6 Reduction, selection, and coupling of models

6.1 Model reduction

Biochemical systems are complex, but in order to understand them, we can use simple mental pictures that neglect many details and show processes as if they happened in isolation. Simplicity can be reached by a change of perspective:

- If we average over many microscopic events, we will obtain a smooth behavior of macroscopic substance concentrations.
- If we observe a fast complex system on a slow time scale, its effective behavior may look simple.

In computational models, we can choose a level of detail that suits our needs: we may consider smaller or larger pathways and simplify, lump, or disregard substances and reactions. We can either do this at the very beginning by *model assumptions*, or we can simplify an existing model by *model reduction*. If a model turns out to be too simple, we may still zoom into the system and acknowledge details that we neglected before, or to zoom out and include more parts of the environment into the model.

Model Simplification Any biochemical model represents a compromise between biological complexity and practical simplicity; its form will depend on data and biological knowledge available and on the questions to be answered. Small models provide several advantages: it is easier to understand them, the effort for simulations is lower, and with fewer parameters, model fitting is easier and more reliable.

Different ways to simplify a given model are shown schematically in Figure 1. Such simplifications can speed up model building and simulations because fewer equations, variables, and parameters are needed, differential equations can be replaced by algebraic equations, and stiff differential equations can be avoided. All simplifications, though, have to be justified: a reduced model should yield a good approximation of the original model for certain quantities of interest, a certain time scale, and certain conditions (a range of parameter values, the vicinity of a certain steady state, or a certain qualitative behavior under study).

Tacit model assumptions Mathematical models describe biological systems in two complementary ways: positively, by how processes *are* modeled and negatively, by which processes are *not* modeled, how mechanisms are simplified, and which quantities are considered constant. The positive facts about the system are stated explicitly, while the negative ones - which are just as important - remain hidden in the model assumptions. Arguably, the most basic negative statement is that a system as a whole can be seen as a module, that is, its environment - e.g. the cell surrounding a pathway - can be neglected. Experiments test both kinds of statements at the same time, and they are often designed from the very beginning such that the simplifying model assumptions will later be justified.

Even the most detailed biochemical model is still a simplified, reduced picture of a much more complex reality. Therefore, considerations about model reduction do not only help to simplify existing models, but also to justify common basic model assumptions and our use of mental models in general.

6.1.1 Reduction of Fast Processes

If processes take place on different time scales, this may allow to reduce the number of differential equations. In gene expression, for instance, binding and unbinding of transcription factors can happen on the order of microseconds, changes in transcription factor activity on the order of minutes, while the culture conditions may change on the order of hours. In a model, we may use a fast equilibrium or time averages for transcription factor binding, a dynamical model for signal transduction and gene expression, and constant values for the culture conditions.

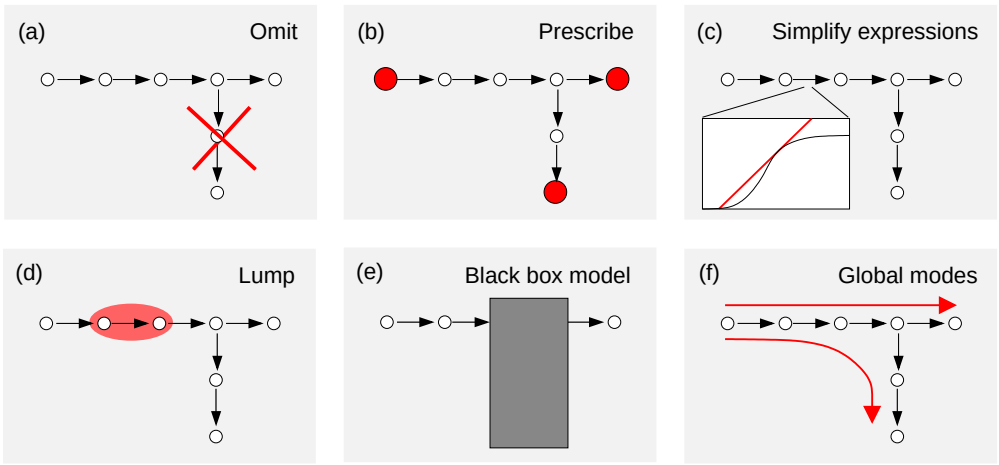


Figure 1: Simplifications in biochemical models. The scheme shows a branched pathway of metabolites (circles) and reactions (arrows). (a) Omitting substances or reactions. (b) Predefining the values of concentrations or fluxes or relations between them. (c) Simplifying the mathematical expressions (e.g. omit terms in a kinetic law, use simplified kinetic laws, , neglect insensitive parameters). (d) Lumping the substances, for instance, similar metabolites, protonation states of a metabolite, or metabolite concentrations in different compartments. Subsequent reactions in a pathway or elementary steps in a reaction can be replaced by a single reaction of the same velocity; for parallel reactions, like the action of isoenzymes, the velocities are summed up; for the two directions of a reaction, the velocities are subtracted. (e) Replacing the model parts by a dynamic black-box model that mimics the input-output behavior. (f) Describing the dynamic behavior by global modes (e.g., elementary flux modes or eigenmodes of the Jacobian).

Cellular processes occur on a wide range of time scales from microseconds to hours, and also the time scale of enzymatic reactions can differ strongly due to the very different enzyme concentrations and kinetic constants.

Time scale separation In numerical simulations, a single fast process, e.g. a rapid conversion between two substances A $\rightleftarrows$ B (as in Figure 2 (b)), can force the numerical solver to use very small integration steps. If the same model also contains slow processes, simulations have to cover a long time scale, and the numerical effort can become enormous. However, fast reactions can be approximated rather easily because the concentration ratio $s_\mathrm{B}/s_\mathrm{A}$ will always be close to the equilibrium constant. If we approximate this by an exact equilibrium in every moment in time, we can replace the reaction by the algebraic relation $s_\mathrm{B}/s_\mathrm{A} = k^\mathrm{eq}$ and get rid of the stiff differential equation that caused the big numerical effort.

The mathematical justification for the effective algebraic equations is illustrated in Figure 2: in state space, fast processes may rapidly move the system state towards a submanifold, on which certain relations hold (e.g. an equilibrium between different concentrations). After an initial relaxation phase, the system state will change more slowly and remain close to this manifold. In general, there may be a hierarchy of such manifolds which are related to different time scales.

6.1.2 Quasi-steady-state and quasi-equilibrium approximation

We shall illustrate two types of approximation, quasi-steady-state and quasi-equilibrium, with a simple model of upper glycolysis

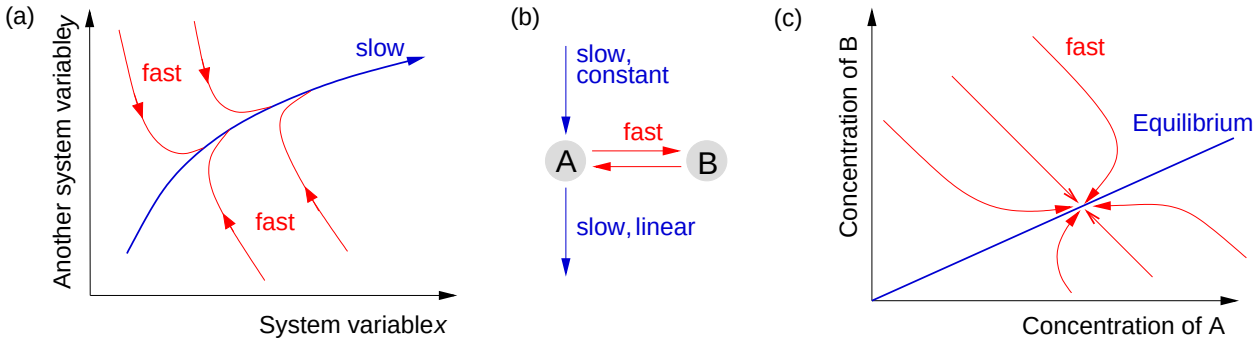
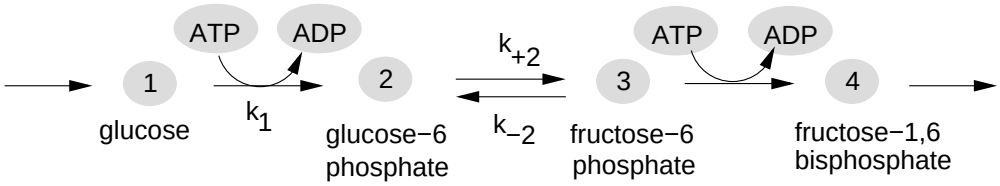


Figure 2: Time scale separation. (a) The dynamics of a system can be illustrated by its trajectories in state space. If the system state is attracted by a submanifold (in the two-dimensional case, a curve), trajectories starting from any point (red) will rapidly approach this manifold (blue). Later, the system will move slowly on the manifold, satisfying an algebraic equation. (b) A small reaction system with different time scales. Fast conversion between metabolites A and B will keep their concentration ratio s_b/s_a close to the equilibrium constant k^{eq}, while slow production and degradation of A only changes the sum $s_a + s_b$. (c) Schematic trajectories for the system shown in (b). For any initial conditions, the concentrations s_{A} and s_{B} will rapidly approach the line $s_{\mathrm{B}}/s_{\mathrm{A}} = k^{\mathrm{eq}}$ and then move slowly along the curve line to the steady state.



Glucose (GLC) is taken up at a rate v_0 and converted subsequently into glucose-6-phosphate (G6P), fructose-6-phosphate (F6P) and fructose-1,6-bisphosphate (FBP), which is then consumed by the following steps of glycolysis. In this model, the cofactors ATP and ADP have fixed concentrations. With mass-action kinetics and a reversible reaction between G6P and F6P, the rate equations read:

$$\begin{aligned}
\mathrm{d}s_1/\mathrm{d}t &= v_0 - k_1\, s_{\mathrm{A}}\, s_1 && (1)\\
\mathrm{d}s_2/\mathrm{d}t &= k_1\, s_{\mathrm{A}}\, s_1 - k_{+2}\, s_2 - k_{-2}\, s_3 && (2)\\
\mathrm{d}s_3/\mathrm{d}t &= k_{+2}\, s_2 - k_{-2}\, s_3 - k_3\, s_{\mathrm{A}}\, s_3 && (3)\\
\mathrm{d}s_4/\mathrm{d}t &= k_3\, s_{\mathrm{A}}\, s_3 - k_4\, s_4 && (4)
\end{aligned}$$

The numbers refer to the metabolites and reactions in the scheme and s_{A} denotes the constant ATP concentration. We first assume that all reactions take place on the same time scale, setting $k_{\pm 2} = 2$ and all other rate constants and the ATP concentration to a value of 1 (arbitrary units). Figure 3 (a) shows simulated concentration curves of GLC, G6P, F6P, and FBP; the initial concentrations are chosen to be zero. For the first 5 time units, the influx has a value of $v_0 = 2$, and the intermediate levels rise one after the other. Then, the influx is reduced to $v_0 = 1$, and the levels decrease again.

How would the system behave if either the first or the second reaction was very fast? The two scenarios can be approximated, respectively, by a quasi-steady state for glucose or a quasi-equilibrium between G6P and F6P.

Quasi-Steady-State Approximation If k_1 is increased to a value of 5 (Figure 3 (b)), glucose is rapidly consumed, so its steady-state level stays low; due to its high turnover, glucose will also adapt almost instantaneously to changes of the input flux. This behavior can be approximated by a quasi-steady-state approximation for the slow time scale: we replace the glucose concentration in each time point by the

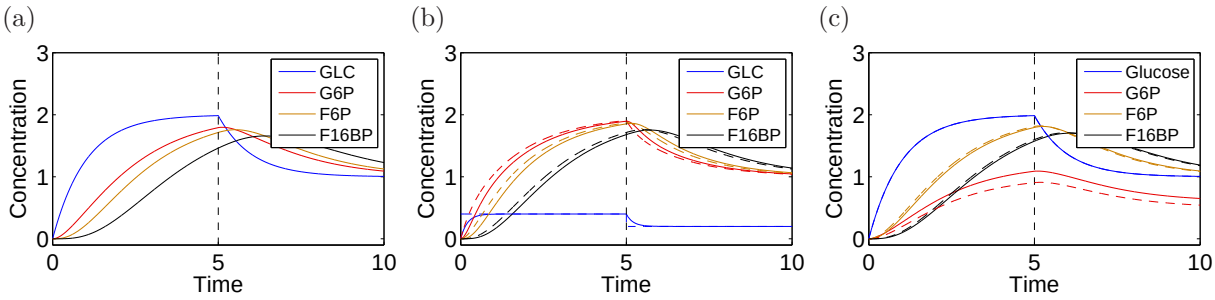


Figure 3: Simulation results for the model of upper glycolysis. (a) Results from the original model, showing levels of GLC, G6P, F6P, and FBP (abbreviations see text, time and concentrations measured in arbitrary units). (b) Results from the model with fast glucose turnover $k_1 = 5$ (solid lines) and the quasi-steady-state approximation (broken lines). (c) Results from the model with fast reversible conversion G6P $\leftrightarrow$ F6P (solid lines), parameters $k_{+2} = 10, k_{-2} = 5$ and the quasi-equilibrium approximation (broken lines).

steady-state value $s_1^{\mathrm{st}}(t) = v_0(t)/(k_1 s_A)$ based on the current value of $v_0(t)$. This algebraic equation replaces the differential equation (1) for s_1. Formally, we could obtain the same result by setting the left-hand side of the differential equation to zero.

Quasi-Equilibrium Approximation Next, we assume a rapid and reversible conversion between the hexoses G6P and F6P. We increase both rate constants at the same time by a large factor ($k_{+2} = 10$ and $k_{-2} = 5$ in Figure 3 (c)) while keeping their ratio $k^{\mathrm{eq}} = k_{+2}/k_{-2}$ fixed: in the simulation, the ratio of F6P to G6P levels rapidly approaches the equilibrium constant [F6P]/[G6P]=$s_3/s_2 = k^{\mathrm{eq}}$. In the quasi-equilibrium approximation, we assume that this ratio is exactly maintained in every moment. By adding equations (2) and (3), we obtain the equation

$$\frac{\mathrm{d}s_{2+3}}{\mathrm{d}t} = \frac{\mathrm{d}(s_2 + s_3)}{\mathrm{d}t} \quad = \quad k_1\, s_{\mathrm{A}}\, s_1 - k_3\, s_{\mathrm{A}}\, s_3. \tag{5}$$

Given s_{2+3} and k^{eq}, we can substitute $s_3 = s_{2+3}\, k^{\mathrm{eq}}/(1 + k^{\mathrm{eq}})$ in Eq. (4) and obtain a simplified differential equation system in which the fast reaction does not appear any more. The two differential equations for s_2 or s_3 are replaced by a single differential equation (for either of the two variables) and an algebraic equation $s_3/s_2 = k^{\mathrm{eq}}$ for the concentration ratio.

6.2 Model Selection

One of the main issues in mathematical modeling is to choose between model structures and to justify this choice. It is often arguable which biological elements need to be considered and there may be a variety of alternative model structures. Models may cover different cellular subsystems, different components or interactions within a subsystem (e.g. feedback interactions), different descriptions of the same process (e.g. different kinetic laws, fixed or variable concentrations), and different levels of detail (subprocesses or time scales). Combining alternative versions of model parts can lead to a combinatorial explosion of model variants, so we need to rule out models that are incorrect or too complicated. With limited and inaccurate data, we will not be able to pinpoint a single very detailed model, but statistical criteria can at least tell us which of the models are supported by the data.

6.2.1 What is a good model?

A good model need not describe a biological system in all details. J. L. Borges writes in a story: "In that empire, the art of cartography attained such perfection that the map of a single province occupied the entirety of a city, and the map of the empire, the entirety of a province. In time, those unconscionable maps no longer satisfied, and the cartographers guilds struck a map of the empire whose size was that of

the empire, and which coincided point for point with it." Systems biology models range from very simple to very complex maps of the cell, but just like usual maps, they never become an exact copy of the biological system. If they did, they would be almost as hard to understand as the biological system itself. Or, as George Box put it, "Essentially, all models are wrong, but some are useful". But - useful for what? Models are made for different purposes and have to meet different requirements (see Figure 4).

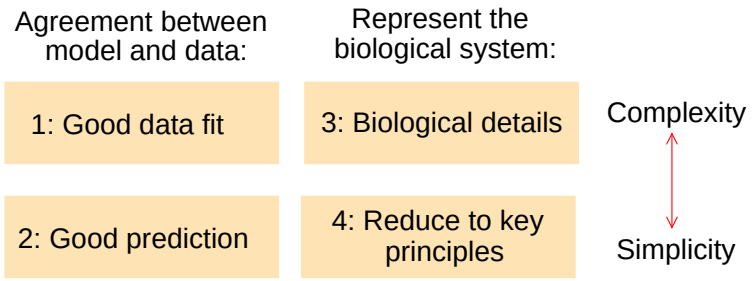


Figure 4: Possible requirements for a good model.

1. In *data fitting*, we aim to describe individual data points by a general mathematical function. Instead of storing many data pairs (x, y) that lie on a curve, we can store a few curve parameters (e.g. offset and slope for a straight line). If the points are not exactly on a curve, we may explain the discrepancy between model and data by statistical errors in the measurement. Given a model structure, we can adjust the model parameters such as to optimize the fit, e.g. by minimizing the sum of squared residuals. Fitting equally applies to dynamical models, which parametrize data curves in an indirect way.

2. In *prediction*, a model is supposed to state general relationships between measured quantities. In contrast to mere data fitting, predictions are supposed to hold for future observations: in the language of statistical learning, the model should generalize well to new data.

3. A *detailed mechanistic model* is supposed to describe processes "as they happen in reality". Of course, the description of an entire cell will never be complete down to molecular or lower levels. In practice, mechanistic models will focus on parts of the cell only and use model assumptions and model reduction to simplify them to a tractable level.

4. To emphasize the *key principles* of a biological process, a model needs to be *as simple as possible*. Simplicity is especially important if a model is supposed to serve as a paradigm; this also holds for experimental model systems, e.g. the Lac operon as a model for microbial gene regulation.

A philosophical principle called *Ockham's razor* (*Entia non sunt multiplicanda praeter necessitatem*) states that a theory should not contain unnecessary elements. Also in statistical model selection, complexity in a model always needs to be supported by data.

A good data fit supports the hypothesis that a model is biologically correct and covers the key features of a system. But - it does not prove it: a complex model - even with an implausible structure - may achieve better fits than a simpler, biologically plausible model. As a rule of thumb, a model with many free parameters may fit given data more easily ("With four parameters I can fit an elephant, and with five I can make him wiggle his trunk." J. von Neumann. But at the same time, the average amount of available experimental information for each parameter decreases, so the estimated parameters and the predictions from the model become poorly determined. Such overfitting is a notorious problem when many free parameters are fitted to few data points or if a large number of possible models is prescreened for good data fits (Freedman's paradox); it can be detected and avoided, though, by making proper use of statistics.

A good fit alone does not count as a support if data are used twice, for parameter estimation and model selection. To find models with reliable parameter estimates and good potential for predictions, we need to give all models equal chances. To correct for the advantage of large models, we may apply the likelihood ratio test or selection criteria, which both favor models with few parameters.

6.2.2 Maximum likelihood estimation and χ^2 test

We can judge the quality of a model by comparing its predictions to experimental data. The structure and parameters of a model can be scored by its *likelihood*, the probability that the model assigns to the observed data. Consider the model "Tomorrow, the sun will shine with 80 % probability": after sunshine has been observed, the observation has a likelihood of 0.8. Mathematically, the likelihood for a model or parameter set θ is defined as $L(\theta|y) = p(y|\theta)$, that is, the conditional probability to observe the data y given the model.

To compute likelihood values for biochemical models, we need to relate the model predictions to experimental data. Often, a measured concentration time series y_i is compared to the results $x_i(\theta)$ of a dynamical model with parameter set θ. The subscript i can refer to both substances and time points. In a simple statistical model, we regard the experimental data as a sum

$$y_i = x_i(\theta) + \xi_i \tag{6}$$

of the model results and measurement errors ξ_i, described by independent Gaussian random variables with mean 0 and width σ_i. The assumption of additive Gaussian errors greatly simplifies calculations, but it need not hold in all cases. With Eq. (6) and the probability density $p_{\xi_i}(\xi)$, the likelihood $L(\theta|y)$ can be written as a function of the model parameters. For the further calculations, we consider the expression

$$-2\log L(\theta|y) \quad = \quad -2\log p(y|\theta) = -2\sum_{i=1}^{n} \log p_{\xi_i}(y_i(t) - x_i(\theta)). \tag{7}$$

By inserting the Gaussian probability density $p_\xi(\xi) \sim \exp(-\xi^2/(2\sigma^2))$, we obtain

$$-2\log L(\theta|y) \quad = \quad \sum_{i=1}^{n} \frac{(y_i - x_i(\theta))^2}{\sigma_i^2} + \text{const.} \tag{8}$$

The quality of the model (6) can be judged from the sum in expression (8), the weighted sum of squared residuals (wSSR). If our model is correct, the y_i will be independent Gaussian random variables with means x_i and variances σ_i^2, and the weighted SSR will follow a χ^2-distribution with n degrees of freedom. On the other hand, if the value of (8) for a given model and given data falls in the upper 5% quantile of the χ_n^2-distribution, the model can be rejected on a 5% confidence level. We would conclude in this case that the model is wrong. In parameter fitting, the number of degrees of freedom in the χ^2-distribution is effectively reduced due to overfitting.

In *maximum likelihood estimation*, we determine a parameter set $\hat{\theta}(y)$ that maximizes the likelihood $p(y|\theta)$ for a given data set y: the resulting likelihood value measures the *goodness of fit*. If the measurement noise has the same width σ for all variables and time points, we obtain

$$-2\log p(y|\theta) = R/\sigma^2 + \text{const.}, \tag{9}$$

where $R = \sum_i (y_i - x_i(\theta))^2$ is the sum of squared residuals. Hence, least squares fitting is an example of maximum likelihood estimation for this particular case.

The likelihood can also be used to choose between different model structures. For instance, the above statement A, "Tomorrow, the sun will shine with 80 % probability" can be compared to the statement B, "Tomorrow, the sun will shine with 50 percent probability". After sunshine has been observed, statement A will have a higher likelihood (Prob(data|A) = 0.8) than statement B, (Prob(data|B) = 0.5), and should be chosen if likelihood is used to select models. Biochemical models can be selected in the same manner. This requires, however, that their parameters have been fixed in advance, as we shall explain now.

6.2.3 Overfitting ("learning by heart") and ways to avoid it

In the following, we assume that a number of alternative models have been proposed for a biological process. We intend to choose between them based on experimental data, in particular, time series of substance

concentrations. In model selection, we compare the two models to experimental data, e.g. a concentration time series for S consisting of triples (t_i, y_i, σ_i) for the i^{th} measurement, each containing a time point t_i, a measured concentration value y_i, and a standard error σ_i.

If models were selected simply by their likelihood, overfitting could severely distort the choice of models. Consider a statistical model with true parameters θ and data y: the maximum-likelihood estimator $\hat{\theta}(y)$ will lead to a higher likelihood $L(\hat{\theta}(y)|y) > L(\theta|y)$ than the true parameters θ just because it was optimized for high likelihood for the realized data. The empirical (maximized) log-likelihood will exceed the log-likelihood of the true parameters, on average, by an amount $\Delta \log L$. This bias depends on how easily the model can fit the noise; usually, it increases with the number of free model parameters. At the same time, the parameters in such models will be poorly determined. Model parameters can be unidentifiable for structural reasons (which should be avoided), but if there are more free parameters than data points, the parameters are unidentifiable for sure. This should be avoided by restricting the number of parameters and by ensuring that they all actually matter for model predictions. We can choose between competing models by statistical tests and model selection criteria.

6.2.4 Likelihood Ratio Test

In *statistical tests*, we compare a more complex model to a simpler background model. According to the null hypothesis, both models perform equally well. In the test, we favor the background model unless it statistically contradicts the observed data. In this case, we would conclude that the data require the more complex model. A test at a confidence level α will ensure that if the null hypothesis is correct, there is only an $\alpha\%$ chance that we wrongly reject it.

The *likelihood ratio test* compares two models A and B (with k_A and k_B free parameters, respectively) by their maximal likelihood values L_A and L_B. The two models have to be nested, that is, model B must be obtained from model A by fixing a number of parameters in advance. In the test, we assume, as a null hypothesis, that both models explain the data equally well. But even if model B is correct, model A will show a higher empirical likelihood because its additional parameters make it easier to fit the noise. For large numbers of data points and independent, Gaussian-distributed measurement errors, the expression $r = 2\ln(L_\text{A}/L_\text{B})$ asymptotically follows a χ^2 distribution, with $k_\text{A} - k_\text{B}$ degrees of freedom. This distribution is used for the statistical test: if the empirical value of r is significantly high, we reject the null hypothesis and accept model A. Otherwise, we accept the simpler model B. The likelihood ratio test can also be applied sequentially to more than two models, provided that subsequent models are nested.

6.2.5 Selection criteria

Alternatively, several candidate models can be compared by a *selection criteria*. Selection criteria are mathematical scoring functions that balance agreement with experimental data against model complexity. To compensate for the advantage of complex models, high numbers of free parameters in the model are punished. Values of the selection criterion can be used to rank the models, choose between them, and to average over them. In model selection, we choose between model structures just as we choose between parameter values in and parameter fitting: in both cases, we intend to find a model that agrees with biological knowledge and that matches experimental data. In parameter estimation, parameter values are determined for a given model structure, while model selection often involves parameter estimation for each of the candidate models.

We saw that the maximal likelihood is biased by a value ΔL, so for model selection, it would be better to score models by an unbiased estimator $\log L(\hat{\theta}(y)|y) - \Delta L$. The value of ΔL is unknown in general, but mathematical expressions for it, so-called *selection criteria*, have been proposed for certain forms of models. By minimizing these objective functions (instead of the likelihood itself), we attempt to find a model that best explains the data, while taking into account the possibility of overfitting. The Akaike information criterion

$$\text{AIC} = -2 \log L(\hat{\theta}(y)|y) + 2k, \tag{10}$$

for instance, directly penalizes the number k of free parameters. If we assume additive Gaussian measurement noise of width 1, the term $-2 \log L(\theta|y)$ in Eq. (10) equals the sum of squared residuals and we obtain

$$\mathrm{AIC} = -2 \log L(\theta|y) + 2k. \tag{11}$$

In some cases, the selection criteria may suggest that none of the models is considerably better than all others. In this situation, we may give up our plan to select a single model and accept that several models should be considered. For example, to estimate a model parameter, we may average over the parameter values obtained from the different models. To give higher weight to parameters from the more reliable models, weighting factors can be constructed from the selection criteria.

6.3 Coupled Systems and Emergent Behavior

All biological systems, from organisms down to cellular pathways, are embedded in larger environments that influence their dynamics. A metabolic pathway, for instance, is part of a larger network and coupled to a transcription network that adjusts its enzyme levels. For the dynamics of such a system, it can make a big difference if the environment is kept fixed or if both systems interact dynamically. To distinguish between these tweo cases, we can say that the system is studied *in isolation* (with fixed or controlled environment) or *coupled* to a dynamic environment. This fundamental distinction does not only hold for models, but also for experimental systems: in an *in vitro* enzyme assay, for instance, conditions like pH or the levels of cofactors can be controlled; in living cells, these values may be regulated dynamically, usually in an unknown manner.

If systems are coupled, new dynamic behavior can emerge. Single yeast cells, for instance, can be coupled by the exchange for chemicals: such interactions can lead, for instance, to synchronized glycolytic oscillations, which are observed both in experiments and in models. The following two examples illustrate the difference between isolated and coupled systems.

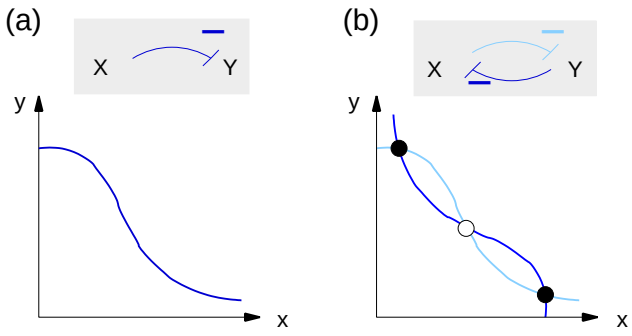


Figure 5: Bistability can emerge from mutual inhibition. (a) A gene level y is modeled in isolation with another gene level x as a regulatory input. The steady-state level y^{st} (blue) depends on the predefined value of x. (b) The coupled system shows bistability as an emergent property, with two stable fixed points (black dots) and one unstable fixed point (white dot) at the intersubsection of the two nullclines.

6.3.1 Feedback through the environment can change a system's behviour: two examples

Example: bistable switch Let us consider two genes X and Y that mutually inhibit each other (Figure 5); we describe their levels x and y by the differential equation model

$$\begin{aligned} \mathrm{d}x/\mathrm{d}t &= f(x,y) \\ \mathrm{d}y/\mathrm{d}t &= g(x,y). \end{aligned} \tag{12}$$

By setting the second equation to zero and solving for y, we obtain the steady-state value of y as a function of x. The curve $y^{\mathrm{st}}(x)$ in Fig. 5 (a) is called the *nullcline* of y. Likewise, we obtain another nullcline $x^{\mathrm{st}}(y)$

from the first equation. These nullclines represent response curves for the individual systems. When both systems are coupled, both steady-state requirements $y^{\mathrm{st}} = f(x^{\mathrm{st}})$ and $x^{\mathrm{st}} = g(y^{\mathrm{st}})$ have to be satisfied at the same time. We obtain three fixed points, two of which are stable. Due to the positive feedback loop, a bistable switch has emerged. The bistability is not a property of the individual genes X and Y - it is an emergent property which is only caused by their coupling.

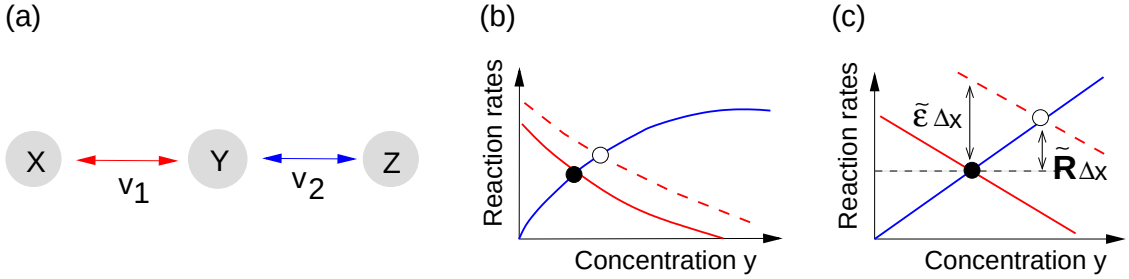


Figure 6: Elasticities and response coefficients describe local and global response to external changes. (a) Chain of two reactions with external metabolites X and Z and intermediate Y. (b) The reaction rates v_1 (red) and v_2 (blue) depend on the intermediate level y. A steady state requires that both rates are identical (black dot). If v_1 is increased - e.g, by an increase of the external substrate X (broken red line), the steady-state flux and concentration are shifted (white dot). (c) The magnified scheme compares the direct increase of the reaction rate $E_S \Delta x$ (depending on the reaction elasticity $\epsilon_x^{v_1}$) to the steady-state flux increase $\tilde{R}_x^{\mathrm{j}} \Delta x$ (with response coefficient $\tilde{R}$).

Example: Reaction velocity and steady-state flux Figure 6 shows two coupled chemical reactions. To study the first reaction in isolation, we fix the concentrations of substrate X and product Y. The reaction rate is then given by the kinetic law $v_1(s_{\mathrm{X}}, s_{\mathrm{Y}}, E_1)$, and the response to a small increase of enzyme activity is described by the elasticity coefficient $\epsilon_{E_1}^{v_1} = \partial v_1 / \partial E_1$. As the enzyme activity increases, the reaction rate can be made arbitrarily large. Alternatively, we can study the stationary flux in the two coupled reactions (with the levels of X and Z fixed and the level of Y determined by a steady-state requirement). Now the rate of the first reaction equals the steady-state flux $j(x, z, E_1, E_2)$ and the effect of an increased enzyme activity is given by a response coefficient $\tilde{R}_{E_1}^{j} = \partial j / \partial E_1$. In this setting, the first enzyme will have a limited effect on the reaction rate: as its activity increases, the enzyme will lose its control and the reaction flux will be mostly controlled by the second enzyme.

6.3.2 Complexity and simplicity

Reductionism and holism The two approaches - considering isolated and coupled dynamics - are characteristic for two contrary views on complex systems. *Reductionism* studies the parts of a system in isolation and great detail. In this view, which is dominant in molecular biology and biochemistry, the global behavior of a system is explained in terms of interactions between the system's parts, and the dynamics is explained in terms of causal chains. *Holism*, on the contrary, emphasizes the fact that new dynamic behavior can emerge from the coupling of subsystems. Instead of tracing individual causal effects, it studies how the global system dynamics responds to changes of external conditions.

Bottom-up and top-down modelling Accordingly, there are two complementary modeling approaches, called bottom-up and top-down modeling; both proceed from simplicity to complexity, but in very different ways. In *bottom-up* modeling, one studies elementary processes in isolation and aggregates them to a model. An example is the glycolysis model of Teusink et al. (2000) that was built from kinetic rate laws measured *in vitro*. *In vitro* measurements of enzyme kinetics allow for an exact characterization and manipulation of supposedly relevant parameters. A metabolic pathway model was constructed by adding the reactions; without further tuning, it yielded a fairly plausible steady-state description of glycolysis. In *top-down*

modeling, on the contrary, a model is built by refining a coarse-grained model of the entire system. If the model structure is biologically reasonable, such a model can be expected to yield fairly good data fits, but there is no general guarantee that it will remains valid as part of a merged model. The two approaches pursue on different goals: bottom-up model is constructed to be locally correct (describe reactions by correct rate laws and parameters), while a top-down model on the other hand, is optimized for a good global fit to *in vivo* behavior. In a model of limited size, it is unlikely that both requirements can be fulfilled at the same time.

6.3.3 Model merging

As more and more models become available, it is a tempting idea to build cell models by merging preexisting models of subsystems. As the models can overlap in their elements (e.g. substances or reactions described), elements from different models have to be compared to each other, as shown in Figure 7.

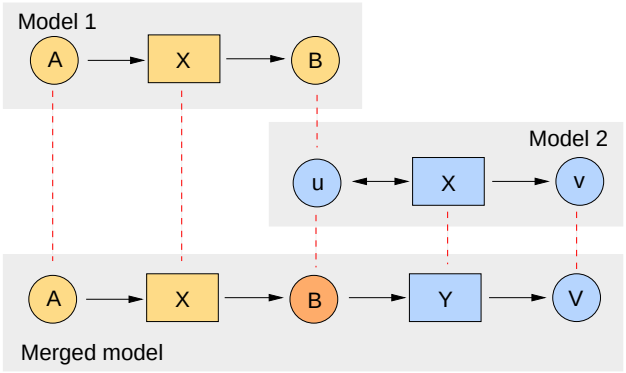


Figure 7: Merging of models. Two models (top and center) are merged to a single model (bottom) containing all model elements. Symbols represent model elements, for instance, substances and reactions. For merging, model elements are aligned (red dashed lines) according to their biological meaning (not shown). A simple name comparison would be unreliable because models can use different naming conventions.

Model merging is based on the reductionist assumption that mechanistic models remain correct in different environments. However, both manual and computer-assisted merging (e.g., with the tool SemanticSBML) poses various kinds of challenges: (i) Model elements (variables, parameters, chemical reactions) have to be compared according to their biological meaning, which requires a clear description by (possibly computer-readable) annotations (e.g. MIRIAM-compliant RDF annotations. (ii) Units must be compared and unified. (iii) Explicit conflicts between the models - e.g. different kinetics for the same reaction - have to be detected and resolved. (iv) Implicit conflicts may arise if the input models make contradicting assumptions or obey contradicting constraints (e.g. thermodynamic relationships between kinetic parameters). (v) If the model parameters have been determined by global fits, they possibly need to be refitted in the merged model. Some of these difficulties can be avoided if submodels are already designed with a common nomenclature and modeling framework. Model merging is greatly facilitated by standardization efforts for experiments and model building.