Systems Identification in Systems Biology

David Šafránek

Masaryk University
Czech Republic

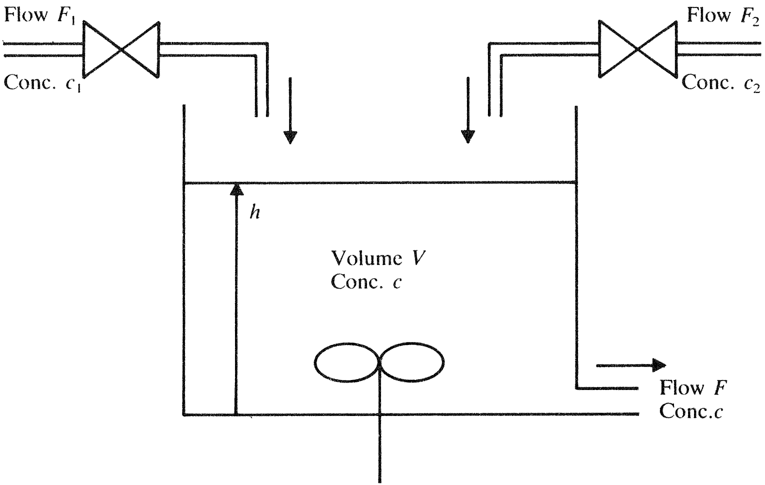
Outline

1. Introduction
2. The Approach: Parametric Identification
3. System Identifiability Problem
4. Overview of Approaches

Outline

1. Introduction
2. The Approach: Parametric Identification
3. System Identifiability Problem
4. Overview of Approaches

A Stirred Tank

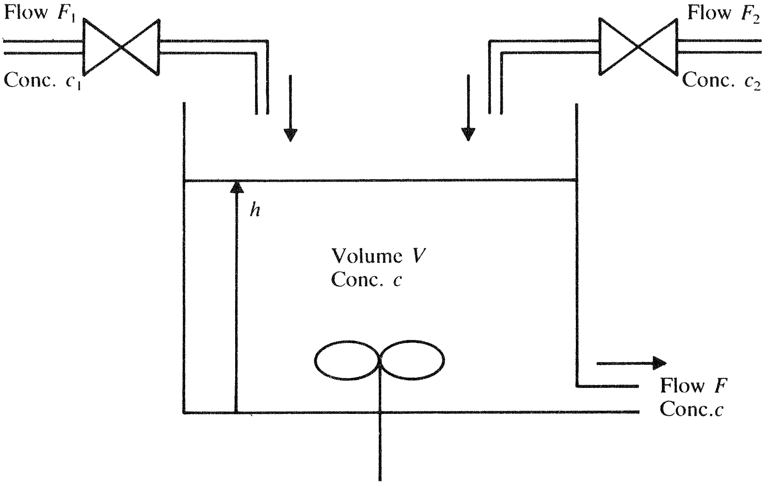


F_1, F_2 ... input flows (controllable)

c_1, c_2 ... input concentrations (uncontrollable)

F, c ... outputs of the system, can be observed (measured)

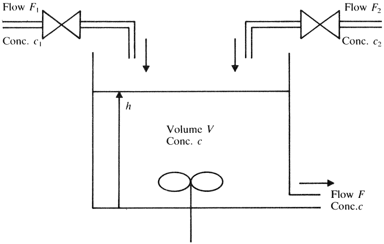
A Stirred Tank



F_1, F_2 ... input flows (controllable)

c_1, c_2 ... input concentrations (uncontrollable)

GOAL: Keep the outputs F, c constant.

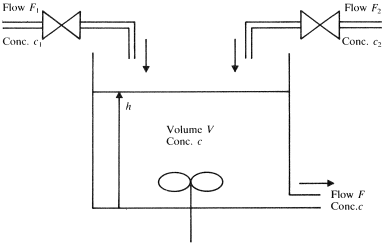


dynamics of the volume: $\frac{dV}{dt} = F_1 + F_2 - F$

$F = a\sqrt{2gh}$ (Torricelli's law of fluid dynamics)
where a ... effective area of the flow, $g \sim 10m/sec^2$

$V = Ah$ where A is tank area (independent on h)

A Stirred Tank – Mathematical Model

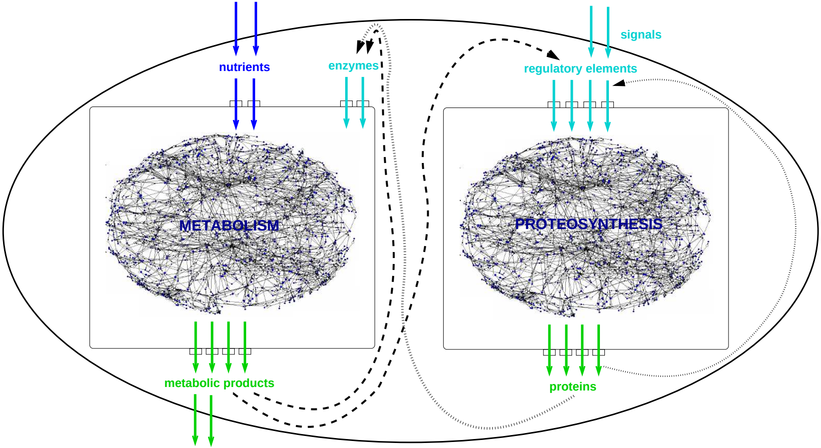


dynamics of the volume: $\frac{dV}{dt} = F_1 + F_2 - F$

$F = a\sqrt{2gh}$ (Torricelli's law of fluid dynamics)
where a ... effective area of the flow, $g \sim 10m/sec^2$

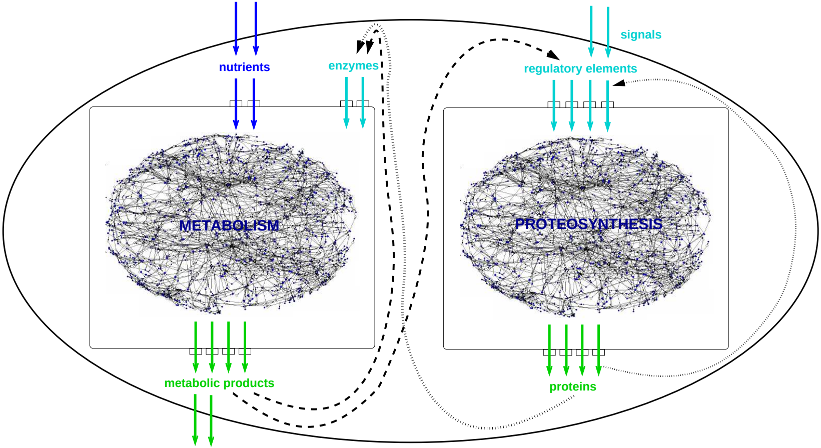
$V = Ah$ where A is tank area (independent on h)

problems: *a* is difficult to obtain, does this form of the Torricelli's law really apply for the real case?


nutrients
enzymes
signals
regulatory elements
METABOLISM
PROTEOSYNTHESIS
metabolic products
proteins

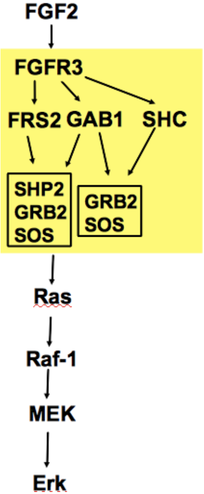
Dynamical Systems in Biology

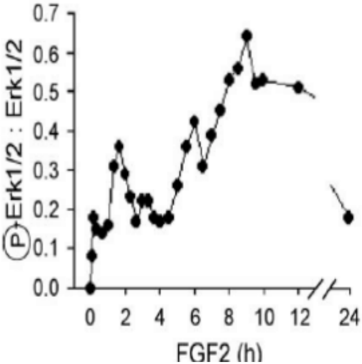
Processes Driving the Living Cell



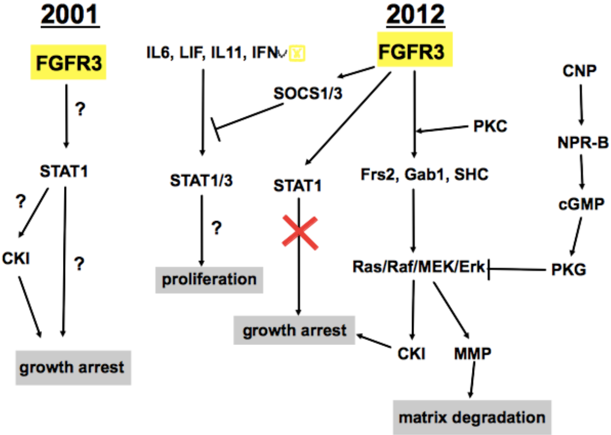
questions: how to control cyanobacteria to gain max ethanol
how to control *E. coli* to gain insuline, ...

Dynamical Systems in Biology
Signalling Pathways





joint work with P. Krejčí, Masaryk University Brno/Medical Genetics Institute, Cedars-Sinai Medical Center, L.A.

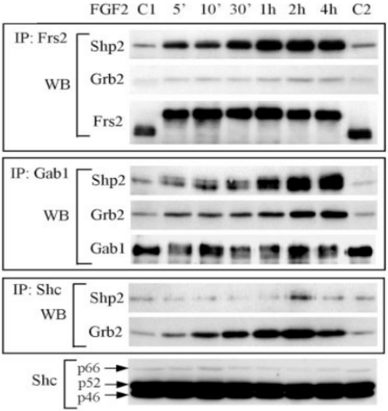


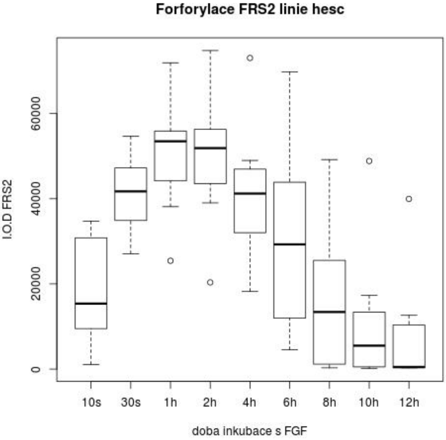
What is the right topology?

Wet-lab Measurements

Western blots/Northern blots

- western blots
- measurements of protein binding (presence of certain proteins)




Forforylace FRS2 linie hesc

Wet-lab Measurements

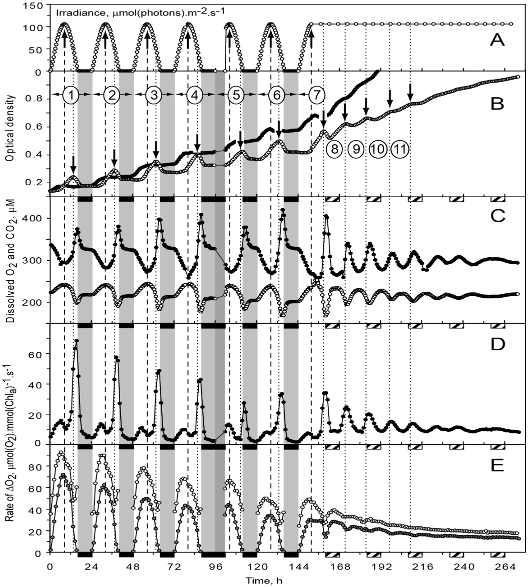
Photobioreactor Data

Optical density as a proxy of chlorophyll content and cell count

Concentration of dissolved O2 and CO2 influenced by photosynthetic activity

Rate of respiration as an indicator of metabolic changes

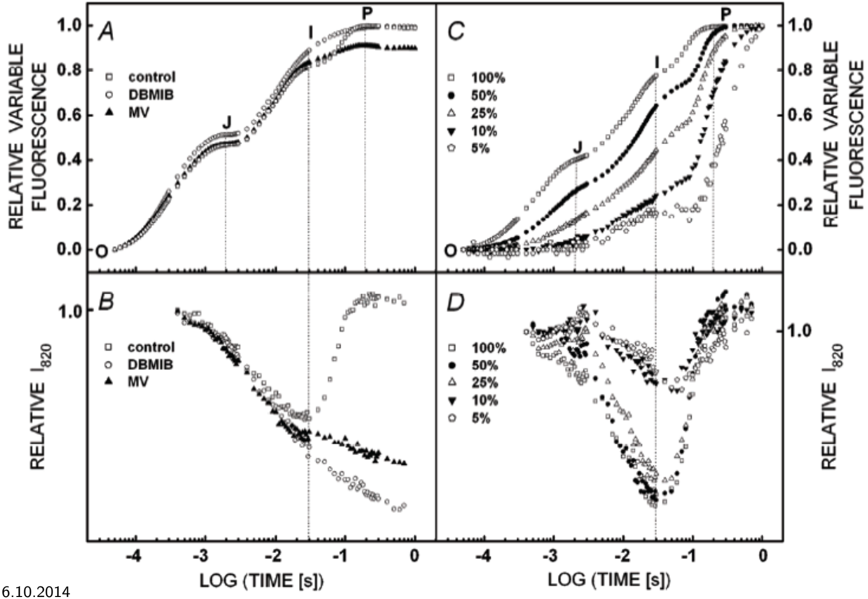
Rate of oxygen evolution



Červený, J., Nedbal, L. (2009) Metabolic rhythms of the cyanobacterium *Cyanothece* sp. ATCC 51142 correlate with modeled dynamics of circadian clock. *J. Biol. Rhythms* 24, 295-303.

Wet-lab Measurements

Fluorometer Data



Outline

1. Introduction

2. The Approach: Parametric Identification

3. System Identifiability Problem

4. Overview of Approaches

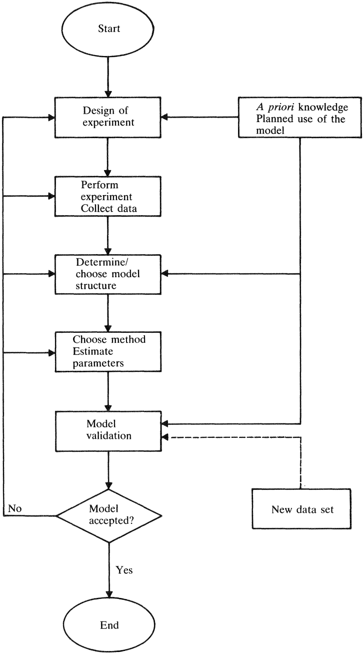
The Approach: System Identification

- INPUT: controlled perturbance of input stimuli
- OUTPUT: measurements of observed variables
- GOAL: find a system that reliably maps INPUT to OUTPUT

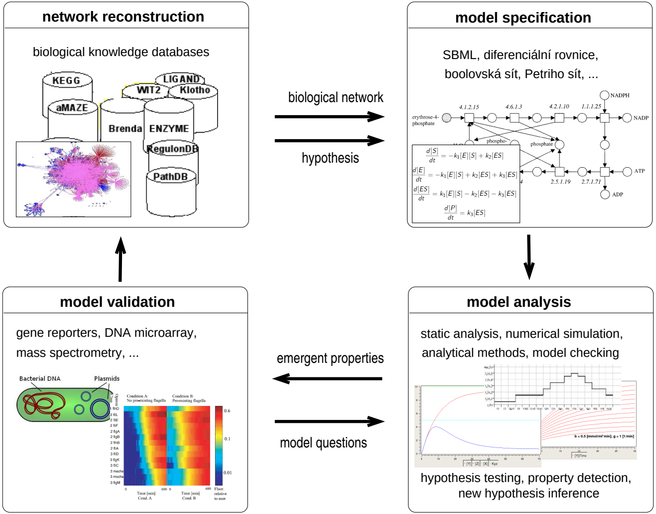
The Approach: System Identification

- INPUT: controlled perturbance of input stimuli
  - typically interesting patterns exploring most of (expected) systems response
  - pulses, oscillations, ...
- OUTPUT: measurements of observed variables
  - time-series or steady state data
  - not all variables might be observed
  - measurements might be very imprecise ⇒ **noisy data**
- GOAL: find a system that reliably maps INPUT to OUTPUT
  - mapping might be non-linear
  - extrinsic noise on both input, output side
  - system might be affected by intrinsic noise (internal stochasticity)

System Identification Workflow



System Identification Workflow



network reconstruction

biological knowledge databases

model specification

SBML, diferenciální rovnice, boolovská síť, Petriho síť, ...

biological network

hypothesis

model validation

gene reporters, DNA microarray, mass spectrometry, ...

model analysis

static analysis, numerical simulation, analytical methods, model checking

emergent properties

model questions

hypothesis testing, property detection, new hypothesis inference

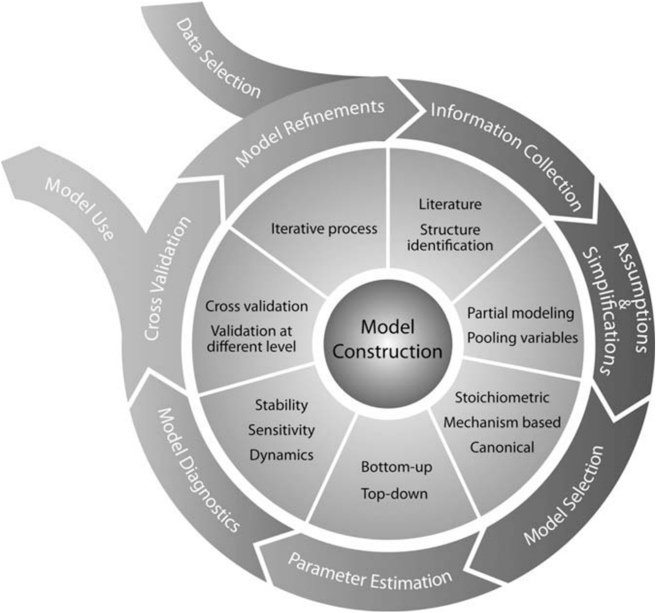
$$\frac{d[S]}{dt} = -k_1[E][S] + k_2[ES]$$

$$\frac{d[E]}{dt} = -k_1[E][S] + k_2[ES] + k_3[ES]$$

$$\frac{d[ES]}{dt} = k_1[E][S] - k_2[ES] - k_3[ES]$$

$$\frac{d[P]}{dt} = k_3[ES]$$

System Identification Workflow
Modelling in Systems Biology



System Identification Concepts

- **system** $\mathcal{S}$
  - mathematical description of the real-world **process**
  - can be an idealization
  - not necessarily required to be known
- **model structure** $\mathcal{M}$
  - non-parametric (table, mapping, frequency diagram, ...)
  - parametric (with a parameter vector θ) $\mathcal{M}(\theta)$
- **identification method** $\mathcal{I}$
  - depends on available data, kind of the process, ...
- **experimental condition** $\mathcal{E}$
  - concrete setting of identification experiment
  - selection and generation of input signals
  - prefiltering of data

Parametric Identification: Problem Statement

Definition

Parametric model $\mathcal{M}(\theta)$ describing n dynamically evolving *autonomous* variables is defined by a set of equations:

$$\begin{aligned}\dot{x}(t) &= f(x(t), u(t), p) \\ y(t) &= g(x(t), s) + \epsilon(t)\end{aligned}$$

where

- $x(t) \in \mathbb{R}^n$ for $t \geq 0$ is a vector of **internal model states**
- $u(t) \in \mathbb{R}^u$ for $t \geq 0$ is a vector of **input stimuli**
- $y(t) \in \mathbb{R}^m$ for $t \geq 0$ is a vector of **observables**
- $\epsilon(t)$ is a normally distributed measurement noise

If $m < n$ we speak about *partially observable* models.
Parameter θ is defined as a vector $\langle p, x(0), s \rangle$.

Parametric Identification: Problem Statement

$$\chi^2(\theta) = \sum_{k=1}^{m} \sum_{l=1}^{d} \left(y_{kl}^D - y_k(\theta, t_l) \right)^2$$

- y_{kl}^D is lth measurement point of the observable y_k taken at time t_l
- $y_k(\theta, t_l)$ is model-predicted y_k at time t_l by employing parameter estimate θ
- parameter estimate $\hat{\theta}$ is obtained as a value that minimizes $\chi^2(\theta)$:

  $$\hat{\theta} = argmin \left[\chi^2(\theta) \right].$$

- objective function and reduction to optimisation problem

- internal states – biochemical substances in the cell
- observables – substances that can be measured in time (e.g., metabolites or fluorescence reporters)
- input stimuli – profile of nutrient support, signalling stimuli or light program
- differential equations define continuous-time deterministic (population-average) evolution of biochemical substances
- autonomity comes from biochemistry and thermodynamics
    - mass-action kinetics, enzyme kinetics, ...
    - in this setting $x(t)$ and p are always positive

Parametric Identification: Problem Statement

Mathematical Models in Biology

- mechanistic models
  - mass-action systems
    - describes rate of any elementary reaction $\sum_{i=1}^{n} X_i \rightarrow \ldots$:

      $$v = k \prod_{i=1}^{n} X_i^{\sigma_i}$$

      where σ_i denotes kinetic order given by stoichiometry
    - easily obtainable model structure if reaction network is known
    - non-linearity is regular if stoichiometry ≤ 1
    - typically leads to over-parametrised models
  - Michaelis-Menten systems
    - enzyme kinetics based on pseudo-steady-state approximation
    - reduces number of variables and parameters
    - but for general case very complicated non-linear equations
    - similar are Hill systems (generalisation of MM)

Parametric Identification: Problem Statement

- canonical models
  - S-systems
    - for each species X_i one set of influxes and one set of effluxes is specified in terms of power-law functions:

      $$\dot{X}_i = \alpha_i \prod_{j=1}^{n} X_j^{\sigma_{ij}} - \beta_i \prod_{j=1}^{n} X_j^{\rho_{ij}}$$

      where n is the number of all system variables, α, β are rate constants for production and degradation, $\sigma, \rho \in \mathbb{R}$ are kinetic orders
  - generalised mass-action (GMA) systems
    - for each species X_i a sum of influxes/effluxes is specified (not aggregated)

      $$\dot{X}_i = \sum_{k=1}^{n_i} \left(\gamma_{ik} \prod_{j=1}^{n} X_j^{f_{ikj}} \right)$$

      where n_j is number of fluxes affecting X_i, γ positive rate constants, and $f \in \mathbb{R}$

Outline

1. Introduction
2. The Approach: Parametric Identification
3. System Identifiability Problem
4. Overview of Approaches

System Identifiability: Theoretical Concept

Define the (theoretical) set of exact parameter values:

$$D_T(\mathcal{S},\mathcal{M}) = \{\theta \mid \mathcal{M}(\theta) \text{ matches the system behaviour } \}$$

Ideally this set should be a singleton. In case of higher cardinality we speak about *overparameterization*.

Assume an estimate $\hat{\theta}(N;\mathcal{S},\mathcal{M},\mathcal{I},\mathcal{E})$ where N is the number of measurements in observed variable y.

System Identifiability: Theoretical Concept

Define the (theoretical) set of exact parameter values:

$$D_T(\mathcal{S}, \mathcal{M}) = \{\theta \mid \mathcal{M}(\theta) \text{ matches the system behaviour } \}$$

Ideally this set should be a singleton. In case of higher cardinality we speak about *overparameterization*.

Assume an estimate $\hat{\theta}(N; \mathcal{S}, \mathcal{M}, \mathcal{I}, \mathcal{E})$ where N is the number of measurements in observed variable y.

Definition

System $\mathcal{S}$ is (parameter) **identifiable under** $\mathcal{M}$, $\mathcal{I}$ and $\mathcal{E}$ iff $\hat{\theta}(N; \mathcal{S}, \mathcal{M}, \mathcal{I}, \mathcal{E}) \to D_T(\mathcal{S}, \mathcal{M})$ as $N \to \infty$.

System Identifiability: Confidence Intervals

$\hat{\theta}_i$ is associated a confidence interval $[\sigma_i^-, \sigma_i^+]$ with the meaning that true value of θ_i is located in $[\sigma_i^-, \sigma_i^+]$ with probability α

- asymptotic confidence

$$\sigma_i^{\pm} = \hat{\theta}_i \pm \sqrt{\Delta_\alpha(\chi^2) \cdot C_{ii}}$$

where

- $\Delta_\alpha(\chi^2)$ is α-quantile for χ^2
- $C = 2 \cdot H^{-1}$
- H is Hessian matrix (describing curvature of χ^2 around $\hat{\theta}_i$ by second-order partial derivatives)

System Identifiability: Confidence Intervals

$\hat{\theta}_i$ is associated a confidence interval $[\sigma_i^-, \sigma_i^+]$ with the meaning that true value of θ_i is located in $[\sigma_i^-, \sigma_i^+]$ with probability α

- asymptotic confidence

$$\sigma_i^{\pm} = \hat{\theta}_i \pm \sqrt{\Delta_\alpha(\chi^2) \cdot C_{ii}}$$

  where

  - $\Delta_\alpha(\chi^2)$ is α-quantile for χ^2
  - $C = 2 \cdot H^{-1}$
  - H is Hessian matrix (describing curvature of χ^2 around $\hat{\theta}_i$ by second-order partial derivatives)

- gives a good approximation of actual uncertainty of $\hat{\theta}_i$ if:
  - data have small error
  - amount of data is large wrt number of parameters
  - exact if $y(t)$ depends linearly on θ

System Identifiability: Confidence Intervals

- finite sample confidence

  $$\{\theta \mid \chi^2(\theta) - \chi^2(\hat{\theta}) < \Delta_\alpha\}$$

  where Δ_α is α-quantile as in the previous case
- gives an approximation of actual uncertainty of $\hat{\theta}_i$ up-to a statistically computed threshold

System Identifiability

Definition

Parameter θ_i is **identifiable** iff the confidence interval $[\sigma_i^-, \sigma_i^+]$ of the estimate $\hat{\theta}_i$ is finite.

System Identifiability

Definition

Parameter θ_i is **identifiable** iff the confidence interval $[\sigma_i^-, \sigma_i^+]$ of the estimate $\hat{\theta}_i$ is finite.

Reasons leading to non-identifiability:

- structural: model structure
- practical: precision of measured data

Structural Identifiability

Definition

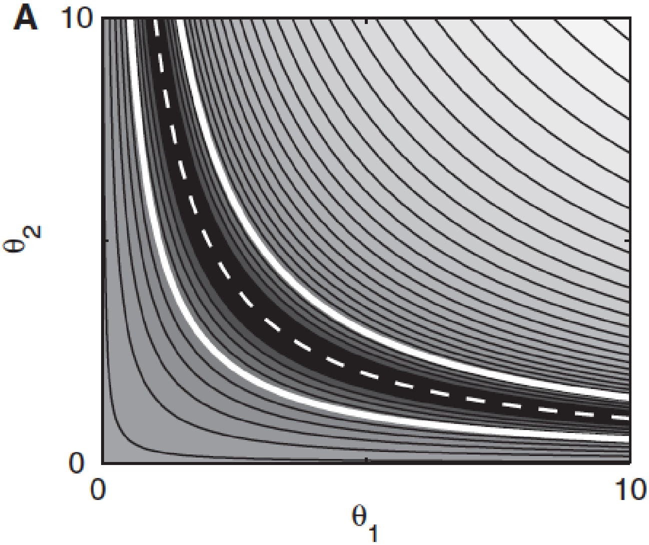
A parameter θ_i is **structurally identifiable** if a unique minimum of $\chi^2(\theta)$ exists with respect to θ_i.

- structural identifiability requires *uniqueness* of the solution
- redundant parameterisation of the model causing insufficient mapping of internal states x to observables y
- denote $\theta_{amb} \subset \theta$ the set of ambiguous parameters
- values of θ_{amb} may be varied without any change in y (and thus $\chi^2(\theta)$ keeps constant)
- in such a case there must be functional relations h among the parameters in θ_{amb} that are invariant wrt $\chi^2(\theta)$, and moreover:

$$\forall i, \theta_i \in \theta_{amb}.\, \sigma_i^- = -\infty \wedge \sigma_i^+ = \infty$$

Structural Identifiability

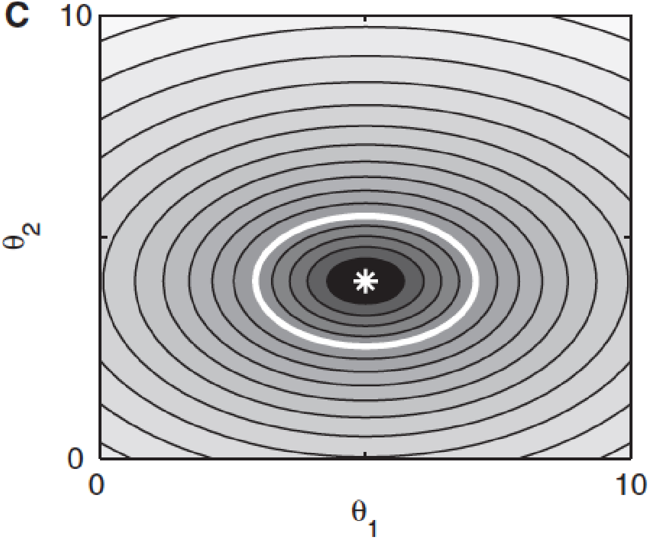
Structuraly Non-identifiable Parameters



functional relation between parameters: $h(\theta_{amb}) = \theta_1 \cdot \theta_2 - 10 = 0$

Structural Identifiability

Structuraly Identifiable Parameters



Practical Identifiability

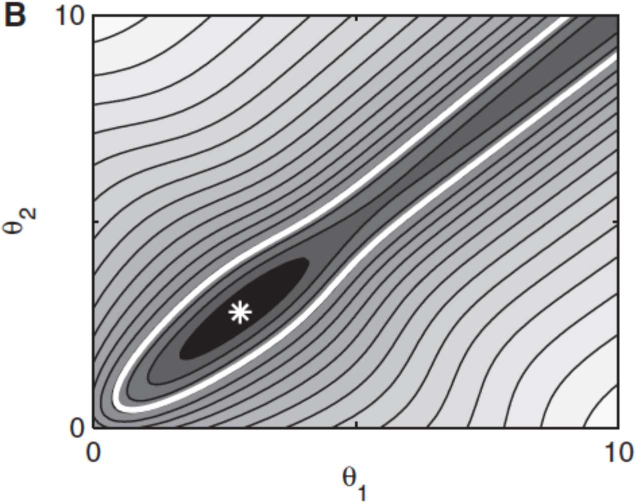
Definition

A parameter estimate $\hat{\theta}_i$ is **practically non-identifiable** if the finite sample confidence interval is infinitely extended in decreasing and/or increasing direction although there exists a unique minimum of χ^2.

- practical identifiability implies structural identifiablitity
- practical non-identifiability does not decide on structural identifiability
- detailed analysis can be used to improved experiment design

Structural Identifiability

Structuraly Non-identifiable System



confidence region is infinitely extended for $\theta_1 \to \infty$ and $\theta_2 \to \infty$

Detecting Identifiability

- differential algebraic methods to analyse the system equations can detect structural identifiability, computionally hard
- detection of χ^2 flateness using simulated and experimental data
  - approximation of curvature measures by quadratic approximation of χ^2 at $\hat{\theta}$
  - computation of Hessian or Fisher information matrix
  - appropriate for linear relations *h* among parameters
  - practical non-identifiability cannot be detected

Detecting Identifiability

Profile Likelihood Method by Raue et al. 2009

- explore the parameter space for each parameter in the direction of least increase in χ^2
- in particular this allows to follow the functional relations $h(\theta_{sub}) = 0$
- for practical identifiability detect crossing of the quantile threshold
- profile likelihood χ^2_{PL} is defined for each parameter θ_i:

$$\chi^2_{PL}(\theta_i) = min_{\theta_{j\neq i}} \left[\chi^2(\theta)\right] .$$

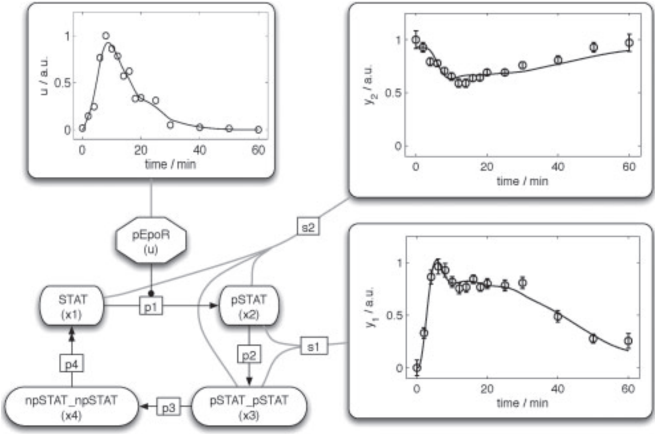
Experiment Design

Profile Likelihood Method by Raue et al. 2009

- suggestion of additional targeted measurements
- need measurements that narrow the confidence interval
- explore trajectories along PL of θ_i to improve estimation of θ_i

Parameter Identification

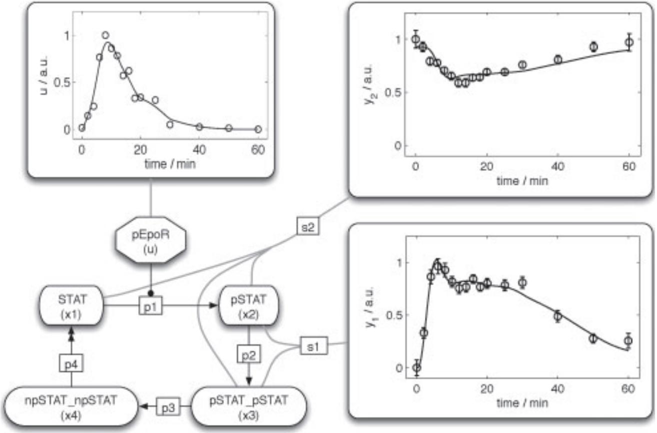
Signalling Pathway Example by Raue et al. 2009



studied system, external stimuli and measured vs. simulated data

Parameter Identification

Signalling Pathway Example by Raue et al. 2009

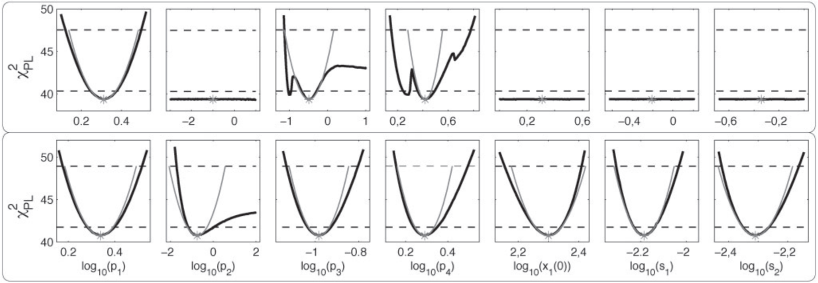


studied system, external stimuli and measured vs. simulated data

$$
\begin{aligned}
\dot{x}_1 &= -p_1 \cdot x_1 \cdot u + 2 \cdot p_4 \cdot x_4^{\tau} & \qquad y_1 &= s_1 \cdot (x_2 + 2 \cdot x_3)\\
\dot{x}_2 &= +p_1 \cdot x_1 \cdot u - p_2 \cdot x_2^2 & \qquad y_2 &= s_2 \cdot (x_1 + x_2 + 2 \cdot x_3)\\
\dot{x}_3 &= +\frac{1}{2} \cdot p_2 \cdot x_2^2 - p_3 \cdot x_3 & &\\
\dot{x}_4 &= +p_3 \cdot x_3 - p_4 \cdot x_4^{\tau} & &
\end{aligned}
$$

Parameter Identification

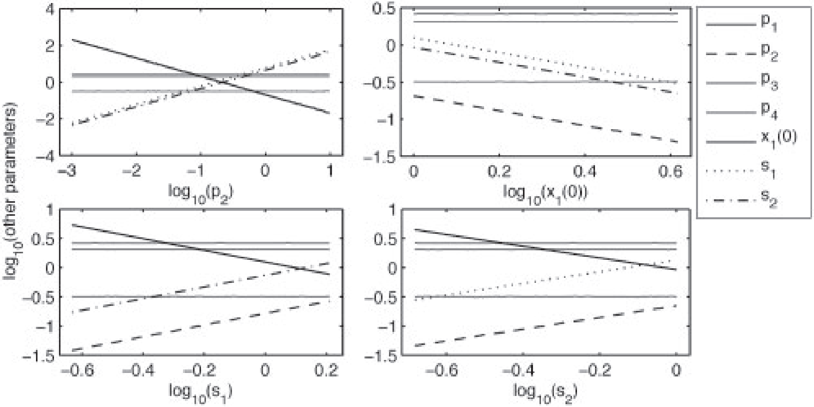
Signalling Pathway Example by Raue et al. 2009



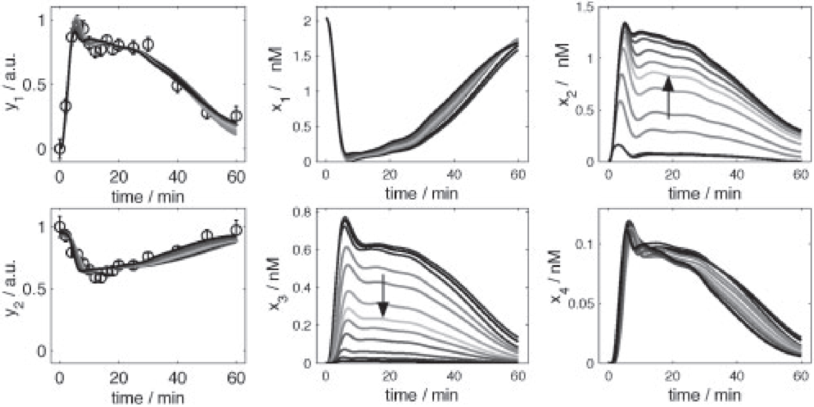
profile likelihood and its quadratic approximation

Parameter Identification

Signalling Pathway Example by Raue et al. 2009



relations among parameters



further PL-based analysis for experimental planning

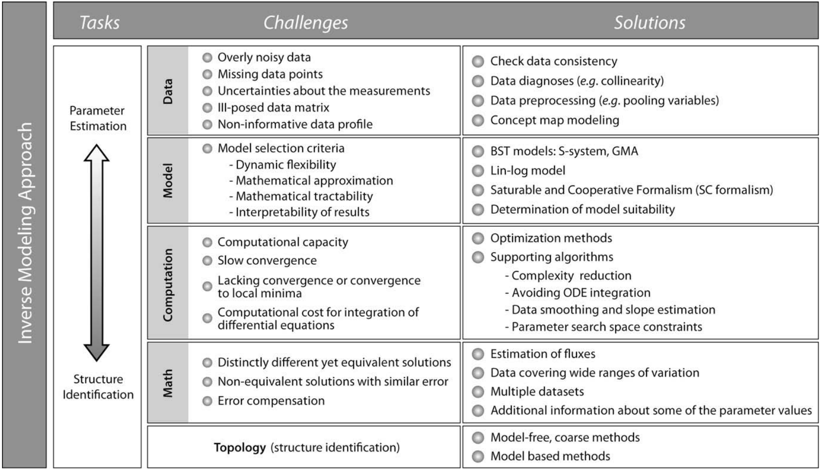
Outline

1. Introduction
2. The Approach: Parametric Identification
3. System Identifiability Problem
4. Overview of Approaches

Parameter Identification: Approaches Overview

- bottom-up vs. top-down modelling
  - bottom-up means detailed reconstruction from first principles
  - top-down (inverse) approach starts from high-throughput data
- steady-state vs. transient modelling
  - steady-state data give simplifying assumption (time is abstracted by long-run view)
  - works well for processes with a unique stable state
  - availability of internal system variables at steady-state (e.g., metabolism)
  - transient analysis more complicated (requires detection of initial states and appropriate time-series resolution is needed to inverse modelling)

Inverse Modelling Approach

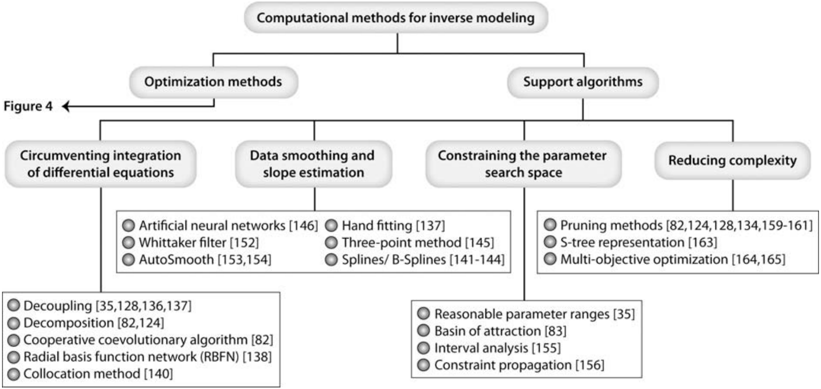


Inverse Modeling Approach

Tasks	*Challenges*		*Solutions*
Parameter Estimation ↕ Structure Identification	**Data**	● Overly noisy data ● Missing data points ● Uncertainties about the measurements ● Ill-posed data matrix ● Non-informative data profile	● Check data consistency ● Data diagnoses (*e.g.* collinearity) ● Data preprocessing (*e.g.* pooling variables) ● Concept map modeling
	Model	● Model selection criteria - Dynamic flexibility - Mathematical approximation - Mathematical tractability - Interpretability of results	● BST models: S-system, GMA ● Lin-log model ● Saturable and Cooperative Formalism (SC formalism) ● Determination of model suitability
	Computation	● Computational capacity ● Slow convergence ● Lacking convergence or convergence to local minima ● Computational cost for integration of differential equations	● Optimization methods ● Supporting algorithms - Complexity reduction - Avoiding ODE integration - Data smoothing and slope estimation - Parameter search space constraints
	Math	● Distinctly different yet equivalent solutions ● Non-equivalent solutions with similar error ● Error compensation	● Estimation of fluxes ● Data covering wide ranges of variation ● Multiple datasets ● Additional information about some of the parameter values
	Topology (structure identification)		● Model-free, coarse methods ● Model based methods

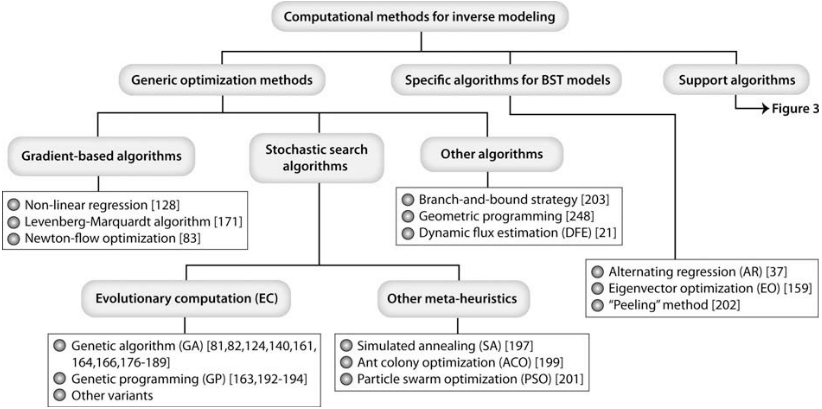
I-Chun Chou, E.O. Voit / Mathematical Biosciences 219 (2009) 57-83

Inverse Modelling Methods


Computational methods for inverse modeling
Optimization methods
Support algorithms
Figure 4
Circumventing integration of differential equations
Data smoothing and slope estimation
Constraining the parameter search space
Reducing complexity
Artificial neural networks [146]
Whittaker filter [152]
AutoSmooth [153,154]
Hand fitting [137]
Three-point method [145]
Splines/ B-Splines [141-144]
Pruning methods [82,124,128,134,159-161]
S-tree representation [163]
Multi-objective optimization [164,165]
Decoupling [35,128,136,137]
Decomposition [82,124]
Cooperative coevolutionary algorithm [82]
Radial basis function network (RBFN) [138]
Collocation method [140]
Reasonable parameter ranges [35]
Basin of attraction [83]
Interval analysis [155]
Constraint propagation [156]

I-Chun Chou, E.O. Voit / Mathematical Biosciences 219 (2009) 57-83

Optimisation Methods



Computational methods for inverse modeling

Generic optimization methods

Specific algorithms for BST models

Support algorithms

→ Figure 3

Gradient-based algorithms

- Non-linear regression [128]
- Levenberg-Marquardt algorithm [171]
- Newton-flow optimization [83]

Stochastic search algorithms

Other algorithms

- Branch-and-bound strategy [203]
- Geometric programming [248]
- Dynamic flux estimation (DFE) [21]

- Alternating regression (AR) [37]
- Eigenvector optimization (EO) [159]
- "Peeling" method [202]

Evolutionary computation (EC)

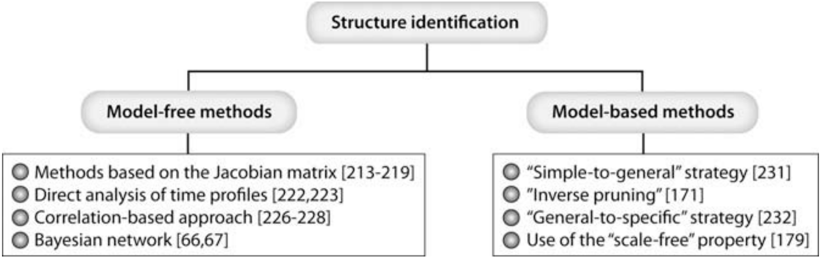
- Genetic algorithm (GA) [81,82,124,140,161, 164,166,176-189]
- Genetic programming (GP) [163,192-194]
- Other variants

Other meta-heuristics

- Simulated annealing (SA) [197]
- Ant colony optimization (ACO) [199]
- Particle swarm optimization (PSO) [201]

I-Chun Chou, E.O. Voit / Mathematical Biosciences 219 (2009) 57-83

Structure Identification



Structure identification

Model-free methods

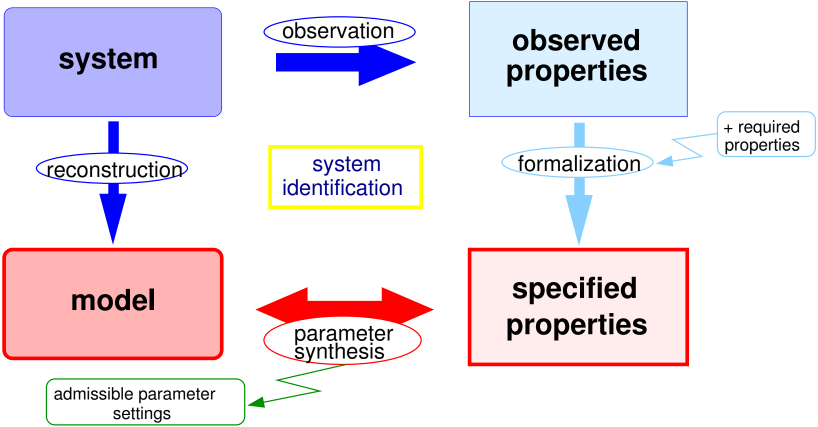
- Methods based on the Jacobian matrix [213-219]
- Direct analysis of time profiles [222,223]
- Correlation-based approach [226-228]
- Bayesian network [66,67]

Model-based methods

- "Simple-to-general" strategy [231]
- "Inverse pruning" [171]
- "General-to-specific" strategy [232]
- Use of the "scale-free" property [179]

I-Chun Chou, E.O. Voit / Mathematical Biosciences 219 (2009) 57-83

Parameter Exploration and Synthesis by Model Checking



Parameter Synthesis from LTL Specifications

Robustness

Given an LTL property φ and a parameterized model $\mathcal{M}$ check if $\mathcal{M}(\theta) \models \varphi$ **holds for all possible parameterizations** $\theta \in \mathcal{P}$ (valuations of parameters), $\mathcal{P}$ is called the **parameter space**.

Parameter Synthesis Problem

Given an LTL property φ and a parameterized model $\mathcal{M}$ **find the maximal set** $P \subseteq \mathcal{P}$ **of parameterizations** such that $\mathcal{M}(\theta) \models \varphi$ for all $\theta \in P$.

Problem Reduction

Robustness is reduced to Parameter Synthesis Problem by taking the set $\mathcal{P}$ of all possible parameterizations as P.

References

- T. Söderström, P. Stoica. System Identification. Prentice-Hall, 1989.
- I-Chun Chou, E.O. Voit. Recent developments in parameter estimation and structure identification of biochemical and genomic systems. Mathematical Biosciences 219 (2009) 57-83
- A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics, Vol. 25 no. 15 2009, pages 1923-1929.
- discussions with Stephan Müller, Jan van Schuppen, Alessandro Abate