

BIOMENG 261

TISSUE AND BIOMOLECULAR ENGINEERING

Module I: Reaction kinetics and systems biology

Oliver Maclaren
oliver.maclaren@auckland.ac.nz

LECTURE 7: SYSTEMS BIOLOGY - A VERY BRIEF OVERVIEW

- Systems biology and understanding complex systems
- Overview of signal transduction, metabolism and gene regulation
- Different formalisms and notation for building up models
- Real modelling example: signal transduction and cardiac hypertrophy
- Models and data: parameter estimation issues

MODULE OVERVIEW

Reaction kinetics and systems biology (*Oliver Maclaren*)
[12 lectures/4 tutorials/2 labs]

1. Basic principles: modelling with reaction kinetics [6 lectures]
Physical principles: conservation, directional and constitutive. Reaction modelling. Mass action. Enzyme kinetics. Enzyme regulation. Mathematical/graphical tools for analysis and fitting.

2. Systems biology I: overview, signalling and metabolic systems
[3 lectures]

Overview of systems biology. Modelling signalling systems using reaction kinetics. Introduction to parameter estimation. Modelling metabolic systems using reaction kinetics. Flux balance analysis and constraint-based methods.

3. Systems biology II: genetic systems [3 lectures]
Modelling genes and gene regulation using reaction kinetics. Gene regulatory networks, transcriptomics and analysis of microarray data.

SYSTEMS BIOLOGY?

Short version:

Traditional biology: Breakdown into pieces

Systems biology:

Put it all back together!
Usually, using *mathematics/computation*

Complementary approaches.

METABOLISM

Short definition:

The consumption and production of chemical substances and energy to sustain life

- Catabolism: breakdown
- Anabolism: build up

Food —→ Life

GENE REGULATION

Short definition:

Control of the levels of enzymes and other proteins (etc) via the regulation and control of transcription and translation of the genetic code

Cells control their ‘production’ of cell ‘machinery’ in response to the environment and other needs.

SIGNAL TRANSDUCTION

Short definition:

How cells sense, translate and respond to external stimuli

i.e. chemical or physical ‘signals’.

E.g.

- Ligands (signal molecules)
- Mechanical forces
- Concentration gradients

FORMALISM? GENERALISED REACTIONS

We have been modelling simple processes in terms of *reactions*

We can use reaction modelling as a *general modelling formalism* for many types of processes and many different levels

- Signal transduction
- Metabolic processes
- Gene regulation
- Etc

ROLE OF FORMALISMS

- Consistent formalisms make it easier to *bridge/connect various scales and process types*
- There are also *many other formalisms*; sometimes you want to combine many different model types for different processes

Here we continue using ‘reactions’ and ODEs as our basic modelling language

BRIEF SIGNAL TRANSDUCTION EXAMPLE

Goal:

- *Exposure* to how more complex models are built up from the ideas introduced so far
- A few *extra details* on signal/membrane modelling
- Some *minor complications*:
 - Area vs volume concentrations and units
 - Varying enzyme levels and MM models

HELPFUL REVIEW: BRIDGING THE LAYERS

Goncalves et al. (2013) ‘Bridging the layers: towards integration of signal transduction, regulation and metabolism into mathematical models’. In: Mol. BioSyst.

On Canvas:

- You should read!
- You won’t necessarily understand it all though!
- I won’t assess you on it!

NOTATION: NET FLUXES AND INDIVIDUAL FLUXES

Often useful to consider the *net flux for a given reaction = difference between forward and backward fluxes*. E.g.

$$J_1^{\text{Net}} = J_1 - J_{-1}$$

Another common notation is

$$J_1 = J_1^+ - J_1^-$$

Here J_1 corresponds to J_1^{Net} . *Careful!*

CARDIAC HYPERTROPHY

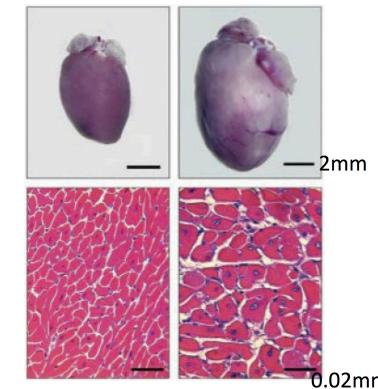
SIGNAL TRANSDUCTION EXAMPLE

Example:

Cooling, Hunter and Crampin (2007) ‘Modelling Hypertrophic IP₃ Transients in the Cardiac Myocyte’. In: Biophys. J.

On Canvas:

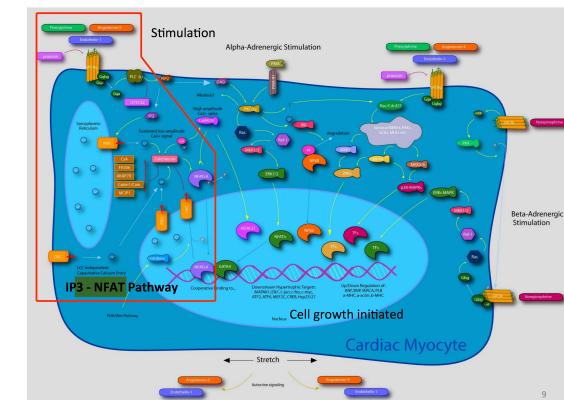
- You should read!
- You won’t necessarily understand it all though!
- I won’t assess you on it!



CARDIAC HYPERTROPHY

CARDIAC HYPERTROPHY

- The heart’s cells *adapt in response to the signalling of the rest of the body*.
- An example is the increase in cell volume: *cardiac hypertrophy*
 - Sometimes *non-pathological* adaptation (athletes, pregnancy)
 - Sometimes *pathological*, maladaptive and can lead to heart disease/failure



Goal: convert to mathematics! See Cooling et al.

EXTRA ELEMENT: SIGNALLING

Simple model (e.g. for IP₃ intracellular signalling molecule)

$$\frac{d[IP_3]}{dt} = J_{\text{prod}} - J_{\text{deg}}$$

where

$$J_{\text{prod}} = a$$

$$J_{\text{deg}} = k_{\text{deg}}[IP_3]$$

Here the production rate is treated as *controlled parameter*. More complex model: see paper.

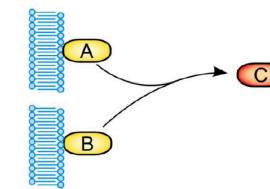
REMAINING PIECES

Use tools from *previous lectures!* E.g.

$$\begin{aligned} J_{11} &= k_{f,11} \times P_g \times Ca - k_{r,11} \times P_{cg} \\ J_{12} &= k_{f,12} \times P_{cg} \\ J_{13} &= k_{f,13} \times P_g \\ J_{14} &= \frac{k_{f,14} \times P_c \times PIP_2}{\left(\frac{k_{m,14}}{C_{pc}} + PIP_2 \right)} \\ J_{15} &= \frac{k_{f,15} \times P_{cg} \times PIP_2}{\left(\frac{k_{m,15}}{C_{pc}} + PIP_2 \right)} \\ \frac{dP}{dt} &= J_{13} - (J_9 + J_8) \\ \frac{dP_g}{dt} &= J_9 - (J_{11} + J_{13}) \\ \frac{dP_c}{dt} &= J_8 + J_{12} - J_{10} \\ \frac{dP_{cg}}{dt} &= J_{10} + J_{11} - J_{12} \\ J_{16} &= k_{f,16} \times IP_3 \\ \frac{dIP_3}{dt} &= (J_{14} + J_{15}) \times C_{pc} - J_{16} \end{aligned}$$

Can you spot the various constitutive models we've seen?

TECHNICAL ISSUE I: MEMBRANE VS CYTOSOLIC CONCENTRATIONS



Remember: *amount is conserved, not concentrations*

- Need conversion factor between ‘per area’ and ‘per volume’ concentrations
 - See handout

TECHNICAL ISSUE II: VARYING ENZYME LEVELS

Here our *enzyme levels are varying* (signalling affects transcription).

Typically *still use MM model but with current E level*:

$$v([S], [E]) = \frac{k[E][S]}{K_M + [S]}$$

instead of

$$v([S]) = \frac{kE_0[S]}{K_M + [S]}$$

(now also need model for [E] variations)

HOW TO DEAL WITH PARAMETERS FOR WHOLE-CELL MODELLING?

Babtie and Stumpf (2017) ‘How to deal with parameters for whole-cell modelling’. In: J. R. Soc. Interface.

On Canvas:

- You should read!
- You won’t necessarily understand it all though!
- I won’t assess you on it!

RELATING MODELS TO DATA: FORWARD AND INVERSE PROBLEMS

Typical modelling deals with (so-called) *forward problems*:

*Given parameters and initial conditions,
predict data*

E.g. heart modelling example.

HOW TO DEAL WITH PARAMETERS FOR WHOLE-CELL MODELLING?

INTERFACE

rsif.royalsocietypublishing.org

Review



Cite this article: Babtie AC, Stumpf MPH.
2017 How to deal with parameters for
whole-cell modelling. *J. R. Soc. Interface*
14: 20170237.
<http://dx.doi.org/10.1098/rsif.2017.0237>

Received: 30 March 2017
Accepted: 22 June 2017

How to deal with parameters for whole-cell modelling

Ann C. Babtie and Michael P. H. Stumpf

Department of Life Sciences, Imperial College London, London, UK

MPHS, 0000-0002-3577-1222

Dynamical systems describing whole cells are on the verge of becoming a reality. But as models of reality, they are only useful if we have realistic parameters for the molecular reaction rates and cell physiological processes. There is currently no suitable framework to reliably estimate hundreds, let alone thousands, of reaction rate parameters. Here, we map out the relative weaknesses and promises of different approaches aimed at redressing this issue. While suitable procedures for estimation or inference of the whole (vast) set of parameters will, in all likelihood, remain elusive, some hope can be drawn from the fact that much of the cellular behaviour may be explained in terms of smaller sets of parameters. Identifying such parameter sets and assessing their behaviour is now becoming possible even for very large systems of equations, and we expect such methods to become central tools in the development and analysis of whole-cell models.

FORWARD AND INVERSE PROBLEMS

However, in science and engineering we are usually confronted with *inverse problems*:

Given data, estimate parameters and/or initial conditions and predict future data

ILL-POSED PROBLEMS

In contrast to forward problems, inverse problems can have

- No solution
- Many solutions
- Unique but unstable solutions

These are called *ill-posed* (c.f. ‘well-posed’) problems (Hadamard, 1902)

IDENTIFIABILITY

Essentially: can we uniquely determine the parameters, e.g. $\theta = (k_1, k_2, \dots)$, given perfect data (model outputs) $f(\theta)$?

Formal requirement, *full identifiability*:

$$\theta_1 \neq \theta_2 \implies f(\theta_1) \neq f(\theta_2)$$

Partial identifiability:

$$g(\theta_1) \neq g(\theta_2) \implies f(\theta_1) \neq f(\theta_2)$$

for a combination (‘group’) of parameters $g(\theta)$.

THINKING ABOUT ILL-POSED PROBLEMS

- *Identifiability and falsifiability* (what is possible to learn about how the world works? E.g. science and philosophy.)
- *Prediction* (can we do a task without necessarily needing to know how the world works? E.g. machine learning.)
- *Preferences* (how can we explicitly incorporate and optimise for additional preferences like ‘simplicity’? E.g. consulting.)

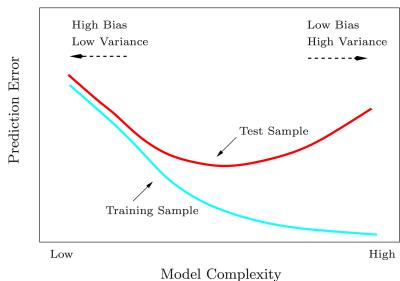
TRADE-OFFS: PREDICTION AND PREFERENCES

One key to solving ill-posed problems is recognising the *trade-offs* involved

- Fit vs complexity
- Overfitting vs underfitting
- Training vs test
- Efficiency vs stability
- Bias vs variance
- Etc

These trade-offs are closely related and *recur throughout science, statistics and engineering*

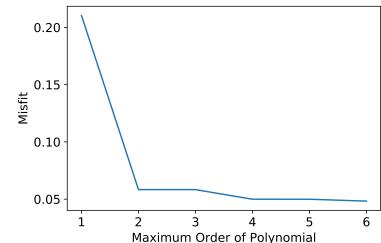
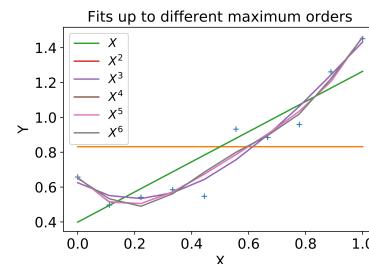
TRADE-OFFS



From Hastie et. al 'Elements of Statistical Learning: Data Mining, Inference and Prediction'.

available at: <http://web.stanford.edu/~hastie/ElemStatLearn/>. See also: 'An Introduction to Statistical Learning' (simplified version of above), available at: <http://www-bcf.usc.edu/~gareth/ISL/>

EXAMPLE: SIMPLE POLYNOMIAL FITTING



EXAMPLE: SIMPLE POLYNOMIAL FITTING

Suppose we have enzyme data and we *want to fit a curve to the double-reciprocal plot*

- Usually use linear (first order) regression (relates to MM)

What if we tried to fit higher order polynomials?

Biomeng 261 Lecture 7

- Overview of 'systems biology'
 - ↳ general ideas
 - ↳ Language & notation
 - ↳ Examples & challenges

Goals we want to keep

- 'building up' (& down!) our models to capture biological complexity
- At the same time we want to understand our models & use them to gain insight, not just numbers

→ Trade-offs!

→ Open problems!

'Systems' Biology ?

Traditional biology:

- break down into pieces
- 'Reductionist'

Systems biology:

- build back up into whole
- 'emergentist'
- Lots of 'interactions' & 'networks' etc.

'Engineering': best practice for

building things up - modules, components, control, hierarchies etc.

These approaches complement each other

→ But engineers typically want to understand, manipulate & build 'wholes' or 'systems'

also leads to synthetic biology
(design & build new biological systems)

Some key pieces of the puzzle

1. Metabolism
2. Signal transduction
3. Gene regulation

1. Metabolism: consumption & production of chemical substances & energy to sustain life

- L Catabolism: Breakdown substances for energy & 'raw materials'
- L Anabolism: Build up components of cells, e.g. proteins/enzymes etc

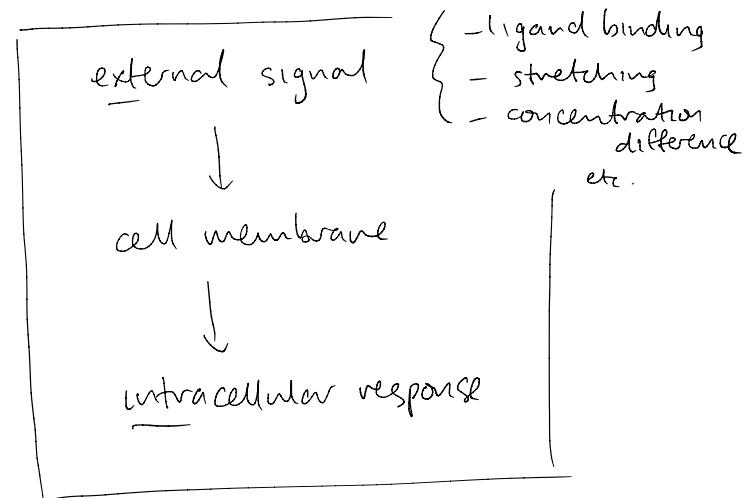
Short version: Food \rightarrow Life

Metabolic pathway

\rightarrow series of steps in some metabolic reaction

2. Signal Transduction

- L How cells sense, respond etc to external stimuli, i.e. chemical or physical 'signals'



Signal transduction pathways & networks

- series of steps in a particular signal transduction process
- involves interaction of series of proteins etc
- often think of in terms of 'modules' & 'components' making up larger 'circuits' & control systems

e.g. a 'switch' module

3. Genetic Regulation

- ↳ control of the levels of enzymes & other proteins via the regulation & control of
 - transcription ($DNA \rightarrow RNA$)
 - translation ($mRNA \rightarrow Proteins$)

- Similarly, can think of in terms of 'circuits' & 'control systems' with various repeating components

Systems biology again:

Metabolism, signal transduction & genetic regulation are all interconnected

Goal: understand (& build) complex biological systems

→ still open problem really!

(some progress....)

→ need languages & formalisms!

Formalisms & Notations - what language?

We have been representing simple processes as 'reactions'

↳ we can use more generally as a modelling language for many types of physical processes & at multiple scales

- ↳ metabolism
- ↳ signal transduction
- ↳ genetic regulation
- etc.

Languages

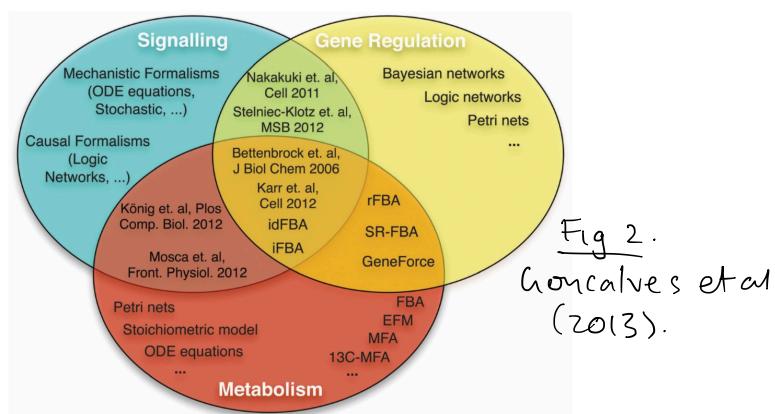
- We will model many things in reaction language that aren't strictly 'chemical reactions' in the usual sense
- think 'networks of processes'
- examples include cell signalling, metabolism & genetic regulation

Other languages exist!

See Gonçalves et al (2013) on Canvas

Bridging the layers: towards integration of signal transduction, regulation and metabolism into mathematical models

Emanuel Gonçalves,^a Joachim Bucher,^b Anke Ryll,^c Jens Niklas,^b Klaus Mauch,^b Steffen Klamt,^c Miguel Rocha^d and Julio Saez-Rodriguez^{*a}



Example : signal transduction

- use basic L1-L6 tools
- add simple stimulus → signal
- a couple of technical details.

Goals :- exposure to general ideas of building models

- a couple of extra details on signal/membrane modelling

Cardiac hypertrophy models

↳ Cooling, Hunter & Crampin (2007)

↳ see Canvas

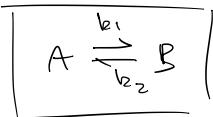
↳ cellML model

↳ IP₃-Calcineurin pathway
stimulates NFAT → binds DNA cooperatively.

[Note on notation :]

[Net fluxes & individual fluxes

Consider



→ we have been using J_1 & J_{-1}
for forward & backward
fluxes, respectively.

As we start to build up
larger systems of reactions we
often 'lump' together into

[net flux] :

$$\boxed{J_1^{\text{Net}} = J_1 - J_{-1}}$$

↑ ↑ ↑
net forward backward

confusingly, another common notation

is

$$\boxed{J_1 = J_1^+ - J_1^-}$$

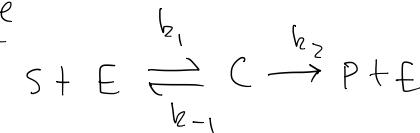
↑ ↑ ↑
net forward backward

[which?]

→ Just be careful!

→ context should make it clear

Example



Notation $\frac{d[S]}{dt} = -J_1^{\text{Net}} = -(J_1 - J_{-1})$
version 1.

—

$$\frac{d[E]}{dt} = -J_1^{\text{Net}} + J_2^{\text{Net}} = -(J_1 - J_{-1}) + J_2$$

etc

$$\& J_1^{\text{Net}} = k_1[S][E] - k_{-1}[C] \text{ etc.}$$

Notation

version 2. $\frac{d[S]}{dt} = -J_1 = -(J_1^+ - J_1^-)$

etc.
↳ $J_1 = k_1[S][E] - k_{-1}[C]$

→ if in doubt, be explicit!

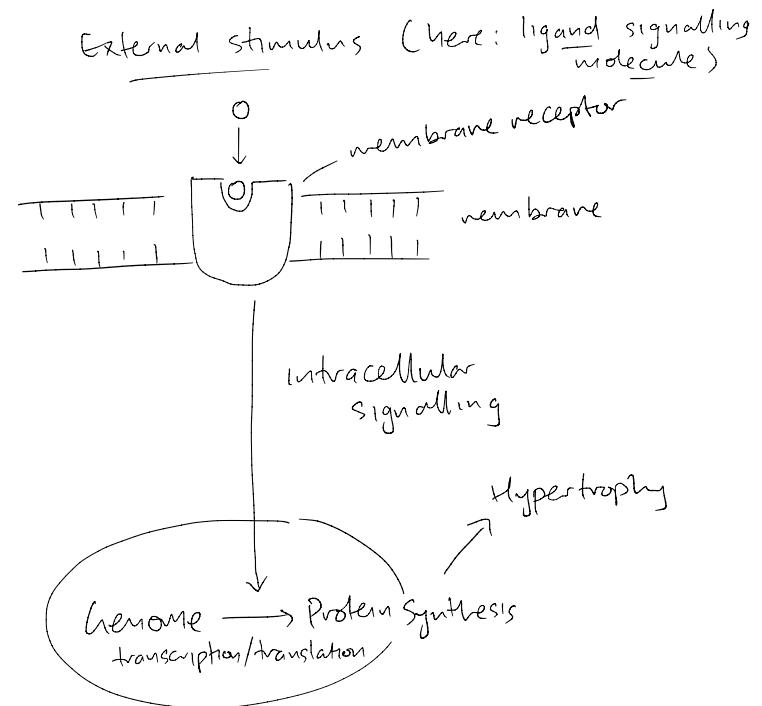
Cardiac Hypertrophy

- cellular hypertrophy: type of adaptation to external signals / load
- cells increase in volume
- risk factor for heart disease/failure
- at cellular level, involves complex interaction of signal transduction pathways

Goal: translate 'cartoon' to a mathematical model



Cartoon of signal transduction pathway



Hypertrophy: excess

trophy: nourishment

Ligand: signalling molecule, binds to receptor

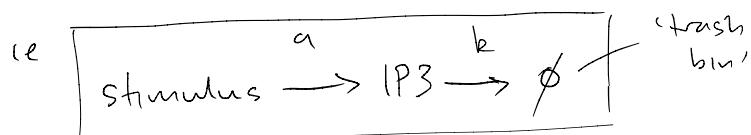
- agonist: ligand that stimulates / activates
- antagonist: ligand that blocks action of an agonist

Simple model of signal activation/decay

$$\text{production of intracellular signalling molecule} = f \left(\begin{array}{l} \text{extracellular} \\ \text{signalling} \\ \text{molecule level} \end{array} \right)$$

+

$$\text{degradation of intracellular signalling molecule} = \text{simple decay}$$



gives

$$\frac{d[IP3]}{dt} = J_{\text{prod}} - J_{\text{deg}}$$

$J_{\text{prod}} = a$ = external/control parameter

$$J_{\text{deg}} = k[IP3]$$

- here we can vary 'a' & see response

- can also model in more detail

↳ see article

Combine signalling/stimulation model

with { Michaelis-Menten
Hill (cooperative)
Mass action

→ get

signal transduction model!

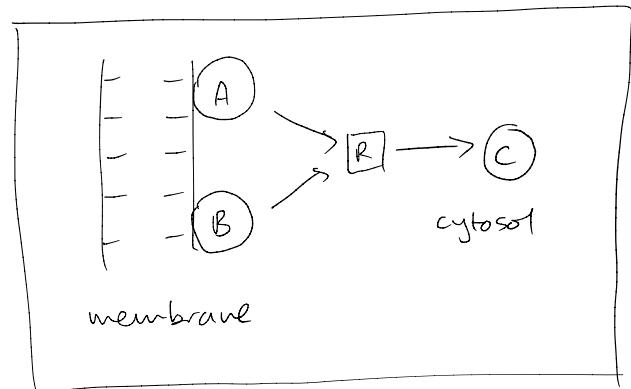
see Cooling, Hunter & Crampin (2007)

"Modelling hypertrophic IP3 transients in the cardiac myocyte"

on Canvas.

Notes : Volume vs Area concentrations
& membranes

some species/proteins are membrane-bound & some cytosolic



→ need to account for when
converting from amounts to
concentrations

$$\frac{d[\text{Area} \times B]}{dt} = -\tilde{J}_1^{\text{net}} [=] \frac{\text{amount}}{\text{time}}$$

$$\frac{d[\text{Volume} \times C]}{dt} = +\tilde{J}_1^{\text{net}} [=] \frac{\text{amount}}{\text{time}}$$



notes --

define $\gamma = \frac{\text{area}}{\text{vol}}$ } conversion factor

recall: $J_1^{\text{net}} = \frac{\tilde{J}_1^{\text{net}}}{\text{vol}} \Rightarrow \tilde{J}_1^{\text{net}} = \text{Vol. } J_1^{\text{net}}$

For now, assume vol constant
(not in general)

$$\frac{d[B]}{dt} = -\frac{1}{\text{area}} \times \tilde{J}_1^{\text{net}} = \frac{\text{Vol. }}{\text{area}} J_1^{\text{net}}$$

$$\frac{d[C]}{dt} = \frac{1}{\text{vol}} \tilde{J}_1^{\text{net}} = J_1^{\text{net}}$$

so

$$\frac{d[B]}{dt} = -\frac{1}{\gamma} J_1^{\text{net}}$$

$$\frac{d[C]}{dt} = J_1^{\text{net}}$$

note conversion factor.

Notes -- Michaelis-Menten

- The levels of enzymes are varying!

- Can't just assume
 $[E] = E_0 - [C]$ etc

- simple approach: replace

$$J = \frac{k E_0 [S]}{K_m + [S]}$$

with

$J = \frac{k[E][S]}{K_m + [S]}$	variable
---------------------------------	----------

+

$\frac{d[E]}{dt}$ model	<i>depends on transcription etc.</i>
-------------------------	--------------------------------------

treat as empirical constitutive equation (fit k , K_m empirically)

Other issues

→ we often end up building models that are as hard to understand as the original system!

Eg Babtie & Stumpf (2017) :

→ Again: many open challenges!

INTERFACE

rsif.royalsocietypublishing.org

Review



Cite this article: Babtie AC, Stumpf MPH. 2017 How to deal with parameters for whole-cell modelling. *J. R. Soc. Interface* 14: 20170237. <http://dx.doi.org/10.1098/rsif.2017.0237>

Received: 30 March 2017
 Accepted: 22 June 2017

How to deal with parameters for whole-cell modelling

Ann C. Babtie and Michael P. H. Stumpf

Department of Life Sciences, Imperial College London, London, UK

(MPHS, 0000-0002-3577-1222)

Dynamical systems describing whole cells are on the verge of becoming a reality. But as models of reality, they are only useful if we have realistic parameters for the molecular reaction rates and cell physiological processes. There is currently no suitable framework to reliably estimate hundreds, let alone thousands, of reaction rate parameters. Here, we map out the relative weaknesses and promises of different approaches aimed at addressing this issue. While suitable procedures for estimation or inference of the whole (vast) set of parameters will, in all likelihood, remain elusive, some hope can be drawn from the fact that much of the cellular behaviour may be explained in terms of smaller sets of parameters. Identifying such parameter sets and assessing their behaviour is now becoming possible even for very large systems of equations, and we expect such methods to become central tools in the development and analysis of whole-cell models.

L See canvas

State of the art:

techniques, and instead start to require computer simulations to explore their behaviour.

1.1. From simple to complex models

Some models aim to capture the essential hallmarks of life—such as metabolism, nutrient uptake, gene expression regulation and replication—but in a simplified representation that does not aim to replicate the true complexity of a whole organism [11–16]. These *coarse-grained* models have shown great promise and allow us to integrate molecular, cellular and population level/scale processes into a coherent—and analytically tractable—modelling framework. While real cells will be much more complicated, these simple model systems have successfully provided insight into fundamental cell physiology, e.g. processes affecting microbial growth rates [12,13,15,16].

Increasingly, there is interest in generating more realistic and *complicated* models that, rather than aiming to provide abstract representations of key features, incorporate extensive details of known components and interactions (or reactions) present in a system. In cell biology, for example, there are now numerous attempts at modelling aspects of metabolism, gene regulation and signalling at cellular level [17–24]. Perhaps the *best established* *first metabolic model*, where a powerful set of tools, based around *flux balance analysis* (FBA) [25], allows us to explore metabolic phenotypes *in silico* at a genomic level for an increasing range of organisms (and some individual cell types) [24,26,27]. However, such models are stoichiometric and thus give us information about biochemical reaction schemes and fluxes, but not details about the system dynamics.

L8-L9

Advances in both high-throughput experimentation and computational power have opened up the possibility of creating and analysing more complex dynamic models of biological systems, including many which represent processes occurring at different scales [28,29]. Numerous models now face the challenge of being *large* (in terms of numbers of species and parameters represented), multi-scale and/or *hybrid* in nature (incorporating multiple different mathematical representations) [23,28–30]. The most ambitious models to date—the WCMs—aim to provide faithful *in silico* representations of real biological cells, including all major cellular processes and components, and are both very large scale and hybrid (figure 1) [31–33].

There are several potential uses for such WCMs:

- (1) To gain mechanistic insights, by serving as an *in silico* ‘blueprint’ through which we study the behaviour of real cells.
- (2) As a rational screening and predictive tool, to explore *in silico* what might be hard or impossible to study *in vivo*.
- (3) To drive new biological discoveries, by showing where we lack sufficient understanding, and identifying promising future directions to pursue experimentally.
- (4) To study *emergent* phenomena which are only apparent when we consider a system as a whole.
- (5) To integrate heterogeneous datasets and amalgamate our current knowledge into a single modelling framework.
- (6) Perhaps eventually to study, via virtual competitions between different cell architectures, evolutionary dynamics in unprecedented detail (but at enormous, currently crippling, computational cost).
- (7) In the meantime, as the community strives to develop viable WCMs, the technological, computational and

Here, we focus on the inference and statistical modelling challenges inherent to developing WCMs (and other complex models), in terms of model construction, parameter estimation, uncertainty and sensitivity analyses, and model validation and refinement. Some of these are generic modelling challenges—but worth reiterating—while others are specific to large-scale, multi-scale and hybrid models.

State of art cont'd.

At present, none of the statistical inference methods outlined above are applicable at the scale of WCMs. However, smaller subsystems, such as individual pathways, regulatory motifs, receptor complexes or systems comprising small sets of metabolic reactions and the associated regulatory processes can be effectively parametrized using such methods [72]. For such systems, we can often estimate parameters, including uncertainty, and we are frequently able to assess parameter sensitivity (typically measured as the change in some model output, e.g. predicted protein abundance, in response to varying a single parameter). In some cases, experimental measurements of species concentrations may allow us to effectively decompose our models into smaller modules for efficient parameter estimation [73]. Bayesian inference methods in particular are limited in terms of scale and are generally only feasible for models with up to tens to hundreds of species and parameters [74,75]. Some optimization approaches are much more scalable though, with recent advances allowing parametrization of ODE models comprising hundreds to thousands of species and parameters [65,76,77]. As always, however, the chance of being trapped in local optima is high for such large-dimensional problems.

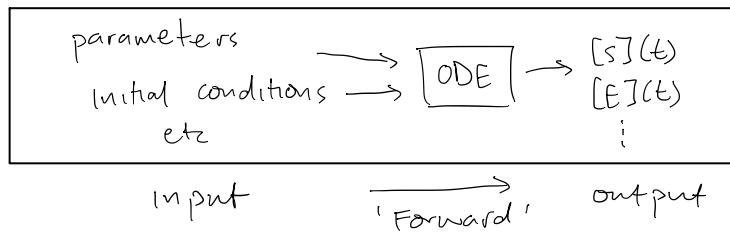
A combination of both inference and experimental estimation will probably be needed to parametrize complex biological models. It is currently impractical to use inference techniques within the context of a full WCM, unless considering very small pre-defined subsets of the parameters and, even then, the computational costs are enormous [67]. We can, however, make use of scalable inference techniques [78] to help us parametrize the component submodels, using experimental information where available as prior knowledge for the inference procedures. This will allow us to avoid some of the potential pitfalls outlined above of experimental estimates, and generate parameter estimates that take into account—to the best of our ability—the influences of cellular and system context, and make use of the most appropriate *in vivo* datasets. Crucially, rigorous statistical inference also enables us to explore the relationships between model parameters and start to understand and quantify the uncertainties inherent to any mathematical model.



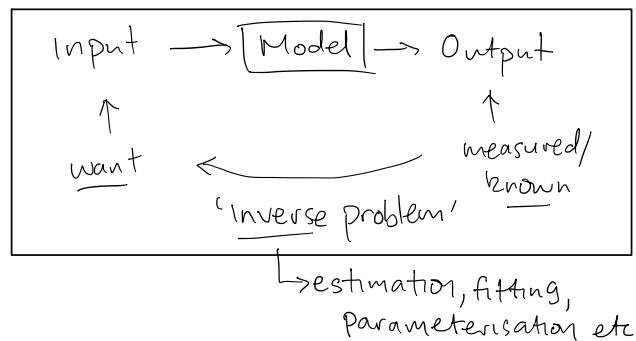
→ brief overview of
some issues --

Parameter estimation : Forward vs Inverse problems.

Usual modelling procedure:

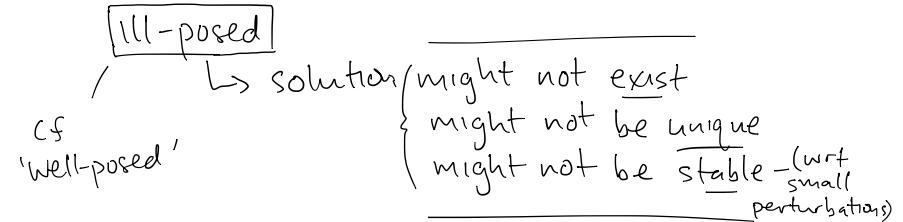


Problem: in reality we have data like (noisy) measurements of S, E concentrations (outputs) while we don't know inputs eg k_1, k_2 etc, ie



What's the problem?

→ in contrast to 'forward' problems, Inverse problems are usually:



Simple example

$$y = f(x) \text{ eg } y = x^2 \quad \left. \begin{array}{l} \text{unique output} \\ \text{for each input} \end{array} \right\}$$

- Observe output $y=4$
- What was input x ?

→ no unique solution!

Solution set: $\{-2, 2\}$

Importantly: parameter fitting is typically ill-posed

Points of view

1. Identifiability / Falsifiability } scientific (phil. oriented)
2. Prediction : training & test (etc) } task oriented
(machine learning)
3. Preferences : simple is 'better'? } multicriteria optimisation / balance

1. Identifiability (model)

unique inputs \xrightarrow{f} unique output
(parameters)
 $\theta = (k_1, k_{-1}, k_2)$
then can invert': f^{-1}

* Key Page *
1/2

$$\boxed{\theta_1 \neq \theta_2 \Rightarrow f(\theta_1) \neq f(\theta_2)}$$

\Rightarrow | "(parameters) are identifiable" | (from data / outputs)

- Often only partial identifiability:

- o groups / combinations of parameters, $g(\theta)$
give different answers for diff. values,
even if diff θ values don't

$$\boxed{g(\theta_1) \neq g(\theta_2) \Rightarrow f(\theta_1) \neq f(\theta_2)}$$

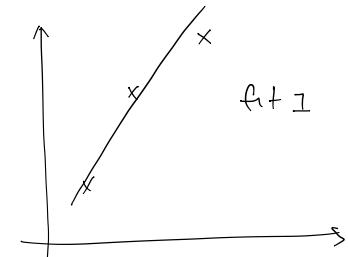
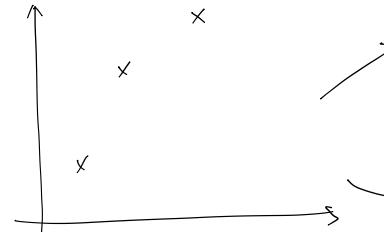
e.g. $K_m = \frac{k_1 + k_2}{k_1}$ in $v = \frac{V_{max} [S]}{K_m + [S]}$

diff. $K_m \Rightarrow$ diff v ,
but multiple k_{-1}
give same K_m & v .

2 & 3 : Prediction & Preferences

Illustration : curve fitting

given data:



Which is better?

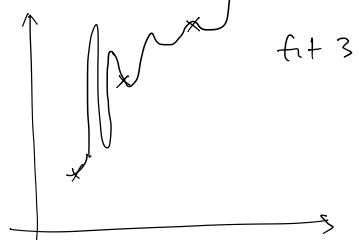
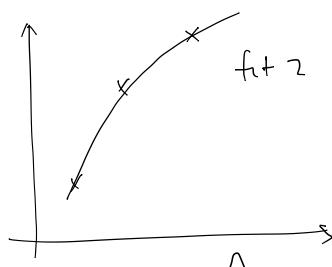
Finite amount of data

\rightarrow many compatible models

\rightarrow more complex 'fits'
given better

\hookrightarrow but are often unstable
& / or fit future data worse

\rightarrow simpler \approx more understandable?



Trade-offs :

- (A) - make more complex
 (B) - make simpler

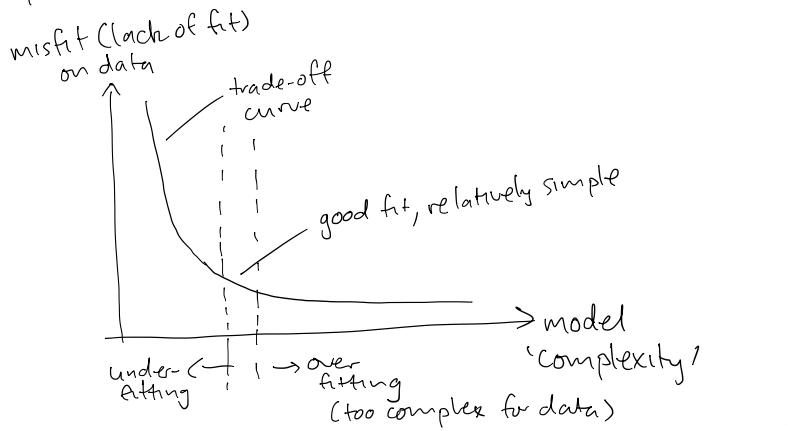
* Key Page *
2/2

- (A) - Fit to given ('training') data
 (B) - Stability / fit to future ('test') data

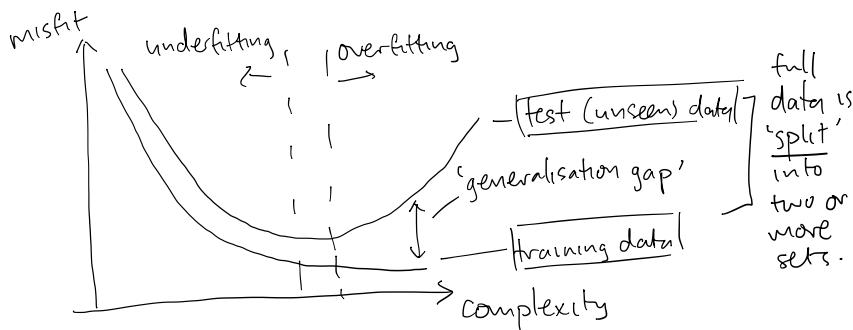
↳ or 'unseen'

'predictive'
or
'empirical'
view

Simple vs complex:

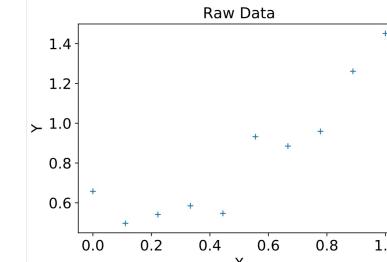


Training/Test ('Predictive' or 'machine learning' view)



Examples

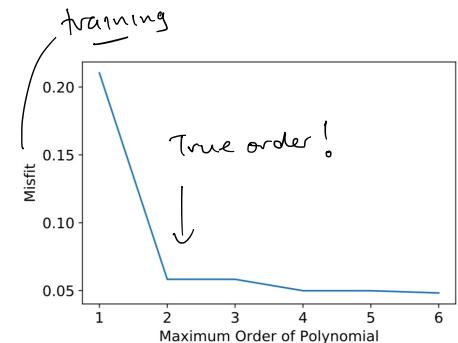
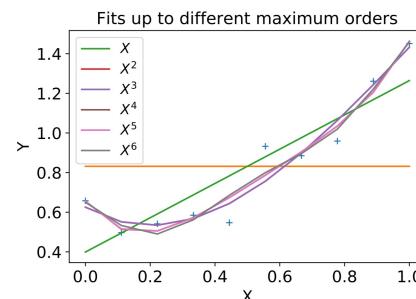
→ Polynomial :



True model

$$Y = X^2 + 0.5 + \text{noise}$$

Fits & trade-offs:

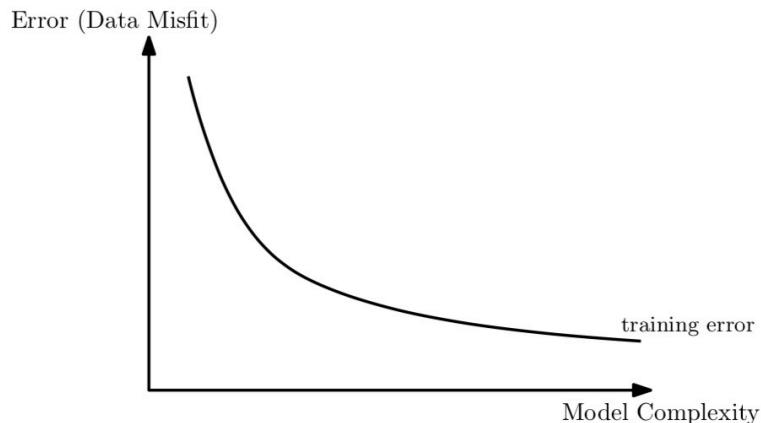


ODEs ? Some basic ideas

- future courses
- research projects

Example test question:

- c) Given the following figure illustrating the training error vs. model complexity for a family of models fitted to a given dataset, add
- A curve illustrating what you would expect the *test error* to look like as a function of model complexity
 - A vertical line indicating what would be a 'good' choice of model complexity to use, given your answer to i.



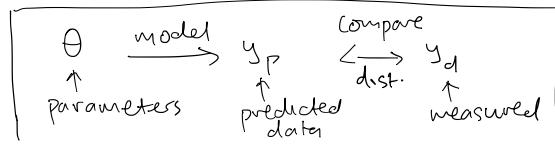
Note:

Measuring data fit/misfit? (eg non-negative, symmetric etc.)

recall:
lab

- Need a norm, distance, metric, cost function etc relating data and model

- Evaluates model by its predicted data



e.g. $d(\underline{y}_p, \underline{y}_d)$ (\underline{y} vector e.g. $\underline{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$)
& $\underline{y}_p = M(\theta)$ leads to

$$\text{'misfit'}(\theta; \underline{y}_d) = d(\underline{y}_d, M(\theta)) = d(\underline{y}_d, \underline{y}_p(\theta))$$

how good are parameters θ at 'predicting' data \underline{y}_d

Typical 'distances' or 'cost' measures:

$$d(\underline{x}, \underline{y}) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \leftarrow \sqrt{\text{sum of squares}}$$

$$d(\underline{x}, \underline{y}) = \sum_{i=1}^n (x_i - y_i)^2 \leftarrow \text{sum of squared diff's.}$$

$$d(\underline{x}, \underline{y}) = \sum_{i=1}^n |x_i - y_i| \leftarrow \text{sum of absolute differences}$$

[not examinable]

Measuring 'complexity'? (Machine learning:
see 'VC dimension')
→ Harder! Ill-defined?
o Simple idea is to use a 'norm' (size)
of model/parameters or 'distance'
from a 'default' or 'null' model
→ clearer for linear models (less so for nonlinear)
Eg $\|\theta\|_d = d(\theta, \theta_0) = \sum_{i=1}^n \theta_i^2$
 $\underbrace{\quad}_{\text{'size' or 'complexity' of parameters}}$ $\underbrace{\quad}_{\text{'distance from' or 'cost relative to' zero (or other refs)}}$

Combining: optimal trade-offs

1. minimise complexity
OR subject to acceptable data fit
 2. minimise data misfit
OR subject to acceptable complexity
 3. minimise data misfit + model complexity
 cost cost
- } can show are equivalent
→ need to choose relative importance (tradeoff parameter) though.

All lead to:

