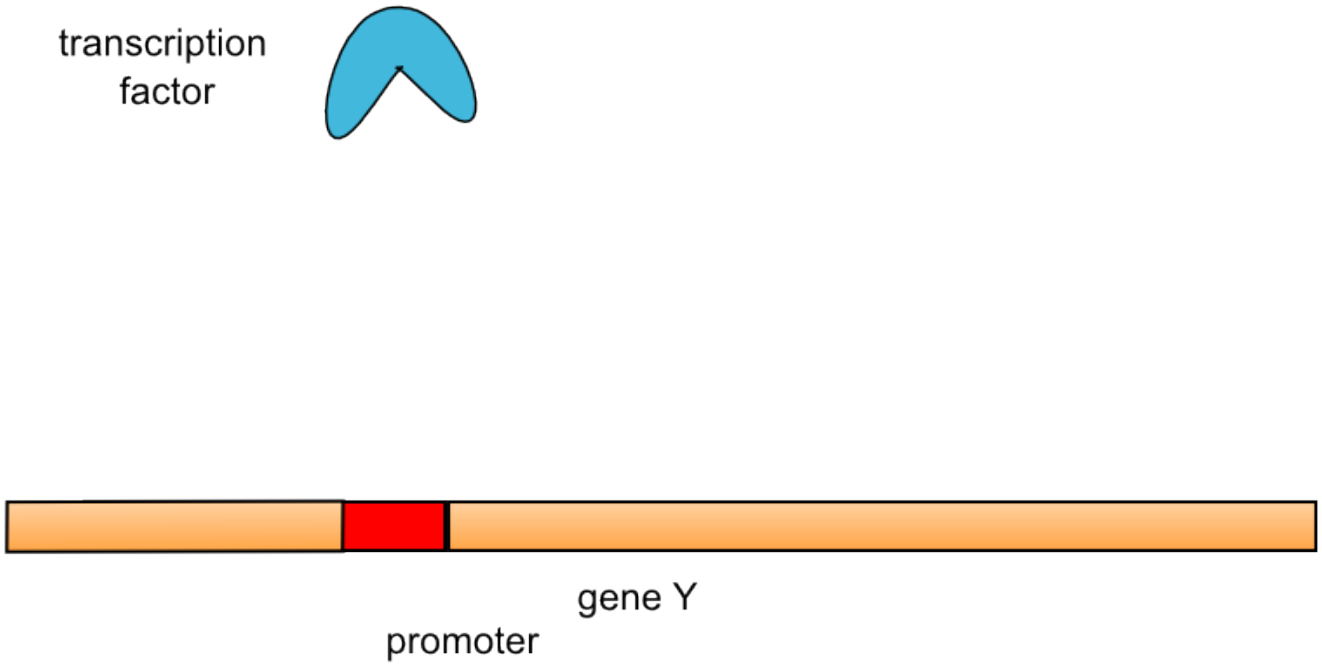
Systems Biology and Scientific Judgement

- In practice a combination of modelling and inference compromises is often required.
- We will present approaches with both modelling and inference compromises.
- Compromises imply we are departing from the pure framework we outlined.
- The degree of departure that's allowable is a matter for careful scientific judgment in the context of a given question.
- We develop new methodologies to try and ensure that compromises are as small as possible.

Roadmap

1. Introduction & Philosophy
2. Biological Systems
3. Modelling the System
4. p53 and SOS Response
5. Signalling Pathway
6. Acknowledgements

Transcriptional regulation of gene expression



RNA Polymerase

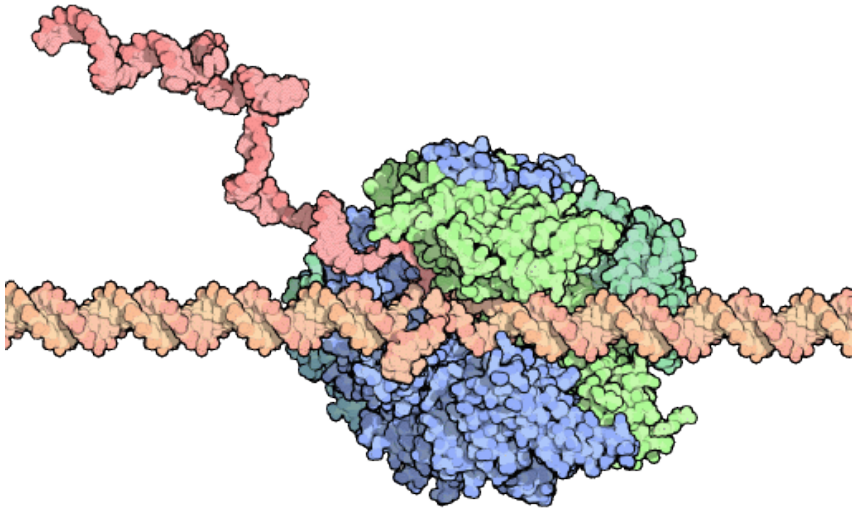
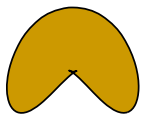


Figure: RNA Polymerase transcribing RNA from DNA. Image from "Molecule of the Month" at the protein data bank: `http://mgl.scripps.edu/people/goodsell/pdb/pdb98/pdb98_1.html`

Repression

Repressing Transcription Factor



The Repressilator

- Real biology involves interaction of several systems.
- The repressilator is the first synthetic biology oscillator (Elowitz and Leibler, 2000).
- Implemented in *E. coli* bacteria.
- How do we model such a system?

The Repressilator

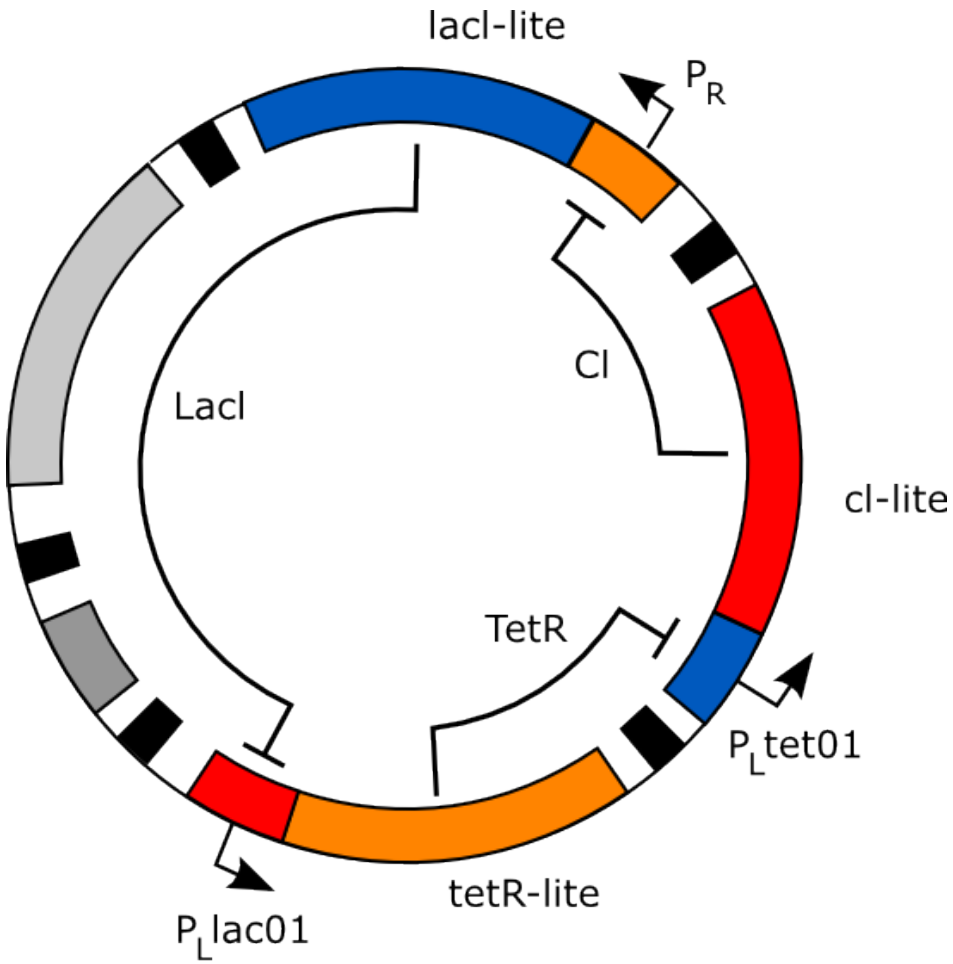


Figure: Repressilator Plasmid. (Elowitz and Leibler, 2000)

Bacteria Plasmids

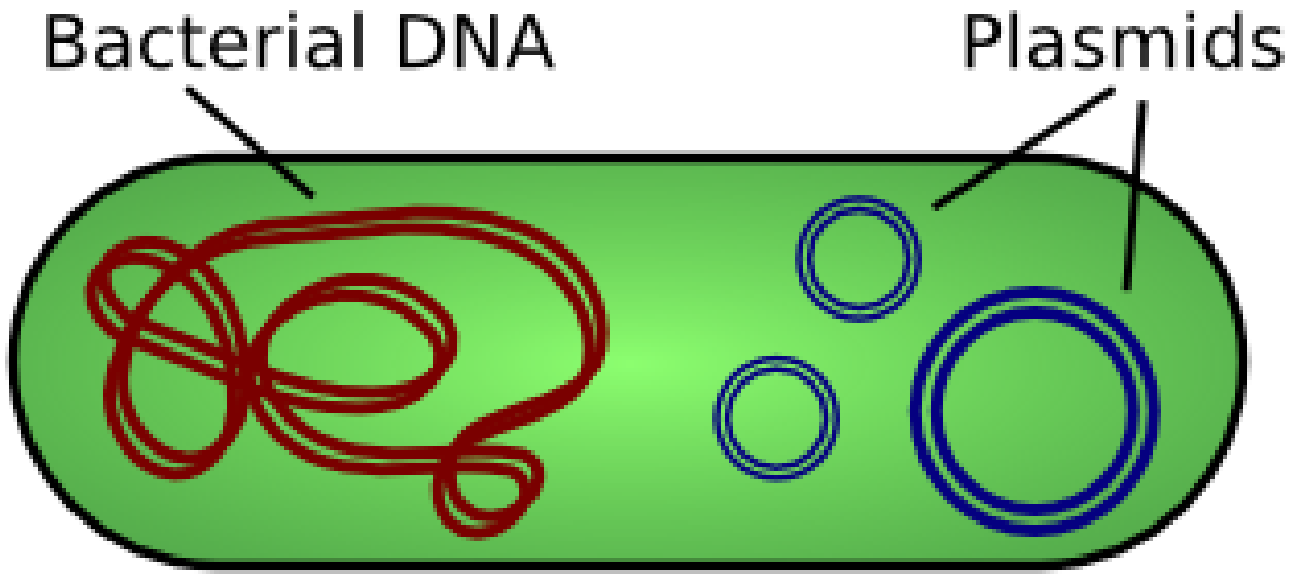
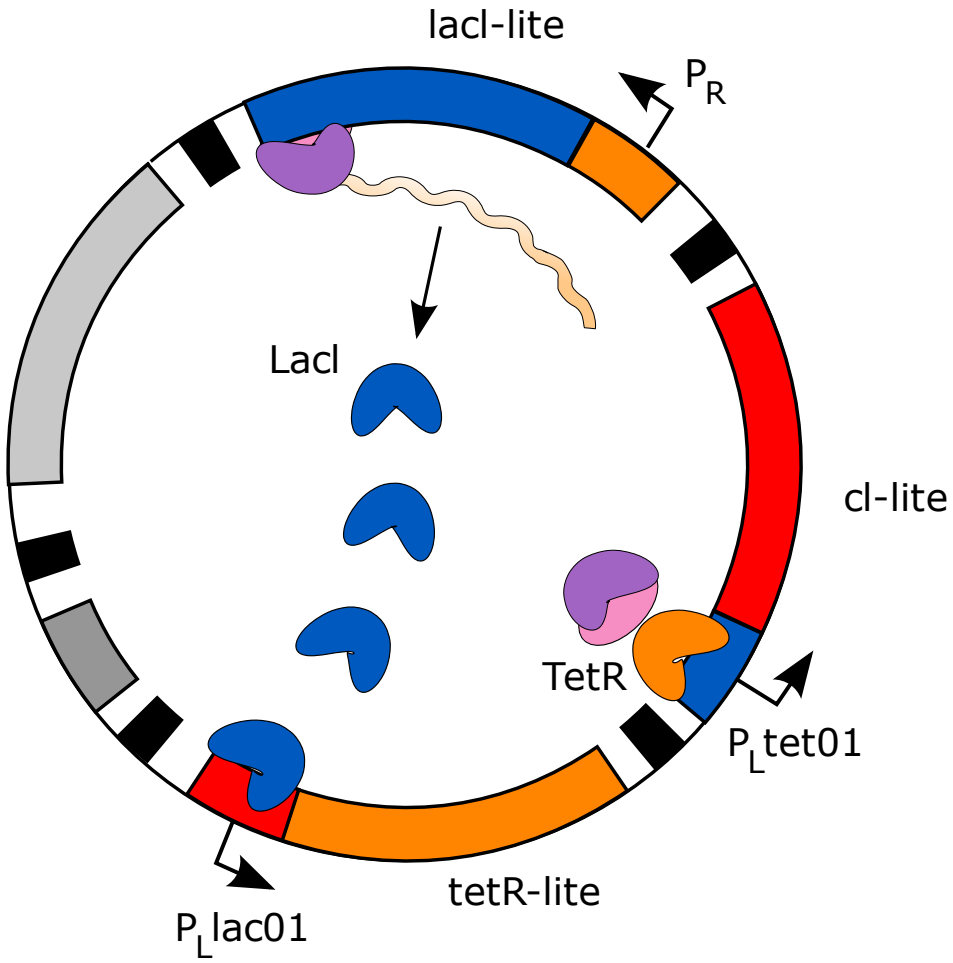
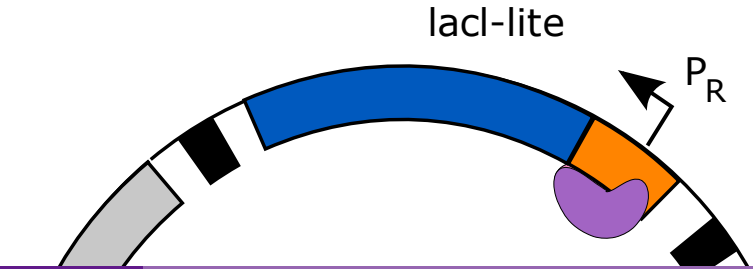


Figure: Schematic of a bacterium with plasmids (Image from wikimedia commons).

Repressilator





Repressilator Results

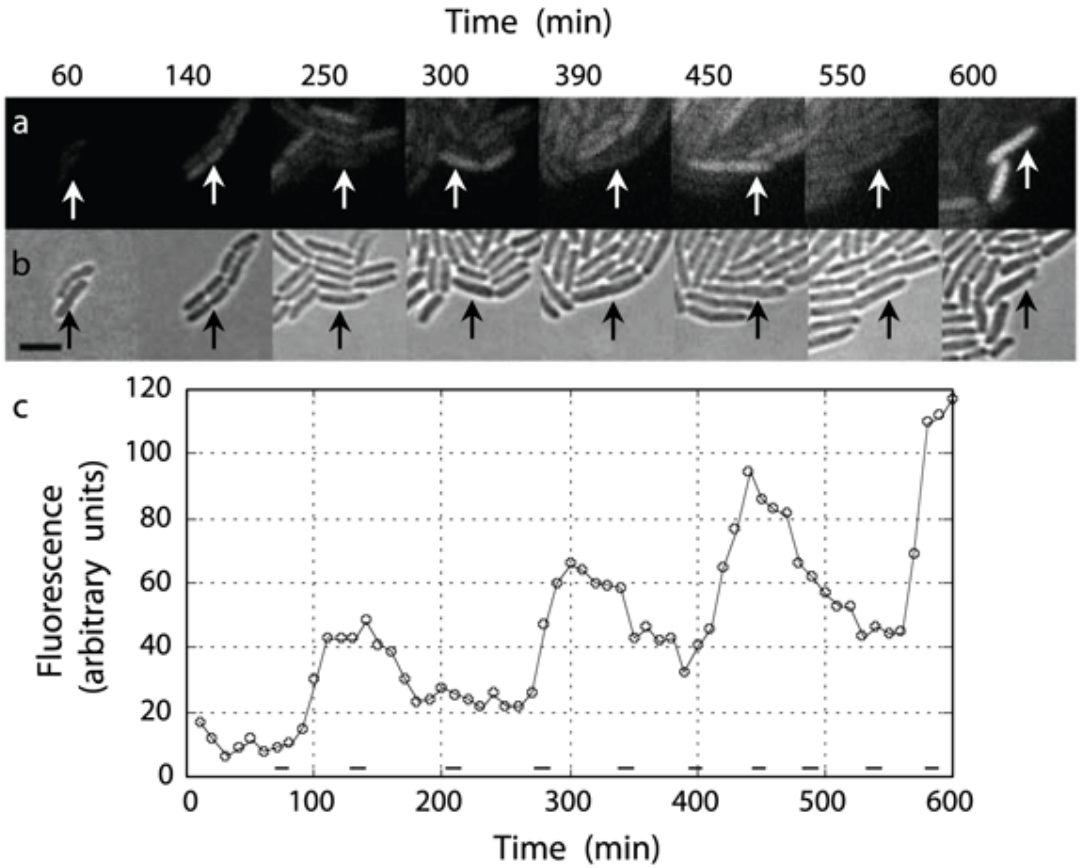


Figure: Observations of GFP. Source `http://en.wikipedia.org/wiki/Image:Repressilator_observations_1.png`

Roadmap

1. Introduction & Philosophy
2. Biological Systems
3. Modelling the System
4. p53 and SOS Response
5. Signalling Pathway
6. Acknowledgements

Stochastic Processes and Mass Action Kinetics

Further reading: Wilkinson (2006, Chapters 1 and 6) Lawrence et al. (2010, Chapters 10, 11, and 12)

- Mass action kinetics — reaction occurs when relevant molecules *collide*.
- Probability of any given reaction, i, occuring in a given instant interval of time $\mathrm{d}t$ is given by $h_i \mathrm{d}t + o(\mathrm{d}t)$.
  - Where h_i is a rate law or hazard function. It is dependent on the current state of the system and c_i a stochastic rate constant.
- Represent a reaction in the form

$$X_1 + X_2 \rightarrow X_3 + X_4$$

where X_1 and X_2 are the *reactants* and X_3 and X_4 are the *products*. Denote numbers of each species by x_1, x_2, x_3 and x_4. State of the system given by vector $\mathbf{x}$.

Stochastic Mass Action Kinetics

- Zeroth order:

$$0 \rightarrow X_1$$

probability of this reaction in interval $\mathrm{d}t$ is $h_i \mathrm{d}t = c_i \mathrm{d}t$

- First order (*e.g.* decay):

$$X_1 \rightarrow 0$$

probability of this rection in interval $\mathrm{d}t$ is $h_i \mathrm{d}t = c_i x_1 \mathrm{d}t$

- Second order:

$$X_1 + X_2 \rightarrow X_3,$$

probability of this reaction in interval $\mathrm{d}t$ is $h_i \mathrm{d}t = c_i x_1 x_2 \mathrm{d}t$.

- For individual reaction, waiting time, τ_i, is sampled from $p(\tau_i) = h_i \exp(-h_i \tau_i)$.

Combining Reactions

- Typical system has multiple reactions at the same time.
- The hazard is a "rate" parameter — if there were no other reactions waiting time until the reaction would be given by an exponential.
- In practice there are other reactions and associated hazards, $\mathbf{h} = \{h_j\}_{j=1}^{M}$.
- Each reaction can affect all other hazard functions, $h_i(\mathbf{x}, c_i)$.
- Sample from the system (Gillespie's *first reaction* method):
  1. Sample time of next reaction from all reactions: $\{\tau_i\}_{i=1}^{M}, \quad \tau_i \sim h_i \exp(-h_i \tau_i)$.
  2. Find next reaction $\mu = \text{argmin}_i \tau_i$.
  3. Update state of system, $\mathbf{x}$, according to rule for that reaction.
  4. Recompute vector of hazards, $\mathbf{h}$.
  5. Repeat

Gillespie's Direct Method

- Previous sampling scheme: M random numbers (1 for each reaction).
- Exploit properties of exponential:
  - τ_j is the minimum value from $\{\tau_i\}_{i=1}^M$ sampled from different exponentials with rates $\{h_i\}_{i=1}^M$.
  - This implies: $\tau_j \sim h_0 \exp(-h_0\tau_j)$ where $h_0 = \sum_{i=1}^M h_i$ and is known as the *combined reaction hazard.*
  - *i.e.* in each small time interval probability of any reaction is $h_0 \mathrm{d}t$.
- The probability of it having arisen from the jth reaction is given by

$$\frac{h_j}{h_0}$$

cf superposition of Poisson processes.

Sampling the System

Gillespie Direct Method

1. Compute the hazards, $\mathbf{h}$.
2. Sample time of next reaction from $\tau_\mu \sim h_0 \exp\left(-h_0 \tau_\mu\right)$
3. Determine which reaction it was: sample μ from a multinomial with probabilities given by $\frac{h_j}{h_0}$.
4. Update the state of the system, $\mathbf{x}$.
5. Increment time $t \rightarrow t + \tau_\mu$.
6. Repeat until simulation time complete.

- This is $O\left(M\right)$.
- Can do in $O\left(\log M\right)$ — use a *dependency* graph to determine when things need calculation Gibson and Bruck (2000).

Repressilator Simulation

Translation:

$$\text{mlac} \longrightarrow \text{Lac} + \text{mlac}$$

First order reaction of mRNA from *lac* gene to protein plus mRNA from *lac* gene.

Repressilator Simulation

mRNA decay:

$$\text{mlac} \longrightarrow 0$$

First order reaction of mRNA from *lac* gene.

Repressilator Simulation

Protein decay:

$$\text{Lac} \rightarrow 0$$

First order reaction of Lac protein.

Repressilator Simulation

Transcription:

$$\text{lac} + \text{RNAP} \rightarrow \text{mlac} + \text{RNAP} + \text{lac}$$

Second order reaction of *lac* gene and RNA polymerase to *lac* mRNA, *lac* gene and RNA polymerase.

Repressilator Simulation

Protein (TF) bound to promoter:

$$\mathrm{CI} + \mathrm{lac} \rightarrow \mathrm{CI} \cdot \mathrm{lac}$$

Second order reaction, TF protein (CI) from another gene binds to *lac* promoter (represented by the gene). This prevents transcription.

Repressilator Simulation

Protein unbinds from promoter:

$$\mathrm{CI} \cdot \mathrm{lac} \rightarrow \mathrm{CI} + \mathrm{lac}$$

First order reaction, TF protein and *lac* promoter region unbind, allowing transcription to take place.

Other Implementation Details

- The effect of each reaction is stored in a matrix $\mathbf{S}$, the stoichiometry matrix.
- A row of this matrix is added to the state vector, $\mathbf{x}$, to account for effects from each reaction.

Simulation Result

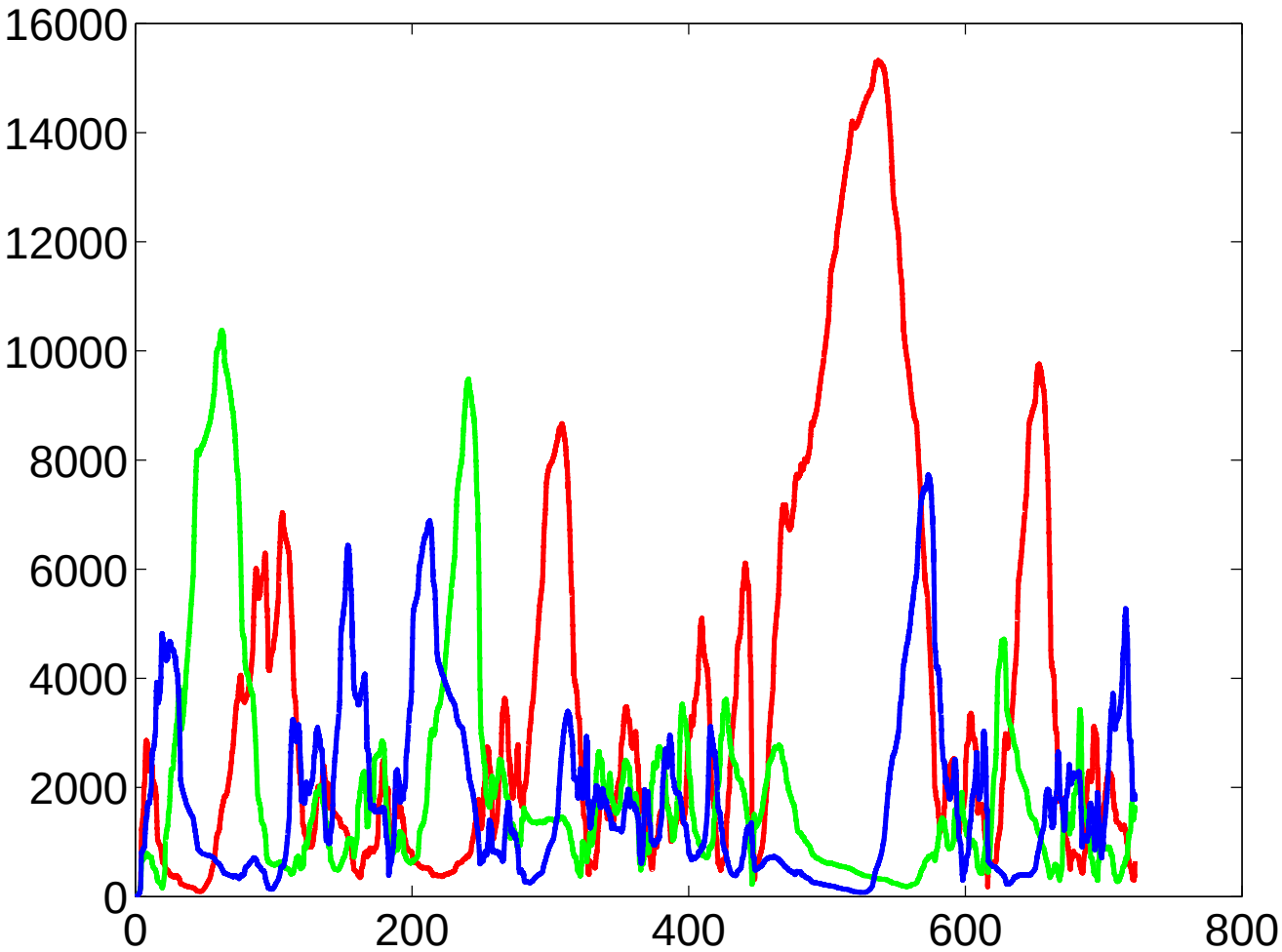


Figure: Simulation of repressilator using Gillespie algorithm. Red is mRNA for lac, Green is mRNA for tetR and Blue is mRNA for cI.

What Next?

- Simulation from the system assumes we know *structure* (stoichiometric matrix, $\mathbf{S}$) and parameters (stochastic rate parameters, $\mathbf{c}$).
- Structure *may* be known or assumed.
- Specifying parameters is more complex.
  - In chemistry *in vitro* measurements can be made.
  - In biology this is more difficult and perhaps less valid.
- Can we do learning? — **this is where ML comes in!**
  - If $\mathbf{x}$ is observed directly in v. high time resolution: yes.
  - In practice it is indirectly observed in lower time resolution.
- Learning in stochastic systems is difficult as marginalisation of these unknowns is required.

Modelling Compromise: A Deterministic Approximation

- Approximate the stochastic system by dealing in *deterministic* concentrations.
- In *chemistry* concentrations involve large numbers, and the approximation is good.
- In *biology* this is generally less true. But often we pool mRNA from many cells.
- For Mass Action Kinetics:

$$X_1 + X_2 \rightarrow X_3$$

$$X_3 \rightarrow 0$$

leads to

$$\frac{\mathrm{d}\left[X_3\right]}{\mathrm{d}t} = k_1\left[X_1\right]\left[X_2\right] - k_2\left[X_3\right]$$

with $[X_i]$ representing concentration of species X_i.

Repressilator Simulation

Translation:

$$\text{mlac} \rightarrow \text{Lac} + \text{mlac}$$

$$\frac{\mathrm{d}\left[\text{Lac}\right]}{\mathrm{d}t} = -k_3\left[\text{Lac}\right] - k_4\left[\text{Lac}\right]\left[\text{mtetR}\right] + k_5\left[\text{mlac}\right] + k_6\left[\text{Lac}\cdot\text{tetR}\right]$$

First order reaction of mRNA from *lac* gene to protein plus mRNA from *lac* gene.

Repressilator Simulation

mRNA decay:

$$\text{mlac} \rightarrow 0$$

$$\frac{\mathrm{d}\,[\text{mlac}]}{\mathrm{d}t} = k_1\,[\text{RNAP}]\,[\text{lacI1}] - k_2\,[\text{mlac}]$$

First order reaction of mRNA from *lac* gene.

Repressilator Simulation

Protein decay:

$$\text{Lac} \rightarrow 0$$

$$\frac{\mathrm{d}\left[\text{Lac}\right]}{\mathrm{d}t} = -k_3\left[\text{Lac}\right] - k_4\left[\text{Lac}\right]\left[\text{mtetR}\right] + k_5\left[\text{mlac}\right] + k_6\left[\text{Lac}\cdot\text{tetR}\right]$$

First order reaction of Lac protein.

Repressilator Simulation

Transcription:

$$\text{lac} + \text{RNAP} \rightarrow \text{mlac} + \text{RNAP} + \text{lac}$$

$$\frac{\mathrm{d}\,[\text{mlac}]}{\mathrm{d}t} = k_1\,[\text{RNAP}]\,[\text{lacI}] - k_2\,[\text{mlac}]$$

Second order reaction of *lac* gene and RNA polymerase to *lac* mRNA, *lac* gene and RNA polymerase.

Repressilator Simulation

Protein (TF) bound to promoter:

$$\mathrm{CI} + \mathrm{lac} \rightarrow \mathrm{CI} \cdot \mathrm{lac}$$

$$\frac{\mathrm{d}\left[\mathrm{CI} \cdot \mathrm{lac}\right]}{\mathrm{d}t} = k_8 \left[\mathrm{CI}\right]\left[\mathrm{lac}\right] - k_{10}\left[\mathrm{CI} \cdot \mathrm{lac}\right]$$

$$\frac{\mathrm{d}\left[\mathrm{CI}\right]}{\mathrm{d}t} = -k_7\left[\mathrm{CI}\right] - k_8\left[\mathrm{CI}\right]\left[\mathrm{lac}\right] + k_9\left[\mathrm{mcI}\right] + k_{10}\left[\mathrm{CI} \cdot \mathrm{lac}\right]$$

$$\frac{\mathrm{d}\left[\mathrm{lac}\right]}{\mathrm{d}t} = -k_8\left[\mathrm{CI}\right]\left[\mathrm{lac}\right] + k_{10}\left[\mathrm{CI} \cdot \mathrm{lac}\right]$$

Second order reaction, TF protein (CI) from another gene binds to *lac* promoter (represented by the gene). This prevents transcription.

Repressilator Simulation

Protein unbinds from promoter:

$$\mathrm{CI} \cdot \mathrm{lac} \rightarrow \mathrm{CI} + \mathrm{lac}$$

$$\frac{\mathrm{d}\left[\mathrm{CI} \cdot \mathrm{lac}\right]}{\mathrm{d}t} = k_8 \left[\mathrm{CI}\right]\left[\mathrm{lac}\right] - k_{10}\left[\mathrm{CI} \cdot \mathrm{lac}\right]$$

$$\frac{\mathrm{d}\left[\mathrm{CI}\right]}{\mathrm{d}t} = -k_7\left[\mathrm{CI}\right] - k_8\left[\mathrm{CI}\right]\left[\mathrm{lac}\right] + k_9\left[\mathrm{mcI}\right] + k_{10}\left[\mathrm{CI} \cdot \mathrm{lac}\right]$$

$$\frac{\mathrm{d}\left[\mathrm{lac}\right]}{\mathrm{d}t} = -k_8\left[\mathrm{CI}\right]\left[\mathrm{lac}\right] + k_{10}\left[\mathrm{CI} \cdot \mathrm{lac}\right]$$

First order reaction, TF protein and *lac* promoter region unbind, allowing transcription to take place.

Simulated Repressilator

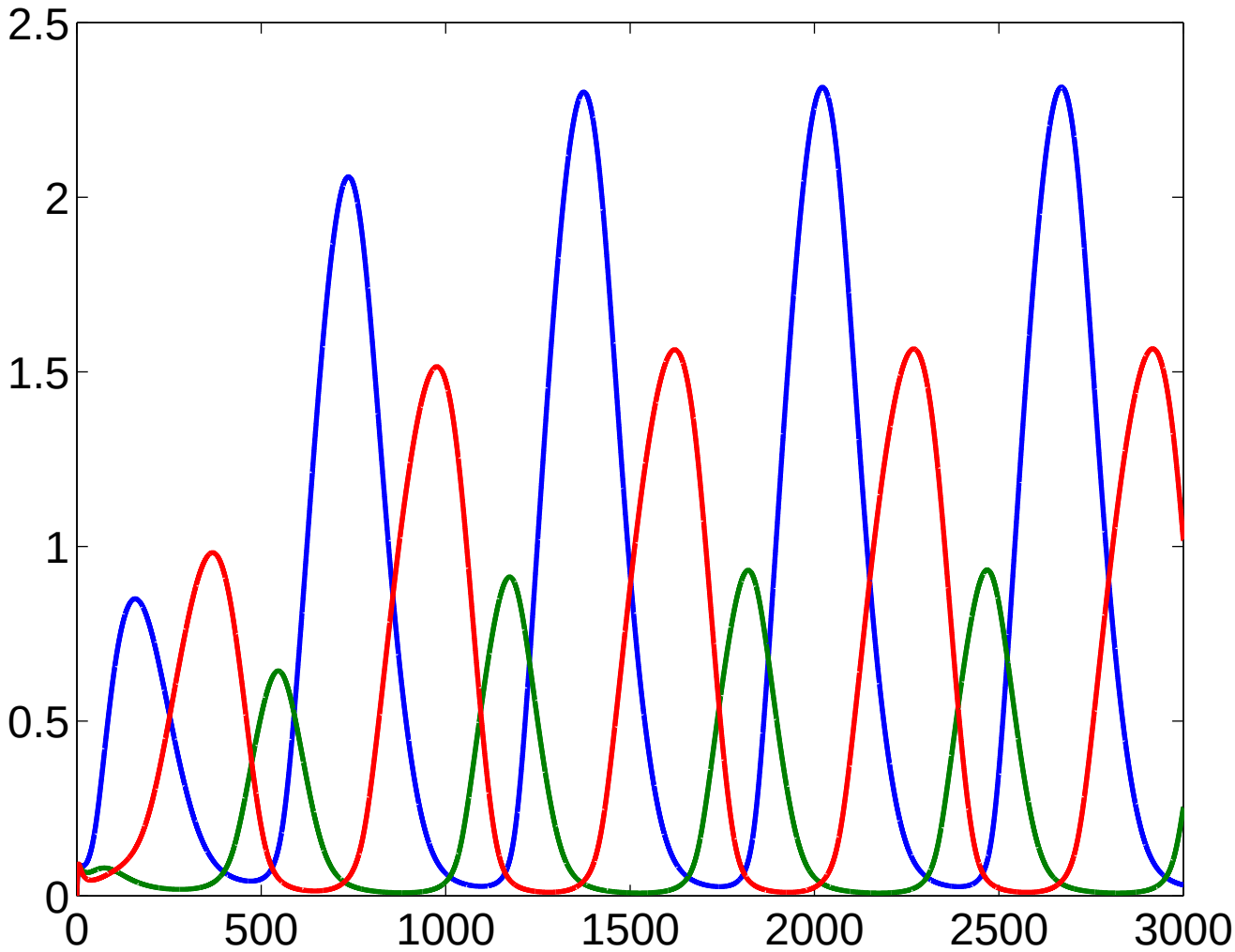


Figure: Simulation of repressilator based on ODEs from COPASI Hoops et al. (2006).

Fitting ODE Models

- Find parameters that allow model to fit a given data set.
- For given parameters and inital conditions solve the system and compare to data.
- Minimise the least squares match to the data with respect to parameters and initial conditions.
- Multimodal optimisation: tools available for fitting (COPASI Hoops et al. (2006)).
- Problems remain:
  1. **How do we deal with a missing chemical species (e.g. TF concentration)?**
     It is common to be missing one or more of the state variables.
  2. **What to do if certain parameters aren't well identified?**
     The system outputs may be insensitive to some parameters.
  3. **If several hypothesised models exist, which should we choose?**
     Bayesian approaches to model ranking: Bayes' factors.

ODE Model of Activation

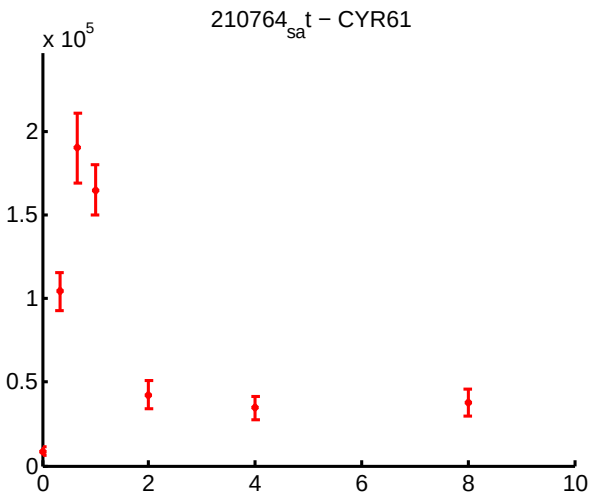
- Linear Activation Model (Barenco et al., 2006, Genome Biology)

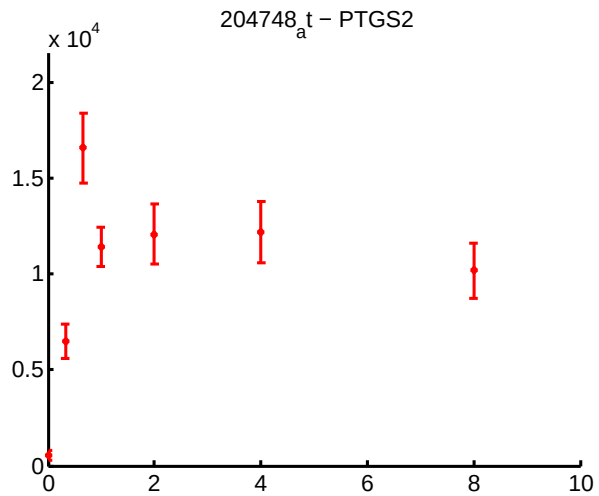
$$\frac{\mathrm{d}x_j(t)}{\mathrm{d}t} = B_j + S_j f(t) - D_j x_j(t)$$

- Slight change in notation:
  - $x_j(t)$ – concentration of gene j's mRNA
  - $f(t)$ – concentration of active transcription factor
  - Model parameters: baseline B_j, sensitivity S_j and decay D_j
  - Application: identifying co-regulated genes (targets)
  - Problem: how do we fit the model when $f(t)$ is not observed?

Why use a model-based approach?

- Model based approach to co-regulated targets ...
  - clustering is often used but,
  - co-regulated genes can differ greatly in their expression profiles





- Clustering cannot be relied on to identify co-regulated genes
- A model-based approach is required

Roadmap

1. Introduction & Philosophy
2. Biological Systems
3. Modelling the System
4. p53 and SOS Response
5. Signalling Pathway
6. Acknowledgements

Cell Damage

- Radiation damages molecules in the cell.
- Most of this damage is quickly repaired — single strand breaks, backbone break.
- Double strand breaks are more serious — a complete disconnect along the chromosome.
- Cell cycle stages:
  - G_1: Cell is not dividing.
  - G_2: Cell is preparing for meitosis, chromosomes have divided.
  - S: Cell is undergoing meitosis (DNA synthesis).
- Main problem is in G1. In G2 there are two copies of the chromosome. In G1 only one copy.

p53 "Guardian of the Cell"

- Responsible for Repairing DNA damage
- Activates DNA Repair proteins
- Pauses the Cell Cycle (prevents replication of damage DNA)
- Initiates *apoptosis* (cell death) in the case where damage can't be repaired.
- Large scale feeback loop with NF-κB.

p53 DNA Damage Repair

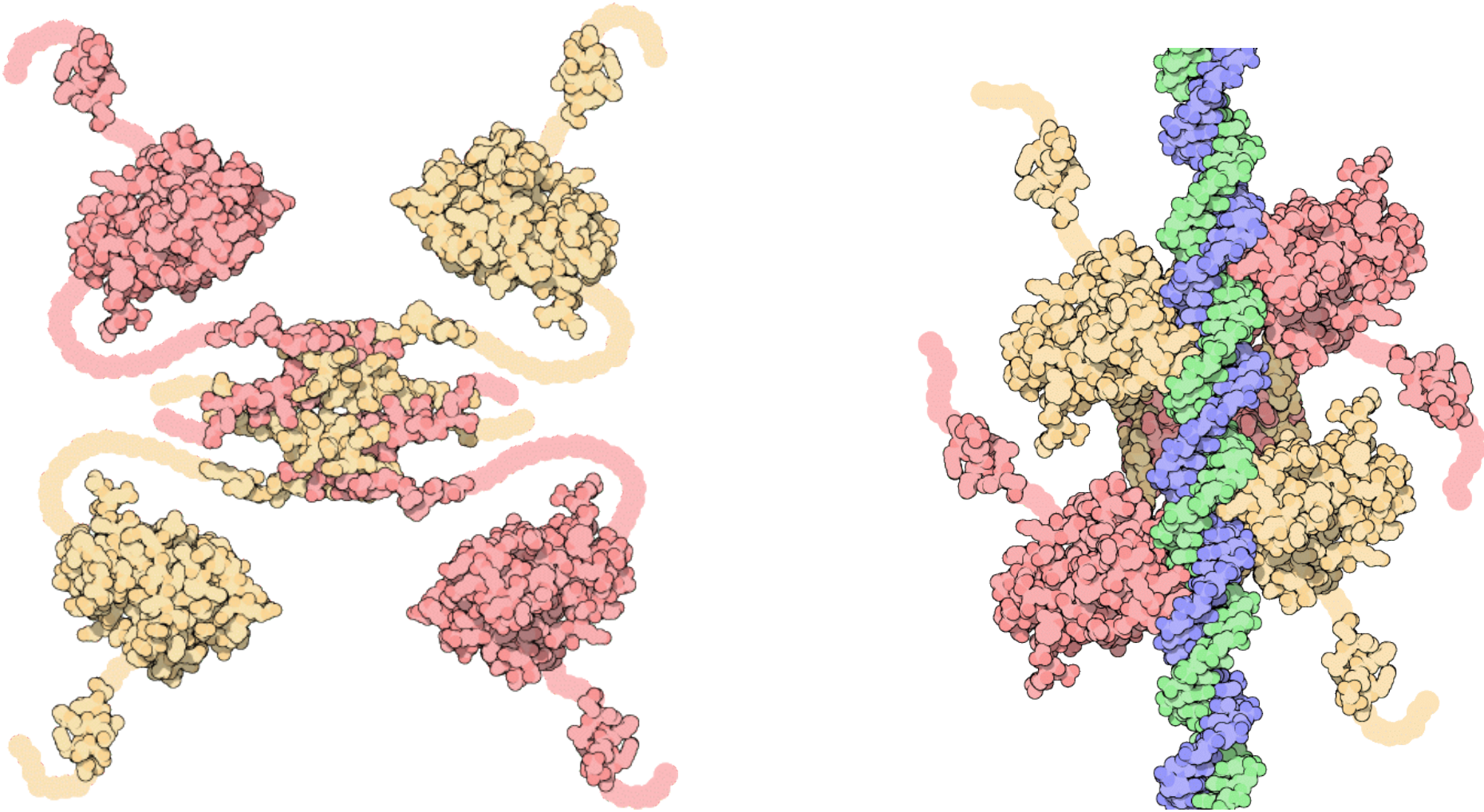


Figure: p53. *Left* unbound, *Right* bound to DNA. Images by David S. Goodsell from `http://www.rcsb.org/` (see the "Molecule of the Month" feature).

p53

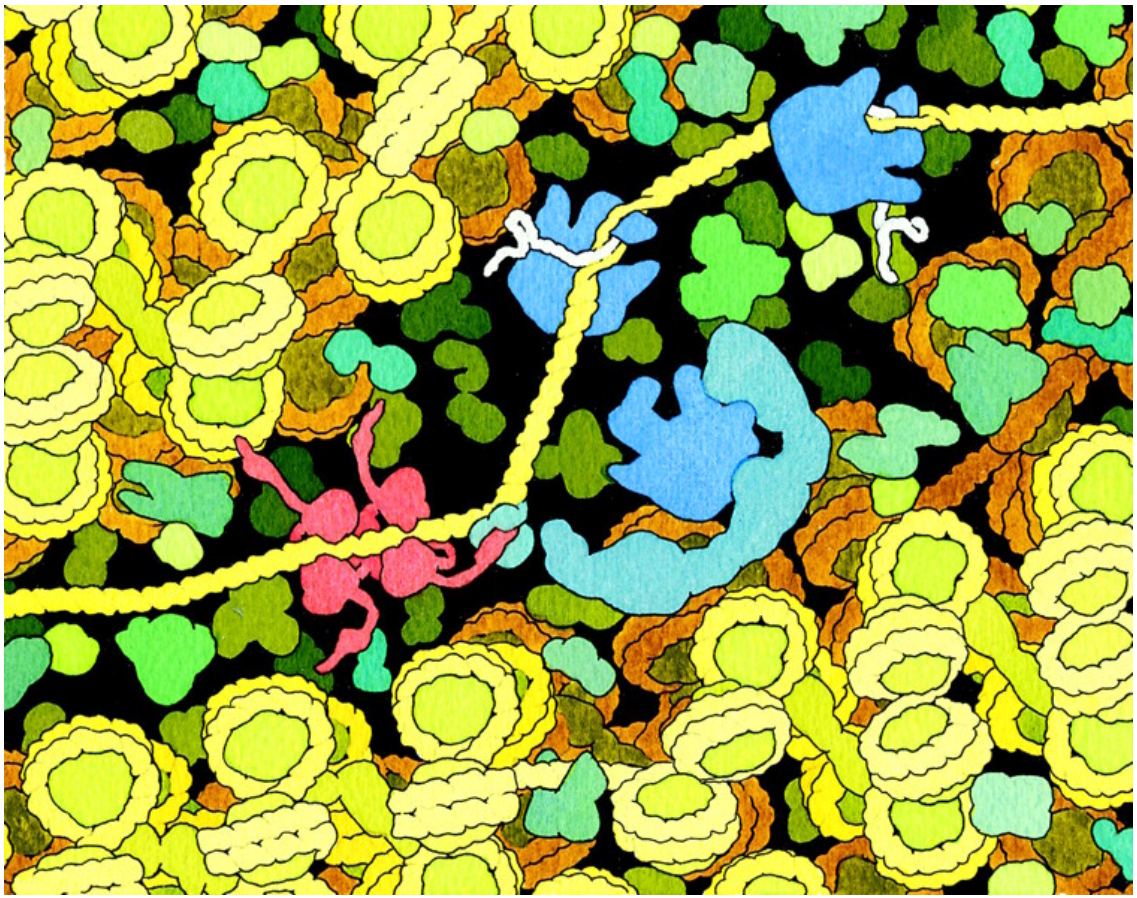


Figure: Repair of DNA damage by p53. Image fromGoodsell (1999).

Some p53 Targets

DDB2 DNA Damage Specific DNA Binding Protein 2. (also governed by C/ EBP-beta, E2F1, E2F3,...).

p21 Cycline-dependent kinase inhibitor 1A (CDKN1A). A regulator of cell cycle progression. (also goverened by SREBP-1a, Sp1, Sp3,...).

hPA26/SESN1 sestrin 1 Cell Cycle arrest.

BIK BCL2-interacting killer. Induces cell death (apoptosis)

TNFRSF10b tumor necrosis factor receptor superfamily, member 10b. A transducer of apoptosis signals.

Modelling Assumption

- Assume p53 affects targets as a single input module network motif (SIM).

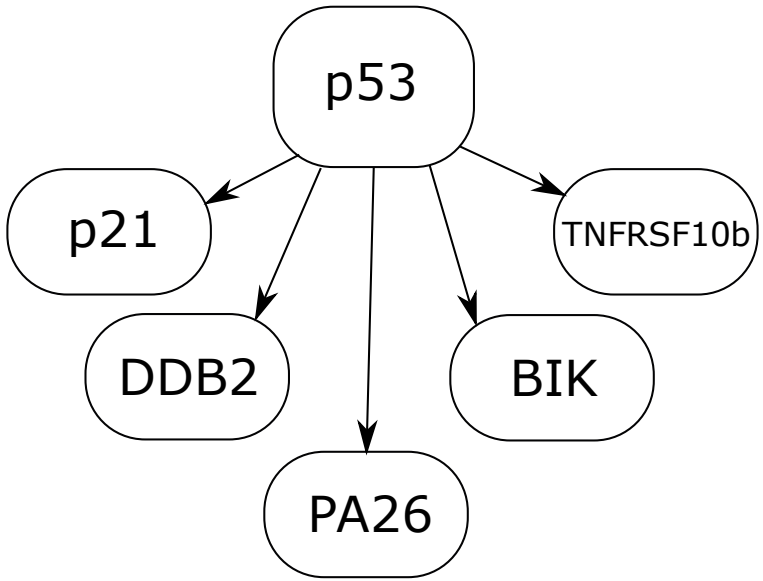


Figure: p53 SIM network motif as modelled by Barenco et al. 2006.

Response of p53 to Ionizing Radiation

- Experiment by Barenco et al. 2006.
- Human leukemia cell line (MOLT4) containing functional p53 and harvested protein and RNA at regular intervals after irradiation.
- The time course was performed in triplicate, and mRNA concentrations measured using Affymetrix U133A microarrays.

Mathematical Model

- Reorder differential equations

$$\frac{\mathrm{d}x_j(t)}{\mathrm{d}t} + D_j x_j(t) = B_j + S_j f(t)$$

- We have observation of $x_j(t)$.
- An estimate of $\frac{\mathrm{d}x_j(t)}{\mathrm{d}t}$ is obtained through fitting polynomials.
- Jointly estimate $f(t)$ at observations of time points along with $\{B_j, D_j, S_j\}_{j=1}^{g}$.
- Use MCMC sampling or maximum likelihood for parameters.

Response of p53

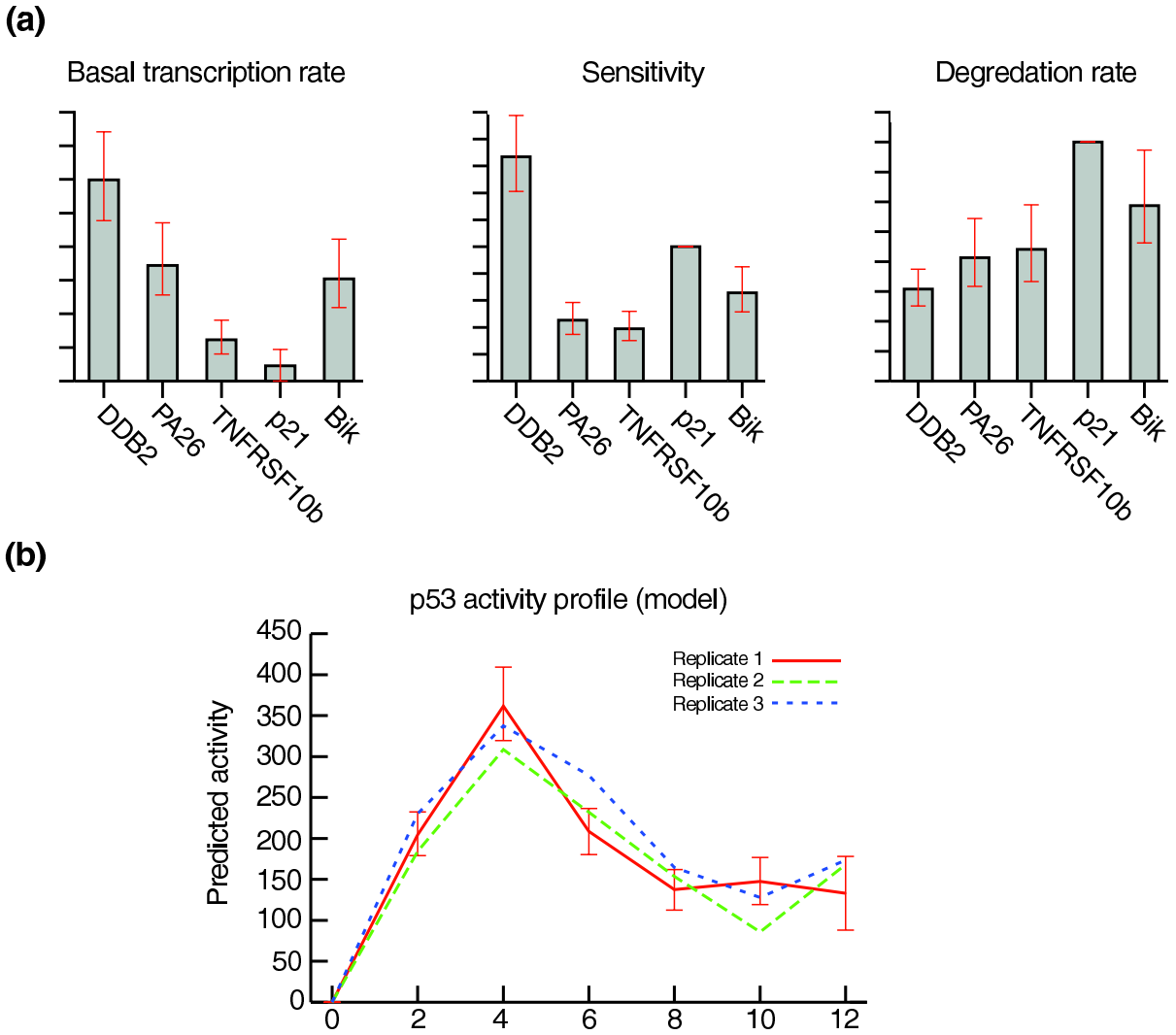


Figure: Results from Barenco et al. (2006). Top is parameter estimates. Bottom is inferred profile.

Respose to p53 ...

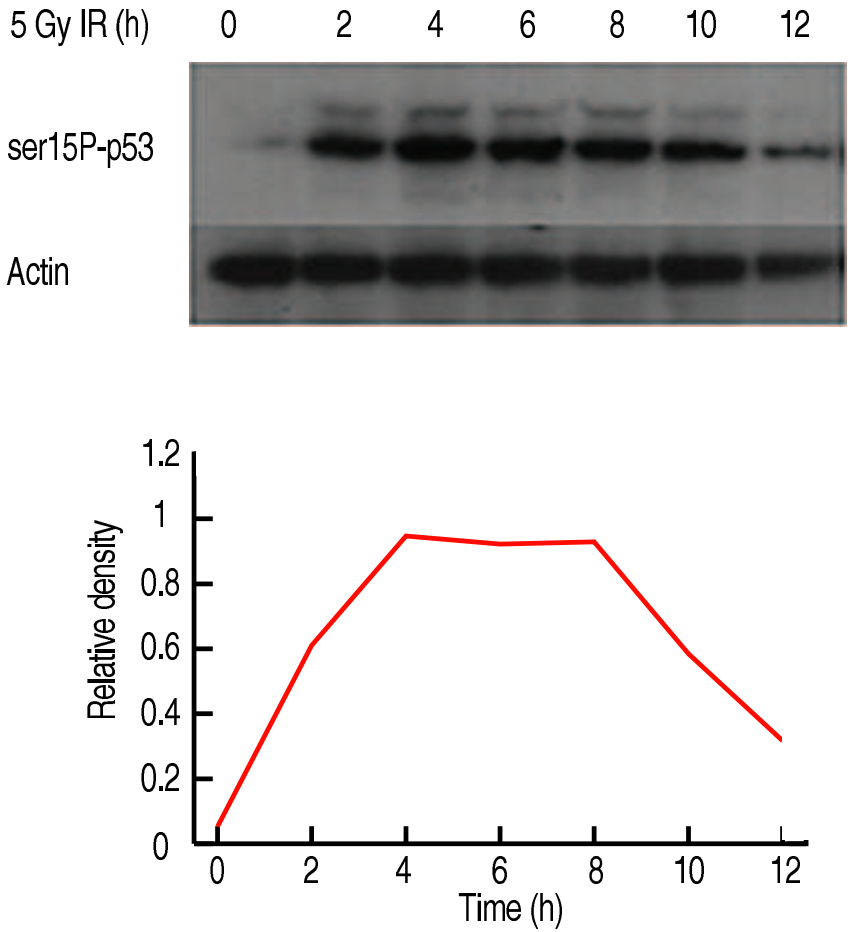


Figure: Results from Barenco et al. (2006). Activity profile of p53 was measured by Western blot to determine the levels of ser-15 phosphorylated p53 (ser15P-p53).

Models of non-linear regulation

- Non-linear Activation: Michaelis-Menten Kinetics

$$\frac{\mathrm{d}x_i(t)}{\mathrm{d}t} = B_i + \frac{S_i f(t)}{\gamma_i + f(t)} - D_i x_i(t)$$

  used by Rogers and Girolami (2006)

- Non-linear Repression

$$\frac{\mathrm{d}x_i(t)}{\mathrm{d}t} = B_i + \frac{S_i}{\gamma_i + f(t)} - D_i x_i(t)$$

  used by Khanin et al., 2006, PNAS 103

SOS Response

- Post replication DNA system: allows DNA replication to bypass errors in the DNA.
- DNA damage may occur as a result of activity of antibiotics.
- LexA is bound to the genome preventing transcription of the SOS genes.
- RecA protein is stimulated by single stranded DNA, inactivates the LexA repessor.
- This allows several of the LexA targets to transcribe.
- The SOS pathway may be essential in antibiotic resistance Cirz et al. (2005).
- Aim is to target these proteins to produce drugs to increase efficacy of antibiotics Lee et al. (2005).

LexA Experimental Description

- Data from Courcelle et al. (2001)
- UV irradiation of *E. coli.* in both wild-type cells and lexA1 mutants, which are unable to induce genes under LexA control.
- Response measured with two color hybridization to cDNA arrays.

Their Model

Given measurements of gene expression at N time points $(t_0, t_1, \ldots, t_{N-1})$, the temporal profile of a gene i, $x_i(t)$, that solves the ODE in Eq. 1 can be approximated by

$$x_i(t) = x_i^0 e^{-\delta_i t} + \frac{B_i}{D_i} + S_i e^{-\delta_i t} \frac{1}{D_i} \sum_{j=0}^{N-2} \left(e^{D_i t_j + 1} - e^{D_i t_j}\right) \frac{1}{\gamma_i + \bar{f}_j}$$

where $\bar{f}_j = \frac{(f(t_j) + f(t_j+1))}{2}$ on each subinterval $(t_j, t_j + 1), j = 0, \ldots, N-2$. This is under the simplifying assumption that $f(t)$ is a piece-wise constant function on each subinterval $(t_j, t_j + 1)$. **One can come up with linear (or higher order) $f(t)$ approximations on each subinterval. This will introduce additional parameters, which will be impossible to infer with any certainty given limited amount of data.**

Khanin et al. (2006)

Their Results

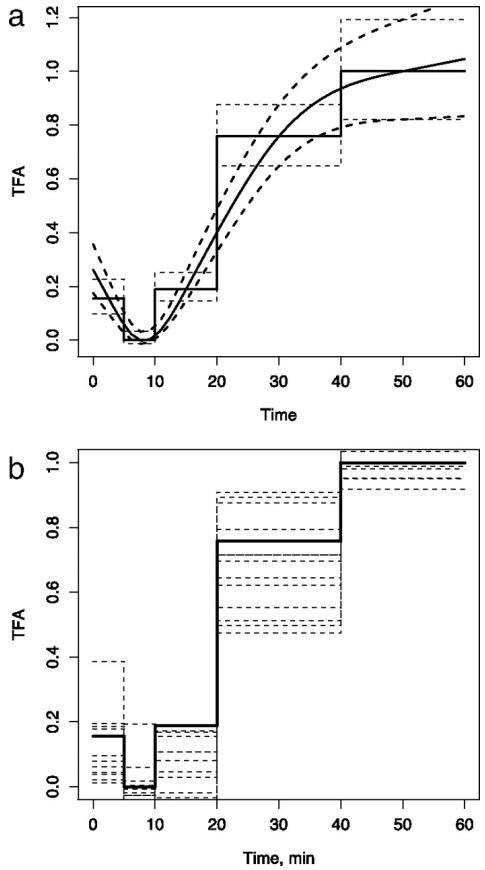


Figure: Fig. 2 from Khanin et al. (2006): Reconstructed activity level of master repressor LexA, following a UV dose of 40 J/m2.

Their Results

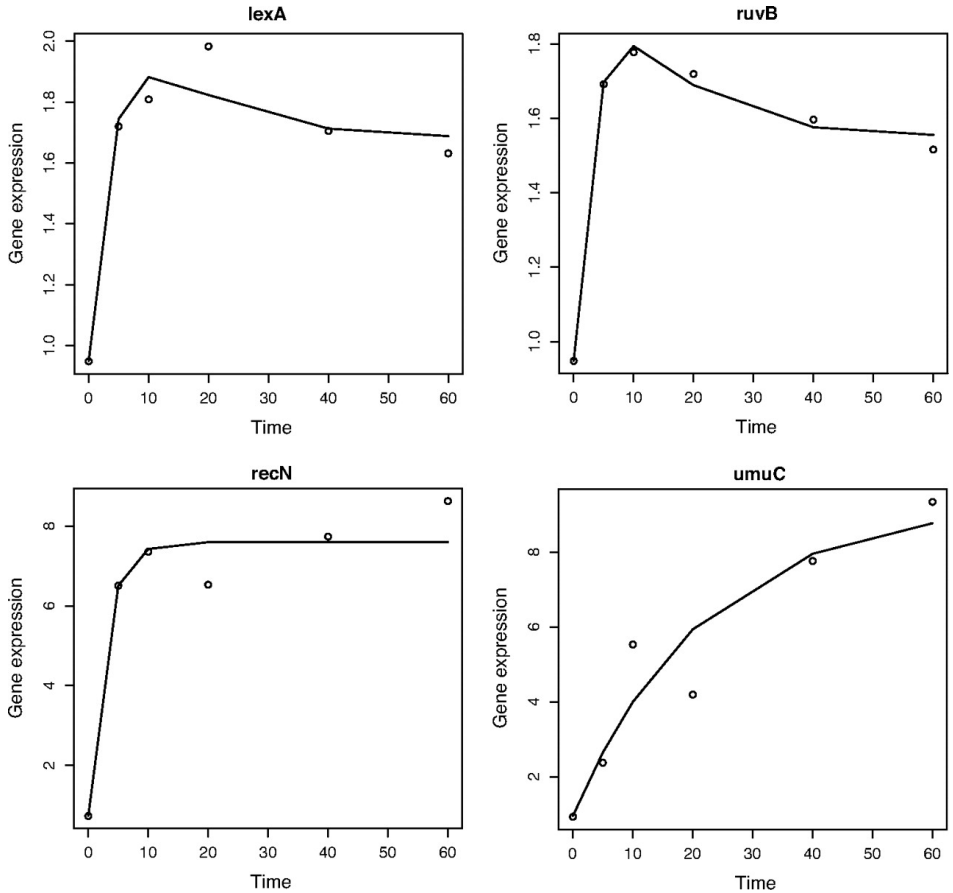


Figure: Fig. 3 from Khanin et al. (2006): Reconstructed profiles for four genes in the LexA SIM.

Actin and Ribosomes

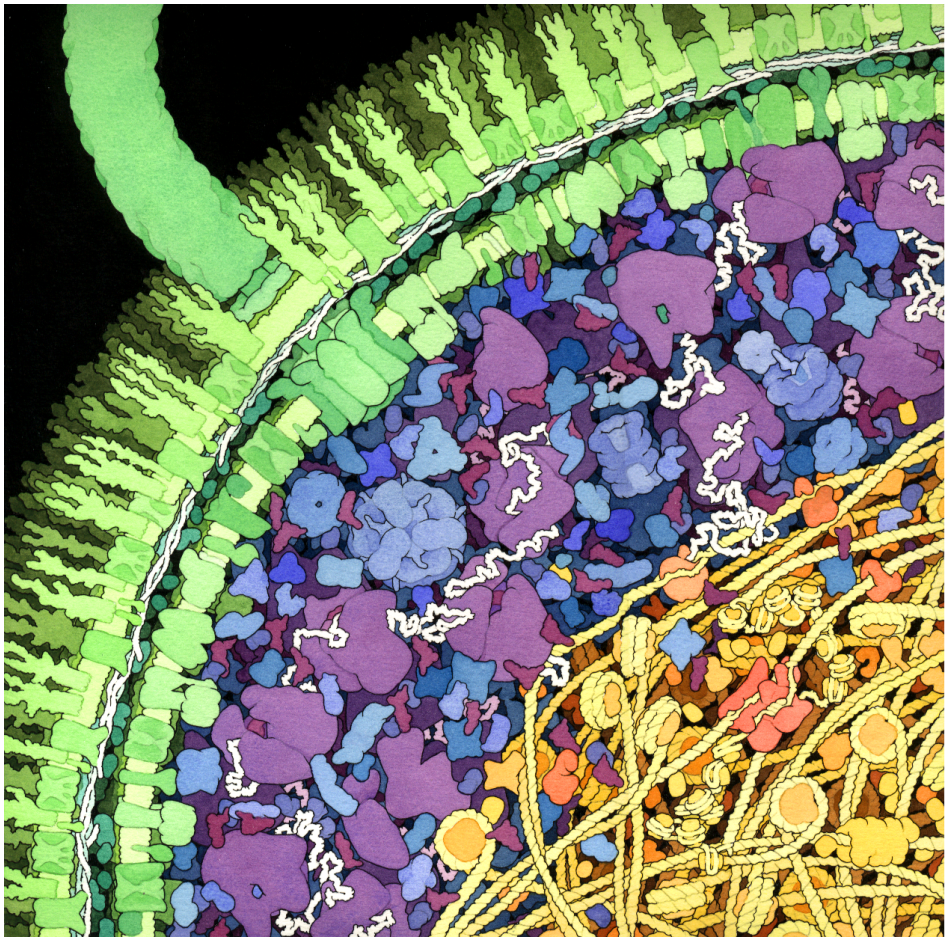


Figure: *E. coli* cell. Illustration courtesy of David S. Goodsell `http://mgl.scripps.edu/people/goodsell/illustration/public`. Confined structure leads to attempts to characterise diffusion in confined spaces, *e.g.* Schuss et al. (2007)

Roadmap

1. Introduction & Philosophy
2. Biological Systems
3. Modelling the System
4. p53 and SOS Response
5. Signalling Pathway
6. Acknowledgements

ERK Signalling Pathway

- Epidermal Growth Factor 40,000-100,000 EGFR per cell.
- Over expressed in tumours — some breast cancer cells 2×10^6 receptors per cell Herbst (2004).
- Over expression leads to an intense signal generation and activation of down stream signalling pathways.

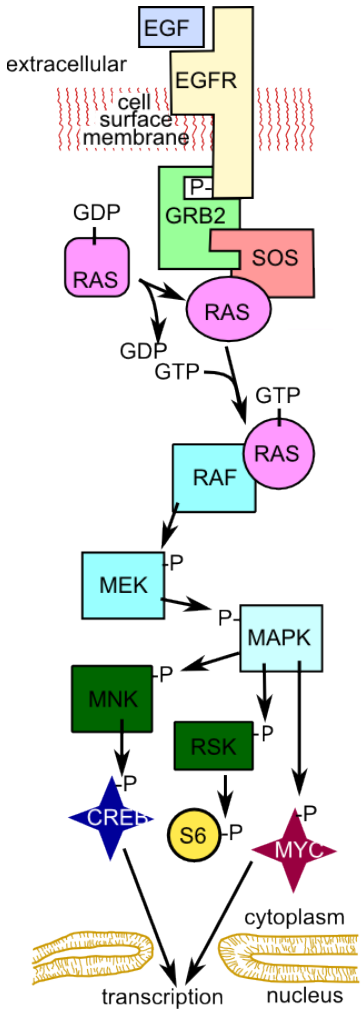


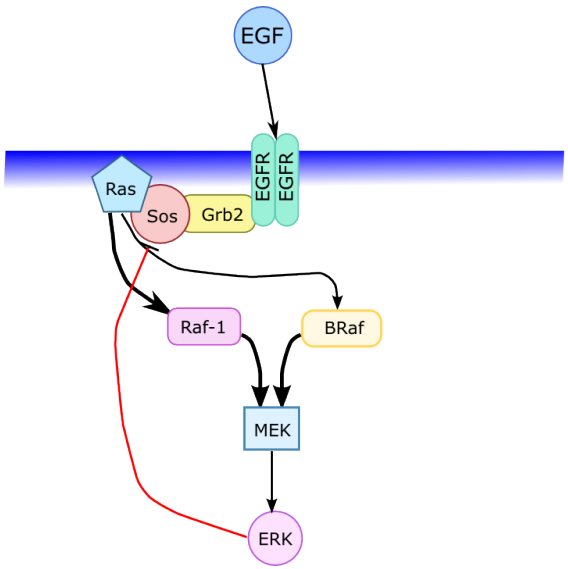
Figure: MAPK Pathway

Multiple Mechanistic Models

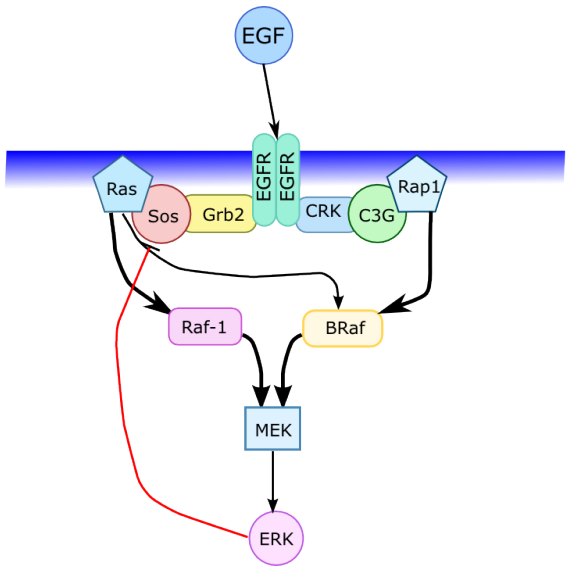
Vyshemirsky and Girolami (2008).

- Multiple mechanistic models describing a pathway.

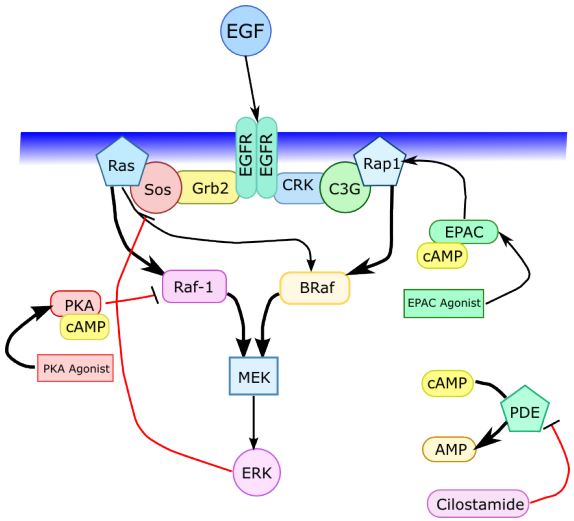
Model 1



Model 2



Model 3



Differential Equations

Models are formally defined using systems of ordinary differential equations:

$$\frac{\mathrm{d}\,[\mathrm{EGF}]}{\mathrm{d}t} = -k_1\,[\mathrm{EGF}]\,[\mathrm{EGFR}]$$

$$\frac{\mathrm{d}\,[\mathrm{Rap1}_a]}{\mathrm{d}t} = \frac{K_{cat12}\,[\mathrm{Rap1}_i]}{K_{m12} + [\mathrm{Rap1}_i]}\,[\mathrm{EPAC}] - \frac{V_{13}\,[\mathrm{Rap1}_a]}{K_{13} + [\mathrm{Rap1}_a]}$$

$$\frac{\mathrm{d}\,[\mathrm{MEK}]}{\mathrm{d}t} = -\frac{K_{cat21}\,[\mathrm{MEK}]\,[\mathrm{Raf}] - 1}{K_{m21} + [\mathrm{MEK}]} - \frac{K_{cat22}\,[\mathrm{MEK}]}{K_{m22} + [\mathrm{MEK}]}\,[\mathrm{BRaf}]$$

Model 1	**Model 2**
50 kinetic parameters	55 kinetic parameters

Bayes' Factors

- Which hypothesised structure is best supported by the data?
- Use Bayes factors: $\frac{P(M_1|D)}{P(M_2|D)}$, ratio of model marginal likelihoods.
- Difficulty is computing $P(M_1|D)$.
- Turn to the *thermodynamic integral* for results.

Thermodynamic Integral

Gelman and Meng (1998)

$$p\left(\boldsymbol{\theta}|\mathbf{x}, M, \alpha\right) = \frac{p\left(\mathbf{x}|\theta, M\right)^{\alpha} p\left(\theta|M\right)}{Z_{\alpha}}$$

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \log Z_{\alpha} = \frac{1}{Z_{\alpha}} \frac{\mathrm{d}}{\mathrm{d}\alpha} Z_{\alpha} = \left\langle \log p\left(\mathbf{x}|\boldsymbol{\theta}\right) \right\rangle_{p(\boldsymbol{\theta}|\mathbf{x}, M, \alpha)}$$

giving

$$\log p\left(\mathbf{x}|M\right) = \int_{0}^{1} \left\langle \log p\left(\mathbf{x}|\boldsymbol{\theta}\right) \right\rangle_{p(\boldsymbol{\theta}|\mathbf{x}, M, \alpha)} \mathrm{d}\alpha$$

Estimation through Sampling

- Construct estimator using samples drawn from prior and all intermediate densities up to the posterior. Gelman and Meng (1998); Friel and Pettit (2008); Lartillot and Philippe (2006).
- Represent integral by discrete α_i values and expectation using trapezoidal rule

$$\log p(\mathbf{y}) = \frac{1}{2}\sum_{i=1}^{L} \Delta_i \left[\left\langle \log p(\mathbf{y}|\boldsymbol{\theta})\right\rangle_{p(\boldsymbol{\theta}|\mathbf{y},\alpha_{i-1})} + \left\langle \log p(\mathbf{y}|\boldsymbol{\theta})\right\rangle_{p(\boldsymbol{\theta}|\mathbf{y},\alpha_i)}\right] + \frac{1}{2}\sum_{i=1}^{L}\epsilon_i, \tag{1}$$

where $\Delta_i = \alpha_i - \alpha_{i-1}$, discretization error is $\epsilon_i = \mathrm{KL}\left(p_{i-1} \,\|\, p_i\right) - \mathrm{KL}\left(p_i \,\|\, p_{i-1}\right)$.

Posterior for Different α

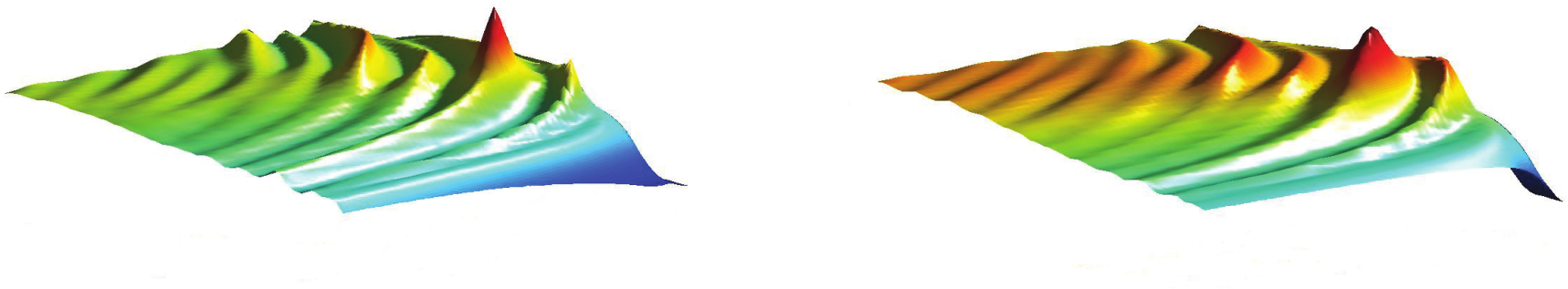


Figure: Annnealing of likelihood. (here $\alpha = 1\alpha = 0.55\alpha = 0.28\alpha = 0.13\alpha = 0.05$ $\alpha = 0$)

Population Monte Carlo

- Further problems from highly multimodal posteriors — use population Monte Carlo methods.

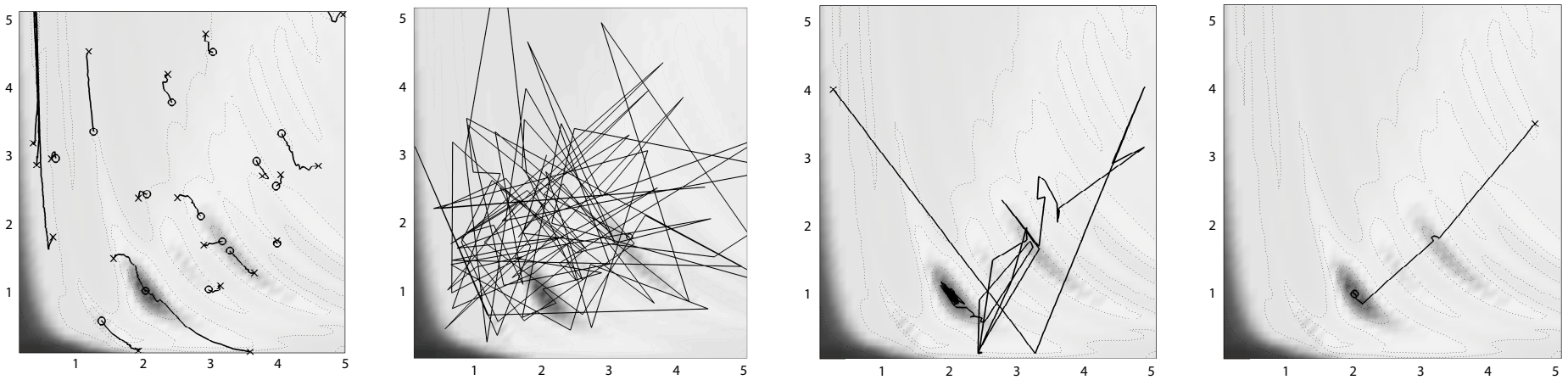
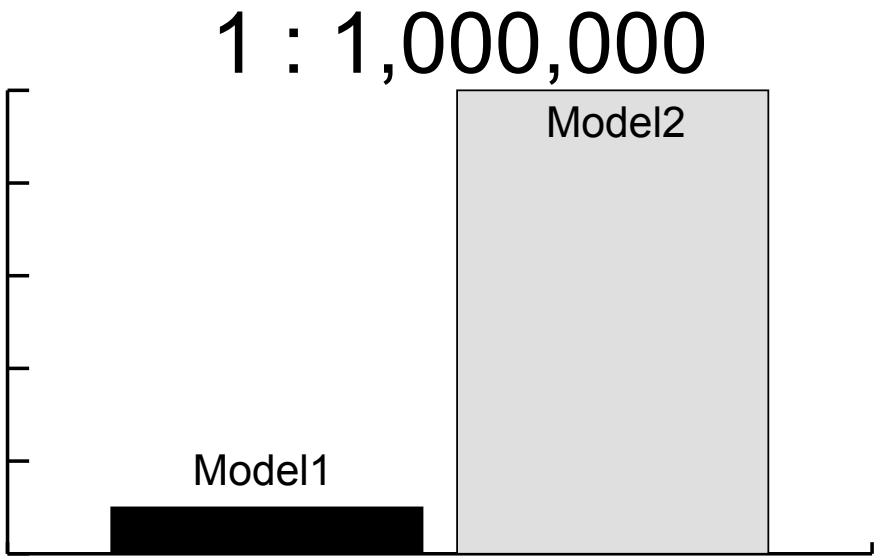
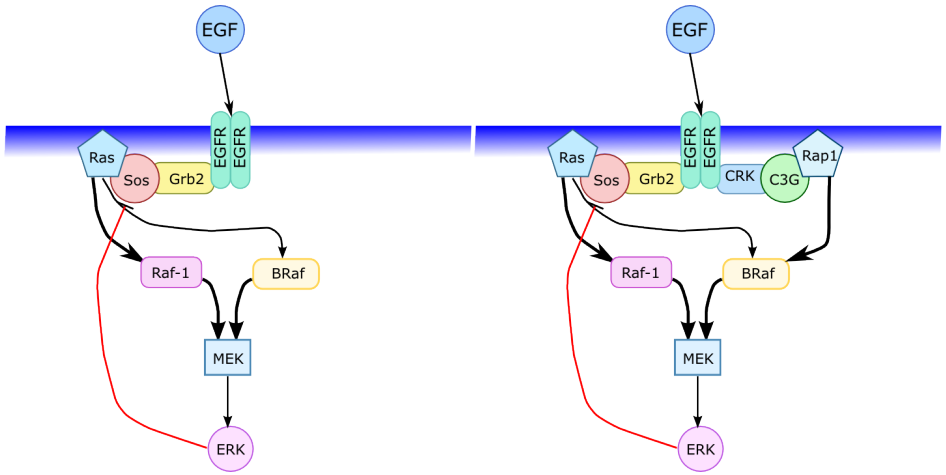


Figure: *Far Left*: standard Monte Carlo gets stuck in different modes. *Middle left*: exploration of space for low α. *Middle right:* intermediate α allows movement between modes. *Far left:* information is exchanged between samples to allow full exploration of posterior.

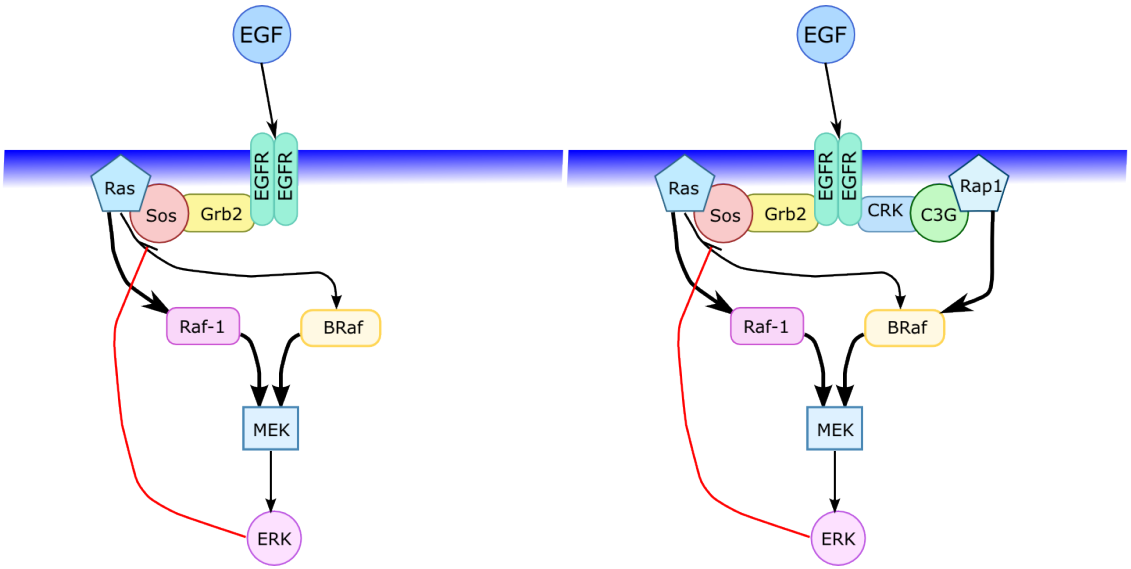
Result

Bayes' Factors for ERK signalling: Result





Hypothesis Implications



- Double branched model has much better support from the experimental evidence: leads to a robust system.
- BRaf was found to be more active than Raf-1. This is confirmed by a number of publications in biochemical journals.
- siRNA Knock-Down experiments have confirmed dual-branch hypothesis (Walter Kolch).

Story So Far

- Systems biology presents us with models and data.
- Challenge for machine learning: introduce our inference techniques to this domain.
- Lots of work on methodological developments necessary still.
- **Next**: an approach to dealing with differential equations with missing chemical species.
  - Gaussian processes allow integration of Bayesian probabilistic inference with differential equations.

Outline

1. Introduction & Philosophy
2. Biological Systems
3. Modelling the System
4. p53 and SOS Response
5. Signalling Pathway
6. Acknowledgements

Acknowledgements

- Investigators: Neil Lawrence and Magnus Rattray
- Researchers: Peo Gao, Antti Honkela, Michalis Titsias and Jennifer Withers
- Charles Girardot and Eileen Furlong of EMBL in Heidelberg (mesoderm development in *D. Melanogaster*).
- Martino Barenco and Mike Hubank at the Institute of Child Health in UCL (p53 pathway).
- Raya Khanin and Ernst Wit of the University of Glasgow and the University of Lancaster (*E. coli* repressor system).

Funded by the BBSRC award "Improved Processing of microarray data using probabilistic models" and EPSRC award "Gaussian Processes for Systems Identification with applications in Systems Biology"

References I

M. Barenco, D. Tomescu, D. Brewer, R. Callard, J. Stark, and M. Hubank. Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biology*, 7(3):R25, 2006. [PDF].

R. T. Cirz, J. K. Chin, D. R. Andes, V. de Crécy-Lagard, W. A. Craig, and F. E. Romesberg. Inhibition of mutation and combating the evolution of antibiotic resistance. *PLoS Biology*, 3(6), 2005.

J. Courcelle, A. Khodursky, B. Peter, P. O. Brown, , and P. C. Hanawalt. Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient *Escherichia coli*. *Genetics*, 158:41–64, 2001.

A. P. Dawid. The well-callibrated Bayesian. *Journal of the American Statistical Association*, 77:605–613, 1982.

A. P. Dawid. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society, A*, 147:278–292, 1984.

M. B. Elowitz and S. Leibler. Synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–338, 2000. [DOI].

R. P. Feynman. There's plenty of room at the bottom: An invitation to enter a new field of physics. Talk at Annual meeting of the American Physical Society, 1959. Available from `http://www.zyvex.com/nanotech/feynman.html`.

N. Friel and A. N. Pettit. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society, B*, 70: 589–607, 2008.

A. Garfinkel. Reductionism. In R. Boyd, P. Gasper, and J. D. Trout, editors, *The Philosophy of Science*, pages 443–459. MIT Press, 1991. [Google Books] .

A. Gelman and X.-L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185, 1998.

M. A. Gibson and J. Bruck. Efficient exact stochastic simulation of chemical systems with many species and many channels. *J. Phys Chem. A*, 104:1876–1889, 2000.

D. S. Goodsell. The molecular perspective: p53 tumor suppressor. *The Oncologist, Vol. 4, No. 2, 138-139, April 1999*, 4(2): 138–139, 1999.

R. S. Herbst. Review of epidermal growth factor receptor biology. *International Journal of Radiation Oncology*, 59(2):S21–S26, 2004. [DOI].

References II

S. Hoops, S. Sahle, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, and U. Kummer. COPASI: a COmplex PAthway SImulator. *Bioinformatics*, 22(24):3067–3074, 2006.

R. Khanin, V. Viciotti, and E. Wit. Reconstructing repressor protein levels from expression of gene targets in *E. Coli*. *Proc. Natl. Acad. Sci. USA*, 103(49):18592–18596, 2006. [PDF]. [DOI].

N. Lartillot and H. Philippe. Computing Bayes factors using thermodynamic integration. *Syst. Biol.*, 55(2):195–207, 2006.

N. D. Lawrence, M. Girolami, M. Rattray, and G. Sanguinetti, editors. *Learning and Inference in Computational Systems Biology*. MIT Press, 2010. In press.

Y. Lazebnik. Can a biologist fix a radio? or, what I learned while studying apoptosis. *Cancer Cell*, 2:179–182, 2002. [PDF].

A. M. Lee, C. T. Ross, B.-B. Zeng, , and S. F. Singleton. A molecular target for suppression of the evolution of antibiotic resistance: Inhibition of the *Escherichia coli* RecA protein by N6-(1-Naphthyl)-ADP. *J. Med. Chem.*, 48(17), 2005.

K. R. Popper. *Science: Conjectures and Refutations*, chapter 1. Routledge, London, 1963. [Google Books] .

S. Rogers and M. Girolami. Model based identification of transcription factor regulatory activity via Markov chain Monte Carlo. Presentation at MASAMB '06, 2006.

Z. Schuss, A. Singer, and D. Holcman. The narrow escape problem for diffusion in cellular microdomains. *Proc. Natl. Acad. Sci. USA*, 104(41):16098–16103, 2007. [DOI].

V. Vyshemirsky and M. A. Girolami. Bayesian ranking of biochemical system models. *Bioinformatics*, 24(6):833–839, 2008. [PDF]. [DOI].

D. J. Wilkinson. *Stochastic Modelling for Systems Biology*. Chapman and Hall/CRC, 2006. [Google Books] .