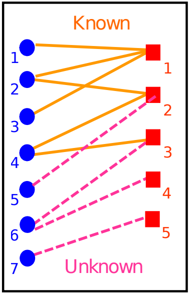
A Bayesian active learning strategy for sequential experimental design in systems biology

Pauwels E., Lajaunie C., and Vert J.P.

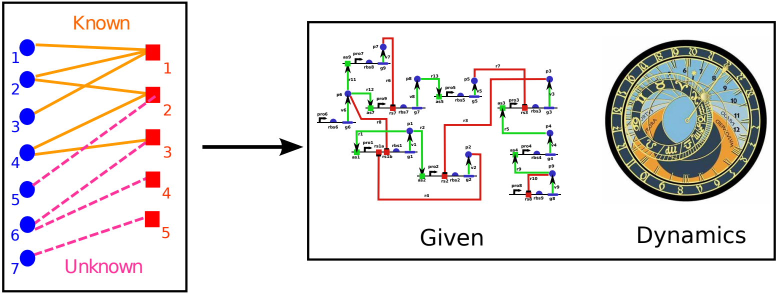
Seminar MIA-T, INRA,
February 14 2014

Sequential experimental design for systems biology

Sequential experimental design for systems biology



Sequential experimental design for systems biology



Context

Many biological problems involve dynamical mechanisms
(regulation, triggering, transport, . . .)

Context

Many biological problems involve dynamical mechanisms (regulation, triggering, transport, ...)

Add one layer of complexity

- "mRNA m is translated into protein p"

Context

Many biological problems involve dynamical mechanisms
(regulation, triggering, transport, ...)

Add one layer of complexity

- "mRNA m is translated into protein p"
- $m \rightarrow p$

Context

Many biological problems involve dynamical mechanisms
(regulation, triggering, transport, ...)

Add one layer of complexity

- "mRNA m is translated into protein p"
- $m \rightarrow p$
- $\frac{d[p]}{dt} = C[m]$

Context

Many biological problems involve dynamical mechanisms (regulation, triggering, transport, ...)

Add one layer of complexity

- "mRNA m is translated into protein p"
- $m \to p$
- $\frac{d[p]}{dt} = C[m]$

A challenging issue

Context

Many biological problems involve dynamical mechanisms (regulation, triggering, transport, ...)

Add one layer of complexity

- "mRNA m is translated into protein p"
- $m \rightarrow p$
- $\frac{d[p]}{dt} = C[m]$

A challenging issue

- Data fit: need to estimate kinetic parameters

Context

Many biological problems involve dynamical mechanisms (regulation, triggering, transport, ...)

Add one layer of complexity

- "mRNA m is translated into protein p"
- $m \to p$
- $\frac{d[p]}{dt} = C[m]$

A challenging issue

- Data fit: need to estimate kinetic parameters
- Hard problem: many interacting species

Context

Many biological problems involve dynamical mechanisms (regulation, triggering, transport, ...)

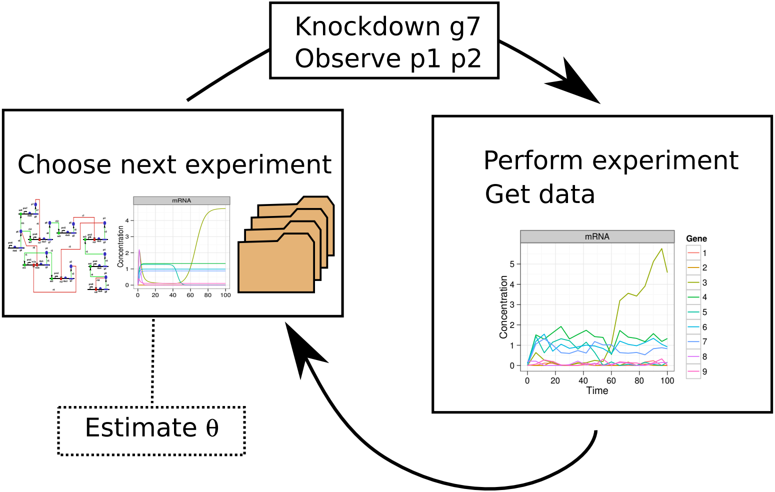
Add one layer of complexity

- "mRNA m is translated into protein p"
- $m \to p$
- $\frac{d[p]}{dt} = C[m]$

A challenging issue

- Data fit: need to estimate kinetic parameters
- Hard problem: many interacting species
- Dream challenge (IBM, EMBL):
  - simulated data
  - molecular perturbation
  - budget constraint

Context


Knockdown g7
Observe p1 p2
Choose next experiment
mRNA
Concentration
Perform experiment
Get data
mRNA
Concentration
Time
Gene
Estimate θ

Outline

1. Context

2. Problem formulation and proposed method

3. Simulation results

Problem set up

Notations

- Model kinetic parameters: $\theta \in \Theta \subseteq \mathbb{R}^p$, unknown θ^*

Problem set up

Notations

- Model kinetic parameters: $\theta \in \Theta \subseteq \mathbb{R}^p$, unknown θ^*
- Experiment: $e \in \mathcal{E}$

  Molecular perturbation: gene deletion, affinity constant decrease, ...
  Observation: protein and mRNA concentration
  Time resolution

Problem set up

Notations

- Model kinetic parameters: $\theta \in \Theta \subseteq \mathbb{R}^p$, unknown θ^*
- Experiment: $e \in \mathcal{E}$

  Molecular perturbation: gene deletion, affinity constant decrease, ...
  Observation: protein and mRNA concentration
  Time resolution
- Model that drives concentration dynamics:

  $\dot{Y} = f(Y, e, \theta)$, unknown θ^*

Problem set up

Notations

- Model kinetic parameters: $\theta \in \Theta \subseteq \mathbb{R}^p$, unknown θ^*
- Experiment: $e \in \mathcal{E}$

  Molecular perturbation: gene deletion, affinity constant decrease, ...
  Observation: protein and mRNA concentration
  Time resolution
- Model that drives concentration dynamics:

  $\dot{Y} = f(Y, e, \theta)$, unknown θ^*
- Experiment: choose $e \in \mathcal{E}$ and get $o \sim P(o|\theta^*; e)$

  $P(o|\theta; e)$ known for any θ and e.

Problem set up

Notations

- Model kinetic parameters: $\theta \in \Theta \subseteq \mathbb{R}^p$, unknown θ^*
- Experiment: $e \in \mathcal{E}$

  Molecular perturbation: gene deletion, affinity constant decrease, ...
  Observation: protein and mRNA concentration
  Time resolution
- Model that drives concentration dynamics:

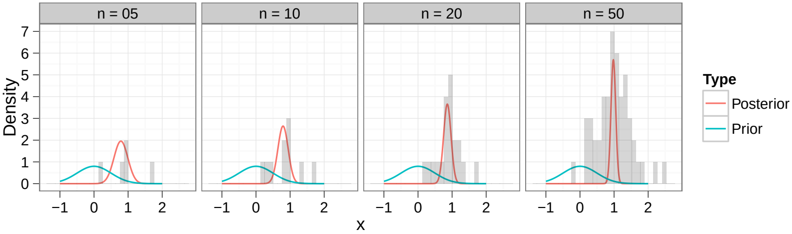
  $\dot{Y} = f(Y, e, \theta)$, unknown θ^*
- Experiment: choose $e \in \mathcal{E}$ and get $o \sim P(o|\theta^*; e)$

  $P(o|\theta; e)$ known for any θ and e.

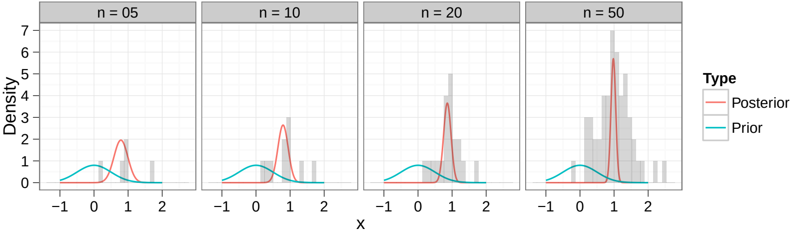
Experimental design: estimate θ^*

- Sequentially choose experiments
- Experimental cost, limited budget

Brief recall on Bayesian update



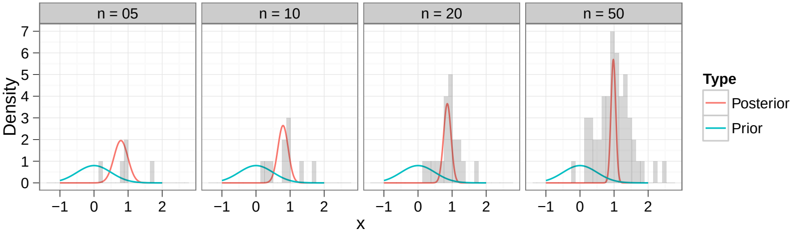
Brief recall on Bayesian update



- Bayes rule: prior π, data o from experiments e

$$P(\theta|o;e) = \frac{P(o|\theta;e)\,\pi(\theta)}{\int_{\theta'} P(o|\theta';e)\,\pi(\theta')d\theta'}$$

Brief recall on Bayesian update



- Bayes rule: prior π, data o from experiments e

$$P(\theta|o;e) = \frac{P(o|\theta;e)\,\pi(\theta)}{\int_{\theta'} P(o|\theta';e)\,\pi(\theta')d\theta'}$$

- Numerical integration

Main idea

- π encodes knowledge about θ^*, loss function $\ell(\theta, \theta^*)$

Main idea

- π encodes knowledge about θ^*, loss function $\ell(\theta, \theta^*)$
- quality of this distribtion

$$E_{\theta\sim\pi}\left[\ell(\theta, \theta^*)\right]$$

Main idea

- π encodes knowledge about θ^*, loss function $\ell(\theta, \theta^*)$
- quality of this distribtion

$$E_{\theta\sim\pi}\left[\ell(\theta, \theta^*)\right]$$

- suppose we choose experiment e and observe o

$$E_{\theta\sim P(\theta|o;e)}\left[\ell(\theta, \theta^*)\right]$$

Main idea

- π encodes knowledge about θ^*, loss function $\ell(\theta, \theta^*)$
- quality of this distribtion

$$E_{\theta \sim \pi} \left[\ell(\theta, \theta^*) \right]$$

- suppose we choose experiment e and observe o

$$E_{\theta \sim P(\theta | o; e)} \left[\ell(\theta, \theta^*) \right]$$

- average over possible observations

$$E_{o \sim P(o | \theta^*; e)} E_{\theta \sim P(\theta | o; e)} \left[\ell(\theta, \theta^*) \right]$$

Main idea

- π encodes knowledge about θ^*, loss function $\ell(\theta, \theta^*)$
- quality of this distribtion

$$E_{\theta\sim\pi}\left[\ell(\theta, \theta^*)\right]$$

- suppose we choose experiment e and observe o

$$E_{\theta\sim P(\theta|o;e)}\left[\ell(\theta, \theta^*)\right]$$

- average over possible observations

$$E_{o\sim P(o|\theta^*;e)}E_{\theta\sim P(\theta|o;e)}\left[\ell(\theta, \theta^*)\right]$$

- average using current state of knowledge

$$R(e;\pi) = E_{\theta'\sim\pi}E_{o\sim P(o|\theta';e)}E_{\theta\sim P(\theta|o;e)}\left[\ell(\theta, \theta')\right]$$

Main idea

- sequence of posteriors

$$\pi_k(\theta) = \frac{P(o_{k-1}|\theta; e_{k-1})\,\pi_{k-1}(\theta)}{\int_{\theta'} P(o_{k-1}|\theta'; e_{k-1})\,\pi_{k-1}(\theta')d\theta'}$$

- reference risk

$$R(\pi_k) = E_{\theta\sim\pi_k} E_{\theta'\sim\pi_k}\left[\ell(\theta,\theta')\right]$$

- next experiment choice

$$e_{k+1} = \arg\max_{e\in\mathcal{E}} \; \frac{R(\pi_k) - R(e;\pi_k)}{C_e}$$

Numerical evaluation of the $R(e;\pi)$

$$R(e;\pi) = \int_{\theta,\theta'} \ell(\theta,\theta') \int_{o} \frac{P(o|\theta;e)\,\pi(\theta)\,P(o|\theta';e)\,\pi(\theta')}{\int_{\theta''} P(o|\theta'';e)\,\pi(\theta'')d\theta''}\,d\theta\,d\theta'$$

Numerical evaluation of the $R(e;\pi)$

$$R(e;\pi) = \int_{\theta,\theta'} \ell(\theta,\theta')\ \int_o \frac{P(o|\theta;e)\,\pi(\theta)\,P(o|\theta';e)\,\pi(\theta')}{\int_{\theta''} P(o|\theta'';e)\,\pi(\theta'')d\theta''}\, d\theta\, d\theta'$$

- draw a sample $\{\theta_i\}_{i=1\ldots N}$ from π;

$$R(e;\pi) \simeq R^N(e;\pi) = \frac{1}{N^2}\sum_{i,j=1}^{N} \ell(\theta_i,\theta_j)\, w_{ij}(e)$$

where $w_{ij}(e) = \int_o \frac{P(o|\theta_i;e)\,P(o|\theta_j;e)}{\sum_{k=1}^{N} P(o|\theta_k;e)}\, do$

Numerical evaluation of the $R(e; \pi)$

$$R(e;\pi) = \int_{\theta,\theta'} \ell(\theta,\theta') \int_o \frac{P(o|\theta;e)\,\pi(\theta)\,P(o|\theta';e)\,\pi(\theta')}{\int_{\theta''} P(o|\theta'';e)\,\pi(\theta'')d\theta''}\, d\theta\, d\theta'$$

- draw a sample $\{\theta_i\}_{i=1\ldots N}$ from π;

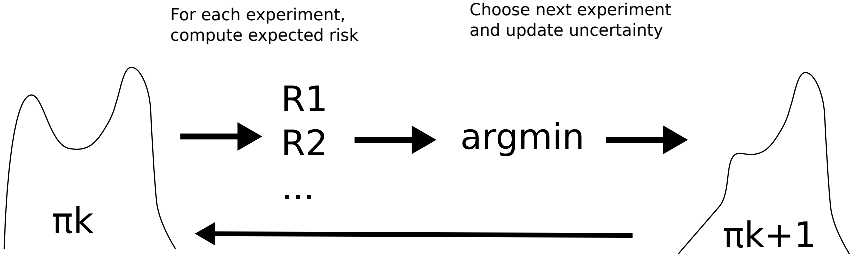
$$R(e;\pi) \simeq R^N(e;\pi) = \frac{1}{N^2}\sum_{i,j=1}^{N} \ell(\theta_i,\theta_j)\, w_{ij}(e)$$

where $w_{ij}(e) = \int_o \frac{P(o|\theta_i;e)\,P(o|\theta_j;e)}{\sum_{k=1}^{N} P(o|\theta_k;e)}\, do$

- draw a sample $\{o_u^i\}_{u=1,\cdots,M}$ from each $P(o|\theta_i;e)$

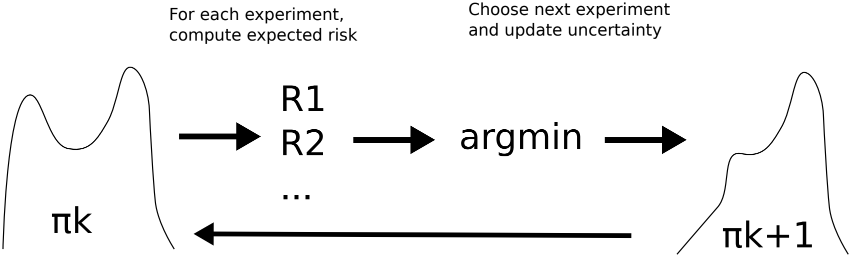
$$w_{ij}(e) \simeq w_{ij}^M(e) = \frac{1}{M}\sum_{u=1}^{M} \frac{P(o_u^i|\theta_j;e)}{\sum_{k=1}^{N} P(o_u^i|\theta_k;e)}$$

Summary


For each experiment,
compute expected risk
Choose next experiment
and update uncertainty
πk
R1
R2
...
argmin
πk+1

Sequential design

Summary



Conceptual advantage

Provides a unique criterion for experimental design

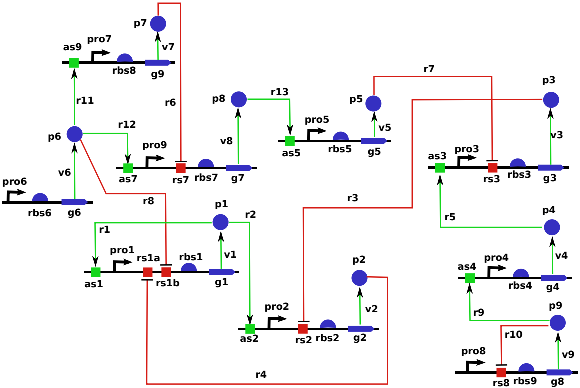
Expectation approximation

Computationally intensive, accuracy difficult to monitor

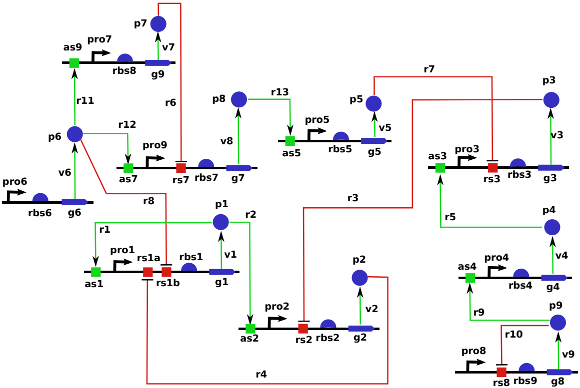
Outline

1. Context
2. Problem formulation and proposed method
3. Simulation results

Dream sub-challenge 1



Dream sub-challenge 1



45 kinetic parameters, 9 genes, 18 molecular species

Limited budget

Estimate true kinetic parameter

Estimate concentration time course for an unseen experiment

Exploring the space of parameters

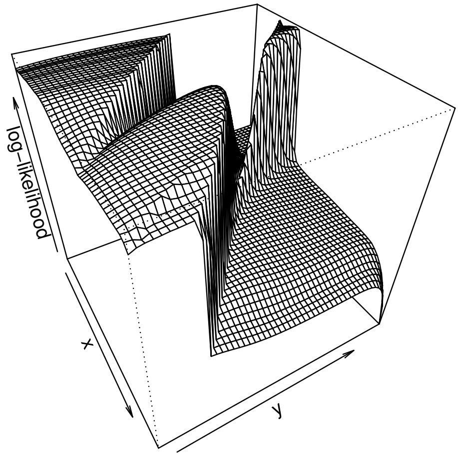
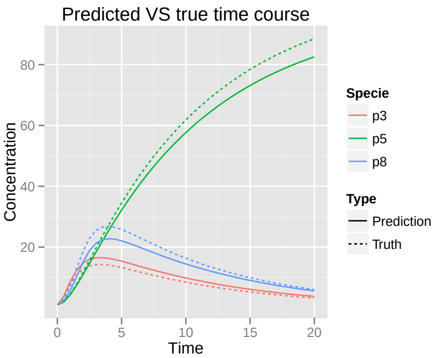


Figure : Posterior surface along a 2D space.

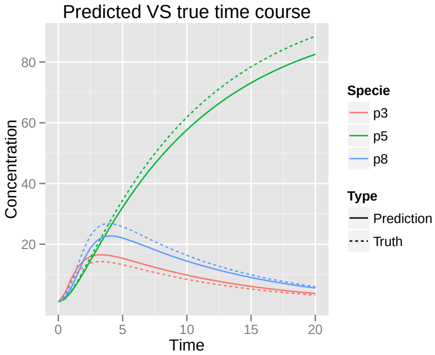
Challenge results

Rank	Dparam	Dprot
1	0.0229	0.0024
2	0.8404	0.0160
3	0.1592	0.0354
4	0.0899	0.0475
5	0.1683	0.0979
6	0.0453	0.1988
7	0.1702	0.3625
8	**0.8128**	**0.3564**
9	0.3766	0.8180
10	0.0699	19.3233
11	0.1883	3.2228
12	5.0278	14.7744



Challenge results

Rank	Dparam	Dprot
1	0.0229	0.0024
2	0.8404	0.0160
3	0.1592	0.0354
4	0.0899	0.0475
5	0.1683	0.0979
6	0.0453	0.1988
7	0.1702	0.3625
8	**0.8128**	**0.3564**
9	0.3766	0.8180
10	0.0699	19.3233
11	0.1883	3.2228
12	5.0278	14.7744



Reproduce global dynamical behaviour

Challenge results

Rank	Dparam	Dprot
1	**0.0229**	**0.0024**
2	0.8404	0.0160
3	0.1592	0.0354
4	0.0899	0.0475
5	0.1683	0.0979
6	0.0453	0.1988
7	0.1702	0.3625
8	**0.8128**	**0.3564**
9	0.3766	0.8180
10	0.0699	19.3233
11	0.1883	3.2228
12	5.0278	14.7744

Winner of dream network parameter inference challenge 6

PLoS ONE | www.plosone.org July 2012 | Volume 7 | Issue 7 | e40052

Experimental Design for Parameter Estimation of Gene Regulatory Networks

Bernhard Steiert[1,2,3]*, Andreas Raue[1,2.], Jens Timmer[1,2,3,4,5], Clemens Kreutz[1,2.]

1 Institute for Physics, University of Freiburg, Freiburg, Germany, 2 Freiburg Center for Systems Biology, University of Freiburg, Freiburg, Germany, 3 Freiburg Institute for Advanced Studies, University of Freiburg, Freiburg, Germany, 4 BIOSS Centre for Biological Signalling Studies, University of Freiburg, Freiburg, Germany, 5 Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden

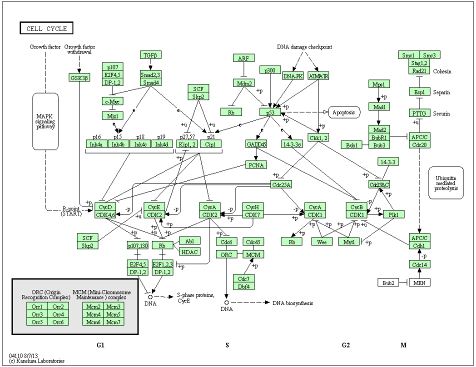
Table 1. Overview of the criteria that were considered for the final decisions.

Abbreviaton	Detailed explanation
(WT)	Wild-type measurements provide the largest data-points to credits ratio.
(P. mRNA)	Protein data has a better data-points to credits ratio than mRNA data. However, the species to be measured have to be specified and choosing the wrong time-courses can yield only little information gain.
(MA)	For microarray data, there is no decision required about which compounds should be measured. This makes the design more robust. If there are fast processes, high-density time resolution is favorable in comparison to low-density measurements.
(OptPerPL)	Perturbation experiments D are selected as maximally informative based on the PL, if the score R(D) in (13) is optimal.
(GelShift)	Because a single time course data set is not informative enough to resolve the practical non-identifiability, this parameter was measured directly by a gel-shift experiment.
(Module)	The parameters to be bought are in a sub-module of bad estimates and therefore there is hope to improve identifiability of the whole module.
(LocMin)	If several local minima have been detected with similar agreement to the data, designs are chosen which optimally discriminate between the local minima.
(SwitchDyn)	The model shows qualitatively different dynamics and a perturbation is able to switch the model's behavior.
(Extra)	The experiment or the parameter values are important for improving the accuracy of the demanded model extrapolation.
(Budget)	Sometimes, experiments are advantageous because the remaining credits allow a more flexible planning or the budget can be spent more comprehensively.

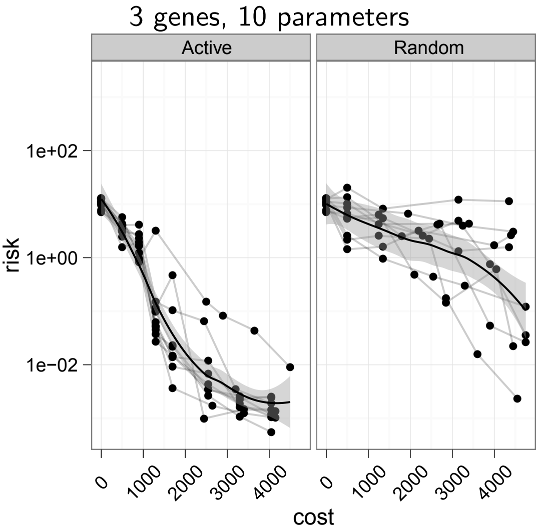
Challenge results

Rank	Dparam	Dprot
1	**0.0229**	**0.0024**
2	0.8404	0.0160
3	0.1592	0.0354
4	0.0899	0.0475
5	0.1683	0.0979
6	0.0453	0.1988
7	0.1702	0.3625
8	**0.8128**	**0.3564**
9	0.3766	0.8180
10	0.0699	19.3233
11	0.1883	3.2228
12	5.0278	14.7744

Kegg cell cycle, 124 genes



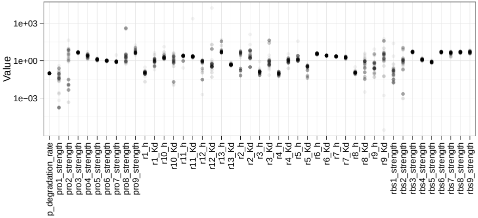
Simulations on a subnetwork

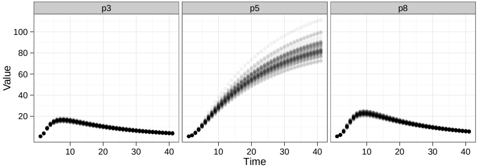


Estimate a single parameter?

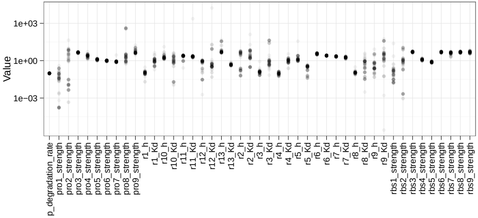
Rank	Dparam	Dprot
2	0.8404	0.0160
10	0.0699	19.3233

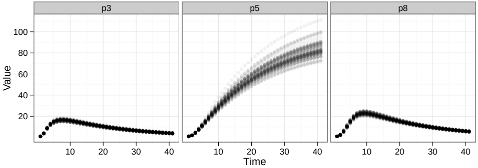
Rank	Dparam	Dprot
2	0.8404	0.0160
10	0.0699	19.3233





Rank	Dparam	Dprot
2	0.8404	0.0160
10	0.0699	19.3233





- Already pointed out in the literature
- Mispecified model, stochastic dynamics, real data . . .

Discussion

- An example of hard small scale problem

Discussion

- An example of hard small scale problem
- Reproducibility is a prerequisite for experimental design
  - Subjectivity, robustness
  - Scale
  - Accessibility to non specialists

Discussion

- An example of hard small scale problem
- Reproducibility is a prerequisite for experimental design
  - Subjectivity, robustness
  - Scale
  - Accessibility to non specialists
- Questions the focus on single parameter estimates

Discussion

- An example of hard small scale problem
- Reproducibility is a prerequisite for experimental design
  - Subjectivity, robustness
  - Scale
  - Accessibility to non specialists
- Questions the focus on single parameter estimates
- Computational challenges
  - Numerical integration in high dimensions
  - Uncertainty propagation in dynamical systems

Submition to BMC systems biology
R packaged code for the subnetwork simulations
Many thanks to Christian Lajaunie and Jean-Philippe Vert.

Submition to BMC systems biology
`R` packaged code for the subnetwork simulations
Many thanks to Christian Lajaunie and Jean-Philippe Vert.

Thank you