Space, noise, and information transmission in mathematical modelling of signalling pathways

Paulina Szymańska-Rożek

extended abstract of the PhD thesis

The theme of this thesis are three aspects of mathematical modelling of signalling pathways - space, noise (stochasticity) and information transmission. All three have been shown on different models known in theoretical biology: reversible cycles of phosphorylation–dephosphorylation, genetic toggle switch, and a model of gene activation by another gene, without and with feedback.

Models and mathematical description

In this thesis I present chosen aspects of mathematical modelling of signalling pathways analyzed on the canvas of models from biochemistry, theoretical biology, and biophysics. Their common mathematical core are Markov Chains with continuous time, moreover other tools such as ordinary differential equations, stochastic processes, and information theory were used.

The results are split into three parts, accordingly to the title of the chapters. The chapter "Space" approaches the aspect of spatiality on the example of a phosphorylation-dephosphorylation cycle. The chapter "Noise" shows the possibilities of benefiting from stochasticity to control a population of cells governed by the genetic toggle switch. In the chapter "Information", I optimize mutual information in a simple regulatory circuit, in which one biological quantity (for example the concentration of sugar in the cell, or the activity of a gene) affects another quantity (for example the concentration of the enzyme metabolizing the sugar, or the activity of another gene), with a feedback or without it.

Space
The spatial aspect of modelling signalling pathways was investigated in the model of a cycle of phosphorylation–dephosphorylation reactions. Substrates are phosphorylated and dephosphorylated by kinzases and phosphotases in the following way:

$$\mathrm{K} + \mathrm{S_u} \quad \xrightarrow{c} \quad \mathrm{K} + \mathrm{S_p}, \tag{1a}$$

$$\mathrm{P} + \mathrm{S_p} \quad \xrightarrow{d} \quad \mathrm{P} + \mathrm{S_u}, \tag{1b}$$

where $\mathrm{S_u}$ and $\mathrm{S_p}$ are unphosphorylated and phosphorylated substrates, respectively. K are the kinases and P – phosphatases. The letters above the arrows mark respective microscopic intensities of reaction - c for phosphorylation, and d for dephosphorylation.

We also consider a model, in which dephosphorylation occurs in the following way:

$$\mathrm{S_p} \xrightarrow{d_0} \mathrm{S_u}, \tag{2}$$

and phosphorylation occurs as previously.
Our main task was to find *effective macroscopic reaction rate constants*: c_{eff} and d_{eff}, that would satisfy the set of ordinary differential equations:

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho_{\mathrm{S_u}} = -c_{\mathrm{eff}}\rho_{\mathrm{K}}\rho_{\mathrm{S_u}} + d_{\mathrm{eff}}\rho_{\mathrm{P}}\rho_{\mathrm{S_p}}, \tag{3}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho_{\mathrm{S_p}} = c_{\mathrm{eff}}\rho_{\mathrm{K}}\rho_{\mathrm{S_u}} - d_{\mathrm{eff}}\rho_{\mathrm{P}}\rho_{\mathrm{S_p}}. \tag{4}$$

Usually, when one considers space in mathematical modelling, the first tool that comes into one's mind, are partial differential equations. However, the approach used in this thesis comprise of probabilitic methods, in particular, we considered a Markov Chain, whose states are defined by the numbers of substrates and their location in the space. The mentioned space is a finite, two-dimensional, triangular (i.e., such that every site has six neighbours) lattice, with no edge. The intensities of transitions between the states of the Markov Chain correspond to the microscopic intensities c and d (when an appropriate pair of substrate-enzyme is in adjacent lattice sites) and the intensity of diffusion, m. We perform numerical simulations a lot (that use the kinetic Monte Carlo algorithm), but in parallel, we conduct analytical reasoning for the cases of $m = 0$ and $m = \infty$.

In short, we want to compute the effective microscopic reaction rate constants defined as follows:

$$c_{\text{eff}} = \frac{n}{\rho_{\text{S}_\text{u}}\rho_{\text{K}}V\Delta t}, \tag{5a}$$

$$d_{\text{eff}} = \frac{n}{\rho_{\text{S}_\text{p}}\rho_{\text{P}}V\Delta t} \tag{5b}$$

where n is the number of reactions that fired in the Δt time interval, and V is the number of lattice sites. The densities of kinases, phosphatases and substrates are marked with ρ with an according subscript: ρ_{K}, ρ_{P}, ρ_{S_u} i ρ_{S_p}. We also calculated the steady state fractions of phosphorylated and dephosphorylated substrates: $\rho_{\text{S}_\text{p}}/\rho_{\text{S}}$ i $\rho_{\text{S}_\text{u}}/\rho_{\text{S}}$.
The mentioned analytical reasoning might be applied in the case of infinite diffusion, i.e., when $m = \infty$. In this case, effective reaction rate constants are $c^{\infty}_{\text{eff}} = 6c$ and $d^{\infty}_{\text{eff}} = 6d$. The second case in which purely analytical reasoning is conducted, is the zero diffusion limit, i.e., when $m = 0$. After longer combinatorics derivations, we obtained formulae for this case as well. We then verified them with numerical simulations. The most "real" case of finite, nonzero diffusion, is analytically a very challenging problem. In the general case, we assumed that the effective macroscopic reaction reate constants comprise of a zero-motility term, c^{0}_{eff} and d^{0}_{eff}, and a term linearly proportional to the diffusion coefficient: $\lambda \cdot m$. Then the ordinary differential equation for the time evolution of the fraction of phosphorylated substrates reads:

$$\tfrac{\text{d}}{\text{d}t}\rho_{\text{S}_\text{p}} = (\lambda m + c^{0}_{\text{eff}})\,\rho_{\text{K}}\,\rho_{\text{S}_\text{u}} - (\lambda m + d^{0}_{\text{eff}})\,\rho_{\text{P}}\,\rho_{\text{S}_\text{p}}, \tag{6}$$

where the coefficient λ had to be established basing on numerical analysis, and only for the symmetric case of $c = d$ and $\rho_{\text{K}} = \rho_{\text{P}}$. The difficulties come from the fact that even for the symmetric case, particularly when enzyme densities are small, the location of molecules on the lattice is nonhomogenous. An unphosphorylated substrate can be found rather in the vicinity of a kinase, and a phosphorylated substrate - near the phosphatase. Consequently, even though the overall probability that the substrate is phosphorylated is 0.5, the effective reaction rate constants are lowered, since the enzyme are surrounded by substrates, with which they already reacted.

In the most general case, the effective macroscopic reaction rate constants depend in a complicated manner on both microscopic reaction intensities, the diffusion coefficient, and the densities of both enzymes. We showed that they decrease with decreasing diffusion and this dependence is more pronounced for the less abundant enzyme. Consequently, the steady-state fraction of phosphorylated substrates can increase or decrease with diffusion, depending on relative concentrations of both enzymes. On top of that, we have analyzed the influence of other factors on the effective macroscopic reaction rate constants:

- the size of the lattice (the number of sites),

- the molecular crowding (adding to the lattice non-reacting, but moving molecules),
- formation of transient enzyme-substrate pairs (preceding or following the reaction).

Feeling some kind of hunger for analytical expressions in the case of finite, non-zero diffusion, we finally considered a model variant in which an appropriate enzyme molecules and a substrate molecule enter the same lattice site in order to react. Thanks to this slight modification we were able to conduct a reasoning about the *mean first passage time* and obtain surprisingly good approximations.

Noise

In this chapter I present the results of the research initiated during the *Quantitative Biology Summer School*, that took place in July 2014, in Albuquerque (New Mexico, USA). The research question raised there was: "can noise be used to drive one cell out of a population of identical cells to exhibit different phenotype (express more proteins) than the other, using a strategy applied equally to the whole population?". The strategy was the UV radiation that enhances protein degradation. In the thesis I used the term "noise", although it should be understood as "fluctuations".

Identical cells in the population are governed by the same genetic mechanism. In order to use fluctuations to drive one cell to a different behaviour, we had to consider such a mechanism, that allows the cell to be in two distinct states, (high and low protein expression). An example of genetic mechanism that leads to two states is the toggle switch, in which two genes inhibit mutually their activity (blocking protein synthesis). I present it in panel (a) of Fig. 1. A simpler model in which the cell ca also attain to different states, is the model of self-inducing gene (the synthesised protein acts as the activator for its own gene) - this scheme is shown in panel (b) of Fig. 1. The deterministic description of the model is one ordinary differential equation for the concentration of proteins in each cell:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = k(x) - R(x), \tag{7}$$

where $k(x)$ is the production propensity of the form: $k(x) = k_0 + k_1 x^m/(x^m+\beta)$, and $R(x)$ is the degradation propensity: $R(x) = (\gamma+u)x$. U is the UV radiation. For a choice of parameters k_0, k_1, β, m, γ and function u, this equation admits three stationary points, for other sets of parameters - only one. An illustration of these variants is shown in panel (c) of Fig. 1.

If the two cells are identical, i.e., the evolution of protein concentration in both of them follows the above ODE, and they have the same initial concentration, then, according to the deterministic prediction they are undistinguishable - at any time, they will contain the same amount of protein. However, if we take in

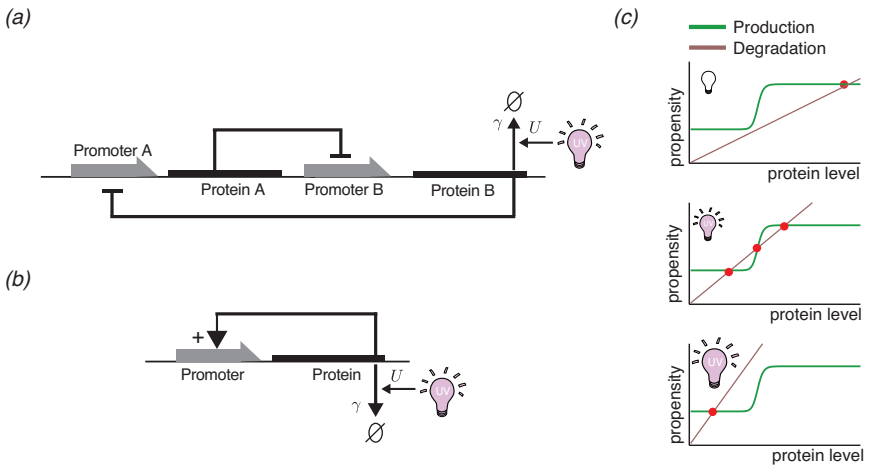

Figure 1: (a) A classical representation of a genetic toggle switch, in which two protein species repress one another and the degradation rate of one protein is enhanced by UV radiation. (b) Self-activating gene, in which the two mutually inhibiting protein dynamics are replaced by a single, self-inducing gene. This is the simplest model that exhibits bistability. Additionally, mRNA production, which is responsible for the majority of noise in protein synthesis, is neglected. Protein degradation can be enhanced by increasing levels of UV radiation. (c) Production and degradation rates versus the number of proteins, x, in the simplified toggle switch model. The production rate, $k(x) = k_0 + k_1 x^2/(x^2 + \beta)$, is plotted in green, and the degradation rate, $R(x) = (\gamma + U)x$, is plotted in brown. In a deterministic representation, protein levels evolve according to $\dot{x} = k(x) - R(x)$. Intersections where $k(x) = R(x)$ provide the stationary points. The three panels correspond to cases where UV is low (top, a single high equilibrium point), moderate (middle, two stable and one unstable equilibrium points) or high (bottom, a single low equilibrium point).

to account stochasticity that underlies all processes involving small number of elements, the number of protein molecules will have to be modelled as a random variable. Depending on the values admitted by this variable, the whole population will be subjected to UV radiation, all cells with the same intensity.

The very first model we considered consisted of one two proteins - one was the chosen cell, and the other was "the rest of the population". We succeeded in building a simple strategy of UV radiation admitting only three values: u_1, u_2 i u_3, depending on the number of protein molecules in the chosen cell and in the rest of the population. This strategy is depicted in Fig. 2 and described in details in the caption. The results for the two-cell model were astonishingly satisfactory - the chosen cell synthesized more proteins than the other cell with

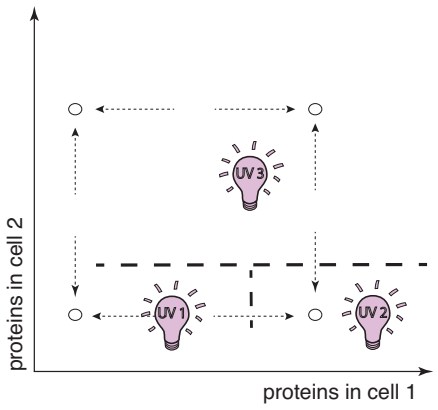

Figure 2: Preliminary control law for the application of UV radiation based upon comparison of the protein content for two cells. When both cells contain little protein (bottom left corner), no UV is applied ($u_1 = 0$). When protein content is high in cell 1 and low in cell 2 (bottom right corner) moderate UV is applied ($u_2 = 0.75$). When protein content in cell 2 is above a threshold ($= 12$) (upper part), high UV is applied ($u_3 = 1.75$). All parameters for the gene regulatory circuit ($k_0 = 5$, $k_1 = 50$, $\beta = 20$, $m = 4$ and $\gamma = 0.5$) are fixed and identical for both cells.

probability 0.99. For a population of 30 cells, with the same UV radiation strategy (UV level was chosen on the basis of protein expression in the chosen cell and the maximum expression in other cells), the chosen cell was within the top 20%. Moreover, we considered the following variants and strategies:

- a model, in which time delays in protein synthesis are taken into account
- a model that accounts for noise in protein synthesis (the intensities of synthesis, k_0 and k_1, in the cells other than the chosen one were multiplied by a coefficient - a normal random variable with mean 1 and standard deviation 0.1. The coefficients k_0 and k_1 in the chosen cell are equal to the mean of k_0 and k_1 in the other cells, or the mean + one standard deviation or the mean − one standard deviation.)
- strategies, in which UV radiation level is chosen based on other than the maximum statistics of the protein expression in other cells: their mean, median, 75the percentile, or based only on the leel of protein in the chosen cell
- full toggle switch model (with two species of “competing” proteins).

Information

This interesting aspect of mathematical modelling of signalling pathways intrigues not only mathematicians, but largely researches of natural sciences - biologists, biophysicist and biochemists. My work on information transmitted in biological circuits were initiated during a doctoral internship in Ecole Normale Supérieure, under the supervision of Aleksandra Walczak. Were were inspired by the simple question if and how biological systems "communicate" both internally, as well as with the environment. The investigated model consists of two binary random variables, between which I calculate Mutual Information (and this quantity is interpreted as the information transmitted in the system). I assume that for information transmission some energy is needed, so a natural question arises how does an energetic constraint influence the amount of information that can be transmitted. Finally I pose a frequently asked question, whether feedback is beneficial; in this case whether it increases the mutual information of the considered random variables. The model is presented in Fig. 3.

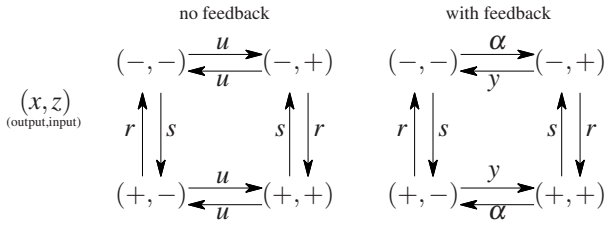

Figure 3: Scheme of the possible states and transitions between them for both models: without feedback (left figure), and with feedback (right figure). Since there are two binary variables there are four states; transition rates are marked next to respective arrows. Note the symmetry between the "pure" ($(-,-)$ and $(+,+)$) states and the "mixed" states ($(-,+)$ and $(+,-)$) in both models.

The main task is to find parameters (u, s, r, α, y) - rates of transition between the states, that maximize mutual information between z_0 - input at time 0 and x_t - output at some time t. Formally, we look for the maximum of the function depending both on the parameters and the time:

$$I[X_t, Z_0] = \sum_{x_t, z_0} p(x_t, z_0) \log \frac{p(x_t, z_0)}{p(x_t)p(z_0)}, \tag{8}$$

where $p(x_t, z_0)$ is the probability, that variables are in states $(-,-)$, $(-,+)$, $(+,-)$, $(+,+)$. This is the definition of mutual information, it measures how the uncertainty we have about one variable is reduced given that we know the other variable. This definition is equivalent to the more common one, that uses *Entropy* of a random variable and *conditional entropy* of random variables, $I[X_t, Z_0] = S(X_t) - S(X_t|Z_0)$.

The energetic constraint we consider while optimizing mutual information is the *entropy production rate*. It is also a function of the parameters and time.:

$$\sigma(t) = \sum_{i,j} p_i(t) w_{ij} \log \frac{p_i(t) w_{ij}}{p_j(t) w_{ji}} \tag{9}$$

In the steady state this quantity is:

$$\sigma(t) \xrightarrow[t\to\infty]{} \quad \sum_{i,j} p_i^{\mathrm{ss}} w_{ij} \log \frac{w_{ij}}{w_{ji}} = \sigma^{\mathrm{ss}} \tag{10}$$

and this is the main energetic constraint considered. However, we also analyzed an additional constraint - *average dissipation*, which is a mean integral of the entropy production rate, calculated up to some time:

$$\Sigma^{\mathrm{avg}}(\tau_p) = \frac{1}{\tau_p} \int_0^{\tau_p} \hat{\sigma}(\tau) d\tau. \tag{11}$$

We refer to this quantity as the *cost* of information transmission.

The considered problem is well posed - for a given time, at which we measure the information about the state of the output, we look for parameters that return the highest value of $I[x_\tau, z_0]$. We also investigate the same kind of problem, assuming that the available energy (steady state entropy production rate, σ^{ss}) is limited. We finally calculate the cost of optimal information transmitted.

We considered four models:

- S - without feedback, with steady state initial condition
- F - with feedback, with steady state initial condition
- $\tilde{\mathrm{S}}$ - without feedback, and the initial condition subjected to optimization
- $\tilde{\mathrm{F}}$ - with feedback, and the initial condition subjected to optimization

The results obtained for the above four models are summarized in Table 1:

	I^{opt}	Expense
S, F	$I(S) < I(F)$	$E(S) = E(F)$
$\tilde{S}$, $\tilde{F}$	$I(\tilde{S}) \leq I(\tilde{F})$	$E(\tilde{S}) > E(\tilde{F})$

Table 1: Comparison between the four models, S, F, $\tilde{S}$, and $\tilde{F}$ in terms of optimal mutual information, I^{opt}, and the average dissipation ("expense"), E.

Steady state initial condition guarantees that the feedback will be more effective in transmitting information that the no-feedback variant, but the costs in both cases are the same. Whereas for the system starting out of steady state the difference in information transmitted is smaller between the no-feedback and feedback variants, but the cost is significantly higher if there is no feedback.

Papers

The results presented in the dissertation were published in three articles.

The results from the chapter about Space are covered in [3] and [1]. In the first article I derived analytical expressions for effective macroscopic reaction rate constants in the limits of zero and infinite motility. For finite, nonzero motility I conducted a reasoning to estimate EMRRCs for a "symmetric" case. I performed stochastic simulations, prepared most of the figures used in the dissertation, and finally wrote the manuscript, which served as a scaffold for the first chapter of the thesis. In the second paper, jointly with the first author I derived more accurate estimates for the EMRRCs for the case of nonzero, finite motility, using the mean first passage time approach. I join hereby authors' contribution declarations.

The research on the stochastic aspect can be found in [2]. In this article I prepared theoretical basis for the numerical simulations conducted by the coauthors. I am the author of the manuscript and of the conducted mathematical reasoning. Authors' contribution declarations are joined.

Finally the chapter on Information transmission contains yet unpublished results obtained during my doctoral internship under the supervision of Aleksandra Walczak. I conducted analytical reasoning, exact (symbolic) computations in Mathematica package and numerical optimization of mutual information.

References

[1] Paweł Nałęcz-Jawecki, Paulina Szymańska, Marek Kochańczyk, Jacek Miękisz, and Tomasz Lipniacki. Effective reaction rates for diffusion-limited reaction cycles. *J Chem Phys*, 143(21):215102, 2015.

[2] Paulina Szymańska, Nicola Gritti, Johannes M. Keegstra, Mohammad Soltani, and Brian Munsky. Using noise to control heterogeneity of isogenic populations in homogenous environments. *Phys Biol*, 12(4):045003, 2015.

[3] Paulina Szymańska, Marek Kochańczyk, Jacek Miękisz, and Tomasz Lipniacki. Effective reaction rates in diffusion-limited phosphorylation-dephosphorylation cycles. *Phys Rev E*, 91:022702, Feb 2015.