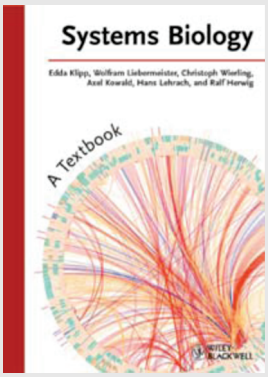



**Systems Biology: A Textbook Answers to Problems**

**Contents**


| | | |
|---|---|---|
| 1 | Introduction | 3 |
| 2 | Modeling of Biochemical Systems | 3 |
| | Answers to Problems | 3 |
| 3 | Specific Biochemical Systems | 10 |
| | Answers to Problems | 10 |
| 4 | Model Fitting | 13 |
| | Answers to Problems | 13 |
| 5 | Analysis of High-Throughput Data | 18 |
| | Answers to Problems | 18 |
| 6 | Gene Expression Models | 20 |
| | Answers to Problems | 20 |
| 7 | Stochastic Systems and Variability | 24 |
| 8 | Network Structures, Dynamics, and Function | 31 |
| 9 | Optimality and Evolution | 34 |
| | Answers to Problems | 34 |
| 10 | Cell Biology | 38 |
| | Answers to Problems | 38 |


*Systems Biology: A Textbook*. Edda Klipp, Wolfram Liebermeister, Christoph Wierling, Axel Kowald, Hans Lehrach, and Ralf Herwig

ISBN: 978-3-527-31874-2

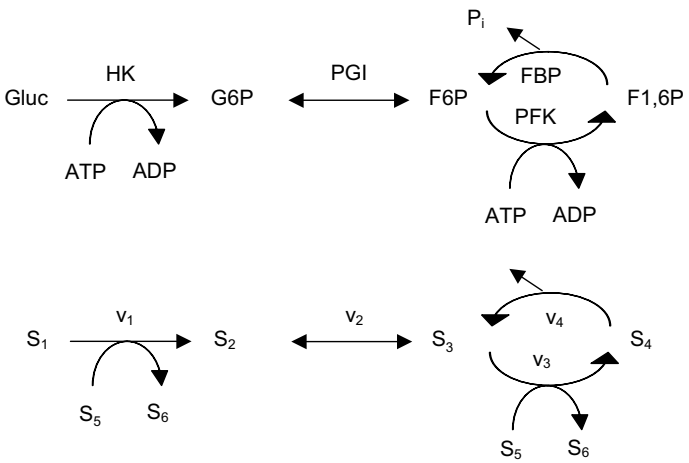
11 Experimental Techniques in Molecular Biology . . . . . 40
Answers to Problems . . . . . 40
12 Mathematics . . . . . 44
13 Statistics . . . . . 44
14 Stochastic Processes . . . . . 44
15 Control of Linear Systems . . . . . 44
16 Databases . . . . . 44
Answers to Problems . . . . . 44
17 Modeling Tools . . . . . 46
Answers to Problems . . . . . 46

## 1 Introduction

## 2 Modeling of Biochemical Systems

Answers to Problems

**Problem 1**
**Problem 1a**




For drawing the network you may either use the biological names or numbered abbreviations. The second version simplifies the mathematical analysis. The ODE system reads:

$$\dot{S}_1 = -\nu_1$$
$$\dot{S}_2 = \nu_1 - \nu_2$$
$$\dot{S}_3 = \nu_2 - \nu_3 + \nu_4$$
$$\dot{S}_4 = \nu_3 - \nu_4$$
$$\dot{S}_5 = -\nu_1 - \nu_3$$
$$\dot{S}_6 = \nu_1 + \nu_3$$

**Problem 1b**

The stoichiometric matrix reads

$$N = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 1 \\ 0 & 0 & 1 & -1 \\ -1 & 0 & -1 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix}.$$

The rank of $\boldsymbol{N}$ is 4. It has 6 rows and 4 columns.

**Problem 1c**

Since the number of columns (4) is equal to the rank of $\boldsymbol{N}$, we find no solution $\boldsymbol{K}$ for the equation $\boldsymbol{N} \cdot \boldsymbol{K} = 0$. Hence, this system has no steady state, except of the trivial steady state where all fluxes vanishes.

Two linear independent solutions $\boldsymbol{G}$ for the equation $\boldsymbol{G} \cdot \boldsymbol{N} = 0$ are $(0, 0, 0, 0, 1, 1)$ and $(1, 1, 1, 1, 0, 0)$. This means that we find two conservation relations for the given reaction system: $ATP + ADP = const.$ and $S_1 + S_2 + S_3 + S_4 = const.$

Please keep in mind that in reality the metabolites of glycolysis are involved in further reactions, which may violate these conservation relations.

**Problem 1d**

The reaction system in Example 2.6 has only three reactions since reaction FBP was neglected, but also six substrates. This implies that Example 2.6 shows a nontrivial steady-state flux and an additional conservation relation. Note: different model formulations can imply different analysis results.

**Problem 2**

**Problem 2a**

N1:

$$\frac{\mathrm{d}}{\mathrm{d}t} S_1 = \frac{\mathrm{d}}{\mathrm{d}t} S_2 = \frac{\mathrm{d}}{\mathrm{d}t} S_3 = -\frac{\mathrm{d}}{\mathrm{d}t} S_4 = -2\frac{\mathrm{d}}{\mathrm{d}t} S_5 = \nu_1$$

N2:

$$\frac{\mathrm{d}}{\mathrm{d}t}S_1 = \nu_1 - \nu_2$$
$$\frac{\mathrm{d}}{\mathrm{d}t}S_2 = \nu_2 - \nu_3$$
$$\frac{\mathrm{d}}{\mathrm{d}t}S_3 = \nu_3 - \nu_4$$
$$\frac{\mathrm{d}}{\mathrm{d}t}S_4 = \nu_4 - \nu_5$$

N3:

$$\frac{\mathrm{d}}{\mathrm{d}t}S_1 = \nu_1 - \nu_2 - \nu_3$$

N4:

$$\frac{\mathrm{d}}{\mathrm{d}t}S_1 = \nu_1 - \nu_2 - \nu_4$$
$$\frac{\mathrm{d}}{\mathrm{d}t}S_2 = 2\nu_2 - \nu_3$$
$$\frac{\mathrm{d}}{\mathrm{d}t}S_3 = \nu_4$$

N5:

$$\frac{\mathrm{d}}{\mathrm{d}t}S_1 = \nu_1 - \nu_2 - \nu_3$$
$$\frac{\mathrm{d}}{\mathrm{d}t}S_2 = -\nu_2 + \nu_3$$
$$\frac{\mathrm{d}}{\mathrm{d}t}S_3 = \nu_2 - \nu_3$$

N6:

$$\frac{\mathrm{d}}{\mathrm{d}t}S_1 = \nu_1 - \nu_2$$
$$\frac{\mathrm{d}}{\mathrm{d}t}S_2 = \nu_4 - \nu_3$$
$$\frac{\mathrm{d}}{\mathrm{d}t}S_3 = \nu_3 - \nu_4$$
$$\frac{\mathrm{d}}{\mathrm{d}t}S_4 = \nu_5$$

**Problem 2b**

Ranks: N1 – 1, N2 – 4, N3 – 1, N4 – 3, N5 – 2, N6 – 3
Independent, nonzero steady-state fluxes:
N1 – none,

$$\text{N2}: \boldsymbol{K} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

$$\text{N3}: \boldsymbol{K} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\text{N4}: \boldsymbol{K} = \begin{pmatrix} 1 \\ 1 \\ 2 \\ 0 \end{pmatrix}$$

$$\text{N5}: \boldsymbol{K} = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}$$

$$\text{N6}: \boldsymbol{K} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$$

Conservation relations

N1 – has four independent conservation relations, for example $\boldsymbol{G} = \begin{pmatrix} 2 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \end{pmatrix}$. The most intuitive solution is the linear combination of the rows of $\boldsymbol{G}$, i.e., $S_1 + S_2 + S_3 + S_4 + S_5 = const.$

N2 – none

N3 – none

N4 – none

N5 – $\boldsymbol{G} = (0 \quad 1 \quad 1)$ or $S_2 + S_3 = const.$

N6 – $\boldsymbol{G} = (0 \quad 1 \quad 1 \quad 0)$ or $S_2 + S_3 = const.$

N1 has only the trivial steady state ($v_1 = 0$) and N6 has only the trivial steady state for reaction $v_5$.

**Problem 3**

Elementary flux modes

N3: $\{v_1, v_2\},\ \{v_1, v_3\}$

N4: $\{v_1, v_2, v_3\}$

**Problem 4**

There are two steady-state solutions for $S_1$: $S_1^{(1)} = -0.270778$ and $S_1^{(2)} = 0.0422064$. Since biological concentrations must be nonnegative, we neglect the negative solution. The flux control coefficients for the steady state with the positive concentration of $S_1$ read:

$$C^J = \begin{pmatrix} 1 & 0 & 0 \\ 0.956 & 0.5 & -0.456 \\ 1.048 & -0.548 & 0.5 \end{pmatrix}$$

**Problem 5**

The equation system reads

$$\frac{\mathrm{d}}{\mathrm{d}t}A = -v_1 + v_3 = -k_1 \cdot A + k_3 \cdot C$$
$$\frac{\mathrm{d}}{\mathrm{d}t}B = v_1 - v_2 = k_1 \cdot A - k_2 \cdot B$$
$$\frac{\mathrm{d}}{\mathrm{d}t}C = v_2 - v_3 = k_2 \cdot B - k_3 \cdot C$$

**Problem 5a**

The Jacobian reads

$$J = \begin{pmatrix} -k_1 & 0 & k_3 \\ k_1 & -k_2 & 0 \\ 0 & k_2 & -k_3 \end{pmatrix}$$

**Problem 5b**

The eigenvalues are

$$\lambda_1 = \frac{1}{2}\left(-5 + i \cdot \sqrt{7}\right), \lambda_2 = -\frac{1}{2}\left(5 + i \cdot \sqrt{7}\right), \text{and } \lambda_3 = 0.$$

The respective eigenvectors are

$$b^{(1)} = \begin{pmatrix} -\frac{3}{2} - \frac{1}{4}\left(-5 + i \cdot \sqrt{7}\right) \\ \frac{1}{2} + \frac{1}{4}\left(-5 + i \cdot \sqrt{7}\right) \\ 1 \end{pmatrix}, b^{(2)} = \begin{pmatrix} -\frac{3}{2} + \frac{1}{4}\left(5 + i \cdot \sqrt{7}\right) \\ \frac{1}{2} - \frac{1}{4}\left(5 + i \cdot \sqrt{7}\right) \\ 1 \end{pmatrix}, \text{and}$$

$$b^{(3)} = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}.$$

**Problem 5c**

The general solution has the form $\boldsymbol{x}(t) = \sum_{i=1}^{n} c_i \boldsymbol{b}^{(i)} e^{\lambda_i t}$ and reads here

$$\begin{pmatrix} A(t) \\ B(t) \\ C(t) \end{pmatrix} = c_1 \cdot \begin{pmatrix} -\frac{3}{2} - \frac{1}{4}\left(-5 + i\cdot\sqrt{7}\right) \\ \frac{1}{2} + \frac{1}{4}\left(-5 + i\cdot\sqrt{7}\right) \\ 1 \end{pmatrix} \cdot e^{\frac{1}{2}\left(-5+i\cdot\sqrt{7}\right)\cdot t}$$

$$+ c_2 \cdot \begin{pmatrix} -\frac{3}{2} + \frac{1}{4}\left(5 + i\cdot\sqrt{7}\right) \\ \frac{1}{2} - \frac{1}{4}\left(5 + i\cdot\sqrt{7}\right) \\ 1 \end{pmatrix} \cdot e^{-\frac{1}{2}\left(5+i\cdot\sqrt{7}\right)\cdot t} + c_3 \cdot \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}$$

**Problem 5d**

For the initial conditions $A(0) = 1, B(0) = 1, C(0) = 0$, we obtain

$$c_1 = -\frac{1}{2},\ c_2 = -\frac{1}{2},\ c_3 = \frac{1}{2}$$

and hence:

$$\begin{pmatrix} A(t) \\ B(t) \\ C(t) \end{pmatrix} = -\frac{1}{2} \cdot \begin{pmatrix} -\frac{3}{2} - \frac{1}{4}\left(-5 + i\cdot\sqrt{7}\right) \\ \frac{1}{2} + \frac{1}{4}\left(-5 + i\cdot\sqrt{7}\right) \\ 1 \end{pmatrix} \cdot e^{\frac{1}{2}\left(-5+i\cdot\sqrt{7}\right)\cdot t}$$

$$-\frac{1}{2} \cdot \begin{pmatrix} -\frac{3}{2} + \frac{1}{4}\left(5 + i\cdot\sqrt{7}\right) \\ \frac{1}{2} - \frac{1}{4}\left(5 + i\cdot\sqrt{7}\right) \\ 1 \end{pmatrix} \cdot e^{-\frac{1}{2}\left(5+i\cdot\sqrt{7}\right)\cdot t} + \frac{1}{2} \cdot \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}$$

Application of Euler's formula ($e^{i\varphi} = \cos\varphi + i\cdot\sin\varphi$) yields

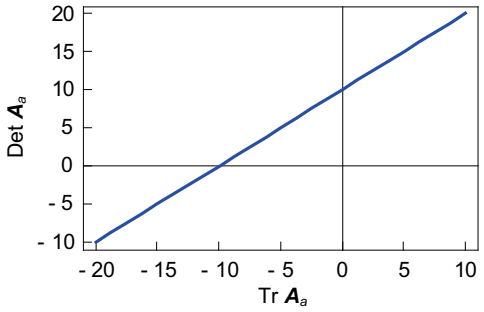
$$\begin{pmatrix} A(t) \\ B(t) \\ C(t) \end{pmatrix} = \frac{1}{14} \cdot e^{-\frac{5}{2}t} \cdot \begin{pmatrix} 7e^{\frac{5}{2}t} + 7\cos\left(\frac{\sqrt{7}}{2}t\right) - 3\sqrt{7}\sin\left(\frac{\sqrt{7}}{2}t\right) \\ 7e^{\frac{5}{2}t} + 7\cos\left(\frac{\sqrt{7}}{2}t\right) + 5\sqrt{7}\sin\left(\frac{\sqrt{7}}{2}t\right) \\ 14e^{\frac{5}{2}t} - 14\cos\left(\frac{\sqrt{7}}{2}t\right) - 2\sqrt{7}\sin\left(\frac{\sqrt{7}}{2}t\right) \end{pmatrix}$$

**Problem 6**

We get Tr $\boldsymbol{A}_a = a$ and Det $\boldsymbol{A}_a = a$.

**Problem 6a**

The plot parametric plot of Det $\boldsymbol{A}_a$ versus Tr $\boldsymbol{A}_a$ looks as follows.




**Problem 6b**

The stability and character of the steady state changes with $a$:

| | |
|---|---|
| $a \leq -10$ | saddle point |
| $-10 < a \leq 2 - 2\sqrt{11}$ | stable node |
| $2 - 2\sqrt{11} < a < 0$ | stable focus |
| $0 < a < 2 + 2\sqrt{11}$ | unstable focus |
| $2 - 2\sqrt{11} \leq a$ | unstable node |

**Problem 7**

An important part of systems biology is data integration. Data used in systems biology is very heterogeneous, e.g., experimental data is coming from different experimental platforms or pathway data differs in the kind of information (e.g., protein–protein interactions, description of the substrates, and products of a reaction, or a detailed kinetic description of a reaction). The definition and standards and its use by systems biology tools is therefore important for the data integration and the reuse of existing data (e.g., quantitative models of biological systems).

## 3 Specific Biochemical Systems

Answers to Problems

**Problem 1**

Problem 1 actually refers to the model of glycolysis synthesis and not of threonine synthesis. The matrix of flux control coefficients (normalized) for the toy model of glycolysis (Section 3.1.2) reads

$$C^J = \begin{pmatrix} 0.75 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.75 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.75 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.75 & 0.5 & 0.5 & 0.5 & 0.648 & 0.352 & 0.5 \\ 0.75 & 0.5 & 0.5 & 0.5 & 0.75 & 0.25 & 0.5 \\ 0.75 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.75 & 0.5 & 0.5 & 0.5 & 0.602 & 0.398 & 0.5 \end{pmatrix}$$

**Problem 2**

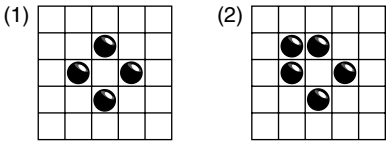
We find a signaling time of $\tau_{Ras_{GTP}} \cong 13.966$ and a signal duration of $\vartheta_{Ras_{GTP}} \cong 15.626$.

**Problem 3**

In the absence of phosphatases, phosphorylated kinases would accumulate due to basal levels of active kinases. Activated kinases cannot be dephosphorylated.

**Problem 4**

(a) There are many solutions, for instance:




(b) The movement of the patterns is shown below (Animations by Rodrigo Camargo).

Animation 1 Animation 2

**Problem 5**

Insert the solution as an ansatz into the diffusion equation and evaluate the derivatives. The solution is $\lambda(k) = Dk^2$.

**Problem 6**

The steady-state condition reads

$$D\nabla^2 s^{st}(x) = \kappa(s^{st}(x))^2. \tag{1}$$

The ansatz

$$s^{st}(x) = a/(x+b)^2 \tag{2}$$

leads to

$$\nabla s^{st}(x) = \frac{-2a}{(x+b)^3} \tag{3}$$

$$\nabla^2 s^{st}(x) = \frac{6a}{(x+b)^4} \tag{4}$$

$$(s^{st}(x))^2 = \frac{a^2}{(x+b)^4}. \tag{5}$$

Inserting Eqs (4) and (5) into Eq. (1) yields $a = 6D/\kappa$, and with the boundary condition $s^{st}(0) = s_0$, the ansatz (2) yields $b = \sqrt{6D/(\kappa s_o)}$. Alternatively, using (3), we can express $b$ by the production rate, i.e. the flux at $x = 0$,

$$j(0) = -D\nabla s^{st}(0) = \frac{2Da}{b^3}$$

$$b = \left(\frac{2Da}{j(0)}\right)^{1/3}.$$

**Problem 7**

When stripe formation is hard-wired, the pattern is reproducible and inheritable, so its details can be optimized by mutation and selection. However, a genetic program for each single stripe would require many additional morphogens and complicated genetic regulation, which would only evolve if there was a strong selection pressure on the exact shape of the pattern. Spontaneously forming stripes (like the zebra patterns) tend to show individual-specific irregularities, which may also be an advantage, e.g. if the pattern serves for camouflage.

**Problem 8**

The cell cycle is divided into the interphase, which is the period between two subsequent cell divisions, and the M phase, during which one cell separates into two. The interphase can be subdivided into $G_1$, S, and $G_2$ phase. A newborn cell

begins in $G_1$ phase where it grows to a certain size, before entering the S phase. During S phase (synthesis phase) the DNA is replicated. When finished the cell enters $G_2$ phase before starting with the nuclear division (mitosis) and subsequent cytoplasmatic division (cytokinesis) during M phase. The cell cycle has three major control points: (1) The "restriction point" (also known as "Start" in yeast) at the end of $G_1$ phase. At this checkpoint the cell determines whether the environment is favorable for a new cell cycle. When passed the cell enters S phase. (2) Checkpoint between $G_2$ phase and M phase. At this checkpoint the cell determines whether the DNA was successfully replicated. (3) Metaphase-to-anaphase transition checkpoint. At this checkpoint, the cell determines whether all chromosomes are attached to the spindle. When passed the cell cycle proceeds with the segregation of the chromosomes.

**Problem 9**

The aggregation of Bax or Bak proteins at the mitochondrial membrane can result in an efflux of cytochrome c and other molecules from the mitochondrial intermembrane space into the cytosol that subsequently leads to the formation of the apoptosome and finally to apoptosis. The aggregation of, e.g., Bax at the mitochondrial membrane is usually inhibited by other molecules, like Bcl2. Bcl2 can also bind to tBid, a protein that is formed by the initiator caspases of the extrinsic pathway. When Bcl2 is inhibited by tBid, it cannot block the aggregation of Bax or Bak anymore and this can results in an activation of the intrinsic apoptotic pathway.

**Problem 10**

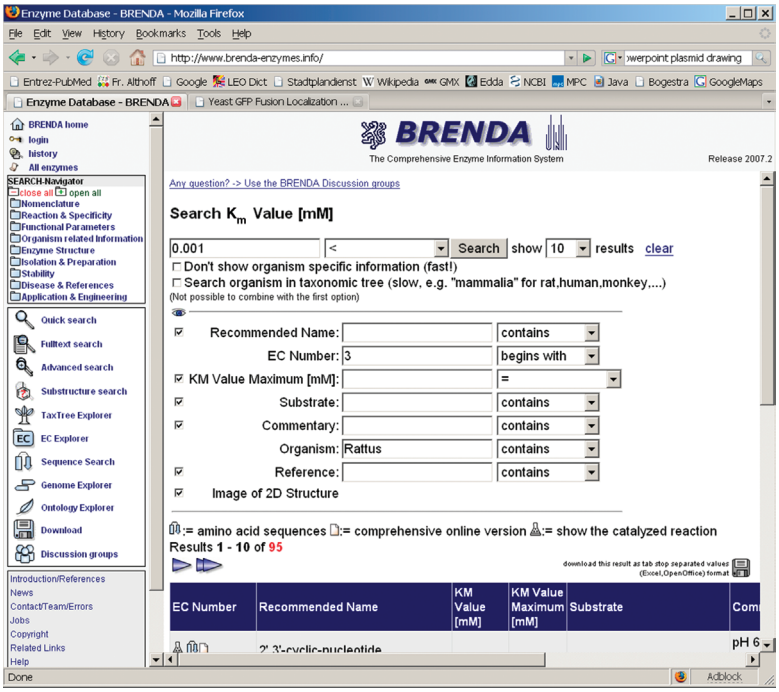
Once a model is established that describes the system of interest (e.g., apoptosis) in sufficient detail, it can subsequently be used for the identification of potential drug targets by the simulation of the inhibitory effects of a potential drug. For example, this can be done by the introduction of a hypothetical drug that can bind a specific model component and by this result in a changed concentration of the active model component.

## 4 Model Fitting

Answers to Problems

### Problem 1

There are 95 enzymes in BRENDA that fulfill the required criteria. Here is a screenshot that shows how to perform the search in BRENDA.




### Problem 2

DNA microarrays are used to measure the amount of large numbers of mRNAs in the cell at a certain time point. These values are taken as indicator for the actual protein concentrations. However, mRNA processing and degradation as well as the

translation process add a large uncertainty to this correlation. The GFP technique avoids this problem because it measures directly the amount of produced protein. Furthermore, GFP measurements can be taken in very short time intervals, and using the appropriate equipment measurements on single cells are possible. Thus, the other main advantage is the high temporal and spatial resolution that the GFP approach can provide.

**Problem 3**

Entering "mitochondrial DNA polymerase" into the quick search field at http://www.yeastgenome.org results in one hit, telling us that the associated gene name is "MIP1". If we then use this gene name for the quick search field at http://yeastGFP.ucsf.edu, we find out that around 377 molecules of the catalytic subunit of the mitochondrial DNA polymerase exist in a single yeast cell.

**Problem 4**

With the assumptions made, maximum-likelihood estimation is equivalent to minimizing the sum of squared residuals (SSR). The SSR reads

$$\begin{aligned} R(\theta) &= \sum_m \left(\theta_1 t_m + \theta_2 - y_m\right)^2 = ||\theta_1 \boldsymbol{t} + \theta_2\, 1 - \boldsymbol{y}||_2 \\ &= ||\boldsymbol{A}\theta - \boldsymbol{y}||_2 = \theta^{\mathrm{T}} \boldsymbol{A}^{\mathrm{T}} \boldsymbol{A}\theta - 2\theta^{\mathrm{T}} \boldsymbol{A}^{\mathrm{T}} \boldsymbol{y} + \boldsymbol{y}^{\mathrm{T}} \boldsymbol{y} \end{aligned}$$

with the vector $\theta = (\theta_1, \theta_2)^{\mathrm{T}}$ and the matrix $\boldsymbol{A} = (t, 1)$ containing the vectors $t$ and $1 = (1, 1, \ldots)^{\mathrm{T}}$ as columns. Minimization of $R(\theta)$ leads to

$$\begin{aligned} 0 &= \nabla_\theta R = 2\boldsymbol{A}^{\mathrm{T}} \boldsymbol{A}\theta - 2\boldsymbol{A}^{\mathrm{T}} \boldsymbol{y} \\ \Rightarrow \theta &= \left(\boldsymbol{A}^{\mathrm{T}} \boldsymbol{A}\right)^{-1} \boldsymbol{A}^{\mathrm{T}} \boldsymbol{y}. \end{aligned}$$

**Problem 5**

(a) The different sample elements $x^{(m)}$ can be seen as independent random variables with mean $\langle X \rangle$ and variance $\mathrm{var}(X)$. Mean and variance of independent random variables are additive, so $\sum_m x_m$ has variance $n\langle X \rangle$ and variance $n\,\mathrm{var}(X)$, respectively. In the estimator $\bar{x}$, the sum is divided by $n$, so we obtain $\langle \bar{x} \rangle = \langle x \rangle$ and $\mathrm{var}(\bar{x}) = 1/n\,\mathrm{var}(X)$ (because the variance scales with the square of the prefactor $1/n$). (d) By computing the empirical mean from independent random samples $(x^{(1)}, \ldots, x^{(n)})$, we effectively draw from the true distribution of $\bar{x}$. However, a finite number of such samples will not suffice to determine exactly the true mean and variance. In bootstrapping, we resample from a given set of data. The distribution of $\bar{x}$ obtained from bootstrap sampling will be centered around the empirical mean of this data set rather than around the true expected value $\langle X \rangle$.

**Problem 6**

(a) With exponentially distributed random errors, the probability to observe a data set $\{(t_1, y_1), (t_2, y_2), \ldots\}$ reads

$$L(\theta|\boldsymbol{y}) \sim \prod_m \exp\left(-\frac{|(y_m - x_m(\theta)|}{a}\right)$$

so the logarithmic likelihood is given by

$$\ln L(\theta|\boldsymbol{y}) = -\frac{1}{a}\sum_m |(y_m - x_m(\theta)| + \text{const.}$$

Instead of maximizing the SSR, we can minimize the 1-norm

$$||\boldsymbol{y} - \boldsymbol{x}(\theta)||_1 = \sum_m |y_m - x_m(\theta)|,$$

that is, the sum of absolute values of the residuals. (b) In comparison to the SSR (which is based on the 2-norm $||\boldsymbol{y} - \boldsymbol{x}(\theta)||_2 = \sum_m (y_m - x_m(\theta))^2$), estimation using the 1-norm will put less weight on points that deviate strongly from the regression curve, so the estimation will be less sensitive to outliers.

**Problem 7**

The maximum-likelihood estimator is defined by a global maximum point of the likelihood and other local maxima do not play a role. If several parameter sets yield the same maximal likelihood value, the model is not identifiable. In Bayesian parameter estimation, on the other hand, broad local maxima of the posterior density may be more important than a narrow global one. Let us consider a local maximum point surrounded by a hill in the posterior landscape $p(\theta|\boldsymbol{y})$. The hill represents a range of similar parameter sets, which together have a posterior probability

$$P_V = \int_{\theta \in V} p(\theta|\boldsymbol{y}) \mathrm{d}\theta$$

where $V$ is the volume in parameter space occupied by the hill. This probability does not only depend on the height of the hill, but also on its width in parameter space. Therefore, a lower, but broad local maximum can represent a more probable ensemble of parameter sets than a higher, but narrow global maximum. This fact is acknowledged in Bayesian estimation.

**Problem 8**

With the equilibrium relation $K_{\text{eq}} = c_{\text{bound}}/c_{\text{free}}$ and the conservation relation $c_{\text{tot}} = c_{\text{free}} + c_{\text{tot}}$, we can solve for

$$\begin{aligned} c_{\text{free}} &= c_{\text{tot}}/(1 + K_{\text{eq}}) \\ c_{\text{bound}} &= c_{\text{tot}} K_{\text{eq}}/(1 + K_{\text{eq}}). \end{aligned}$$

The stoichiometric coefficients for the free concentrations $c_{\text{free}}$ (from the old model) can be used for the total concentrations $c_{\text{tot}}$ (in the new model). Within the kinetic laws, $c_{\text{free}}$ has to be replaced by $c_{\text{tot}}/(1 + K_{\text{eq}})$. Formally, this is equivalent to a rescaling of some kinetic parameters (e.g. Michaelis constants) in the kinetic laws by a factor of $1 + K_{\text{eq}}$.

**Problem 9**

(a) One way to interpret the statement is as follows: even if the behavior of several elements (i.e., their internal dynamics and their potential response to external influences) is known, the behavior of the coupled system is not obvious: although it may be predictable in principle, we would maybe not be able to guess it. This difficulty can be partially overcome by the use of mathematical models and computer simulations.

(b) An important task in systems biology is to pinpoint the relevant elements of a system and to understand which global behavior follows from their interactions. It is usually acknowledged that the elements of a system are systems themselves, but for simplicity or lack of knowledge, their inner structure is not resolved in the model. This attitude combines holism and reductionism. The pure reductionist approach - probing the parts under conditions where they are more or less uncoupled - would provide detailed information about their properties, but it is limited to somewhat artificial, non-physiological situations. A pure holistic approach tests the parts as they are embedded in the living system; such data will reflect a realistic, natural situation, but it is much harder to obtain detailed, high quality data, and it is also much harder to analyze them because their interpretation requires reliable models of the cell. Therefore, the interpretation may be biased towards the mental models that we assumed in first place.

**Problem 10**

For bacteria, a comprehensive model would comprise thousands of genes, chemical reactions, and metabolites (for numbers in a current *E. coli* model, see section 8.1 in the book). Each gene corresponds to at least one mRNA and one protein species. Considering individual sorts of glycoproteins would add tens of thousands of variables. In eukaryotes, the number of genes is on the order of 6000 (for yeast) and 30000 (for mammals). When modelling alternative splicing and other sorts of RNA, we need additional mRNA species. If we consider organelles, ubiquitous substances including most metabolites possibly have to be described by variables for individual compartments. In a particle-based model, we consider the positions of all relevant molecules. With a protein concentration in the range of mM, we would obtain about 1000 copies of a protein in a bacterium, i.e. 3000 degrees of freedom for their positions in the cell. For the atoms inside these proteins, the number of degrees of freedom would be thousands of times higher (with typical protein weights on the order of 50 kD, corresponding to the weight of 50000 hydrogen atoms).

**Problem 11**

Obviously, a biochemical model cannot describe a system in all microscopic details (at atomic resolution), all physical aspects (e.g. quantum mechanical effects), and under all conditions (a moth being burned in a candle light). Therefore, one should require a weaker form of correctness, e.g. “agreement with general physical laws, biological knowledge about the system, and observations made in the system”. This

kind of correctness can indeed be achieved, but it may not hold any more once new data become available. A helpful guideline is keeping the following questions in mind: for what purposes can/should a model be used? What models will actually be reused by other people?

**Problem 12**

In order to determine an individual parameter, one should, as a rule of thumb, measure variables that (i) respond strongly to this parameter and (ii) show little correlation with variables that have been measured before. A computational method for optimal experimental design is as follows: infer a parameter distribution (e.g. a posterior distribution) from the current model fit, draw parameter sets from this distribution, and simulate the future experiment with these parameters. Then apply the planned statistical analysis to the artificial data and compare the resulting estimators to the "true" sampled parameters. This approach will indicate which kinds of experiments can provide, on average, the most useful information.

**Problem 13**

The punishment terms for the different selection criteria read:

| | Criterion | A | B | C |
|---|---|---|---|---|
| AIC | $2k$ | 4 | 6 | 8 |
| AICc | $2k + \frac{2k(k+1)}{n-k-1}$ | $40/7 \approx 5.71$ | 10 | 16 |
| BIC | $k \log n$ | $2\ln 10 \approx 4.61$ | $3\ln 10 \approx 6.91$ | $7.16\ln 10 \approx 9.21$ |

By adding these terms to the log-likelihood, one obtains the selection criteria

| Criterion | A | B | C |
|---|---|---|---|
| AIC | 14.0 | 11.0 | **10.0** |
| AICc | 15.71 | **15** | 18.0 |
| BIC | 14.61 | 11.91 | **11.21** |

where the best solutions are highlighted in red.

Answers to Problems

**Problem 1**

The recursive formula that computes the possible combinations that result in a value $z$ of $T$ can be described as (example in *C* programming language):

```
/** function calculates number of combinations that the
Wilcoxon rank-sum **/
/** test statistic gets a value of z if the treatment series is
of size **/
/** n and if the control series is of size m **/
int w_combi(int z, int n, int m)
{
        /** if sum is beyond possible bounds **/
        if(z < n*(n+1)/2 || z > (m+n)*(m+n+1)/2-m*(m+1)/2)
        return(0);
        /** if we have the rank of only one datum **/
        else if(n == 1 && z < m+2)
        return(1);
        /** if we have no control datum **/
        else if(m == 0 && z == n*(n+1)/2)
        return(1);
        else
        return(w_combi(z-(m+n),n-1,m)+w_combi(z,n,m-1));
}
```

In the next step we use the fact that the Wilcoxon distribution is symmetric around its expectation, $E(T) = n(n+m+1)/2$. We compute the lower and upper boundaries of possible values of T that are more extreme than the observed value and sum all combinations:

```
/* lower tail */
for(i=min;i<=lower;i++)
        *P += (double)w_combi(i,n,m);
```

```
/* upper tail */
for(i=upper;i<=max;i++)
        *P += (double)w_combi(i,n,m);
```

To derive the final P-value, we divide by the number of all possible values for *T*.

**Problem 2**

The P-value for Student's t-test is 0.963, thus the result is not significant at the 0.05 level. The P-value for Wilcoxon's test is 0.028, thus the result is significant at the 0.05 level. The ratio (group 2 mean divided by group 1 mean) of the values is 1.02, the ratio of the ranks is (188/112) 1.68. Thus, judging significance by values is less successful than judging significance by ranks. This results from the fact that in group 1 there are two outlier values (i.e. 5599 and 14820) which corrupt the ratio of the values but less the ratio of the ranks. Thus, Wilcoxon's test is less sensitive against outlier values. To robustify Student's t-test in that respect we can remove the outliers from the sample.

**Problem 3**

We use the *Cauchy-Schwartz inequality* to show the inequality given by the hint. Then we define $a_i = x_{ni} - x_{mi}$ and show the inequality.

**Problem 4**

This is a practical exercise.

**Problem 5**

Use the formulas
$E(X) = \sum_{i=0}^{n} ip(i)$ and $Var(X) = E(X^2) - E(X)^2$to derive the expectations and variances. These are $E(X) = np$ and $Var(X) = np(1-p)$ for the Binomial distribution and $E(X) = n\frac{K}{N}$ and $Var(X) = \frac{N-n}{N-1} n \frac{K}{N}\left(1 - \frac{K}{N}\right)$ for the Hypergeometric distribution.

Answers to Problems

**Problem 1**

We first compute the ratio

$$Z_1/Z_0 = \frac{\binom{N}{n-1} e^{-(n-1)\beta E_0} e^{-\beta E_1}}{\binom{N}{n} e^{-n\beta E_0}} = \frac{N!}{(n-1)!(N-n+1)!} \frac{n!(N-n)!}{N!} e^{\beta(E_0 - E_1)}$$

$$= \frac{n!(N-n)!}{(n-1)!(N-n+1)!} e^{-\beta\Delta E} = \frac{n}{N-n+1} e^{-\beta\Delta E}$$

where $\Delta E = E_1 - E_0$ Assuming that $N \gg n$, we can approximate the first term by $n/N$, and we obtain

$$Z_1/Z_0 \approx \frac{n}{N} e^{-\beta\Delta E}.$$

Calculating $Z_1/(Z_1 + Z_0)$ is now straightforward.

**Problem 2**

If $\mu$ and $c(t)$ are known, the synthesis rate can be expressed as

$$v(t) = \frac{dc(t)}{dt} + \mu c(t). \tag{6}$$

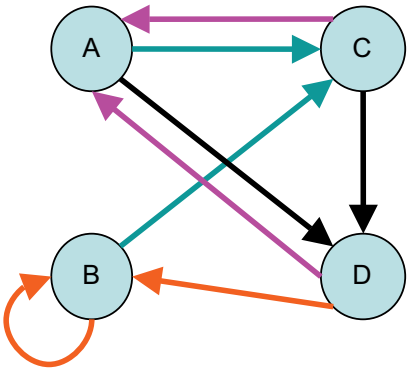
If $\mu$ has an unknown finite value, the relative weighting of both terms is unknown. However, we can still consider the limiting cases "fast turnover" ($\mu \to \infty$), which yields $v(t) = \mu c(t)$ and "no degradation" ($\mu = 0$), which yields $v(t) = \frac{dc(t)}{dt}$. If data are only available for a couple of time points, the time derivative $\frac{dc(t)}{dt}$ has to be estimated. The results, and therefore the estimation of $v(t)$ may become unreliable, especially if the data are noisy.

**Problem 3**

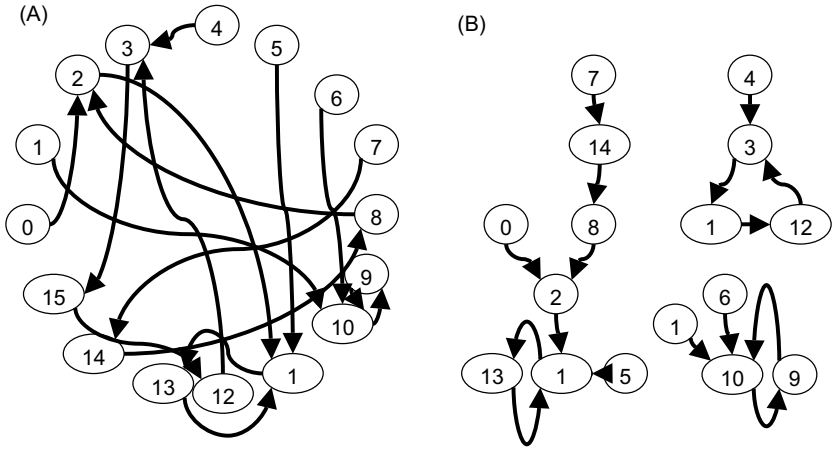
Summary of states and successive states

| | Predecessor state | | | | | Successor state | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| # | A | B | C | D | → | A | B | C | D | # |
| 0 | 0 | 0 | 0 | 0 | → | 0 | 1 | 0 | 0 | 2 |
| 1 | 1 | 0 | 0 | 0 | → | 0 | 1 | 0 | 1 | 10 |
| 2 | 0 | 1 | 0 | 0 | → | 1 | 1 | 0 | 1 | 11 |
| 3 | 1 | 1 | 0 | 0 | → | 1 | 1 | 1 | 1 | 15 |
| 4 | 0 | 0 | 1 | 0 | → | 1 | 1 | 0 | 0 | 3 |
| 5 | 1 | 0 | 1 | 0 | → | 1 | 1 | 0 | 1 | 11 |
| 6 | 0 | 1 | 1 | 0 | → | 0 | 1 | 0 | 1 | 10 |
| 7 | 1 | 1 | 1 | 0 | → | 0 | 1 | 1 | 1 | 14 |
| 8 | 0 | 0 | 0 | 1 | → | 0 | 1 | 0 | 0 | 2 |
| 9 | 1 | 0 | 0 | 1 | → | 0 | 1 | 0 | 1 | 10 |
| 10 | 0 | 1 | 0 | 1 | → | 1 | 0 | 0 | 1 | 9 |
| 11 | 1 | 1 | 0 | 1 | → | 1 | 0 | 1 | 1 | 13 |
| 12 | 0 | 0 | 1 | 1 | → | 1 | 1 | 0 | 0 | 3 |
| 13 | 1 | 0 | 1 | 1 | → | 1 | 1 | 0 | 1 | 11 |
| 14 | 0 | 1 | 1 | 1 | → | 0 | 0 | 0 | 1 | 8 |
| 15 | 1 | 1 | 1 | 1 | → | 0 | 0 | 1 | 1 | 12 |

Sketch of the network




Possible state transitions (for numbering see the above table): (A) states sorted in a circle, (B) states rearranged to make attractors and basins of attraction visible




The system moves to one of the three periodic solutions (attractors) $11 \to 13 \to 11$, $9 \to 10 \to 9$, or $3 \to 15 \to 12 \to 3$.

**Problem 4**

(i) ODE system describing direct mutual inhibition:

$$\frac{\mathrm{d}}{\mathrm{d}t} A_{\mathrm{Gene}} = \frac{A_{\mathrm{Gene}}}{B_{\mathrm{Gene}}} \cdot k_1$$
$$\frac{\mathrm{d}}{\mathrm{d}t} B_{\mathrm{Gene}} = \frac{B_{\mathrm{Gene}}}{A_{\mathrm{Gene}}} \cdot k_2$$

(ii) Mutual inhibition of mRNA formation

$$\frac{\mathrm{d}}{\mathrm{d}t} A_{\mathrm{mRNA}} = \frac{A_{\mathrm{Gene}} \cdot k_1}{1 + B_{\mathrm{mRNA}}/k_{1r}}$$
$$\frac{\mathrm{d}}{\mathrm{d}t} B_{\mathrm{mRNA}} = \frac{B_{\mathrm{Gene}} \cdot k_2}{1 + A_{\mathrm{mRNA}}/k_{2r}}$$

(iii) Mutual inhibition including gene, mRNA, and protein levels.

$$\frac{\mathrm{d}}{\mathrm{d}t} A_{\mathrm{mRNA}} = \frac{A_{\mathrm{Gene}} \cdot k_1}{1 + B_{\mathrm{protein}}/k_{1p}}$$
$$\frac{\mathrm{d}}{\mathrm{d}t} A_{\mathrm{protein}} = A_{\mathrm{mRNA}} \cdot k_{1r}$$
$$\frac{\mathrm{d}}{\mathrm{d}t} B_{\mathrm{mRNA}} = \frac{B_{\mathrm{Gene}} \cdot k_2}{1 + A_{\mathrm{protein}}/k_{2r}}$$
$$\frac{\mathrm{d}}{\mathrm{d}t} B_{\mathrm{protein}} = B_{\mathrm{mRNA}} \cdot k_{2r}$$

Although the above equations are very simple, they are not unique and you may find other ways of descriptions. These equations consider only the production of

compounds, not their degradation. Degradation should be included to prevent unlimited growth.

Including more detail into the analysis refines the description and can make it easier compatible to experimental data (e.g., for mRNA or protein abundance). On the other hand, it increases the number of differential equations to be solved and the number of parameters to be estimated.

## 7 Stochastic Systems and Variability

### Problem 1

Assuming a volume of $1\,\mu m^3 = 10^{-18} m^3$ for a small prokaryotic cell and a concentration of $1 mM = 1\,mol/m^3$, we obtain an amount of $10^{-18}\,N_A \approx 6 \cdot 10^5$ molecules. If we assume a Poisson distribution (e.g. due to random diffusion across the cell membrane), the standard deviation is about $\sqrt{6 \cdot 10^5} \approx 8 \cdot 10^2$, corresponding to a relative deviation of about 0.1 percent. With a smaller concentration of 1 nM, we obtain about $0.6 \pm 0.8$ molecules per cell. In this case, a stochastic modeling approach should be used. For the eukaryote (example *S. cerevisiae*), the cell volume is 125 times biggger, so we obtain about $7.5 \cdot 10^7 \pm 7.7 \cdot 10^3$ (metabolite) and $75 \pm 8.8$ (mRNA) molecules, with much smaller relative deviations.

### Problem 2

The linearized chemical Langevin equation can be written as

$$dx/dt \approx \boldsymbol{A}x + V\boldsymbol{B}\xi$$

where $V$ denotes the compartment volume. For a molecule species with particle number $x_i$, the matrices (for a system containing this single molecule species only) read

$$\boldsymbol{A} = \sum_l n_{il}\tilde{\varepsilon}_i^l, \qquad \boldsymbol{B} = V^{-1/2}\sum_l n_{il}\sqrt{\nu_l}. \tag{7}$$

By solving the Lyapunov equation (7.11), one obtains the variance

$$\boldsymbol{Q} = \mathrm{var}(x_i) = -V\frac{(\sum_l n_{il}\sqrt{\nu_l})^2}{2\sum_l n_{il}\tilde{\varepsilon}_i^l}. \tag{8}$$

### Problem 3

We start with the Langevin equation for $z = (g, x, y)^T$,

$$\frac{dz}{dt} = \boldsymbol{N}\boldsymbol{W}z + \boldsymbol{N}Dg(\boldsymbol{W}z)^{1/2}\xi, \tag{9}$$

with the stoichiometric matrix $\boldsymbol{N}$ and the unscaled elasticity matrix $\boldsymbol{W}$. The matrices $\boldsymbol{N}$ and $\boldsymbol{W}$ and the Jacobian $\boldsymbol{A} = \boldsymbol{N}\boldsymbol{W}$ read

$$\boldsymbol{N} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}, \qquad \boldsymbol{W} = \begin{pmatrix} w_{+\mathrm{x}} & 0 & 0 \\ 0 & w_{-\mathrm{x}} & 0 \\ 0 & w_{+\mathrm{y}} & 0 \\ 0 & 0 & w_{-\mathrm{y}} \end{pmatrix},$$

$$\boldsymbol{A} = \begin{pmatrix} 0 & 0 & 0 \\ w_{+\mathrm{x}} & -w_{-\mathrm{x}} & 0 \\ 0 & w_{+\mathrm{y}} & -w_{-\mathrm{y}} \end{pmatrix}.$$

To compute the average molecule numbers, we disregard the noise term in (9), solve the stationarity condition $\boldsymbol{A}z = 0$ while fixing $g = 1$, and obtain the average amount vector $\langle z \rangle$ and the corresponding propensity vector $\langle \boldsymbol{a} \rangle$:

$$\langle z \rangle = \begin{pmatrix} 1 \\ w_{+x}/w_{-x} \\ (w_{+x}/w_{-x})(w_{+y}/w_{+y}) \end{pmatrix}, \qquad \langle \boldsymbol{a} \rangle = \boldsymbol{W}\langle z \rangle = \begin{pmatrix} w_{+x} \\ w_{+x} \\ w_{+x}w_{+y}/w_{-x} \\ w_{+x}w_{+y}/w_{-x} \end{pmatrix}.$$

To compute the covariance matrix $\boldsymbol{Q} = \mathrm{cov}(z)$, we solve the Lyapunov equation

$$\boldsymbol{A}\boldsymbol{Q} + \boldsymbol{Q}\boldsymbol{A}^{\mathrm{T}} + \hat{B}\hat{B}^{\mathrm{T}} = 0 \tag{10}$$

where

$$\hat{B} = \boldsymbol{N}\,\mathrm{Dg}(\langle \boldsymbol{a} \rangle)^{1/2} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \sqrt{w_{+\mathrm{x}}} & -\sqrt{w_{+\mathrm{x}}} & 0 & 0 \\ 0 & 0 & \sqrt{w_{+\mathrm{x}}w_{+\mathrm{y}}/w_{-\mathrm{x}}} & -\sqrt{w_{+\mathrm{x}}w_{+\mathrm{y}}/w_{-\mathrm{x}}} \end{pmatrix}.$$

As the covariance matrix $\boldsymbol{Q}$ is symmetric and the auxiliary variable $g = 1$ is fixed (no variance or covariances), $\boldsymbol{Q}$ must have the form

$$\boldsymbol{Q} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \alpha & \beta \\ 0 & \beta & \gamma \end{pmatrix}$$

with $\alpha = \mathrm{var}(x)$, $\beta = \mathrm{cov}(x, y)$, $\gamma = \mathrm{var}(y)$. We insert the matrices into the Lyapuov equation (10), omit the vanishing first row and column (corresponding to the auxiliary variable $g$), and obtain

$$0 = \begin{pmatrix} -\alpha w_{-\mathrm{x}} & -\beta w_{-\mathrm{x}} \\ \alpha w_{+\mathrm{y}} - \beta w_{-\mathrm{y}} & \beta w_{+\mathrm{y}} - \gamma w_{-\mathrm{y}} \end{pmatrix} + \begin{pmatrix} -\alpha w_{-\mathrm{x}} & -\beta w_{-\mathrm{x}} \\ \alpha w_{+\mathrm{y}} - \beta w_{-\mathrm{y}} & \beta w_{+\mathrm{y}} - \gamma w_{-\mathrm{y}} \end{pmatrix}^{T} + \begin{pmatrix} 2w_{+\mathrm{x}} & 0 \\ 0 & \dfrac{2w_{+\mathrm{x}}w_{+\mathrm{y}}}{w_{-\mathrm{x}}} \end{pmatrix} \tag{11}$$

$$= \begin{pmatrix} -2\alpha w_{-x} + 2w_{+x} & \alpha w_{+y} - \beta w_{-y} - \beta w_{-x} \\ \alpha w_{+y} - \beta w_{-y} - \beta w_{-x} & 2\beta w_{+y} - 2\gamma w_{-y} + \dfrac{2w_{+x}w_{+y}}{w_{-x}} \end{pmatrix} \tag{12}$$

By solving Eq. (11) for $\alpha$, $\beta$, and $\gamma$, we obtain

$$\alpha = \mathrm{var}(x) = w_{+x}/w_{-x}$$
$$\beta = \mathrm{cov}(x,y) = \frac{w_{+x}w_{+y}}{w_{-x}} \frac{1}{w_{-x} + w_{-y}}$$
$$\gamma = \mathrm{var}(y) = \frac{w_{+x}w_{+y}}{w_{-x}w_{-y}} \left(1 + \frac{w_{+y}}{w_{-x} + w_{-y}}\right)$$

**Problem 4**

(a) The spectral response coefficient matrix for the parameter $\xi$ reads

$$\begin{aligned} \boldsymbol{H}(i\omega) &= \boldsymbol{C}(i\omega\,\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{B} = \begin{pmatrix} \beta + i\omega & 0 \\ -\alpha_2 & \beta + i\omega \end{pmatrix}^{-1} \begin{pmatrix} \alpha_1 \\ 0 \end{pmatrix} \\ &= \frac{1}{(\beta + i\omega)^2} \begin{pmatrix} \beta + i\omega & 0 \\ \alpha_2 & \beta + i\omega \end{pmatrix} \begin{pmatrix} \alpha_1 \\ 0 \end{pmatrix} \\ &= \frac{1}{(\beta + i\omega)^2} \begin{pmatrix} \alpha_1(\beta + i\omega) \\ \alpha_1\alpha_2 \end{pmatrix} = \begin{pmatrix} \alpha_1/(\beta + i\omega) \\ \alpha_1\alpha_2/(\beta + i\omega)^2 \end{pmatrix}. \end{aligned}$$

We obtain the spectral densities for white noise input

$$\Phi_1(\omega) = H_1(i\omega)H_1(i\omega)^\dagger = \frac{\alpha_1}{\beta + i\omega}\frac{\alpha_1}{\beta - i\omega} = \frac{\alpha_1^2}{\beta^2 + \omega^2}$$

$$\Phi_2(\omega) = H_2(i\omega)H_2(i\omega)^\dagger = \frac{\alpha_1\alpha_2}{(\beta + i\omega)^2}\frac{\alpha_1, \alpha_2}{(\beta - i\omega)^2} = \frac{\alpha_1^2\alpha_2^2}{(\beta^2 + \omega^2)^2}.$$

(b) The covariance function for gene 1 can be computed by inverse Fourier transformation

$$C_1(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\alpha_1^2}{\beta^2 + \omega^2} e^{i\omega\tau} d\omega = \frac{\alpha_1^2}{2\beta} e^{-\beta|\tau|}. \tag{13}$$

Alternatively, we can obtain it from the definition of the covariance function, using the pulse response function $K_1 = e^{-\beta t}\alpha_1$ (see Sigal *et al.* (2006) Nature 444, 643-646)

$$
\begin{aligned}
C_1(\tau) &= \langle x_1(t)x_1(t+\tau)\rangle = \left\langle \left( \int_{-\infty}^{t} K_1(t-t')\xi(t')\mathrm{d}t' \right) \left( \int_{-\infty}^{t+\tau} K_1(t+\tau-t'')\xi(t'')\mathrm{d}t'' \right) \right\rangle \\
&= \left\langle \left( \int_{-\infty}^{t} \mathrm{e}^{-\beta(t-t')}\alpha_1\xi(t')\mathrm{d}t' \right) \left( \int_{-\infty}^{t+\tau} \mathrm{e}^{-\beta(t+\tau-t'')}\alpha_1\xi(t'')\mathrm{d}t'' \right) \right\rangle \\
&= \alpha_1^2\mathrm{e}^{-\beta(2t+\tau)} \int_{-\infty}^{t}\int_{-\infty}^{t+\tau} \mathrm{e}^{\beta t'}\mathrm{e}^{\beta t''}\langle\xi(t')\xi(t'')\rangle\mathrm{d}t''\mathrm{d}t' \\
&= \alpha_1^2\mathrm{e}^{-\beta(2t+\tau)} \int_{-\infty}^{t} \mathrm{e}^{2\beta t'}\mathrm{d}t' = \frac{\alpha_1^2}{2\beta}\mathrm{e}^{-\beta\tau}
\end{aligned}
$$

where $\tau$ is assumed to be non-negative. In this calculation, we used the autocorrelation function of white noise, $\langle\xi(t_1)\xi(t_2)\rangle = \delta(t_1-t_2)$, and the definition of Dirac's $\delta$ distribution, $\int_{-\infty}^{\infty} f(t')\delta(t-t')\mathrm{d}t' = f(t)$.

The spectral density $\Phi_2$ of gene 2,

$$
\Phi_2(\omega) = \frac{\alpha_1^2}{\beta^2+\omega^2}\frac{\alpha_2^2}{\beta^2+\omega^2},
$$

is a product of two terms. Each of them has the same form as the spectral density of gene 1. Therefore, the covariance function for gene 2 can be written as a convolution of the inverse Fourier transforms,

$$
C_2(\tau) = \int_{-\infty}^{\infty} \left(\frac{\alpha_1^2}{2\beta}\mathrm{e}^{-\beta|\tau-\tau'|}\right)\left(\frac{\alpha_2^2}{2\beta}\mathrm{e}^{-\beta|\tau'|}\right)\mathrm{d}\tau' = \frac{(\alpha_1\alpha_2)^2}{4\beta^2}\int_{-\infty}^{\infty} \mathrm{e}^{-\beta(|\tau-\tau'|+|\tau'|)}\mathrm{d}\tau'.
$$

For positive time lags $\tau > 0$, the integral can be split into three parts,

$$
\begin{aligned}
\int_{-\infty}^{\infty} \mathrm{e}^{-\beta(|\tau-\tau'|+|\tau'|)}\mathrm{d}\tau' &= \int_{-\infty}^{0} \mathrm{e}^{-\beta(\tau-2\tau')}\mathrm{d}\tau' + \int_{0}^{\tau} \mathrm{e}^{-\beta\tau}\mathrm{d}\tau' + \int_{\tau}^{\infty} \mathrm{e}^{-\beta(2\tau'-\tau)}\mathrm{d}\tau' \\
&= \mathrm{e}^{-\beta\tau}\left[\frac{1}{2\beta}\mathrm{e}^{2\beta\tau'}\right]_{-\infty}^{0} + \tau\mathrm{e}^{-\beta\tau} + \mathrm{e}^{\beta\tau}\left[\frac{-1}{2\beta}\mathrm{e}^{-2\beta\tau'}\right]_{\tau}^{\infty} = \left(\frac{1}{\beta}+\tau\right)\mathrm{e}^{-\beta\tau}.
\end{aligned}
$$

As the covariance function must be symmetric, we obtain

$$
C_2(\tau) = \frac{(\alpha_1\alpha_2)^2}{4\beta^2}\left(\frac{1}{\beta}+|\tau|\right)\mathrm{e}^{-\beta|\tau|}.
$$

We normalize the covariance functions by their values at time lag $\tau = 0$ and obtain the correlation functions

$$
R_1(\tau) = \frac{C_1(\tau)}{C_1(0)} = \mathrm{e}^{-\beta|\tau|}, \quad R_2(\tau) = \frac{C_2(\tau)}{C_2(0)} = \mathrm{e}^{-\beta|\tau|}(1+\beta|\tau|).
$$

**Problem 5**

(a) Let $x$ denote a possible parameter value and $p(x)$ a probability density. The entropy is given by the functional

$$S[p] = -\int_a^b p(x)\log p(x)\mathrm{d}x, \tag{14}$$

which has to be maximized under the normalization condition

$$1 = Z[p] = \int_a^b p(x)\mathrm{d}x.$$

Maximization of (14) implies that for any variation curve $\Delta p(x)$, it must hold that

$$\begin{aligned} 0 &= \frac{\partial}{\partial\alpha}\left(S[p+\alpha\Delta p] + \lambda Z[p+\alpha\Delta p]\right) \\ &= \frac{\partial}{\partial\alpha}\left(-\int_a^b [p(x)+\alpha\Delta p(x)]\log[p(x)+\alpha\Delta p(x)] + \lambda[p(x)+\alpha\Delta p(x)]\mathrm{d}x\right)_{\alpha=0} \end{aligned}$$

where $\lambda$ is a Langrangian multiplier. This yields

$$\begin{aligned} 0 &= \int_a^b \Delta p(x)\log p(x) + p(x)1/p(x)\Delta p(x) + \lambda\Delta p(x)\mathrm{d}x \\ &= \int_a^b \Delta p(x)[\log p(x) + 1 + \lambda]\mathrm{d}x \end{aligned}$$

for any variation $\Delta p(x)$. The condition is only satisfied if the term in brackets vanishes; this implies that the probability density must be constant over the entire interval.

(b) Same proof idea as for (a).

**Problem 6**

The count number $n$ for positive outcomes follows a binomial distribution with probabilities $(n) = \binom{N}{n} q^n (1-q)^{N-n}$, mean value $\langle n\rangle = qN$, and variance $Nq(1-q)$. The relative number $\hat{q} = n/N$ can be used as an estimator for the probability $q$. It has the mean value $\langle\hat{q}\rangle = q$ and variance var $(\hat{q}) = \dfrac{q(1-q)}{N}$, so its standard deviation decreases as $1/\sqrt{N}$.

**Problem 7**

In steady state, there is a single stationary flux through all reactions, which transports a phosphate group from ATP to inorganic phosphate. In particular, the fluxes

$$J_{1'} = k_{1'}(u)[\mathrm{X}\cdot\mathrm{ATP}] \tag{15}$$

$$J_{3'} = k_{3'}[X\cdot Y_P\cdot\mathrm{ATP}] \tag{16}$$

must be equal. Here X · ATP denotes the complex formed by X and ATP. The steady-state concentration of the complex X·$Y_P$·ATP can be computed from its balance equation

$$\frac{\mathrm{d}[\mathrm{X}\cdot \mathrm{Y}_P\cdot \mathrm{ATP}]}{\mathrm{d}t} = k_{+3}[\mathrm{X}\cdot \mathrm{ATP}][\mathrm{Y}_P] - (k_{-3}+k_{3'})[\mathrm{X}\cdot \mathrm{Y}_P\cdot \mathrm{ATP}]. \tag{17}$$

It reads

$$s^{\mathrm{st}}_{X\cdot Y_P\cdot \mathrm{ATP}} = \frac{k_{+3}}{(k_{-3}+k_{3'})}[\mathrm{X}\cdot \mathrm{ATP}][\mathrm{Y}_P]. \tag{18}$$

By inserting this expression into Eq. (16) and equating the result to Eq. (15), we obtain

$$k_{3'}\frac{k_{+3}}{(k_{-3}+k_{3'})}[\mathrm{X}\cdot \mathrm{ATP}][\mathrm{Y}_P] = k_{1'}(u)[\mathrm{X}\cdot \mathrm{ATP}].$$

Thus either $[\mathrm{X}\cdot \mathrm{ATP}] = 0$, i.e. the flux has to vanish, or the output concentration $[\mathrm{Y_P}]$ will read

$$[\mathrm{Y_P}] = \frac{(k_{-3}+k_{3'})}{k_{+3}}\frac{k_{1'}(u)}{k_{3'}}.$$

**Problem 8**

The values $a$, $b$, and $x$ are measured in 1/s, 1/(mM s), and mM, respectively. The behavior of the system must be independent of the choice of physical units; when rescaling the time units by a factor $1/\lambda$, we replace $a \to \lambda\, a, b \to \lambda\, b, x \to x$. Thus, if $x = f(a, b)$ is a solution in the original units, then $x = f(\lambda\, a, \lambda\, b)$, with the same mathematical function $f$, must be a solution as well. This does not only hold for a change of time units, but also for an actual rescaling of time - which in turn is equivalent to an increase of all enzyme activities because each reaction scales proportionally with an enzyme activity. Therefore, $x$ will not be changed if both enzyme concentrations are multiplied by a positive factor $\lambda$ and is therefore precisely robust against coupled relative fluctuations of both enzymes.

**Problem 9**

Compute the steady state of the resulting closed-loop system: from equation (7.45), we obtain

$$\frac{\mathrm{d}z}{\mathrm{d}t} = -(y - y_0) = -k\,u - k'\,z + y_0. \tag{19}$$

For arbitrary, but constant input $u$ and gain $k$, the steady-state condition $\mathrm{d}z/\mathrm{d}t = 0$ yields

$$z^{\mathrm{ss}} = \frac{y_0}{k'} - \frac{k}{k'}\,u. \tag{20}$$

By equating Eq. (19) again to zero (steady state) and inserting Eq. (20), it follows that $y^{ss} = y_0$.

**Problem 10**

If the enzyme concentrations appear as prefactors in the rate laws, then $y_i^*$ scales with $k = 0$, while $t_i^*$ scales with $k = 1$, so we obtain the summation theorems

$$\sum_l C_{il}^{y^*} = 0, \quad \sum_l C_{il}^{t^*} = 1. \tag{21}$$

## 8 Network Structures, Dynamics, and Function

**Problem 1**

(a) The adjacency matrices for the three graphs read

$$A_a = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & 1 & 1 & 1 & \cdot & 1 \\ \cdot & \cdot & 1 & 1 & 1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & 1 & \cdot & 1 \\ \cdot & 1 & \cdot & \cdot & \cdot & \cdot \end{pmatrix} \quad A_b = \begin{pmatrix} \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\ 1 & 1 & 1 & 1 & \cdot & 1 \\ \cdot & 1 & 1 & 1 & 1 & \cdot \\ \cdot & 1 & 1 & \cdot & 1 & \cdot \\ \cdot & \cdot & 1 & 1 & \cdot & 1 \\ \cdot & 1 & \cdot & \cdot & 1 & \cdot \end{pmatrix}$$

$$A_c = \begin{pmatrix} \cdot & \cdot & 1 & \cdot & \cdot & \cdot \\ \cdot & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & \cdot & 1 & \cdot \\ \cdot & 1 & \cdot & \cdot & 1 & 1 \\ \cdot & 1 & 1 & 1 & \cdot & \cdot \\ \cdot & 1 & \cdot & 1 & \cdot & \cdot \end{pmatrix}$$

If self-edges are not counted, the degrees for the six nodes read 1, 4, 3, 3, 3, 2, respectively, for both graph (b) and graph (c). This yields the numbers of potential three-loops, 0, 6, 3, 3, 3, 1 for the six nodes. The actual numbers of three-loops are 0, 1, 2, 2, 1, 0 for graph (b) and 0, 3, 1, 2, 2, 1 for graph (c). The clustering coefficient for node 1 is not defined. The clustering cofficients for the remaining nodes read 1/6, 2/3, 2/3, 1/3, 0 for graph (b) and 1/2, 1/3, 2/3, 2/3, 1 for graph (c).

Graph (a) contains three feed-forward loops, $2 \to 3 \to 4$, $3 \to 5 \to 4$ and $5 \to 3 \to 4$.

The shortest way from node 6 to node 5 contains three edges (via nodes 2 and 3), while the shortest way from node 5 back to node 6 consists of a single edge. Hence, the topological distance is not symmetric and cannot be a distance in the mathematical sense.

**Problem 2**

As the degree $k$ changes by a factor of 2 (from 10 to 20), the corresponding percentage changes by a factor of $2^{-\gamma} \approx 0.22$, so the percentage of nodes with degree $k = 20$ is about 0.2 percent.

**Problem 3**

To test if self-inhibition appears as a network motif, we compare the network to a random graph $G_E(n, m)$ with the same number of nodes ($n$) and edges ($m$). In this background model, the probability to find a self-edge at a specific gene is approximately $q = m/n^2$. Checking every gene for a self-edge corresponds to $n$ (approximately independent) trials, so we expect a number of $nq \pm \sqrt{nq}$ self-edges in total. With $n = 424$ nodes and $m = 519$ edges, this would yield about $1.2 \pm 1.1$ self-inhibitions. The number 42 deviates from the expected value by about 37 standard deviations and is therefore highly significant. This conclusion depends on the choice of the background model. Self-inhibitions could have evolved in this high number by active selection because they can stabilize protein levels and speed up responses.

**Problem 4**

Network motifs are local patterns in a graph that appear significantly more often than in a random graph, which represents the background model. Self-inhibition is a motif in transcription networks. It can stabilise expression levels, which may help to make the network robust against external perturbations and against varying expression of other genes in the cell. In addition, self-inhibition can speed up responses to external stimuli without requiring a fast protein turnover. This allows the cell to adapt rapidly to environmental changes at a relatively low energetic price. The evolution of network motifs can be explained by active selection for network motifs that increase the cell's fitness, or by neutral evolutionary mechanisms like gene duplication. Note, again, that the "network motif" property depends on the definition of the random graph that is used as the background model.

**Problem 5**

The gene groups $X_1$ and $X_2$ show a pulse; $\sigma^K$ is switched on with a delay, leading to another delayed pulse of $X_3$ and a sustained activation of $X_4$.

**Problem 6**

Homology: (i) the skeleton structure of different mammalian species; (ii) Sequence similarity between genes of common evolutionary origin; (iii) Evolutionarily conserved master regulators. Analogy: (i) Eyes of insects and vertebrates; (ii) Evolutionarily unrelated signalling systems in bacteria (two-component systems) and in eukaryotes (MAP kinase cascades); (iii) Self-inhibition of evolutionarily unrelated regulatory proteins.

**Problem 7**

Let us consider three limiting cases: (i) If two gene products can completely compensate for each other, a double deletion will strongly affect the genes' function, while single deletions will have little effect (aggravating epistasis). (ii) If two gene products need to be present to exert their function, then this function will be lost after a single deletion and the second deletion will have little further impact (buffering epistasis). (iii) If there is no functional relation between the genes at all, we may assume that each single deletion decreases the fitness by a certain factor, irrespective of the remaining genetic background. With the usual definitions of epistasis, this would yield an epistasis value of zero. If we just consider these extreme cases, then the epistasis value of two genes would allow to predict their functional relation ("compensation", "cooperativity", or "functional independence"). In reality, genes may also show partial compensation or cooperation, leading to intermediate epistasis values.

**Problem 8**

If an engineered gene circuit shows a predicted dynamics or exerts a predicted function, this indicates that its function is relatively independent of the rest of the cell, e.g., that it is only weakly affected by fluctuations in cell variables like protein production or growth rate. A successful implementation of a gene circuit makes it more likely that other circuits built from the same elements will also exert their predicted function.

**Problem 9**

We collect the vectors $\boldsymbol{y}_\alpha, \boldsymbol{y}_\beta, \ldots$ for all modules in a vector $\boldsymbol{y}$ for the entire system. For this vector, we obtain the response matrix

$$\begin{aligned} \tilde{\boldsymbol{R}}_{\mathrm{p}}^{\mathrm{Y}} &= \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{p}} = \frac{\partial \boldsymbol{s}}{\partial \boldsymbol{p}} + \left.\frac{\partial \boldsymbol{s}}{\partial \boldsymbol{x}_\mu}\right|_{\boldsymbol{x}=\boldsymbol{y}} \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{p}} \\ &= \tilde{\boldsymbol{R}}_{\mathrm{p}}^{\mathrm{S}} + \tilde{\boldsymbol{R}}_{S}^{S} \tilde{\boldsymbol{R}}_{p}^{Y}. \end{aligned} \tag{22}$$

If $\tilde{\boldsymbol{R}}_{\mathrm{S}}^{\mathrm{S}}$ is invertible (which we need to assume), then solving for $\tilde{\boldsymbol{R}}_{\mathrm{p}}^{\mathrm{Y}}$ yields Eq. (8.13).

## 9 Optimality and Evolution

Answers to Problems

**Problem 1**

(a) The linear programming problem for the flux vector $\boldsymbol{\nu}$ reads

$$(0 \quad 0 \quad 1)\boldsymbol{\nu} \stackrel{!}{=} \max$$

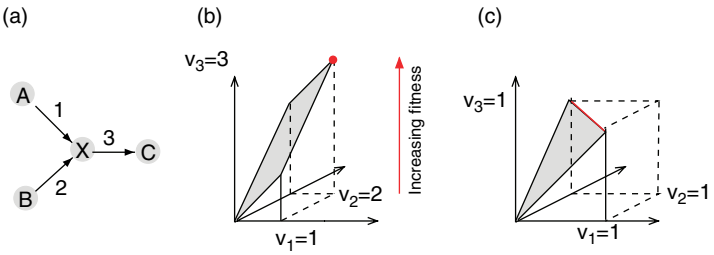
$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \nu \leq \begin{pmatrix} 1 \\ 2 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\boldsymbol{N}\boldsymbol{\nu} = 0.$$

with the stoichiometric matrix

$$\boldsymbol{N} = (1 \quad 1 \quad -1)$$

for the balanced metabolite X. The optimal solution is $\boldsymbol{\nu} = (1 \quad 2 \quad 3)^{\mathrm{T}}$ (red dot in Figure (b)). (b) The constraint $\nu_3 = 1$ already determines the optimal value of the objective function. The optimum can be achieved with different flux distributions $\boldsymbol{\nu} = (\alpha \quad 1-\alpha \quad 1)^{\mathrm{T}}$, where $\alpha$ can only assume values between 0 and 1 (red line in Figure (c)).



**Problem 2**

(a) A forward reaction flux would require differences of the chemical potentials $\mu_A > \mu_B$, $\mu_B > \mu_C$, $\mu_C > \mu_A$, which leads to the contradiction $\mu_A > \mu_A$.

(b) The stoichiometric matrix reads $\boldsymbol{N} = \begin{pmatrix} 1 & -1 & -1 & 0 \\ 0 & 1 & 1 & -1 \end{pmatrix}$. The stationary fluxes can be written as linear combinations of the fluxes $\boldsymbol{u} = (1 \quad 1 \quad 0 \quad 1)^{\mathrm{T}}$ and $\boldsymbol{w} = (1 \quad 0 \quad 1 \quad 1)^{\mathrm{T}}$. As a circular flux between B and C would be thermodynamically unfeasible, fluxes 2 and 3 must have the same sign (or at least one of them has to vanish). This leaves as possibilities $\boldsymbol{v} = \alpha\boldsymbol{u} + \beta\boldsymbol{w}$ where $\alpha$ and $\beta$ have either the same sign or at least one of them is zero.

**Problem 3**

(a) and (b) The energy balances for the (hypothetical) uncoupled reactions and for the coupled reaction with production of $n$ ATP molecules read

| | |
|---|---|
| 2 ADP $\rightarrow$ 2 ATP | 2·49 kJ/mol |
| glucose $\rightarrow$ 2 lactate | $-$ 205 kJ/mol |
| $n$ ADP + glucose $\rightarrow$ $n$ ATP + 2 lactate | 49 $n$ $-$ 205 kJ/mol |

For different values of $n$, we obtain the energy values $-205$ kJ/mol ($n = 0$), $-156$ kJ/mol ($n = 1$), $-107$ kJ/mol ($n = 2$), $-58$ kJ/mol ($n = 3$), $-9$ kJ/mol ($n = 4$), 40 kJ/mol ($n = 5$). The process is feasible for all negative energy balances, i.e., for $0 \leq n \leq 4$.

(c) We assume that the flux $j$ can be written as $j = k(205 - 49n)$ mol/s, with a dimensionless proportionality constant $k$. The production rate of ATP reads $nj = k(205n - 49n^2)$. The condition for maximal ATP production rate is

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}n} k(205n - 49n^2) = k(205 - 2 \cdot 49n) \\ \Rightarrow n &= 205/98 \approx 2 \end{aligned}$$

The efficiency (ATP production per glucose molecule) is given by $n$ itself. The maximal possible value is $n = 4$, because for $n \geq 5$, the process would be thermodynamically infeasible.

**Problem 4**

$$J = \frac{S_0 \prod_{j=1}^{r} q_j - S_r}{\sum_{l=1}^{r} \frac{1}{k_l} \prod_{m=l}^{r} q_m} = \frac{1 \cdot 5^4 - 1}{\frac{1}{1} \cdot 5^4 + \frac{1}{1} \cdot 5^3 + \frac{1}{1} \cdot 5^2 + \frac{1}{1} \cdot 5^1} = \frac{624}{780} = 0.8$$

**Problem 5**

The steady-state flux is maximal, if all enzyme concentrations are maximal, here $E_i = 2$ for $i = 1, \ldots, 4$. We get

$$J = \frac{S_0 \prod_{j=1}^{r} q_j - S_r}{\sum_{l=1}^{r} \frac{1}{E_l k_l} \prod_{m=l}^{r} q_m} = \frac{1 \cdot 5^4 - 1}{\frac{1}{2} \cdot 5^4 + \frac{1}{2} \cdot 5^3 + \frac{1}{2} \cdot 5^2 + \frac{1}{2} \cdot 5^1} = \frac{1048}{780} = 1.6$$

**Problem 6**

If restrictions apply only to the sum of enzyme concentrations, we use the relation

$$E_i^{\text{opt}} = E_{\text{total}} \cdot \sqrt{Y_i} \cdot \left( \sum_{l=1}^{r} \sqrt{Y_l} \right)^{-1} \text{ with } Y_l = \frac{1}{k_l} \prod_{m=l}^{r} q_m, \text{ hence}$$

$$Y_1 = 5^4,\ Y_2 = 5^3,\ Y_3 = 5^2,\ Y_4 = 5^1$$

$$E_1^{\text{opt}} = \frac{8 \cdot 25}{6\left(5 + \sqrt{5}\right)} = 4.606,\ E_2^{\text{opt}} = \frac{8 \cdot 5 \cdot \sqrt{5}}{6\left(5 + \sqrt{5}\right)} = 2.060,\ E_3^{\text{opt}} = \frac{8 \cdot 5}{6\left(5 + \sqrt{5}\right)}$$

$$= 0.921,\ E_2^{\text{opt}} = \frac{8 \cdot \sqrt{5}}{6\left(5 + \sqrt{5}\right)} = 0.412$$

**Problem 7**

The resulting optimal steady-state flux is

$$J^{\text{opt}} = \frac{\left(1 \cdot 5^4 - 1\right) \cdot 8}{\left(\frac{1}{25} \cdot 5^4 + \frac{1}{5 \cdot \sqrt{5}} \cdot 5^3 + \frac{1}{5} \cdot 5^2 + \frac{1}{\sqrt{5}} \cdot 5^1\right) \cdot 6 \cdot \left(5 + \sqrt{5}\right)}$$

$$= \frac{624 \cdot 8}{\left(6 \cdot \left(5 + \sqrt{5}\right)\right)^2} = 2.648$$

**Problem 8**

$$\text{It } \textit{holds } \frac{\mathrm{d}S_0}{\mathrm{d}t} = -k_1 \cdot E_1 \cdot S_0 \text{ and } \frac{\mathrm{d}S_1}{\mathrm{d}t} = k_1 \cdot E_1 \cdot S_0 = -\frac{\mathrm{d}S_0}{\mathrm{d}t}.$$

$$S_0(t) = S_0(0) \cdot e^{-k_1 E_1 t},\ S_1(t) = S_0(0) \cdot e^{-k_1 E_1 t}\left(e^{-k_1 E_1 t} - 1\right)$$

$$\tau = \frac{1}{S_0(0)} \int_{t=0}^{\infty} (S_0(0) - S_1(t))\mathrm{d}t = \frac{1}{k_1 E_1}$$

For two reactions:

$$S_0(t) = S_0(0) \cdot e^{-k_1 E_1 t},\ S_1(t) = S_0(0) \cdot \frac{k_1}{k_2 - k_1}\left(e^{-k_1 E_1 t} - e^{-k_2 E_2 t}\right)$$

$$S_2(t) = S_0(0) \cdot \frac{1}{k_1 - k_2}\left(-e^{-k_2 E_2 t} k_1 + k_1 - k_2 + e^{-k_1 E_1 t} k_2\right)$$

$$\tau = \frac{1}{S_0(0)} \int_{t=0}^{\infty} (S_0(0) - S_2(t))\mathrm{d}t = \frac{k_1 E_1 + k_2 E_2}{k_1 E_1} S_0(0)$$

**Problem 9**

In stationary state, both strategies have the same fitness value $f_1 = f_2$, so

$$f_{11}x_1 + f_{12}x_2 = f_{21}x_1 + f_{22}x_2.$$

Solving for the ratio $x_1/x_2$ yields

$$\begin{aligned} 0 &= (f_{11} - f_{21})x_1 + (f_{12} - f_{22})x_2 \\ \Rightarrow x_1/x_2 &= -(f_{12} - f_{22})/(f_{11} - f_{21}) \\ &= -\frac{(c-v)/2}{-v/2} = \frac{c-v}{v}. \end{aligned}$$

**Problem 10**

The rate equations for the resource concentration $s$ and the population sizes $n_i$ read

$$\begin{aligned} \mathrm{d}s/\mathrm{d}t &= v - \sum_i n_i J_i^{\mathrm{S}}(s) \\ \mathrm{d}n_i/\mathrm{d}t &= a J_i^{\mathrm{ATP}}(s) n_i - b n_i \end{aligned}$$

with constants $a$ and $b$.

(a) For a single strain, the steady-state equations read

$$\begin{aligned} 0 &= v - n J^{\mathrm{S}}(s) \\ 0 &= a J^{\mathrm{ATP}}(s) n - b n \end{aligned}$$

Solving these equations for $n$ yields $n = \eta v a/b$ where $\eta = J^{\mathrm{ATP}}(s)/J^{\mathrm{S}}(s)$.

(b) In a direct competition, the net growth rate for the $i^{\mathrm{th}}$ strain reads $J^{\mathrm{ATP}}(s) - b/a$, so the strain with the largest ATP production rate (at the current level $s$) will grow at the highest rate and outcompete the other strains.

## 10 Cell Biology

Answers to Problems

**Problem 1**

Proteins can have either a filamentous structure or a globular structure. The exact protein structure is defined by the amino acid sequence and interaction with the molecules in the protein's environment. Electrostatic interactions between the individual amino acids themselves and the protein environment (e.g., the pH or other interaction partners, like other proteins, lipids or ions) determine the exact three-dimensional protein structure. Proteins are described by different structures. The primary structure is simply the sequence of the amino acids. Very regular molecular arrangements constitute the secondary structure, e.g., an α-helix or a β-sheet. The elements of the secondary structure are fold further into a specific three-dimensional structure, the so called tertiary structure. The tertiary structure might be influenced by posttranslational modifications or interactions with ions that stabilize specific conformations. Assemblies of several proteins determine the quaternary structure. It is controlled by interaction and aggregation of individual protein monomers. Many proteins like tubuline or superoxide dismutase are only functional as multimers in such quaterny structure.

**Problem 2**

There are four different nucleotide bases used by the genomic DNA. With two bases it would be possible to code only $4^2 = 16$ different states. Since there are 20 different amino acids used in protein biosynthesis plus a stop signal for translation, at least 21 different states must be able to be represented by the genetic code. This makes it necessary to use at least a combination of three nucleotides ($4^3 = 64$ combinations) for the representation of 21 different states.

**Problem 3**

Covalent cross links between protein residues, in particular disulfide bridges between cysteine residues can stabilize their three-dimensional structure. Moreover, a high rate in protein synthesis can guarantee that sufficient functional proteins for the maintenance of cellular processes are present.

**Problem 4**

After translation, a protein can obtain new properties from posttranslational modifications of the amino acids. Different functional groups can be attached to the amino acids, like lipids, carbohydrates, acetate and phosphate. Moreover, disulfide bridges between cysteine residues can be established or proline residues can be modified to hydroxyproline by addition of a hydroxyl group.

**Problem 5**

Prokaryotes are evolutionary prior to eukaryotes. Since prokaryotes do not have an efficient compartmentalization, an mRNA molecule can undergo translation already while it is still transcribed. During the evolution of eukaryotic cells transcription and translation became spatially separated and processes for mRNA sequence modification became possible, e.g. splicing. An advantage of splicing is that regions of the coding mRNA template can be removed or alternatively be used. This introduces a greater variability of proteins that are transcribed from a single gene.

**Problem 6**

A compartment provides a local reaction space and thus substrates and products of a reaction or of a reaction sequence are in close proximity. This enables sequential reaction processes, in which products of a previous reaction can easily be used as substrates of another reaction. This is also a very important precondition for the development of life. Eukaryotic cells benefit from compartmentalization as individual processes of the cell can be separated from each other, e.g., highly reactive substances can be separated to protect other cellular structures (e.g., DNA) from getting damaged. Thus, compartmentalization allows the establishment of local reaction spaces and the separation of cellular processes.

**Problem 7**

Proteins that have a signaling sequence that roots them to the membrane are synthesized by ribosomes of the rough ER. All proteins are synthesized from the N-terminus to the C-terminus and those proteins that are transmembrane proteins of the cell membrane have their N-terminus in the ER lumen and their C-terminus remains outside. Subsequently, post-translational modification can take place in the ER lumen and the Golgi complex. Finally, the vesicles containing the newly synthesized transmembrane proteins fuse with the cell membrane and the N-terminus that is inside the vesicles will face the outer cellular space.

**Problem 8**

Mitochondria have their own DNA and can only be derived from an existing mitochondrion. Thus, if a new cell looses all of its mitochondria it probably will die, since mitochondria cannot regrow anymore.

## 11 Experimental Techniques in Molecular Biology

Answers to Problems

**Problem 1**
Under the assumption that all four nucleotides appear with the same probability in the target DNA there is on average one Bam HI recognition site for every $4^6 = 4096$ nucleotides. So we can expect to find around 12 recognition sites in bacteriophage λ. Under the same assumptions we expect for a restriction enzyme with an 8 bp long recognition sequence approximately $4600000/4^8 = 70.19$ cutting sites. For real sequences the actual number of restriction sites can be quite different since the nucleotide frequencies are often unequal.

**Problem 2**
100 ml medium can contain $10^9$ bacteria and so to know how many generations it takes to reach this number we have to solve the equation: $2^x = 10^9$. After 29.89 generations, that means after 9 h, 57 min and 48 s the bacterial population has reached this size. However, in reality this calculation would be too simplistic, since the growth rate declines with increasing population density and decreasing nutrient content of the medium.

**Problem 3**
The gel matrix not only provides structural stability for the gel but also represents an obstruction for the moving macromolecules. If the pores are too small the macromolecules cannot migrate at all. To separate large fragments of DNA or proteins the pore size has therefore to be increased. There is of course a limit to this strategy since reducing the matrix content leads to very fragile gels.

**Problem 4**
The amino acids of proteins contain functional groups that carry charges depending on the pH. Under a very acidic pH the carboxyl groups are neutral while the amino groups carry a positive charge leading to a positive net charge for the protein. At a very high pH the amino groups are neutral and the carboxyl groups are negatively charged, resulting in a negative net charge. Consequently, there exists a pH value where negative and positive charges are exactly equal so that the net charge of the

protein is zero. This pH value is the isoelectric point of the protein. This phenomenon is the basis for an electrophoresis variant called isoelectric focusing where proteins migrate through a pH gradient until they reach their isoelectric point (because neutral molecules don't move in an electric field).

**Problem 5**

There is a technical and a practical reason for the use of a secondary antibody in Western blotting. The technical reason is signal amplification. Normally, several secondary antibodies can bind to each primary antibody resulting in an amplification of the fluorescence signal. Furthermore, the same secondary antibody can be used in many experiments. It is therefore more practical and time saving to label large quantities of the secondary antibody with a fluorescent dye instead of small quantities of primary antibodies for each experiment.

**Problem 6**

A His-tag is a short sequence of usually six histidine amino acids that is introduced at the N- or C-terminus of proteins. This makes it possible to detect or purify the modified protein with high specificity using commercially available antibodies against the His-tag.

**Problem 7**

The sedimentation coefficient "*S*" is specified in Svedberg units. The larger the sedimentation coefficient, the faster the sedimentation rate of the macromolecule. *S* depends on the mass ($m$) and density of the particle ($\rho_{par}$), as well as on the density ($\rho_{sol}$) and friction ($f$) of the medium. The ribosomal subunits, for instance, got their name from their sedimentation coefficient (40S subunit and 60S subunit). Because the friction is controlled not only by the size of the particle, but also by its shape, *S* values are not additive. The complete ribosome (40S plus 60S) sediments at 80S and not at 100S.

$$S = \frac{m\left(1 - \rho_{\mathrm{sol}}/\rho_{\mathrm{par}}\right)}{f}.$$

**Problem 8**

High-performance liquid chromatography, also known as high-pressure liquid chromatography, uses pressures of several hundred atmospheres to force the protein solution through the column material. This leads to higher flow rates, which leaves the molecules less time for diffusion and thus results in a higher resolution of the separation process. Since the column material is packed more densely than in conventional columns, the columns can be much smaller to achieve the same resolution. This, in turn, means that very small probe volumes can be used for HPLC.

**Problem 9**

Proteins pose two main problems for high-throughput techniques that are much less pronounced for DNA-based techniques. First, proteins have to be synthesized in

sufficient quantities in appropriate organisms. However, overexpressing proteins can lead to problems because they might be toxic or precipitate in the cell. The second obstacle is the purification of the synthesized proteins. Since proteins are chemically much more diverse than DNA, the optimal purification procedure often varies from protein to protein. His-tagging the desired protein can help to reduce this problem, since it allows us to use one type of antibody for the purification of all tagged proteins.

**Problem 10**

The mass of the amino acids of the first peptide is $134 + 76 + 133 + 147 + 132 + 156 = 778$ Da. But we have to subtract the masses of five water molecules that are released by the condensation process. So the mass of the first peptide is 688 Da. Similarly we calculate the mass of the second peptide to $132 + 76 + 134 + 147 + 132 + 156 - 5 \times 18 = 687$ Da. The difference is approx. 1455 ppm ($10^6 \times 1/687$), an accuracy easily achieved by modern mass spectrometers.

**Problem 11**

This is classical genetics. If an animal heterozygous for the transgene "A" is crossed with another heterozygous animal the fraction of homozygous offspring is: $\mathrm{Aa} \times \mathrm{Aa} = \mathrm{AA} + 2\mathrm{Aa} + \mathrm{aa}$. This means 25% of the offspring are homozygous for AA. If crossed with a wild type animal there will be no homozygous AA offsprings at all since: $\mathrm{Aa} \times \mathrm{aa} = 2\mathrm{Aa} + 2\mathrm{aa}$.

**Problem 12**

The RNAi technique uses short double stranded pieces of RNA to trigger the degradation of mRNA containing the sequence of this dsRNA. The method has two main advantages over knockout animals. The synthesis of the required dsRNA is fast and cheap so that more experiments can be performed in the same time, or the same experiment can be completed much faster. The second advantage is that genes can be studied, whose knockout is lethal by applying the technique only after the animal is born. But the method also has disadvantages. The major problem is that the effect is only transient. The transfected RNAi molecules are degraded or diluted so that the knock down effect finally disappears. Another problem is the variability of the effect. Knockouts completely destroy the activity of a gene, while the RNAi suppression level varies from sequence to sequence.

**Problem 13**

If the binder has two binding sites for the target the following reactions are possible:

$$\begin{aligned} T + B &\underset{k_{\mathrm{off}}}{\overset{k_{\mathrm{on}}}{\rightleftharpoons}} TB \\ TB + T &\underset{k_{\mathrm{off}}}{\overset{k_{\mathrm{on}}}{\rightleftharpoons}} T_2B \end{aligned},$$

where $k_{\mathrm{on}}$ and $k_{\mathrm{off}}$ of both reactions are equal since the binding sites are identical in case of an antibody. This means we now have two species with bound target ($TB$ and $T_2B$) and thus we need two differential equations. For pedagogical reasons the terms

of the first equation have not been simplified, so that the reader can more easily follow the derivation of the individual expressions.

$$\frac{\mathrm{d}TB}{\mathrm{d}t} = k_{\mathrm{on}} \cdot (B_0 - TB - T_2B) \cdot T_0 - k_{\mathrm{off}} \cdot TB + k_{\mathrm{off}} \cdot T_2B - k_{\mathrm{on}} \cdot (B_0 - TB - T_2B) \cdot TB$$

$$\frac{\mathrm{d}T_2B}{\mathrm{d}t} = k_{\mathrm{on}} \cdot (B_0 - TB - T_2B) \cdot TB - k_{\mathrm{off}} \cdot T_2B$$

During the washing period both species are released from the surface ($TB \xrightarrow{k_{\mathrm{off}}} B$, $T_2B \xrightarrow{k_{\mathrm{off}}} TB$). After solving the differential equation for $T_2B$, we can replace $T_2B$ in the equation for $TB$ with this expression and then also solve this differential equation. $TB_0$ and $T_2B_0$ are the concentrations of $TB$ and $T_2B$ at the start of the washing step.

$$\frac{\mathrm{d}TB}{\mathrm{d}t} = -k_{\mathrm{off}} \cdot TB + k_{\mathrm{off}} \cdot T_2B$$

$$\frac{\mathrm{d}T_2B}{\mathrm{d}t} = -k_{\mathrm{off}} \cdot T_2B$$

$$T_2B(t) = T_2B_0 \cdot e^{-k_{\mathit{off}} \cdot t}$$

$$TB(t) = \left(k_{\mathit{off}} \cdot t \cdot T_2B_0 + TB_0\right) \cdot e^{-k_{\mathit{off}} \cdot t}$$

## 12 Mathematics

## 13 Statistics

## 14 Stochastic Processes

## 15 Control of Linear Systems

## 16 Databases

Answers to Problems

**Problem 1**
Databases can provide information about components and reactions, including stoichiometry and reaction properties (e.g. reversibility of a reaction), information about the kinetic laws and their respective parameters, as well as experimental data from individual small scale or large scale experiments that can be used, e.g., for parameter fitting.

**Problem 2**
For instance, the Reactome database is a valuable starting point for the development of different models. E.g., it provides detailed information about individual reactions

of gycolysis. Once the individual reactions of the system are defined, kinetic parameters of the respective enzymes can be found in BRENDE or SABIO-RK.

**Problem 3**

The content information of the ConsensusPathDB website (Version 10, April, 4th 2009) indicates that 4792 reactions from Reactome are present in the ConsensusPathDB and only 296 of them can be mapped to the 1629 reactions that were imported from the KEGG database. Both databases have 1025 physical entities in common.

**Problem 4**

Bcl-XL can be found in 10 different databases that have imported into ConsensusPathDB (Version 10, April, 4th 2009). Selecting only the entry Bcl-XL from the search results and preceding to the interaction listing indicates that Bcl-XL has 326 interactions annotated in the different databases present in ConsensusPathDB from which 98 are distinct. Now several interactions can be selected and visualized within ConsensusPathDB. Subsequently, the interaction networks can also be exported into several formats.

## 17 Modeling Tools

Answers to Problems

**Problem 1**
The smaller the number of the modeled entities, the more important it is to use stochastic techniques. If only few molecules of a certain type exist it becomes increasingly unrealistic to treat this number as continuous variable. Furthermore, under those conditions it is important to take random fluctuations into account, since they lead to very large relative changes in numbers (a jump from two to three molecules is a 50% increase). However, small numbers are a necessary, but not a sufficient condition to see differences between a deterministic and a stochastic simulation of a system. Only if additional conditions exist that are not easy to identify in advance, the two types of simulation will differ. This would be the case if different steady states exist, which are so closely together that they can be crossed by random fluctuations, if self-replicating entities are modeled (which exist in 0 or more copies) or if the rare species acts as activating switch for a genetic program.

**Problem 2**
In the last years it was recognized that the lack of portability is an important stumbling block for the development and re-use of large models. To solve this problem SBML, the Systems Biology Markup Language, has been developed. Over 100 software tools now support SBML so that researchers can easily exchange equations, parameter, and initial concentration settings as well as auxiliary information like boundary conditions and compartment information.

**Problem 3**
libSBML is a library that can be called from many programming languages like C/C++, Java, Python, Perl or List and is used to manipulate SBML files. libSBML provides functions to read, write, and create models that conform to the SBML standard.

**Problem 4**
Yes. Matlab can use the libSBML library to create SBML models and for Mathematica the package MathSBML is freely available, which provides the same functionality as libSBML.

**Problem 5**
See the movie CellDesigner4_Intro_ax01.wmv (also available at http://www2.hu-berlin.de/biologie/theorybp/video_tutorials.php).

**Problem 6**
See the movie Copasi4B24_TimeCourse_ax02.wmv (also available at http://www2.hu-berlin.de/biologie/theorybp/video_tutorials.php).

**Problem 7**
See the movie Copasi4B24_ParameterEstimation_ax03.wmv (also available at http://www2.hu-berlin.de/biologie/theorybp/video_tutorials.php). The fitted value for $V_{max}$ is 23.36 mmol/l and for $K_m$ we get 197.17 mmol/l.

**Problem 8**
See the movie Dizzy_TimeCourse_ax04.wmv (also available at http://www2.hu-berlin.de/biologie/theorybp/video_tutorials.php).