Univerzita Komenského v Bratislave

Fakulta matematiky, fyziky a informatiky

Michal Hojčka

Autoreferát dizertačnej práce

Dynamical models in gene expression

na získanie akademického titulu philosophiae doctor

v odbore doktorandského štúdia:

9.1.9 aplikovaná matematika

Bratislava 23.4.2018

Dizertačná práca bola vypracovaná v dennej forme doktorandského štúdia na Katedre aplikovanej matematiky a štatistiky Fakulty matematiky, fyziky a informatiky Univerzity Komenského.

Predkladateľ: **Mgr. Michal Hojčka**
Katedra aplikovanej matematiky a štatistiky
Fakulta matematiky, fyziky a informatiky
Univerzita Komenského
Mlynská dolina
842 48, Bratislava 4

Školiteľ: **doc. Mgr. Pavol Bokes, PhD.**
Katedra aplikovanej matematiky a štatistiky

Študijný odbor: **9.1.9 aplikovaná matematika**, študijný program: **aplikovaná matematika**

Predseda odborovej komisie:

prof. RNDr. Daniel Ševčovič, DrSc.
Katedra aplikovanej matematiky a štatistiky
FMFI UK, 842 48 Bratislava

Introduction

According to [36], a gene is defined as a hereditary unit of DNA that is required to produce a functional product. This process of transforming the information from a gene to create further gene products is known as gene expression. The essential part in gene expression as well as in virtually every process on the cellular level is carried out by proteins, large biomolecular objects consisting mainly from the amino acids. In most mathematical models concerning gene expression, we focus on the first and the most important step of gene expression, the transcription. A special type of proteins, called transcription factors, possesses crucial functionality in this process as they activate transcription by binding to specific DNA sequences. The result of transcription is a primary RNA transcript; following the next steps of gene expression we ultimately obtain a functional protein. For further biological insight we refer to [28].

Biochemical reactions can be studied using a number of different mathematical formalisms and we can distinguish between the two main approaches. The first, deterministic, approach exploits deterministic ODE models to describe the dynamics of biochemical reactions. An alternative way to study biological systems is through stochastic models which consider each reaction as a single random event. The advantage of this approach is the fact that these models describe the behavior of the system well also at lower numbers of the involved species, as is the case for the number of proteins and other species present in the biological processes inside the cells such as gene expression [12, 48]. Therefore, deterministic modelling of such reactions can be quite inaccurate and we turn instead to stochastic methods [32]. As they work with discrete number of molecules, they can easily be simulated through stochastic simulation algorithms, in particular the Gillespie algorithm [19, 20].

Being a very timely topic, gene expression sparked a new wave of interest in Markovian models of chemical kinetics, e.g. [44]. In this thesis we present a simplified model in which we neglect the intermediary processes associated with mRNA creation and focus solely on protein production. We assume that the protein is produced with a constant rate and that the rate of its decay is proportional to the number of proteins. We study the protein dynamics in presence of so-called decoy binding sites [52, 31] on the DNA. Our model further takes into account protein binding/unbinding reactions with these binding sites. Similar models have already been studied previously; in particular [18] investigated the model with protected complexes, i.e. the case when bound proteins were immune to degradation, showing that the steady-state distribution is Poissonian. Our model allows bounded proteins to degrade, which introduces additional noise into the model [5, 7]. For simplicity, we ignore effects of burst-like protein synthesis or transcriptional auto-regulation [5, 8, 46]. Unfortunately, as is often the case, the solution of free protein probability distribution cannot be obtained in a closed form. However, we can employ the fact that biochemical reactions often operate on different timescales [40, 47] to address the issue. History of applying these assumptions in stochastic modelling is rather new [9], but in recent years were thoroughly investigated in works such as [25, 23, 24]. Particularly, in the context of our model, the interactions between the protein and its binding sites occur on a substantially faster timescale than the production and decay of protein does [2]. Therefore we can successfully use singular perturbation methods [9, 38, 39] to obtain the quasi-steady-state solution to our problem.

Goals of the Thesis

In the first chapter we summarise all the useful definitions from the fields of probability and differential equations, which we use in later parts of the thesis. In chapter two we focus on the theory regarding deterministic and stochastic modelling of biochemical reactions together with a couple of illustrative examples in order to provide an introduction into the topic. We also present a concept of the Master equation. We would like to address a common problem in the field, which is the inability to solve the Master equation for all but very simple cases. In the third chapter we present a simplified non-bursting model motivated by the gene expression in which we consider protein creation/degradation and the interaction with the decoy binding sites. Our goal is to employ different techniques in order to obtain the solution for the free protein distribution (or some reasonable approximation). We would like to study the Fano factor and the difference of the obtained distribution from the Poisson distribution. In the fourth chapter we aim to obtain the Fano factor in a closed form for the large-system-size case. In the last chapter we study another model also associated with gene expression, which considers the interactions between mRNA and microRNA molecules and the silencing effect on the population of mRNA. We aim to utilize similar techniques as for the previous model in order to derive the distribution of mRNA and the associated Fano factor.

Results for a gene expression model in the presence of decoy binding sites

We describe the system of chemical reactions motivated by the gene expression in which we focus just on the level of protein in the system and the interactions with the decoy binding sites. We are taking into account simplified, non-bursting, regime of protein production.

We introduce the following variables in our system:

X - total protein, X_f - free protein,
Y - all binding sites, Y_f - free binding sites,
C - complex (protein bound to the binding site),

and we assume that three reversible reactions can take place:

1) Protein production/decay.

$$\emptyset \underset{\gamma}{\overset{k}{\rightleftharpoons}} X_f$$

2) Protein binding/unbinding reaction.

$$X_f + Y_f \underset{k_-}{\overset{k_+}{\rightleftharpoons}} C$$

3) Decay of the complex (a free binding site is vacated).

$$C \xrightarrow{\gamma} Y_f$$

In order to avoid confusion with X, we use N instead of X_f as the number of free protein.

Our main goal is to investigate the distributions of free and total protein in the system. In the beginning we write down the associated Master equation. Then we focus on the distribution of total protein. Using the method of generating functions we transform Master equation into partial differential equation which we solve and obtain solution in the form of the Poisson distribution with the mean $\langle X \rangle = \frac{k}{\gamma} \cdot (1 - e^{-\gamma t})$. The main part consists of performing singular perturbation reduction to obtain quasi-steady-state solution for the free protein distribution. We prove the correctness of the solution by mathematical induction. The formula for the probability distribution of the free protein reads

$$P_N = \sum_{X=N}^{N+Y} P_X \cdot \frac{k_b^N}{N!(X-N)!(Y-X+N)!} \cdot \left(\sum_{i=\max\{0,X-Y\}}^{X} \frac{k_b^i}{(X-i)!(Y-X+i)!i!} \right)^{-1}, \tag{1}$$

where P_X is the probability for total protein amount which follows the Poisson distribution with mean $\frac{k}{\gamma}$. Probability P_N when $Y = 0$ follows the same distribution. We also derive results for large values of Y which yields an approximated Poisson distribution with a different mean:

$$\langle N \rangle = Var(N) = \frac{k_b \langle X \rangle}{k_b + Y}.$$

We use the quasi-steady-state solution to investigate the statistical characteristics of obtained distribution.

Then we introduce and compare three different ways how to obtain the number of free protein using numerical simulations:

1) Using Gillespie algorithm. In this case we simulate each reaction as it really happens through time.

2) Using explicit formula for free protein count in quasi-steady state (calculated earlier in the thesis). In this case we assume that $k_- \gg \gamma$ in order to justify the approximation.

3) Solving the system of ODEs given by the Master equation. As the problem contains infinite number of equations, we have to set maximal value of the number of all protein at which the system is truncated.

We report very good agreement between the distributions and thus we justify using quasi-steady-state solution as a very good approximation for the free protein distribution.

In the next chapter we try to further simplify the distribution of free protein species and its Fano factor for some reasonable specific case; we aim to obtain a closed-form formula for statistical moments. In order to do that we perform an expansion of the Master equation in a linear-noise scenario in the similar manner as presented in Chapter X in [49]. We focus on the limit case when the size Ω of the system is large enough ($\Omega \gg 1$) and we identify the system size Ω with the dissociation constant $k_b = k_-/k_+$ in our model. This is a standard approach and it guarantee that the binding and unbinding reaction rates are of the same order. We divide this procedure into three main stages. First we derive deterministic mean of the examined variables, which has a single non-negative solution, which we refer to as $\bar{n}$, $\bar{y}_f$, $\bar{c}$, and has the form

$$\begin{aligned} \bar{c}(x,y) &= \frac{x+y+1-\sqrt{x^2+y^2+1+2x+2y-2xy}}{2}, \\ \bar{n}(x,y) &= x-\bar{c}, \\ \bar{y}_f(x,y) &= y-\bar{c}. \end{aligned} \tag{2}$$

Secondly we include noise from binding/unbinding reactions and finally we add noise from total protein number fluctuation. The main challenge here is to combine these individual results here in a correct way. In the end we obtain the formula for the Fano factor in the form

$$F = 1 + \frac{\bar{n}\bar{y}_f(1+\bar{n})}{(1+\bar{n}+\bar{y}_f)^2}.$$

We also justify our results by comparing them to formulae from quasi-steady-state results using numerical simulations. We report a very good fit between quasi-steady-state solution for large Ω and approximated large-system-size results. We confirm that the Fano factor is greater than one for the intermediate levels of binding sites in contrast with Poissonian character (the Fano factor equals one) for no binding sites or their excess as presented on Figure 0.1.

These results were also published in [21].

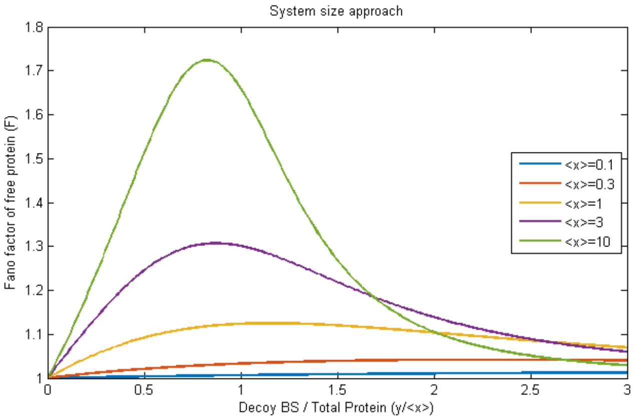

Figure 0.1: *Fano factor for large system size with $y/\langle x \rangle$ as an independent variable.*

Results for mRNA – microRNA system

A microRNA is a small, non-coding RNA and contains about 22 nucleotides (abbreviated form miRNA is sometimes used instead of microRNA). It can be found mainly in some viruses, plants and animals. MicroRNA was discovered for the first time in 1993 on the lin-4 gene, which was repressing another, lin-14 gene [30, 51]. This process of gene repression is a member of broad class known as RNA silencing.

In our model, we consider two reactants in the system: X (mRNA) and Y (microRNA). They are subject to two slow reactions

$$\emptyset \xrightarrow{\kappa} \mathrm{X}+\mathrm{Y}, \quad \mathrm{X} \xrightarrow{1} \emptyset,$$

and three fast reactions

$$\mathrm{X}+\mathrm{Y} \xrightarrow{\frac{\alpha q}{\varepsilon}} \mathrm{Y}, \quad \mathrm{X}+\mathrm{Y} \xrightarrow{\frac{\alpha(1-q)}{\varepsilon}} \emptyset, \quad \mathrm{Y} \xrightarrow{\frac{1}{\varepsilon}} \emptyset.$$

We formulate the Master equation, and use generating functions to transform the Master equation into a partial differential equation of the second order. In a specific parametric regime, this partial differential is reduced to an ordinary differential equation, which is solved using the hypergeometric functions. This solution is used to construct non-trivial approximations to the probability mass function in the form

$$P_{M,N} = \frac{\delta_{N,0}\left(\frac{\kappa}{\alpha}\right)^M}{M!\left(\kappa q + \frac{1}{\alpha} + 1\right)_M} \times \frac{{}_0F_1\left(\kappa q + \frac{1}{\alpha} + 1 + M, -\frac{\kappa q}{\alpha}\right)}{{}_0F_1\left(\kappa q + \frac{1}{\alpha} + 1, \frac{\kappa}{\alpha}(1-q)\right)} + O(\varepsilon).$$

We calculate the factorial moments of the distribution as

$$\mu_{(M)} = \frac{\left(\frac{\kappa}{\alpha}\right)^M}{\left(\kappa q + \frac{1}{\alpha} + 1\right)_M} \times \frac{{}_0F_1\left(\kappa q + \frac{1}{\alpha} + 1 + M, \frac{\kappa}{\alpha}(1-q)\right)}{{}_0F_1\left(\kappa q + \frac{1}{\alpha} + 1, \frac{\kappa}{\alpha}(1-q)\right)} + O(\varepsilon).$$

The Fano factor can be expressed in terms of the first two factorial moments as

$$\mathrm{F} = 1 + \frac{\mu_{(2)}}{\mu_{(1)}} - \mu_{(1)}. \tag{3}$$

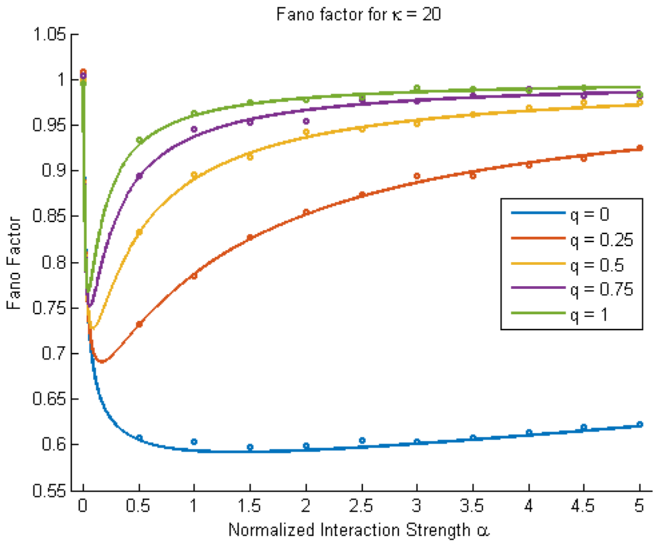

Figure 0.2: *Fano Factor of X as a function of the interaction strength with Y.*

Finally, we compare our results with stochastic simulations (using Gillespie algorithm) and discuss the numerical observations. We report good fit between stochastic simulation results and our approximated solution. Here we display the Fano factor for different choice of parameters on Figure 0.2. These results (Chapter 5 in the thesis) are a part of article [6], which is currently submitted for publication.

Summary

We introduced a simplified gene expression model in the presence of (decoy) binding sites. We presented its Master equation, which does not have closed-form solution. We derived the distribution of total protein and then we employed singular-perturbation reduction techniques to obtain a quasi-steady-state approximation. Using this approximation we were able to obtain explicit formula for the free protein distribution. In addition to quasi-steady-state approximation, we introduced and compared two other methods to obtain free protein distribution. First one was the stochastic simulation through Gillespie algorithm and the second one numerical solving of the stiff system of ODEs. Comparing with other methods, we justified the correctness of quasi-steady-state formula. Then we employed this formula to observe statistical moments for a wide range of input parameters. We focused on the Fano factor, which yielded substantially different results from Poissonian case. Then we extended our model with the assumption of large system size, using the dissociation constant as its measure. With linear noise approximation we obtained simple expression for the Fano factor of free protein distribution. With the help of numerical simulation we showed the consistency with results from previous chapter. In the final chapter we applied similar methods on mRNA – microRNA system of reactions. We obtained explicit formula for mRNA distribution and compared it with numerical simulations. Finally we studied the distribution for many input parameters and demonstrated the differences between the calculated Fano factor and the benchmark Poissonian case. Although we applied our methodologies on relatively simple models, we expect that it can be helpful to employ analogous approaches in other stochastic models of gene expression or more general biological systems.

Bibliography

[1] Abramowitz, M., Stegun, I.: *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, National Bureau of Standards, Washington, D.C., 1972.

[2] Alon, U.: *An Introduction to Systems Biology: Design Principles of Biological Circuits*, Chapman & Hall/CRC, 2007.

[3] Bartel, D.P.: *MicroRNAs: genomics, biogenesis, mechanism, and function*, Cell 116 (2): 281-297, 2004.

[4] Bhattacharyya, A.: *On a measure of divergence between two statistical populations defined by their probability distributions*, Bulletin of the Calcutta Mathematical Society 35: 99–109, 1943.

[5] Bokes, P., Singh, A.: *Protein copy number distributions for a self-regulating gene in the presence of decoy binding sites* PLoS ONE Vol. 10, No. 3: 1-19, 2015.

[6] Bokes, P., Hojčka, M., Singh, A., : *Buffering gene expression noise by microRNA based regulation*, manuscript submitted for publication

[7] Burger, A., Walczak, A.M., Wolynes, P.G.: *Abduction and asylum in the lives of transcription factors*, P. Natl. Acad. Sci. USA Vol. 107: 4016-4021, 2010.

[8] Burger, A., Walczak, A.M., Wolynes, P.G.: *Influence of decoys on the noise and dynamics of gene expression*, Physical Review E Vol. 86(4): 041920, 2012.

[9] Cao, Y., Gillespie, D.T., Petzold, L.R.: *The slow-scale stochastic simulation algorithm*, Journal of Chemical Physics, 122(1), 014116, 2005.

[10] Casella, G., Berger, R. L.: *Statistical inference: Second edition*, Thomson Learning, Ann Arbor, 2002.

[11] Chen, W.W., Neipel, M., Sorger, P.K.: *Classic and contemporary approaches to modeling biochemical reactions*, Genes Dev 24 (17): 1861–1875, 2010.

[12] Elowitz, M.B., Levine, A.J., Siggia, E.D., Swain, P.S.: *Stochastic gene expression in a single cell*, Science Vol. 297: 1183-1186, 2002.

[13] Erban, R., Chapman, S.J., Maini, P.K.: *A practical guide to stochastic simulations of reaction-diffusion processes*, arXiv:0704.1908, 2007.

[14] Euler, L.: *Institutiones Calculi Integralis, vol. 1 of Opera Omnia Series*, 1769.

[15] Feinberg, M.: *Lectures onChemical Reaction Networks*, Lectures delivered at the Mathematics Research Center, University of Wisconsin Madison, 1979.

[16] Fritz, J.: *Partial differential equations (4th ed.)*, Springer, 1991.

[17] Gauss, C.F.: *Disquisitiones Generales Circa Seriem Infinitam, vol. 3*, Werke, 1813.

[18] Ghaemi, R., Del Vecchio, D.: *Stochastic analysis of retroactivity in transcriptional networks through singular perturbation*, Americal Control Conference, 2012: 2731-6.

[19] Gillespie, D.T.: *A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions*, Journal of Computational Physics 22 (4): 403–434, 1976.

[20] Gillespie, D.T.: *Exact Stochastic Simulation of Coupled Chemical Reaction*, The Journal of Physical Chemistry 81 (25): 2340–2361, 1977.

[21] Hojčka, M., Bokes, P.: *Non-monotonicity of Fano factor in a stochastic model for protein expression with sequesterisation at decoy binding sites*, Biomath 6, 1710217, 2017.

[22] Johnson, N., Kotz, S., Kemp, A.: *Univariate Discrete Distributions, 3rd ed.*, Wiley-Interscience, 2005.

[23] Kang, H-W., Kurtz, T.G.: *Separation of time-scales and model reduction for stochastic reaction networks*, The Annals of Applied Probability Vol. 23, No. 2, 529-583, 2013.

[24] Kang, H-W., Kurtz, T.G., Popovic, L.: *Central limit theorems and diffusion approximations for multiscale Markov chain models*, The Annals of Applied Probability Vol. 24, No. 2, 721-759, 2014.

[25] Kim, J.K., Sontag, E.D.: *Reduction of multiscale stochastic biochemical reaction networks using exact moment derivation*, PLOS Computational Biology Vol.13(6): e1005571, 2017.

[26] Knuth, D.E., Graham R.L., Patashnik O.: *Concrete mathematics*, Addison Wesley, 1989.

[27] Koshland, D.E., Némethy, G., Filmer, D.: *Comparison of experimental binding data and theoretical models in proteins containing subunits*, Biochemistry 1966 Jan; 5(1):365-85, 1966.

[28] Latchman, D.S.:*Transcription factors: an overview*, The International Journal of Biochemistry & Cell Biology 29 (12): 1305–12, 1997.

[29] Laurenzi, I.J.: *An analytical solution of the stochastic master equation for the reversible bimolecular reaction kinetics*, The Journal of Chemical Physics 08/2000; 113(8): 3315-3322.

[30] Lee, R.C., Feinbaum, R.L., Ambros, V.: *The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14*, Cell 75: 843-854, 1993.

[31] Lee, T.H., Maheshri, N.: *A regulatory role for repeated decoy transcription factor binding sites in target gene expression*, Mol. Syst. Biol. Vol. 8: 576, 2012.

[32] McAdams, H.H., Arkin, A.: *Stochastic mechanisms in gene expression*, P. Natl. Acad. Sci. USA Vol. 94: 814-819, 1997.

[33] Michaelis, L., Menten, M.L.: *Die Kinetik der Invertinwirkung*, Biochem Z 49: 333–369, 1913.

[34] Murray, J.D.: *Mathematical biology: An introduction*, Springer, 2002.

[35] Nordsieck, A., Lamb, W.E., Uhlenbeck, G.E.: *On the theory of cosmic-ray showers in the furry model and the fluctuation problem*, Physica 7: 344-360, 1940.

[36] Nussbaum, R.L., McInnes, R.R.; Willard, H.: *Thompson & Thompson Genetics in Medicine (8 ed.)*, Elsevier, 2016.

[37] Øksendal, B.: *Stochastic Differential Equations*, Springer, 2000.

[38] Feigelman, J., Marr, C., Popovic, N.: *A Case Study on the Use of Scale Separation-Based Analytic Propagators for Parameter Inference in Stochastic Gene Regulation*, J. Coupled Syst. Multiscale Dyn. 3(2): 164-173, 2015.

[39] Popovic, N., Marr, C., Swain, P.S.: *A geometric analysis of fast-slow models for stochastic gene expression*, Journal of Mathematical Biology 72(1): 87-122, 2016.

[40] Schauer, M., Heinrich, R.:*Quasi-steady-state approximation in the Mathematical Modeling of Biochemical Reaction Networks*, Mathematical Biosciences Vol. 65: 155-170, 1983.

[41] Ševčovič, D.: *Parciálne deferenciálne rovnice a ich aplikácie*, IRIS, 2008.

[42] Shampine, L. F., Gear, C. W.:*A user's view of solving stiff ordinary differential equations*, SIAM Review 21 (1), 1–17, 1979.

[43] Shampine, L. F., Reichelt, M. W.: *The MATLAB ODE Suite*, SIAM Journal on Scientific Computing, Vol. 18, 1–22, 1997.

[44] Shahrezaei, V., Swain, P.S.: *Analytical distributions for stochastic gene expression*, P. Natl. Acad. Sci. USA Vol. 105: 17256–17261, 2008.

[45] Singh, A., Hespanha, J.P., *Moment closure techniques for stochastic models in population biology*, Proc. Amer. Control Conf., 4730-4735, 2006.

[46] Soltani, M., et al.: *Nonspecific transcription factor binding can reduce noise in the expression of downstream proteins*, Physical Biology 12(2015), 055002.

[47] Stiefenhofer, M.: *Quasi-steady-state approximation for chemical reactional networks*, Journal of Mathematical Biology Vol. 36: 593-609, 1998.

[48] Taniguchi, Y., et al.: *Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells*, Science Vol. 329: 533-538, 2010. doi.org/10.2142/biophys.51.136

[49] Van Kampen N.G.: *Stochastic processes in physics and chemistry*, Elsevier Science B.V., 1992.

[50] Weiss, N.A.: *A Course in Probability*, Addison–Wesley, 2005.

[51] Wightman, B., Ha, I., Ruvkun, G.: *Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans*, Cell 75: 855-862, 1993.

[52] Wunderlich, Z., Mirny, L.A.: *Different gene regulation strategies revealed by analysis of binding motifs*, Trends Genet. 25(10): 434-440, 2009.

[53] Yu, J., Xiao, J., Ren, X., Lao, K., Xie, X.S.: *Probing gene expression in live cells, one protein molecule at a time*, Science Vol. 311: 1600-1603, 2006. doi.org/10.1126/science.1119623

List of articles related to the PhD thesis

1. Hojčka, M., Bokes, P.: *Non-monotonicity of Fano factor in a stochastic model for protein expression with sequesterisation at decoy binding sites*, Biomath 6, 1710217, 2017.
2. Bokes, P., Hojčka, M., Singh, A., : *Buffering gene expression noise by microRNA based regulation*, manuscript submitted for publication

Contributions at international conferences related to the PhD thesis

1. Hojčka, M., Bokes, P.: *Stochastic modelling of the interaction between protein and decoy binding sites on the DNA*, Poster on ECMTB/SMB 2016 Conference, Nottingham