Key concepts in mechanistic systems biology modeling for the course 8BKG45

Elin Nyman, Department of Biomedical Engineering, Linköping University

This document is intended to explain key concepts in mechanistic or small-scale systems biology taught in the course 8BKG45 at Linköping University. Please let me know if you need more explanations for any of the concepts or if you find errors in this document.
Good luck at the exam! /Elin

1. What is mechanistic/small-scale systems biology?

In mechanistic/small-scale systems biology, we use the current knowledge about a biological system to build mathematical models. We are usually interested in the understanding of key mechanisms that give rise to behaviors we observe in data. Common approaches in mechanistic systems biology is either to keep the model as small as possible and test one mechanism/one hypothesis at a time, or, the other way around, to start with all known mechanisms and reduce the model until only the necessary components are there. Even though we use the term ”small-scale”, the methods can be scaled to handle rather large models.

Mechanistic modeling can be more hypothesis driven or more data driven. In a hypothesis driven approach, we start with both data and knowledge about the system when we formulate the models, while in a data driven approach we do not use/need the full knowledge about the biological system to formulate the models. Instead we let data decide which biological interactions to include. The outcome of the analysis in a hypothesis driven approach is described in ”the modeling cycle” (see **8. The modeling cycle** below). In a data driven approach, the outcome is usually new hypothesis/ideas about the biological system under study. In both hypothesis driven modeling and in hypothesis driven modeling, we need to test and validate the findings of the model-based approach with new data. In this part of the course, we focus on hypothesis driven modeling.

2. Model formulation

We use ordinary differential equations (ODEs) to formulate models of the following format:

$$\frac{dx}{dt} = \dot{x} = f(x,u,p)$$

where x is a vector that contains model states and these states are derived with respect to time and therefore are allowed to change over time, u contains the model input which can both be constant and time-varying, and p are the parameters that are constant over time. The function f is a nonlinear function that depends on x, u and p.

We only use the simplest form of kinetics: mass-action kinetics (e.g. $v1 = k1 \cdot x1$) to describe the rate of the reactions. There are also more complex kinetic equations, such as saturated Michaelis-Menten expressions.

In this course, we need to be able both to go from a drawing of the biological system, a so called interaction graph, to the fully specified model ODEs, and vice versa. There is a general recipe to follow:

1. Identify model states, x
2. Identify reaction rates, v, including assumptions / what we know about parameters
3. Formulate ODEs, $d/dt(x)$
4. Identify what is measured, $\hat{y}$
5. Include all parameters, $p = (k, x(0), ky)$, and their values

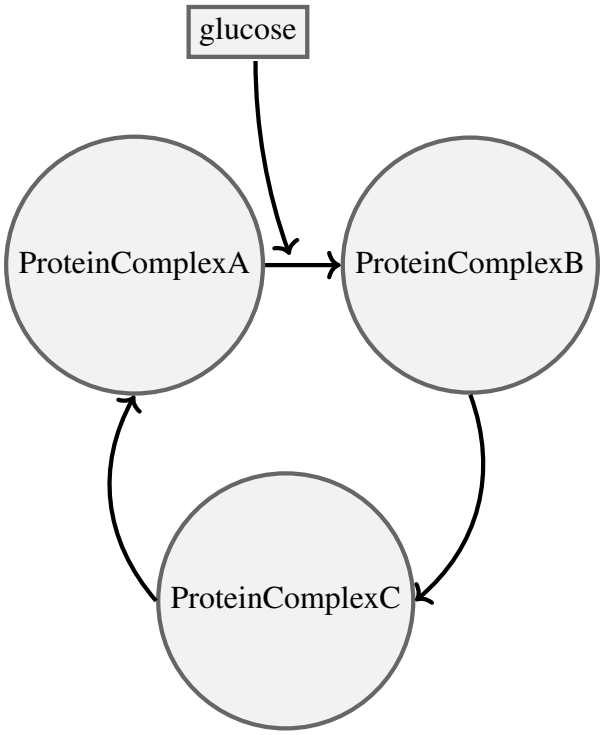
In this course, we need to know the name and function of all model parts:

- States, x, are derived with respect to time (denoted $d/dt(x)$ or $\dot{x}$), and are therefore allowed to change with time.
- Parameters, p are constant with respect to time. There are different kinds of parameters: rate constants, usually denoted $k1, k2$, etc, initial conditions, $x(0)$, and measurement parameters, ky. The values for parameters are usually not known, and have to be guessed or estimated based on data.
- Reaction rates, v, determines the rate of reactions. In this course we only use the simplest form of kinetics (mass-action kinetics), e.g./ $v1 = k1 \cdot x1 \cdot x2$.
- Measurement equation, e.g. $\hat{y} = ky \cdot x2$, and the meaning of this equation. Here: we cannot measure $x2$ directly, but something proportional to $x2$.

Let us look at an example. If you have a protein complex that can take transition between three different states with this set of reactions,

$$ProteinComplexA \rightarrow ProteinComplexB$$
$$ProteinComplexB \rightarrow ProteinComplexC$$
$$ProteinComplexC \rightarrow ProteinComplexA$$

and you know glucose is added to the system and involved in the first reaction, which is equivalent to this interaction graph



You also know that you can measure something that is proportional to the concentration of ProteinComplexB.

1. Identify model states:

$$x1 = [ProteinComplexA]$$
$$x2 = [ProteinComplexB]$$
$$x3 = [ProteinComplexC]$$

2. Identify reaction rates, including assumptions / what we know about parameters:

$$v1 = k1 \cdot x1 \cdot u$$
$$v2 = k2 \cdot x2$$
$$v3 = k3 \cdot x3$$

   Assumtions: mass-action kinetics, that glucose is the input $u = glucose$

3. Formulate ODEs:

$$d/dt(x1) = -v1 + v3$$
$$d/dt(x2) = v1 - v2$$
$$d/dt(x3) = v2 - v3$$

4. What is measured?

$$\hat{y} = ky \cdot x2$$

5. Parameters and their values:
   $k1 = 3$, $k2 = 1$, $k3 = 2$
   $ky = 0.5$
   $x1(0) = 0$, $x2(0) = 100$, $x3(0) = 10$
   All parameter values are assumed, since they are not given. Note that we need these values to be able to simulate the model. Also, we need to assume a value for the input strength: $u = 1$.

The full model formulation is given below:

$$
\begin{aligned}
d/dt(x1) &= -v1 + v3 \\
d/dt(x2) &= v1 - v2 \\
d/dt(x3) &= v2 - v3 \\
v1 &= k1 \cdot x1 \cdot u \\
v2 &= k2 \cdot x2 \\
v3 &= k3 \cdot x3 \\
\hat{y} &= ky \cdot x2
\end{aligned}
$$

$x1(0) = 0$, $x2(0) = 100$, $x3(0) = 10$
$k1 = 3$, $k2 = 1$, $k3 = 2$
$ky = 0.5$
$u = 1$

To instead go from a model formulated as ODEs to an interaction graph or model reactions, look at the all terms at the right hand side of the ODEs, including their sign (negative or positive). The terms represent the reaction rates ($v1, v2$, etc). For example,

$$
\begin{aligned}
d/dt(x1) &= -v1 + v2 \\
d/dt(x2) &= v1 - v2 \\
d/dt(x3) &= -v3 + v4 \\
d/dt(x4) &= v3 - v4
\end{aligned}
$$

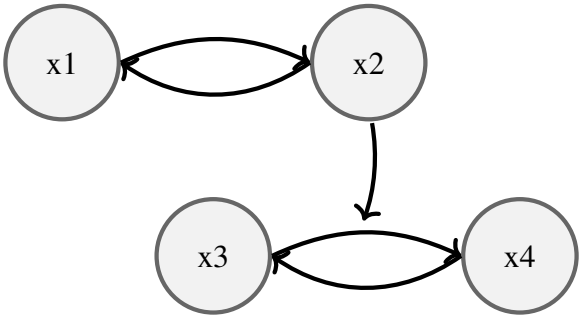
Here we see that $v1$ goes from $x1$ to $x2$ and $v2$ in the other direction, and that $v3$ goes from $x3$ to $x4$ and $v4$ in the other direction. To know more, we need to know the equations for the reaction rates:

$$
\begin{aligned}
v1 &= k1 \cdot x1 \\
v2 &= k2 \cdot x2 \\
v3 &= k3 \cdot x3 \cdot x2 \\
v4 &= k4 \cdot x4
\end{aligned}
$$

Here we see that $v3$ contains both $x3$ and $x2$ and therefore $x2$ must be involved in the transition from $x3$ to $x4$. We therefore conclude that we have these reactions:

$$\begin{aligned} x1 &\rightarrow x2 \\ x2 &\rightarrow x1 \\ x3 + x2 &\rightarrow x4 \\ x4 &\rightarrow x3 \end{aligned}$$

Which is equivalent to this interaction graph:



3. Model simulation

The models formulated as ODEs can usually not be solved analytically. Instead, we use numerical solvers to simulate the model. The most simple such solver to use is called the Euler method or the forward Euler method. The Euler method uses the following formula to compute the values for x after one time step (Δt):

$$x(\Delta t) = x(0) + d/dt(x(0)) \cdot \Delta t$$

i.e. we use the initial value for x and add the time-derivative for x at time = 0 and multiply this value with the time step (Δt). In this way we take a step in the direction of the slope of the ODE.

In reality, more sophisticated methods are used in numerical solvers to simulate models. E.g. several slopes can be computed and averaged, the second derivative can be used in the computation, and the time step can be varied depending on if we are in a region with a large slope, or in a region where we approach steady-state. are used to simulate models. In the computer exercise, we use such more sophisticated solvers to simulate the models for many time steps. These solvers are implemented in the used library.

Let us look at an example of how to use the simple Euler method. If we have this model,

$$\begin{aligned} d/dt(x1) &= -v1 + v2 \\ v1 &= k1 \cdot x1 \\ v2 &= k2 \\ x1(0) &= 4 \\ k1 &= 0.5 \\ k2 &= 1 \end{aligned}$$

and want to calculate $x1$ after a time-step of 0.1, $x1(0.1)$ using the Euler method:

$x(\Delta t) = x(0) + d/dt(x(0)) \cdot \Delta t = x(\Delta t) = x(0) + (-k1 \cdot x1(0) + k2) \cdot \Delta t = 4 + (-0.5 \cdot 4 + 1) \cdot 0.1 = 4 + (-1) \cdot 0.1 = 3.9$

4. Parameter estimation

Model parameters, especially the kinetic parameters that decides the rate of reactions, are usually not possible to measure experimentally within biology. Therefore, we use methods to estimate parameter values based on the available data. We use a cost function (also known as objective function or loss function) to evaluate the agreement between model simulations and data for each set of parameter values that we simulate. A cost function can look like this:

$$v(p) = \sum \left(\frac{y(t) - \hat{y}(t,p)}{SEM(t)} \right)^2$$

where the sum is over all measured time points, t; p is the parameters; $y(t)$ is the measured data and $\hat{y}(t)$ is the model simulations that corresponds to the data; $SEM(t)$ is the uncertainty given as standard error of the mean for the data. The expression is squared so that negative values become positive.

The residuals are the difference between data and model simulations $(y(t) - \hat{y}(t,p))$ and we want them to be as small as possible. We therefore minimize the cost function. To do so, there are global and local minimization functions to use. Global minimization functions aim to find the global minimum and therefore search both uphill and downhill in the landscape of possible parameter values to not get stuck in local minima. Local minimization functions, on the other hand, search only downhill, and needs to be combined with global minimization or multiple starting points to be effective.

In this course, we practically try out parameter estimation in the computer exercise, and look at examples where we go from model simulations with big residuals, i.e. a bad fit with data, to model simulations in agreement with data.

5. Statistical tests

Statistical tests are used to evaluate the agreement between model simulations and data. Usually, a visual inspection where you look at model simulations and data in the same graph, gives a hint on which test to use to see if you can reject the model/hypothesis or not. The statistical tests we need to know about in this course are:

- $\chi 2$-test for the size of the residuals – is the model in good enough agreement with data when you account for the data uncertainty?
- Whiteness test for correlation between residuals – is there a systematic error in the model that give rise to correlated residuals?
- Likelihood ratio test to compare models that all are in agreement with data – is one of the models significantly better than the others?
- Cross validation for model complexity – is the model too complex in relation to data and therefore over-fitted to the data used in parameter estimation?

We need to know how to formulate the null hypothesis and the alternative hypothesis for these tests, and we need to know when to use them and which conclusions we can draw using these tests. Lecture slides are rich in information on statistical tests, use them!

6. Model predictions with uncertainty

When we have a model that has passed statistical tests, and therefore is in good enough agreement with data, we want to use the model to predict the outcome of a new experiment. A single model simulation do not contain all possible model behaviors, instead we need to take into account the uncertainty of the model predictions. This uncertainty comes both from the model complexity in relation to estimation data, and the uncertainty of the estimation data. In practise, what we do is that we gather many parameter values that all give rise to simulations in agreement with data and use these parameter values to simulate the model predictions. Model predictions with uncertainty will look like an area or many lines, instead of a single line. Model predictions with low uncertainty (i.e. with a small area or overlapping lines) are more useful since these can be tested experimentally.

7. Experimental design

Mechanistic systems biology can be used to guide which experiment to do next, to get the most new knowledge about the biological system. The model is used to simulate different potential experiments in the computer before the experiment is performed, and in this way the most informative experiment can be selected. Important is to take into account the uncertainty of model predictions in the experimental design (see above).

8. The modeling cycle

The modeling cycle describes the work process in a mechanistic systems biology project. To start with, we need experimental data and biological knowledge to be able to formulate at least the first hypothesis. The hypothesis is formalized into ordinary differential equations, and the parameters of the model are estimated using a cost function that is being minimized. From this step there are two different outcomes: 1) The model is **not** in agreement with data according to a statistical test, and the model must be rejected. In this case, we need to formulate a new hypothesis. 2) The model is in agreement with data and can thus not be rejected with a statistical test. In this case, we collect an approximation of all parameters that give rise to model simulations that are in agreement with data and use for model predictions with uncertainty. These predictions are used to design new experiments to perform, and with the outcome of the experiment we see if the model can be validated (if the model prediction agrees with the new experiment) or has to be rejected (if the model prediction do not agree with the new experiment). In the latter case, we start over again with the formulation of a new hypothesis, using the gained insights from previous tested models. For each circuit, we update our knowledge about the biological system under study.