www.elsevier.com/locate/yjtbi

Influence of gene copy number on self-regulated gene expression

Jakub Jędrak, Anna Ochab-Marcinek

PII: S0022-5193(16)30252-1
DOI: http://dx.doi.org/10.1016/j.jtbi.2016.08.018
Reference: YJTBI8786

To appear in: *Journal of Theoretical Biology*

Received date: 9 February 2016
Revised date: 15 July 2016
Accepted date: 11 August 2016

Cite this article as: Jakub Jędrak and Anna Ochab-Marcinek, Influence of gen copy number on self-regulated gene expression, *Journal of Theoretical Biology* http://dx.doi.org/10.1016/j.jtbi.2016.08.018

Influence of gene copy number on self-regulated gene expression

Jakub Jędrak and Anna Ochab-Marcinek*
Institute of Physical Chemistry, Polish Academy of Sciences, ul. Kasprzaka 44/52, 01-224 Warsaw, Poland

(Dated: August 11, 2016)

Using an analytically solvable stochastic model, we study the properties of a simple genetic circuit consisting of multiple copies of an self-regulating gene. We analyse how the variation in gene copy number and the mutations changing the auto-regulation strength affect the steady-state distribution of protein concentration.

We predict that one-reporter assay, an experimental method where the extrinsic noise level is inferred from the comparison of expression variance of a single and duplicated reporter gene, may give an incorrect estimation of the extrinsic noise contribution when applied to self-regulating genes.

We also show that an imperfect duplication of an auto-activated gene, changing the regulation strength of one of the copies, may lead to a hybrid, binary+graded response of these genes to external signal.

The analysis of relative changes in mean gene expression before and after duplication suggests that evolutionary accumulation of gene duplications may non-trivially depend on the inherent noisiness of a given gene, quantified by maximal mean frequency of bursts.

Moreover, we find that the dependence of gene expression noise on gene copy number and auto-regulation strength may qualitatively differ, e.g. in monotonicity, depending on whether the noise is measured by Fano factor or coefficient of variation. Thus, experimentally-based hypotheses linking gene expression noise and evolutionary optimisation may be ambiguous as they are dependent on the particular function chosen to quantify noise.

I. INTRODUCTION

Gene copy number variation is an ubiquitous phenomenon that manifests itself in multiplication of gene fragments, single genes, groups of genes up to whole genome. Duplicated genes contribute to gene evolution; subsequent mutations may turn one of gene copies into an inactive pseudo-gene, which may accumulate further mutations without affecting the phenotype [1–3]. Gene copies may be parts of either chromosomal or extra-chromosomal DNA. In bacterial cells, low-copy plasmids appear in the numbers of copies characteristic to plasmid type, and the numbers are conserved during cell division [1]. Bacterial plasmids, as well as the circular molecules of DNA found in mitochondria and chloroplasts may also appear in high numbers of copies (e.g., 20-40 for chloroplasts of higher plants [1]).

Variation in the number of copies of a particular gene in a living cell may strongly affect the concentration of protein encoded for by that gene. This in turn may have a profound impact on the phenotype, and hence on the fitness of the organism. The relationship between copy number variation and phenotype is of great interest in higher eukaryotes such as mammals, including humans, where gene copy number variation is known to be related not only to differences in concentrations of some enzymes (e.g., starch amylase, [4]) but also to several genetic diseases [5, 6] as well as cancer [7]. However, it is usually easier to study experimentally the effects of copy number variation in model unicellular organisms, such as *E. coli* or *S. cerevisiae*; strains of such organisms differing by gene copy number may be relatively easily constructed [8, 9]. Yet, within the existing mathematical models of gene expression [10–17], usually a single gene copy is considered, and the influence of gene copy number on gene expression is neglected. To the best of our knowledge, there are only few papers providing a theoretical description of the influence of copy number variation on gene expression [9, 18–23]

In particular, in Ref. [18], the influence of copy number variation on the gene expression level was studied in the case of four different network motifs, from a simple auto-activated gene (positive feedback) to more complicated, two- and three-gene circuits. This analysis, although thorough and throwing much light on the subject, was nonetheless based on deterministic approach so it neglected the molecular noise, inherent to as small biochemical systems as living cells. In the present paper, we will focus on how the noise produced by self-regulating gene depends on the copy number of that gene.

The dependence of gene expression noise on the strength of negative self-regulation of two gene copies was analysed in Refs. [20, 21]. It was concluded that gene expression noise, measured there by Fano factor, may prevent the evolution of strong negative auto-regulation in diploid cells, and this was proposed as a possible explanation of the observed difference in abundance of negative auto-regulation between *E. coli* (where negative auto-regulation is a frequently appearing network motif) and *S. cerevisiae* and other eukaryotic species (where it is much less frequent). The authors pointed out that it may also account for the fact that duplicated copies of negatively self-regulating genes are relatively rare in *E. coli*, despite the fact that roughly half of all known transcription factors of *E. coli* take part in negative auto-

*electronic address: ochab@ichf.edu.pl

regulation [21]. We will show, however, that the widely used quantitative measures of noise, Fano factor and coefficient of variation, may behave in a different way as the gene copy number is varied, so any conclusions about evolutionary selection based on gene expression noise are highly speculative as long as it is not known how the natural evolution measures the noise to select for its most advantageous amount.

Volfson et al. studied gene expression variability as depending on the gene copy number [9] in five strains of *S. cerevisiae* differing by the number of gene-promoter inserts of the GAL system. They used a simple scaling argument to determine whether the fluctuations in protein concentration were of intrinsic or extrinsic origin. According to the standard distinction between the two types of noise, *intrinsic noise* is defined as a side effect of the specific reactions that result in gene expression, when a small number of molecules takes part in these reactions. On the other hand, *extrinsic noise*, also affecting these reactions, is that produced by some unspecified external processes, e.g., fluctuations in the accessibility of transcriptional machinery or fluctuations of the environment. However, it should be noted that in [9] a tacit assumption was made that in order for the simple scaling to hold, the gene of interest should not be self-regulating, i.e., if there are any fluctuations of TF concentration affecting the state of the promoter, they are of extrinsic origin. In such a case, the mean protein concentration scales linearly with the gene copy number G, whereas the coefficient of variation (standard deviation divided by the mean) scales as $G^{-1/2}$ for purely intrinsic fluctuations and is independent on G for purely extrinsic fluctuations. In the present paper we will show, however, that this scaling cannot be assumed in the case of self-regulating genes because intrinsic noise in their products affect, at the same time, their promoters as the fluctuations in TF concentration.

We study how the expression of positively or negatively self-regulating gene (cf. Fig. 1) depends on gene copy number. We assume that this number does not change during the cell's life time and that the gene copies are coupled only by their protein products, being their own transcription factors (TFs). Another assumption is fast on/off switching of the promoter state that allows to describe its regulation by TFs in terms of Hill kinetics; recent experimental observations seem to support this assumption [24, 25]. We use the analytical framework proposed in Ref. [11]: The protein is assumed to be produced in exponentially distributed stochastic bursts [26, 27], whereas mRNA, whose dynamics is much faster than that of the protein, is not explicitly present in the model. Analytical expressions for the steady-state distribution of protein concentration can be derived for an arbitrary number of gene copies, not necessarily identical in terms of their affinity for TF, provided all copies are coding for the same protein. We analyse the influence of the mutations changing gene copy number and auto-regulation strength on the shape of the steady-state protein probability distribution.

The model analysed here is one of few analytically tractable stochastic models of multiple gene copies that can be constructed from the single-gene models currently known in the literature. Although the system we study is one of the simplest genetic circuits, we will show further on that it can still produce some behaviours that are unintuitive or have not been associated with this type of a gene system, and that the interpretation of its behaviours still gives rise to some confusion when experimental data are analysed in terms of the amount of noise present in the circuit.

In Section III A we model n identical copies of a self-regulating gene. We note that the two measures of noise in the system, Fano factor and coefficient of variation, may behave in a qualitatively different manner as functions of gene copy number. We also point out that experimental data acquired from the one-colour assay [28], performed on two gene copies, may be interpreted incorrectly in the case of self-regulating genes. In Section III B we study two non-identical copies of a self-regulating gene, which differ in their auto-regulation strength. We show that such a gene pair can show a mixed, binary-graded response to external signal, an effect that has not been, to date, associated with gene duplication. We show that mean expression of two gene copies can scale in a rather unintuitive way as compared to the mean expression of a single gene copy, depending on how much the two copies differ in their auto-regulation strength and depending on the maximal mean frequency of protein bursts, which may have an impact on evolutionary accumulation or extinction of gene duplications. We also point at possible qualitative differences in behaviour of Fano factor and coefficient of variation in the case of non-equivalent gene copies.

II. THEORY

A. Model

We model G copies of a gene in a cell, G being a fixed parameter. The copies may not be identical, due to mutations in the operator or promoter region of each copy. Still, we assume that the gene product (protein) is identical for all of them. We start from the following scheme:

$$\begin{array}{l} (\mathrm{DNA})_1 \xrightarrow{\tilde{k}_{11}(x)} \\ (\mathrm{DNA})_2 \xrightarrow{\tilde{k}_{12}(x)} \\ \qquad\quad \ldots \quad \mathrm{mRNA} \xrightarrow{k_2} \mathrm{Protein} \\ (\mathrm{DNA})_G \xrightarrow{\tilde{k}_{1G}(x)} \end{array} \tag{1}$$

$$\mathrm{mRNA} \xrightarrow{\gamma_1} \emptyset, \quad \mathrm{Protein} \xrightarrow{\gamma_2} \emptyset. \tag{2}$$

mRNA production takes place on each of G gene copies.

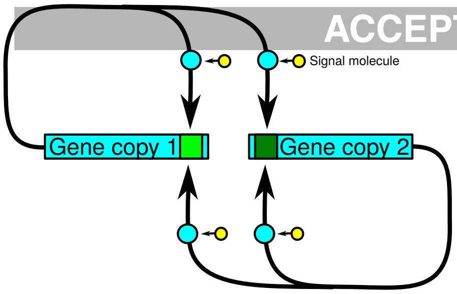

FIG. 1: Schematic representation of the system consisting of two self-regulating and mutually regulating gene copies. The copies can differ in their operator-TF affinity. The strength of TF binding to operators can be modified by signal molecules. Here, positive auto-regulation is shown (as depicted by large arrows), but we also consider the case of negative auto-regulation. More than two gene copies are also considered.

Transcription and translation adds mRNA and protein molecules to the common pool because they are assumed to be identical (1). Similarly, degradation processes of both mRNA and protein are common for the products of all gene copies (2).

Both translation as well as mRNA and protein degradation processes (2) are treated here as simple first-order reactions, with the rate constants k_2, γ_1 and γ_2, respectively (see Table I in Appendix A). However, due to auto-regulation (Fig. 1), transcription rates depend on the protein concentration x; the effective rate constants are given by $\tilde{k}_{1j}(x) = k_{1j} h_j(x)$, where k_{1j} is the bare rate constant and $h_j(x)$ is the transfer function of the j-th gene copy

$$h_j(x) = (1 - \epsilon_j) H_j(x) + \epsilon_j, \qquad j = 1, 2, \ldots, G. \qquad (3)$$

$\epsilon_j = k_{1j\epsilon}/k_{1j}$ is the measure of transcriptional leakage, and thus $\epsilon_j < h_j(x) < 1$; k_{1j} has an interpretation of the transcription rate constant for a fully active operator of j-th copy, whereas $k_{1j\epsilon}$ is the corresponding quantity for inactive operator (basal transcription), cf. Appendix A and Ref. [17].

The present model assumes that TF binding to the operator is governed by Hill kinetics, i.e. the binding/unbinding rates of a TF molecule to the operator are fast compared to the time scales of other reactions [13, 29]. In the case of cooperative TF binding, the regulatory function $H_j(x)$ in Eq. (3) is given by

$$H_j(x) = \left[1 + \left(\frac{x}{K_j} \right)^{n_j} \right]^{-1}. \qquad (4)$$

Cooperative TF binding means that the TFs effectively activate/repress the gene only when n_j of them are bound to the operator, e.g. when the TF occurs as multimer or when there are n_j TF binding sites on the operator and each of them, when occupied, makes it more probable for the TF to bind to other binding sites. Cooperativity n_j thus governs the steepness of $H_j(x)$, whereas K_j measures the regulation strength of the j-th gene copy. $n_j > 0$ denotes negative auto-regulation and for $n_j < 0$ the feedback is positive. Note that the Hill function $H(x)$ in Eq. (3) is multiplied by the $(1 - \epsilon)$ factor. This is in contrast with the formulation of Ref. [11], where nonzero leakage introduced only the additive term (ϵ). According to the rules of chemical kinetics, the present formulation is universal (its derivation being explained in detail in [17]) whereas that of Ref. [11] is only valid for small leakage.

Because usually both the mRNA production and degradation reactions are much faster than the corresponding processes for the protein, mRNA concentration is assumed to be a fast degree of freedom and is eliminated from the model [11, 12] (see also [14, 30–33] for detailed studies of time scale reduction from the full kinetic scheme in related models). In effect, protein production takes the form of stochastic bursts of a random size [11]. In the case of G gene copies, the probability $p(x,t)$ that at the time t the protein concentration is equal to x satisfies the Master equation

$$\begin{aligned} \frac{\partial p(x,t)}{\partial t} &= \gamma_2 \sum_{j=1}^{G} a_j \int_0^x w(x - x') h_j(x') p(x', t) dx' \\ &+ \gamma_2 \frac{\partial}{\partial x} [x p(x,t)]. \qquad (5) \end{aligned}$$

In the above, the protein concentration $x \geq 0$ is a continuous variable, u is the burst size, $w(u) = \nu(u) - \delta(u)$, where $\nu(u) = (1/b)\exp(-u/b)$ is the burst size probability distribution (note that the burst sizes are identically distributed for each gene copy), whereas a_j and b are defined by

$$a_j \equiv \frac{k_{1j}}{\gamma_2}, \qquad b \equiv \frac{k_2}{\gamma_1}, \qquad (6)$$

whereas $\delta(u)$ is Dirac delta distribution [11].

The stationary solution of Eq. (5), with the normalisation constant A, follows from Eq. (8) of Ref. [11]:

$$p(x) = A x^{-1} e^{-x/b} \prod_{j=1}^{G} \exp\left[a_j \int \frac{h_j(x)}{x} dx \right]. \qquad (7)$$

In the case of cooperative TF binding, $H_i(x)$ is given by Eq. (4) and from Eq. (7) we obtain

$$p(x) = A x^{-1} e^{-x/b} \prod_{j=1}^{G} x^{a_j} H_j(x)^{\frac{a_j(1-\epsilon_i)}{n_j}}. \qquad (8)$$

(The functional form of $p(x)$ (7) for non-cooperative TF binding is given in Appendix B.)

It should be noted that, in the present model, the bursting of each gene copy is a Poisson process, independent from the bursting of all other copies. Thus, their protein production rates are coupled only by the common pool of proteins that regulate the genes as their TFs.

B. Terminology

In this subsection we briefly explain the meaning of the terms that will be used further on in the paper.

Influence of external signal on gene regulation. TF can bind one or more signalling (effector) molecules (Fig. 1) or undergo phosphorylation, which changes the TF affinity to operator [17, 29]. In our model, the presence of signalling molecules is taken into account only implicitly, by assuming that the value of K_j in Eq. (4) depends not only on the TF-operator affinity but also on the fraction of active TFs that are able to bind the operator. This fraction depends on the concentration of the signalling molecules (see Appendix A in this paper and Ref. [17], Appendix A therein, for details). In other words, K_j, which quantifies the steepness of the Hill function, can be used as a measure of the intensity of an external signal that activates or deactivates the TFs.

Unimodal vs. bimodal distributions. A distribution of concentrations of a protein in cell population is unimodal when it has a single maximum and it is bimodal when it has two maxima.

Graded response vs. binary response. This concept concerns the changes of the distribution's shape due to variation of the signal intensity that defines the fraction of active TFs able to control the promoter. As the signal level is varied, gene expression varies between its minimal and maximal values. When the protein distribution is unimodal for all signal intensities, such that the signal level only defines the position of the single peak of the distribution, then the response of the gene is graded. On the other hand, the response is binary when the protein distribution changes its shape from unimodal at minimal expression to bimodal at intermediate expression level, and then it settles down again to a fixed unimodal distribution at maximal expression [17]. Further on, we will show that a mixed response is also possible, if, after the unimodal-bimodal-unimodal transition, the distribution does not become fixed but it shifts, now in a graded manner, towards some higher maximum of gene expression.

III. RESULTS

In this Section we present the results obtained by numerical evaluation of the probability distributions (8) or their moments.

A. Identical gene copies

The assumption of equivalent gene copies is legitimate when the differences between local genetic context (neighbourhood of each gene copy) are negligible, and in the case of some engineered genetic circuits [8].

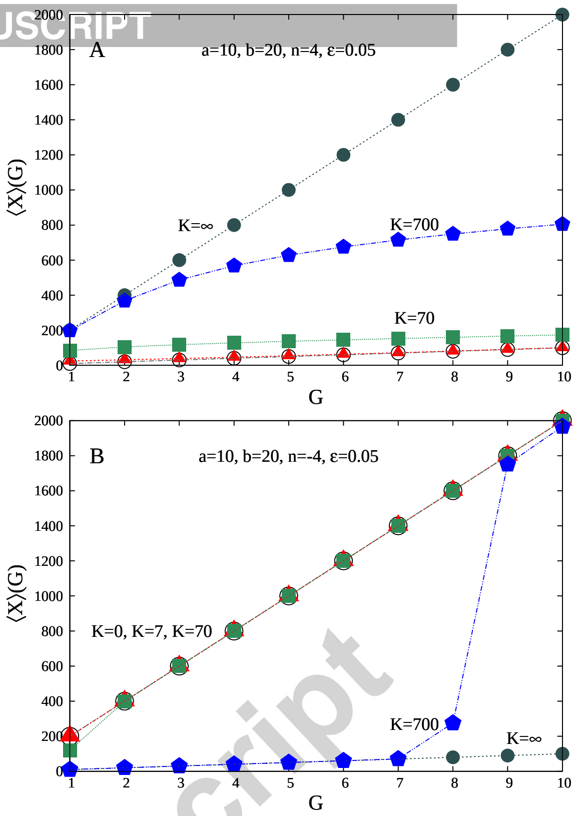

FIG. 2: Average protein number may depend on gene copy number in a nonlinear manner in self-regulating genes. A: Negative auto-regulation ($n = 4$). B: Positive auto-regulation ($n = -4$). The abrupt increase for $K = 700$ and $G = 8$ is due to the transition of the protein number distribution through bimodality, cf. Fig. 10 in Appendix C. Feedback strength parameter $K = 0$ (empty circles), $K = 7$ (triangles), $K = 70$ (squares), $K = 700$ (pentagons), and $K = \infty$ (full circles). Maximum mean burst frequency $a = 10$. Mean burst size $b = 20$. Leakage $\epsilon = 0.05$. Lines provide guide for the eye only.

1. *The maximum burst frequency scales linearly with gene copy number*

From Eq. (7) it follows that if the regulatory functions of each gene are identical, $h_j(x) = h(x)$, then their burst frequencies simply add up. In particular, if the whole gene (i.e. its protein-coding and regulatory parts) is present in G copies such that the maximum burst frequency $a_j = a$, then the system is equivalent to a single copy of a gene, with the parameter re-scaling:

$$a \to Ga. \tag{9}$$

For self-regulating genes, the probability density function for the protein number reads therefore

$$p_G(x) \;=\; Ax^{Ga-1}e^{-x/b}\left[1+\left(\frac{x}{K}\right)^n\right]^{-\frac{Ga(1-\epsilon)}{n}}. \tag{10}$$

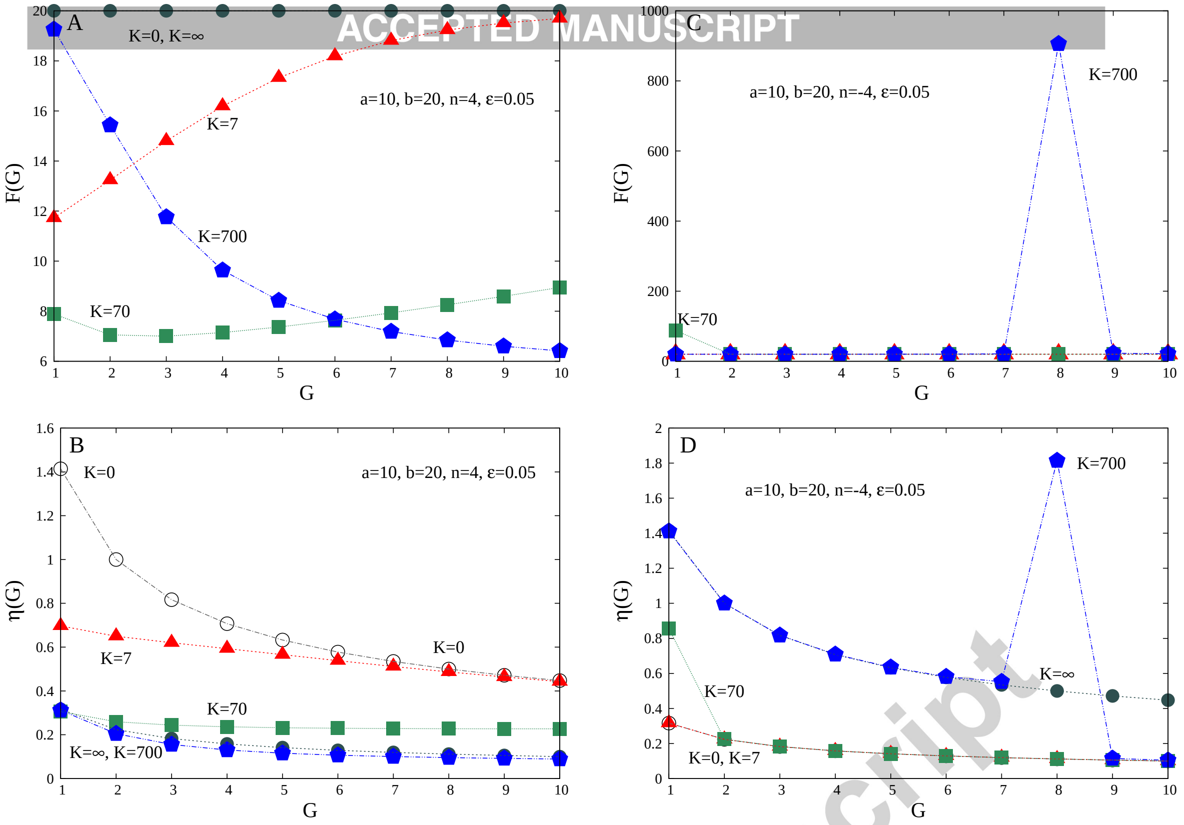

FIG. 3: In self-regulating genes, Fano factor and by coefficient of variation may depend on gene copy number in a qualitatively different manner. A, B: Negative auto-regulation, $n = 4$. A: Depending on the feedback strength parameter K, Fano factor $F = \sigma^2/\langle x\rangle$ may both decrease, increase or vary in a non-monotonous manner as gene copy number G is varied. B: Coefficient of variation $\eta = \sigma/\langle x\rangle$ is a monotonically decreasing function of gene copy number G. C, D: Positive auto-regulation, $n = -4$. Here, for $K = 700$, Fano factor $F(G)$ has just one maximum (C), whereas the coefficient of variation $\eta(G)$ has two clear maxima (D). The sharp maximum for $K = 700$ and $G = 8$ is due to the transition of the protein number distribution through bimodality, cf. Fig. 10 in Appendix C. In absence of gene regulation ($K = 0$ and $K = \infty$), $F = b$ and $\eta \sim G^{-1/2}$. For negatively self-regulating genes, $F(G) < b$ and for positive auto-regulation, $F(G) > b$. Parameters and the corresponding symbols are same as in Fig. 2.

2. *Non-linear scaling of the average protein number, Fano factor and coefficient of variation with gene copy number*

Cell fitness may depend not only on protein concentration, but also on the level of gene expression noise. Two standard quantitative measures of gene expression noise, Fano factor $F = \sigma^2/\mu_1$ and coefficient of variation $\eta = \sigma/\mu_1$, are used interchangeably in the literature [9, 11, 26, 27]. F is a natural quantity to measure deviation of a given probability distribution from Poisson distribution, for which $F = 1$. Under the assumption of no extrinsic noise, for unregulated gene expression $h(x) = const.$ and $p_G(x)$ is then a gamma distribution [11]. The mean protein number and the measures of gene expression noise have then a simple dependence on G: $\langle x\rangle \sim G^1$, $\eta \sim G^{-\frac{1}{2}}$, $F \sim G^0$ [9].

Self-regulation leads to deviations from the above scaling. The dependence of the average protein number $\langle x\rangle$ on gene copy number G is in general non-linear (Fig. 2). Fano factor F is no longer independent on G (Figs. 3A,C); coefficient of variation η no longer scales like $G^{-1/2}$ (Figs. 3B,D).

What is striking, the influence of gene copy number G on gene expression noise depends on the particular measure of noise (Fig. 3). In some of the considered cases F and η exhibit different qualitative dependences on G: For example, in negatively self-regulating genes, η may decrease while at the same time F is an increasing, decreasing or even non-monotonous function of G (Fig. 3A). For positive auto-regulation, $\eta(G)$ may have two maxima whereas $F(G)$ has just one clear maximum (Fig. 3C,D, $K = 700$). In the case of positive auto-regulation, it is in accordance with intuition that the abrupt changes shared by both measures of noise, $F(G)$ and $\eta(G)$, are associated with transition of the protein number distribution through bimodality (cf. Fig. 10 in Appendix C). However, other cases of non-monotonic behaviours of $F(G)$ and $\eta(G)$, including those for negative auto-regulation, are non-intuitive. Since the behaviours of the two measures of noise may differ quite significantly, statements like 'gene duplication increases noise of protein distribution' are meaningless until a particular measure of noise is chosen.

3. *Interpretation of one-reporter assay.*

In a large-scale experiment, Stewart-Ornstein et al. [28] measured the contribution of extrinsic noise in the expression of *S. cerevisiae* genes using the one-reporter assay. The classical two-reporter assay [34] consists in measurement of expression of two reporter genes that produce fluorescent proteins of different colours and the correlation between their fluorescence levels provides the information about the intensity of extrinsic noise that affects globally both promoters. The concept of the one-reporter assay, instead, consists in a comparison of statistics of the expression level x_1 of a single reporter gene with the statistics of the expression level $x_1 + x_2$ of two copies of that same reporter gene, both producing identical fluorescent proteins. Extrinsic noise was defined in [28] as $[cov(x_1, x_2)/(\langle x_1\rangle\langle x_2\rangle)]^{1/2}$, where the covariance is defined by the variances of expression of a single gene copy and two identical gene copies [9]:

$$cov(x_1, x_2) = [var(x_1 + x_2) - 2var(x_1)]/2, \qquad (11)$$

assuming that $\langle x_1\rangle = \langle x_2\rangle$ and $var(x_1) = var(x_2)$, because the products of the two gene copies, and the copies themselves, are identical. However, this definition of extrinsic noise becomes problematic if applied to *self-regulating* genes. Here, $var(x_1 + x_2) \neq 2var(x_1)$ even in absence of any extrinsic factors affecting globally both gene copies. Moreover, it is possible that $cov(x_1, x_2) < 0$, which would yield the square root of a negative number as the value of extrinsic noise, as defined according to [28]. In Fig. 11, Appendix D, we show examples of negative covariance produced by negatively and positively self-regulating genes.

It should be noted that the occurrence of negative covariance is, in itself, nothing unusual. What is problematic, is the definition of extrinsic noise as measured by the covariance, because it implies that zero covariance should indicate zero extrinsic noise. We therefore argue that interpretation of the experimental results from one-reporter assay in terms of intrinsic and extrinsic noise must be done with caution: A distinction is needed between (i) extrinsic noise as a factor *external to the promoter only*, which affects the state of the promoter, e.g. by the concentration of TFs [13], even if these TFs are produced by the gene they regulate, and (ii) extrinsic noise as a global factor affecting both gene copies *independently of their expression* (e.g., the variability in concentration of RNA polymerases or ribosomes).

One can, for example, imagine that two gene copies are self-regulating, which causes negative covariance of their expression, but simultaneously they are affected by a global noise source that increases the covariance, in such a way that the covariance sums out to zero. The interpretation of this result using the definition of extrinsic noise as measured by covariance, which was proposed by Stewart-Ornstein et al. [28], would lead to an erroneous conclusion that these genes are not affected by extrinsic noise whatsoever.

We show an example of such a situation in Fig. 4, where one and two copies of a positively self-regulating gene are additionally affected by global fluctuations in the mean size of a protein burst b, e.g. due to varying ribosome concentration in cells. We assume that b is gamma distributed,

$$g(b) = \frac{b^{k-1}e^{-b/\theta}}{\theta^k\Gamma(k)}, \qquad (12)$$

with $\langle b\rangle = k\theta$ and $var(b) = k\theta^2$. Then, the distribution of proteins in cell population [35]

$$q_G(x) = \int_0^\infty p_G(x, b)g(b)db. \qquad (13)$$

At certain width of the distribution of b, the covariance $cov(x_1, x_2)$ between the expression of one and two gene copies is zero, because the global fluctuations in ribosome concentration compensate the negative covariance that was the result of self-regulation. The fact that $cov(x_1, x_2) = 0$ does not imply here that extrinsic noise is absent.

B. Two non-equivalent gene copies

We now turn to the situation when the promoters or operators of different gene copies are not identical. This may happen due to mutational changes in one of the initially identical copies or due to mutations leading to duplication of an incomplete gene, with missing fragments of the regulatory parts. Gene duplication may also result in two copies which are nonequivalent due to their different neighbourhood (different genetic context).

For simplicity, we confine our attention to two gene copies ($G = 2$). We are interested in the effects of mutations affecting TF affinity to the operator region of one of the two copies, such that $K_1 \neq K_2$. For cooperative TF binding, assuming identical b for both copies, the steady-state distribution of protein concentration is given by Eq. (8) for $G = 2$:

$$p(x) = Ae^{-x/b}x^{a_1+a_2-1}[H_1(x)]^{\frac{a_1(1-\epsilon_1)}{n_1}}[H_2(x)]^{\frac{a_2(1-\epsilon_2)}{n_2}}. \qquad (14)$$

In contrast to the case of equivalent gene copies, $p(x)$ (14) cannot be obtained from the single gene copy case $p_1(x)$ (10) just by the simple scaling (9), even for $n_1 = n_2 \equiv n$, $a_1 = a_2 \equiv a$, and $\epsilon_1 = \epsilon_2$ (these equalities will hold further on). In the following considerations, we have chosen example values of parameters, $n = \pm 4$ and $b = 20$.

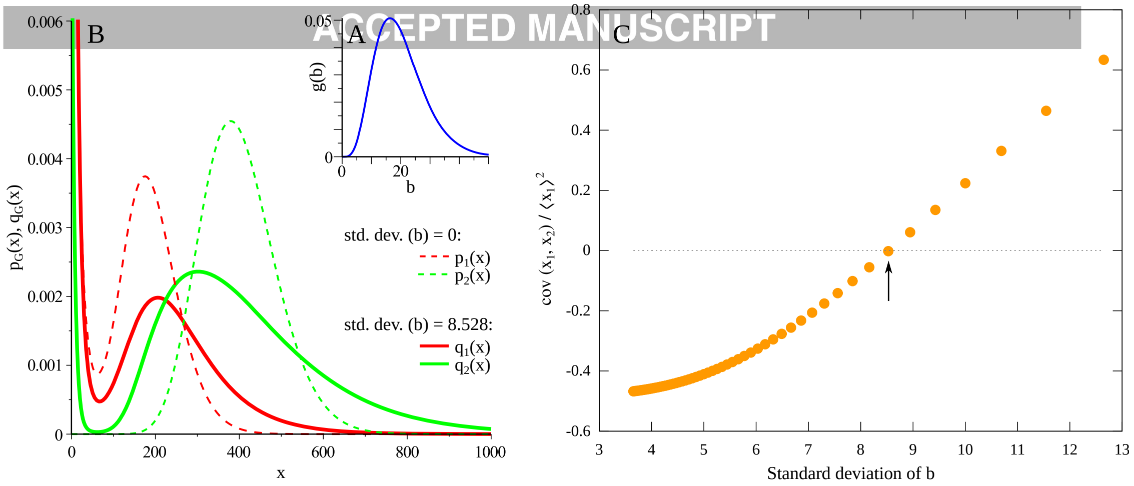

FIG. 4: Combined effect of transcription-factor noise in self-regulating genes and global extrinsic noise may cause zero covariance between the expression of one and two gene copies, which means that the covariance is not a good indicator of the presence of extrinsic noise, e.g. in one-reporter assay as presented in [28]. A: Extrinsic noise modelled by gamma-distributed fluctuations in mean protein burst size, b, e.g. due to a variable concentration of ribosomes. The distribution $g(b)$ (Eq. (12)) has parameters at which the covariance (11) is approximately equal to 0: $\langle b \rangle = 20$, $k = 5.5$, $\theta = \langle b \rangle / k$, $(var(b))^{1/2} = \theta k^{1/2}$. B: Protein distributions with the contribution of the extrinsic noise $g(b)$ for a single and duplicated self-regulating gene ($q_1(x)$ and $q_2(x)$, Eq. (13), solid lines). The clear bimodality of $q_2(x)$ is the effect of strongly bimodal contributions for some values of $b < 20$. For comparison, protein distributions for zero extrinsic noise are shown ($p_1(x)$ and $p_2(x)$ with non-fluctuating $b = 20$, dashed lines). Parameters: $a = 10$, $K = 70$, $\epsilon = 0.05$, $n = -4$. C: Covariance between the expression of one and two gene copies (Eq. (11), re-scaled by the mean protein number squared), as a function of varying parameters k and θ in $g(b)$, such that the mean value of b is fixed: $\langle b \rangle = k\theta = 20$. The arrow indicates the level of extrinsic noise shown in Fig. A and by solid lines in Fig. B, where the fluctuations of b compensate the transcription-factor noise, so that the covariance is very close to zero. This example shows that zero covariance does not imply that the extrinsic noise is zero.

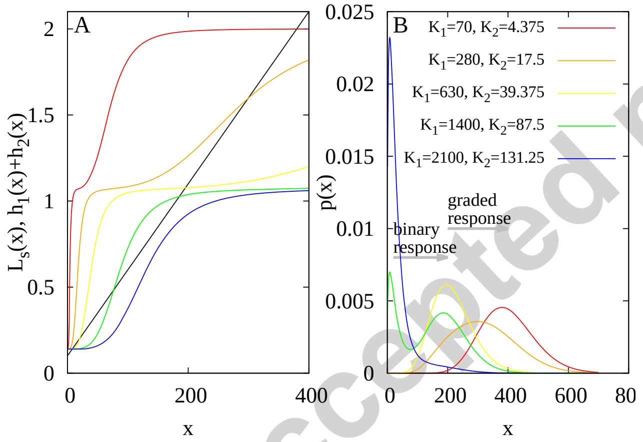

FIG. 5: When two positively self-regulating gene copies have different sensitivities to TF, the geometric construction (A) may predict a mixed, binary+graded, response (B). Binary response is seen for the distribution peaks in the range $0 < x < 200$, and graded response for $200 < x < 400$. Parameters: $n = -4$, $a = 10$, $b = 20$, $\epsilon_1 = \epsilon_2 = 0.07$.

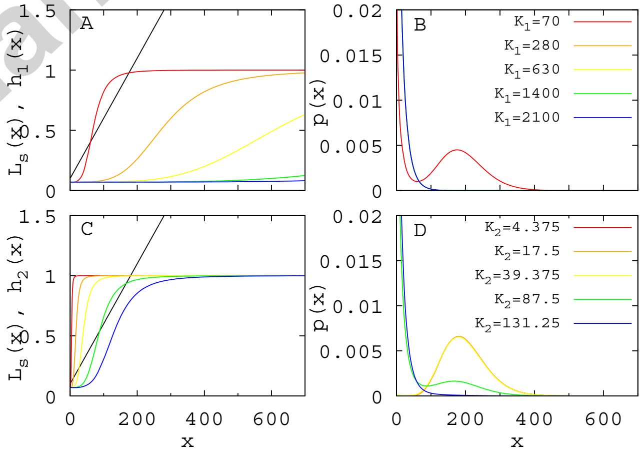

FIG. 6: Each of the genes, whose collective behaviour was shown in Fig. 5, has a binary response when present in the cell in a single copy. Parameters are the same as in Fig. 5. In Fig. B, the orange, yellow, green and blue curves overlap. In Fig. D, the yellow, orange and red curves overlap.

1. Non-equivalent copies of a self-regulating gene may have a mixed binary+graded response to a signal.

A well-known fact is that the distribution of protein concentrations produced by a positively self-regulating gene can be bimodal [11, 17, 36, 37]. And thus, the conditions for bimodality also hold for G identical copies of such a gene, with the re-scaling (9). For easier visualisation of the parameter regions where the bimodality is present, one can use a geometric construction that indicates the number and positions of the extrema of $p_G(x)$ [17]:

$$h(x) = \frac{1}{abG}x + \frac{1}{aG}. \tag{15}$$

The positions of the intersections of the transfer function $h(x)$ and the straight line corresponds to the positions of the extrema of the distribution. If the construction shows two intersections of $h(x)$ and the straight line, then there is one additional maximum in $x = 0$. It should be noted, however, that if the geometric construction predicts the mathematical fact of existence of multiple extrema, they may still not always be clearly visible on the plot of distributions: One maximum may be much smaller than the other even if the points of the intersection seem to be well separated on the plot of the geometric construction. A particular example illustrating such a situation of apparent unimodality is presented in Fig. 10, Appendix C. Yet, the geometric construction is a convenient tool to gain a qualitative understanding of the system's behaviour (see also [37] for a more detailed analysis of distribution properties based on the construction). The construction turns out to be especially instructive for the case of two non-equivalent gene copies: Now, it contains two regulatory functions of both genes,

$$\frac{1}{ab}x + \frac{1}{a} \ = \ h_1(x) + h_2(x), \tag{16}$$

and it allows us to visualise an example of a nontrivial response of the two-gene system to a varying signal that modifies the TF binding strength [17]. In Fig. 5, the sensitivities of both gene copies to TF differ by the factor of 16, which is reflected by the corresponding ratio of K_1 to K_2. An external signal of a certain intensity, e.g. the presence of certain concentration of ligand that binds to TF, or phosphorylation of a certain fraction of TFs, changes proportionally both coefficients K_1 and K_2, which causes the change of the steepness of the regulatory functions, $h_1(x) + h_2(x)$. For positive self-regulation, the geometric construction predicts that when both gene copies have different sensitivities to TF, a mixed response to the signal is possible, which is a combination of binary and graded responses. First, the more sensitive copy responds in a binary manner to the signal: The probability mass is transferred between two peaks of $p(x)$. But when the binary response is over, i.e., $p(x)$ becomes again unimodal, then gene expression does not saturate at the fixed level. Instead, the single peak moves towards an even higher expression level, now reflecting the (graded) response of the second, less sensitive, gene copy. One might naively expect that this hybrid behaviour occurs because one of the genes has graded response and the other has binary response. However, this is not the case: In our example, each gene has a binary response when present in the cell in just one copy (Fig. 6). The mixed type of cellular response was experimentally found in different contexts [38, 39], but, to date, it has not been associated with gene duplication.

2. Evolutionary accumulation of gene duplications may non-trivially depend on the inherent frequency of bursting of a given gene

Since both copies of the self-regulating gene also regulate each other, the mean expression $\langle x_1 + x_2 \rangle$ of the two copies is, in general, not equal to the double mean expression $\langle x_1 \rangle$ in cells where just a single gene copy is present. The ratio $\langle x_1 + x_2 \rangle / \langle x_1 \rangle$ depends on the regulation strengths K_1, K_2 of both genes but also on the parameter a that describes the maximum mean burst frequency and can be considered, alongside Fano factor and coefficient of variation, as another measure of the noise present in the system – the larger a, the closer is the behaviour of the system to the deterministic model [17]. Fig. 7 shows the behaviour of $\langle x_1 + x_2 \rangle / \langle x_1 \rangle$ as a function of K_2/K_1 for an example set of parameters. For a given a, the function increases (for auto-repression) or decreases (for auto-activation), but the magnitude and threshold of this changes for each value of a in a rather unintuitive way. One can, on the other hand, see that $\langle x_1 + x_2 \rangle / \langle x_1 \rangle$ depends non-monotonically on the noisiness the system, measured by $1/a$. In Fig. 8 we plot cross-sections of Fig. 7 as functions of a. Based on Fig. 7, we can discuss two scenarios of gene duplication:

1. The gene copies are identical (green lines in Fig. 8). In general, gene duplication increases total gene expression, but by a different factor, depending on regulation type and a. Duplication of a negatively self-regulating gene (Fig. 8A, green line) causes a smaller change in its expression than duplication of a positively self-regulating gene (Fig. 8B, green line). This effect is easy to explain intuitively: in the former case, additional repressors are produced, in contrast to additional activators in the latter case. In the case of auto-activation, there is a value of a at which the increase of expression is maximal. This corresponds to activation of the previously inactive gene due to duplication. The effect can be intuitively visualised using the geometric construction, even though it shows extrema of the protein number distributions and not their means: Since the slope of $x/(ab) + 1/a$ in Eq. (16) depends on a, this parameter changes the relative distance between the points of intersection of the straight line with $h_1(x)$ and with $h_1(x) + h_2(x)$ (see Appendix E, Fig. 13).

If there exists a threshold of natural selection above which gene expression is too high and the cell is eliminated, then only those duplications will remain, for which the expression is increased the least. Thus, duplications of those auto-repressed genes, whose noisiness measured in $1/a$ is low, have a greater chance to survive (Fig. 8; see also the geometric construction in Appendix E, Fig. 14). On the other hand, for auto-activation, the duplications of genes with a small or large, but not intermediate, a have a greater chance to survive; this corresponds to situations when either the duplication does not suffice to induce the genes, or when the gene has already been induced before duplication.

2. The gene copies differ in their operator-TF affinity parameters K_i, which can occur when the new copy is placed in a different genetic context than the original (e.g. where the exposition of the operator to TFs is better/worse) or when the promoter is not fully copied. The changes lowering the operator-TF affinity ($K_2 > K_1$) are more probable. Intuitively, one can predict that the gene copy with such a defective operator will increase the total expression in the case of auto-repression, or decrease it in the case of auto-activation, as compared to the perfect duplication. However, interestingly, when the copies of a negatively self-regulating gene differ in their affinities to TF in a sufficiently large extent, then an optimal value of a occurs at which their expression increases the least as compared to the expression of a single gene copy (Fig. 8A, blue and magenta lines; see also the geometric construction in Appendix E, Fig. 15). This means that survival of a defective duplication is more probable for a rather small a (around its optimal value), differently than in the case of a perfect duplication, where large a increased the probability of survival.

On the other hand, in the case of auto-repressed genes, evolution may lead to accumulation of those rare cases of duplications in which the operator-TF affinity is increased (i.e., K_2 decreased). For auto-activated genes, such increase would not be evolutionarily preferred.

Accumulation of gene duplications may thus depend not only on the type of regulation (negative/positive) but also on the amount of noise in the system, measured by the maximum mean burst frequency a. This dependence is, however, non-trivial because it may be different for perfect and imperfect duplications.

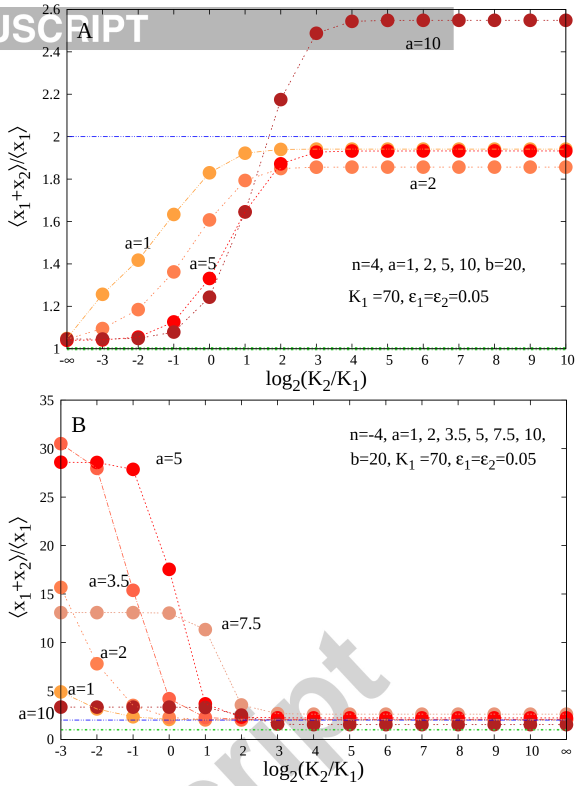

FIG. 7: Relative change in the average protein concentration before and after gene duplication, as a function relative affinity of both genes for TF, K_2/K_1. A: Negative auto-regulation, $n = 4$. B: Positive auto-regulation, $n = -4$. Parameters: $b = 20$, $n = 4$, $K_1 = 70$, and $\epsilon_1 = \epsilon_2 = 0.05$. Horizontal dashed lines mark the level of 1 (green) and 2 (blue) for comparison.

3. Fano factor and coefficient of variation behave differently as the function of TF-operator affinity

According to earlier findings, two identical copies of negatively self-regulating genes are characterised by a smaller gene expression noise than heterozygotes ($K_2 \neq K_1$) [20]. However, a careful analysis of the behaviour of our model at some exemplary values of parameters (Figs. 9) shows that this rule is not universal. We observe that Fano factor F and coefficient of variation η may behave in a different way as the functions of the relative promoter sensitivities K_1/K_2 and depending on the maximal burst frequency a. In the case of two copies of a negatively self-regulating gene (Fig. 9A,B), for large a, Fano factor has minima in the vicinity of the homozygous case, $K_2 = K_1$. For small a, the minimum of F occurs when there is a two-fold difference in sensitivity between the promoters (for $a = 1$, $K_1 = 70$, $K_2 \approx 35$). On the other hand, coefficient of variation shows shallower minima and their positions are different than those for Fano factor. For small noise, the minimum of η occurs when there is a two-fold difference in sensitivity between the promoters but at different values than in the case of F: (for $a = 10$, $K_1 = 70$, $K_2 \approx 140$). If gene pairs with $K_2 \ll K_1$ and $K_2 \gg K_1$ are compared, large differences in Fano factor occur for large a, but, at the same time, large differences in coefficient of variation occur for small a. In the case of two copies of a positively self-regulating gene (Fig. 9C,D), both measures of noise have maxima (roughly corresponding to the transition through bimodal distributions, see Fig. 16 in Appendix F) but again their positions differ for F and η. In the case of coefficient of variation, the maxima are less pronounced and

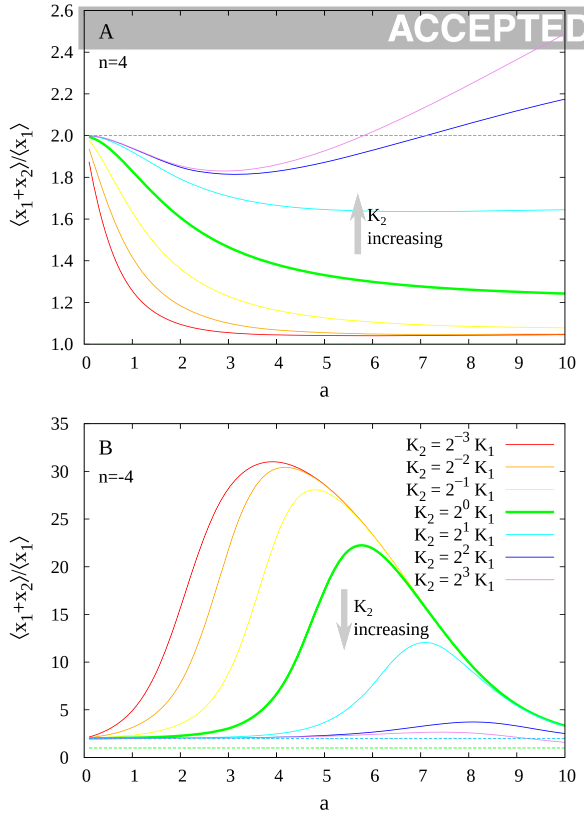

FIG. 8: Relative change in the average protein concentration before and after gene duplication, as a function of maximum mean burst frequency a, for various values of K_2/K_1. A: Negative auto-regulation, $n = 4$. B: Positive auto-regulation, $n = -4$. Parameters and the meaning of horizontal dashed lines are same as in Fig. 7.

they disappear above some value of a, which is however different than for the maxima of F. We also note that, for both negatively and positively self-regulating genes, coefficient of variation varies monotonically with a, whereas the behaviour of Fano factor is non-monotonic.

And therefore, if we attempted to draw any conclusions from these examples about evolutionary optimisation of promoter sensitivity with respect to noise, then these conclusions would differ depending on whether they are based on the behaviour of Fano factor or the coefficient of variation. If one of two initially identical genes undergoes mutation, these two measures of noise would give different predictions as to what type of mutation is more beneficial, the one increasing or the one decreasing the sensitivity of the promoter.

IV. DISCUSSION

A. Conclusions

In the present paper, we have studied the influence of gene copy number, auto-regulation strength, and transcriptional leakage on the properties of the simplest genetic circuit, a self-regulating gene.

Although this genetic circuit is extremely simple, the analysis of how it behaves depending on the number of gene copies may be crucial for correct interpretation of experimental results. In a large-scale experiment (456 genes), Stewart-Ornstein et al. [28] used the one-reporter assay to measure covariance between the expression of single genes and two identical copies of those genes in *Saccharomyces cerevisiae*. Adopting the ideas of Volfson et al. [9] who studied non-auto-regulated genes, the authors of [28] interpreted the covariance as a measure of extrinsic noise affecting the genes. However, if the studied genes were self-regulating, this interpretation would break down because, as we have shown in the present paper, the transcription-factor noise may cause negative covariance, whereas global extrinsic noise, e.g. due to cell-to-cell differences in ribosome concentration, may compensate it, such that the total covariance is zero. In that case, the interpretation proposed in [28] would lead to an erroneous conclusion that the genes are not affected by extrinsic noise.

An obvious observation is that, within the studied model, a system of multiple *identical* gene copies can be equally well interpreted as a single "super-gene", whose transfer function $h(x)$ is the same as in a single gene copy whose transcription rate is accordingly multiplied. On the other hand, when the gene copies are *non-identical* due to promoter mutation affecting the TF binding, the effective transfer function has a nontrivial shape (a case that is rather difficult to interpret in terms of some molecular mechanisms affecting a single "super-gene"). We have shown that this may lead to a mixed, binary+graded response of the gene system to external signal modulating the TF activity: In a certain range of the signal, the histogram of gene expression is bimodal with the height of the peaks varying as the signal is varied, but when that range is exceeded, the gene expression does not saturate. Instead, a single peak gradually changes its position as the signal intensity is further increased. This behaviour is the result of mutual regulation of both genes: It may occur even if each of the genes alone has a binary response when present in the cell in a single copy. The hybrid response was observed in different cellular contexts (nuclear phosphorylated ERK as well as Egr1 and its mRNA, induced by gonadotropin-releasing hormone in LβT2 mouse cells [38], phosphorylated Stat5 induced by erythropoietin in foetal erythroblasts [39]). However, to date, this type of response has not been associated with gene duplication.

Our analysis of the relative change in gene expression before and after gene duplication suggests that the evo-

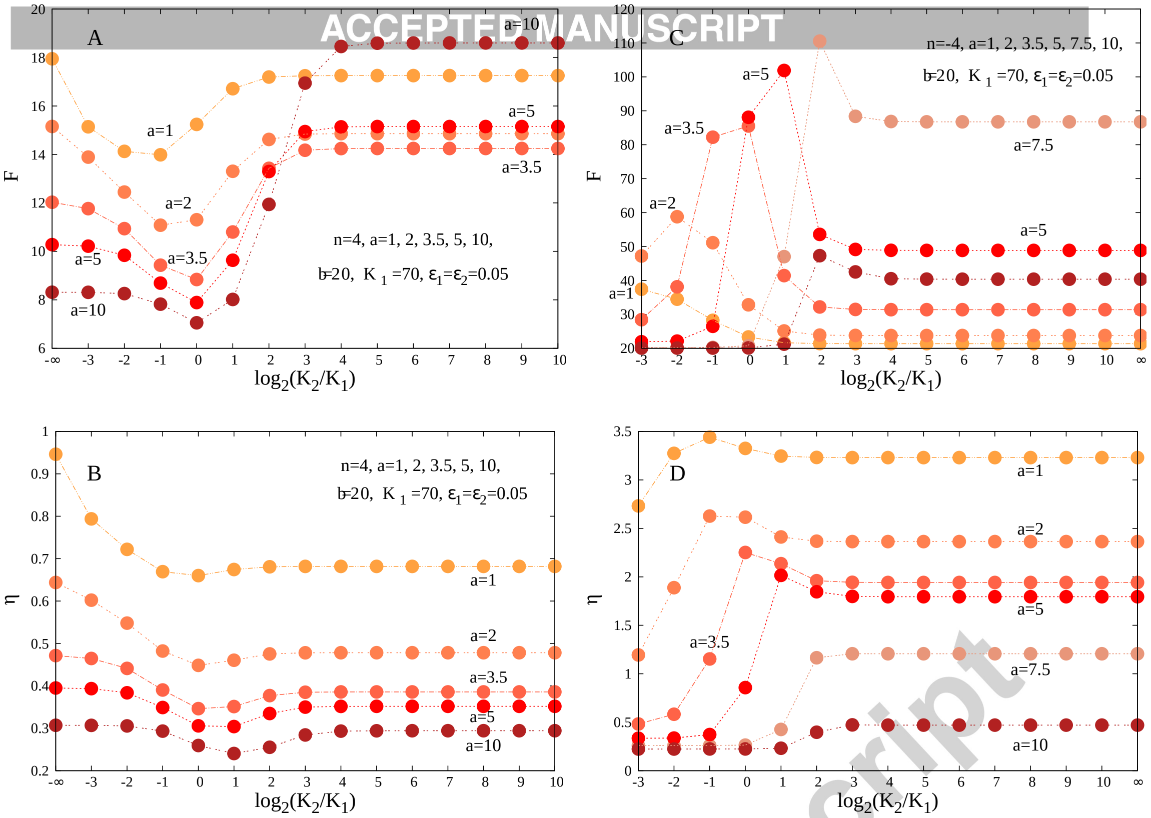

FIG. 9: Two non-equivalent copies of a negatively (A,B) and positively (C,D) self-regulating gene: Different measures of noise, Fano factor F and coefficient of variation η, may show differences in their behaviour as functions of the relative sensitivity K_2/K_1 of both promoters to auto-regulation. For negative auto-regulation, $n = 4$ (A,B), the positions and depth of minima are different for F and η. For positive auto-regulation, $n = -4$ (C,D), the maxima of both measures of noise roughly correspond to the transition through bimodal distributions, see Fig. 16. The exact positions and height of the maxima are, however, different for F and η. Additionally, for both positive and negative auto-regulation, F varies non-monotonically with a, whereas the dependence of η on a is monotonic. Parameters: $b = 20$, $K_1 = 70$, $\epsilon_1 = \epsilon_2 = 0.05$.

lutionary survival of additional gene copies may not only depend on whether the auto-regulation is negative or positive, but also on the amount of noise in the system, measured by the inherent maximal mean burst frequency a of a given gene. The dependence for perfect duplications (identical gene copies) may be different than for defective duplications (the operator of the new copy having a lower affinity for TF): In the case of perfect duplications of auto-repressed genes, there may be a preference for accumulation of such duplications when the genes are characterised by high burst frequency a. On the other hand, some cases of defective duplications may survive when the genes have an optimal, low burst frequency a. In the case of auto-activated genes, evolution may avoid accumulation of duplications of those gene for which a is in an intermediate regime because such duplications of an uninduced gene may lead to exceeding of the induction threshold. Finally, there may also be a (more obvious) preference for those rare cases of duplications of auto-repressed genes, in which the operator-TF affinity is increased, whereas such an increase would not be preferred in the case auto-activated genes. The above predictions can be tested experimentally by checking which types of gene duplications (perfect or imperfect) and in what types of genes (the ones with frequent or infrequent protein bursts) tend to accumulate in the course of evolution.

In order to investigate gene expression noise, we have computed two standard measures of noise (Fano factor F and the coefficient of variation η). It turns out that F and η behave differently as functions of gene copy number, and, in the case of two non-identical gene copies, as functions of the relative auto-regulation strengths of the two genes. Consequently, in any analysis of gene expression noise, the outcome depends on which measure of noise is used. This makes any statements on the influence of gene expression noise on cell fitness ambiguous. On one hand, it seems that coefficient of variation, as a dimensionless quantity, may be a more reasonable choice. On the other hand, even if the qualitative behaviour of $F(G)$ and $\eta(G)$ is similar, it is not guaranteed that a definite conclusion regarding the selective role of gene expression noise can be drawn. For example, in the case of bet-hedging strategy, cell fitness depends on the shape of the protein concentration distribution (bimodal vs. unimodal) in a way that not always can be captured by simple measures of noise like η or F (e.g., it is possible that a bimodal distribution with strongly defined peaks can have the same η as a wide unimodal distribution). We do not know which measure of gene expression (if any) is used by Nature to quantify the influence of noise on cell fitness, and it is likely that such measure is to be found individually for each system of interest.

B. Limitations of the model

The main limitation of the present approach is that it allows to treat only the case of gene copies coding for identical gene product. Since the gene copies are assumed to be coupled only by the total protein concentration, the model does not take into account other coupling mechanisms affecting the burst rates of all genes, e.g. general DNA remodelling. Also, the present formalism cannot be used to investigate more complicated genetic circuits (e.g., toggle switch). Moreover, it is likely that in real systems, the same point mutation may affect both the transcription rate (hence, burst frequency), auto-regulation strength, and basal transcription level. In such a case, the model here considered is only a first approximation; within a more involved description, some model parameters (a, ϵ, and K) should not be treated as independent. Another simplification is that the model is one-dimensional. This may cause neglection of some effects that are possible only in higher dimensions, e.g. oscillations. We have also assumed here that the gene copy number G is identical for all cells in the population. However, this may be not the case and we may deal with a distribution of G values, e.g. when high-copy plasmids are used to construct a multi-copy strains. In such a situation our model may be easily generalised by introducing a probability distribution $p(G)$ for different values of G, a conditional probability for finding x protein in a cell containing G gene copies, $p(x|G)$, and finally the joint probability $p(x,G) = p(x|G)p(G)$, $G \in \mathbb{N}$. The marginal probability distribution $p(x) \equiv \sum_G p(x,G)$ should be then used as a correct protein number probability distribution in the population. Since nuclear transport is neglected within the present model, our results seem to be more relevant to prokaryotes than to eukaryotes. However, in most papers devoted to copy number variation in eukaryotes, the division of the cell into nucleus and cytoplasm is not taken into account, and exactly the same models are used to describe gene expression in both groups of organisms. Most of proteins in *E. coli* appear in relatively high concentrations [40, 41] and therefore the discreteness of the protein number is not taken into account within the present approach. Our model with a continuous x variable may thus incorrectly describe systems containing small numbers of proteins. Note that this may also include the cases where the total number of TFs is large but the number of active TFs (not taken explicitly into account in our model) is very small. The presence of discrete states of the promoter is here taken into account only in an effective manner by making use of the Hill function. On the other hand, the discrete counterpart of the analytical framework proposed in Ref. [11] is known [15], and it seems adaptable to study the system of multiple copies of a self-regulating gene. Finally, it should be noted that the present model does not allow to study the changes of gene copy number G in time. We only compare stationary expression of gene systems containing different fixed numbers of gene copies. However, in some cases, the number of gene copies may change on the time scales as short as a fraction of a cell cycle (10^3-10^4 s), the most obvious example being chromosome replication during the replication cycle. In rapidly growing and dividing bacteria, DNA replication leads to a more than two-fold increase of the copy number of some genes (multi-forked chromosomes) [1] . Another example of a rapid copy number variation is the change of a viral genome copy number during multiple bacteriophage infections of bacteria [42, 43]. Modelling of time-dependent gene expression in such cases would require a different theoretical approach.

Acknowledgments

We would like to thank Marek Skoneczny and Marcin Tabaka for helpful discussions. The research was supported by the Ministry of Science and Higher Education Grant No. 0501/IP1/2013/72 (Iuventus Plus).

Transcription factor binding:

$${}_{\mathrm{j}}\mathrm{O} + \mathrm{nX} \overset{K_j}{\rightleftharpoons} {}_{\mathrm{j}}\mathrm{OX}_{\mathrm{n}}$$

mRNA synthesis/degradation:

Repressor	Activator
${}_{\mathrm{j}}\mathrm{O} \xrightarrow{k_{1\mathrm{j}}} \mathrm{Y} + {}_{\mathrm{j}}\mathrm{O}$	${}_{\mathrm{j}}\mathrm{O} \xrightarrow{k_{1\mathrm{j}\epsilon}} \mathrm{Y} + {}_{\mathrm{j}}\mathrm{O}$
${}_{\mathrm{j}}\mathrm{OX}_{\mathrm{n}} \xrightarrow{k_{1\mathrm{j}\epsilon}} \mathrm{Y} + {}_{\mathrm{j}}\mathrm{OX}_{\mathrm{n}}$	${}_{\mathrm{j}}\mathrm{OX}_{\mathrm{n}} \xrightarrow{k_{1\mathrm{j}}} \mathrm{Y} + {}_{\mathrm{j}}\mathrm{OX}_{\mathrm{n}}$

$$\mathrm{Y} \xrightarrow{\gamma_1} \varnothing$$

Transcription factor synthesis/degradation:

$$\mathrm{Y} \xrightarrow{k_2} \mathrm{X} + \mathrm{Y}$$

$$\mathrm{X} \xrightarrow{\gamma_2} \varnothing$$

TABLE I: Kinetic scheme that can be used to derive the effective kinetic description of (1) and (2). Y: mRNA, k_{1j}: rate of mRNA synthesis from the operator of the j-th gene copy in the active state, $k_{1j\epsilon}$: rate of mRNA synthesis from the operator of the j-th gene copy in the inactive state (leakage), γ_1: rate of mRNA degradation, k_2: rate of protein synthesis, γ_2: rate of protein degradation.

APPENDIX A: DETAILED SCHEME OF THE REACTIONS

In this Appendix and in Table I we present the detailed list of biochemical reactions used to derive effective kinetic description of transcription, translation and degradation or dilution of mRNA and protein as given by (1) and (2), making use of Hill kinetics.

For $|n| \geq 1$ cooperative TF binding corresponds to the situation where, for each gene copy, the probability of finding the j-th operator ${}_jO$ in any of the intermediate states, ${}_jOX_1$, ..., ${}_jOX_{n-1}$, is negligible. This assumption leads to the Hill function of the form (4), with terms proportional to x^i, $i = 1, \ldots, n-1$ in denominator being absent. Alternatively, $H_j(x)$ (4) may be obtained if we assume that TFs rapidly form a n-molecule complex.

For simplicity, reactions such as binding of a signalling molecule to TF, phosphorylation, multimerisation are not explicitly taken into account here. Generalisations of the present model with the signalling molecules explicitly included would be too complicated to be analytically solvable.

Yet, we can bypass this difficulty by allowing for the dependence of the parameters K_j of the present model on the concentration of the signalling molecules. Examples of such dependence can be found in Refs. [17, 29]. We make a general assumption that, when the cooperativity of TF binding is strong and binding of inactive TFs is negligible, then the parameter K_j in the Hill function for the $j-$th gene copy can be approximated as

$$K_j = \left(\frac{\kappa^{(1)}_{j,off}\kappa^{(2)}_{j,off}\cdots\kappa^{(n)}_{j,off}}{\kappa^{(1)}_{j,on}\kappa^{(2)}_{j,on}\cdots\kappa^{(n)}_{j,on}} \right)^{1/n} \frac{1}{f_a}, \tag{A1}$$

where f_a is the active fraction of TFs, $\kappa^{(i)}_{j,on}$ is the rate of binding of an active TF to the $i-$th binding site on the operator, and $\kappa^{(i)}_{j,off}$ is the corresponding unbinding rate (see Ref. [17], Appendix therein, for detailed derivation). Therefore, K_j contain both the information about TF binding affinity (which can change due to mutations) as well as the information about the active TF fraction (which can change due to varying signal level). If there is more than one gene gene copy and each of them has different TF-operator affinity, an external signal changes globally the active fraction of TF f_a, which means that

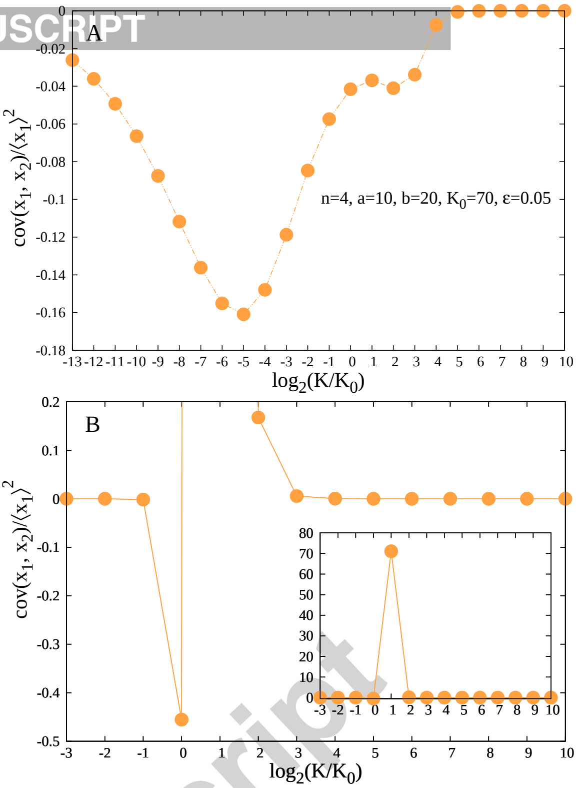

FIG. 11: Example normalised covariances (Eq. 11) of the expression of two identical copies of a negatively ($n = 4$, A) and positively ($n = -4$, B) self-regulating genes, depending on the parameter K that measures TF binding strength. A: For negative auto-regulation, the covariance is negative, and it tends to zero in the limit of "saturated" regulation, where both genes behave as independent copies that do not regulate each other. B: For positive self-regulation, the covariance can be negative or positive, which corresponds to the transition through bimodality of the distribution produced by a single gene copy or by two gene copies, see Fig. 12. Parameters: $a = 10$, $b = 20$, $\epsilon = 0.05$, $K_0 = 70$, $K = 2^{\xi}K_0$, $\xi = -13..10$ (K_0 and ξ being auxiliary variables that scale the value of K).

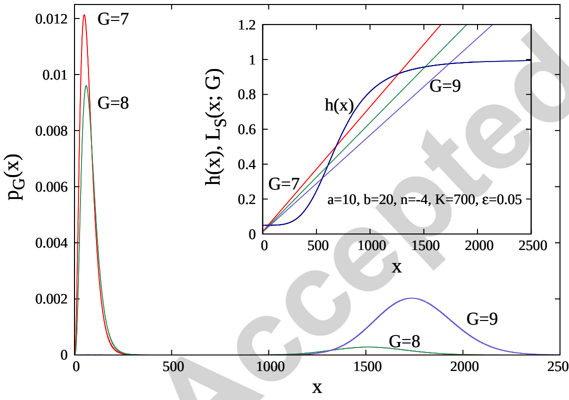

FIG. 10: Probability distributions $p_G(x)$ for $G = 7$, $G = 8$, and $G = 9$ in Fig. 3C,D, where the sharp maxima of F and η correspond to the transition through a strongly bimodal distribution. Parameters are same as in Fig. 3C,D. Inset: The corresponding geometric construction, revealing that all probability distributions are, strictly speaking, bimodal, but for both $G = 7$ and $G = 9$ the height of one of the maxima is orders of magnitude smaller than in the other. This shows that the geometric construction provides the necessary, but not sufficient criterion for visually distinguishable bimodality.

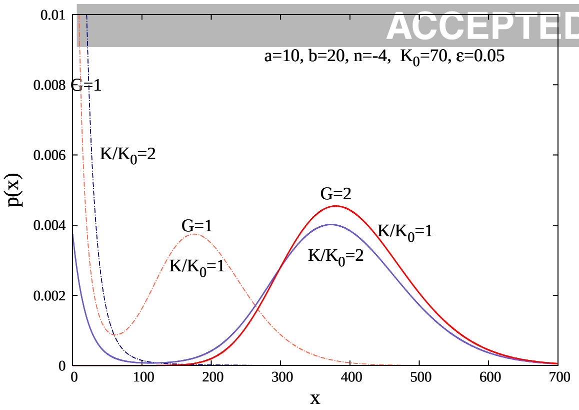

FIG. 12: Plots of $p_1(x)$ (dashed lines) and $p_2(x)$ (solid lines) for $K/K_0 = 1$ (red) and for $K/K_0 = 2$ (blue), i.e., corresponding to minimum and maximum of normalised covariance shown in Fig. 11B. $p_1(x)$ for $K/K_0 = 2$, denoted by blue dashed line, is unimodal, with maximum in 0.

the signal varies the parameters K_j in the same proportion for each gene.

Transcriptional leakage is modelled by

$$\epsilon_j = \frac{k_{1j\epsilon}}{k_{1j}}. \qquad \text{(A2)}$$

APPENDIX B: NON-COOPERATIVE TRANSCRIPTION FACTOR BINDING

We present here an analytical form of the steady-state distribution of protein concentration for the case of a non-cooperative TF binding. This case was not analysed in [11]. However, in some situations the assumption that TFs bind to operator independently may be more realistic than the limit of strongly cooperative TF binding. In the present case, the Hill function of a j-th gene copy reads

$$H_j(x) = \left[1 + \left(\frac{x}{K_j}\right)\right]^{-n_j}. \qquad \text{(B1)}$$

In order to find an explicit form of (7) for $h_j(x)$ given by (3) and (B1), we need the following result

$$\begin{aligned}\int \frac{h(x)}{x} dx &= \int \left(\frac{(1-\epsilon)}{x\,(1+x/K)^n} + \frac{\epsilon}{x}\right) dx \\ &= (1-\epsilon) \sum_{i=2}^{n} \frac{1}{i-1} \left(1 + \frac{x}{K}\right)^{1-i} \\ &+ \ln(x) - (1-\epsilon) \ln\left(1 + \frac{x}{K}\right). \qquad \text{(B2)}\end{aligned}$$

Eq. (B2) follows from the identity

$$\frac{1}{(s-1)s^m} = \frac{1}{(s-1)} - \sum_{l=1}^{m} \frac{1}{s^l}, \qquad \text{(B3)}$$

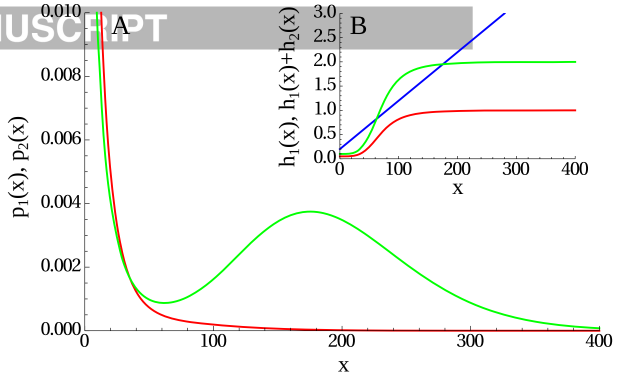

FIG. 13: Duplication of the non-induced positively self-regulating gene increases the number of TFs, which leads to induction of both genes in a subpopulation of cells (Fig. 8B, green line therein, $K_2 = K_1$), which is depicted by the bimodal distribution of protein number. A: Distributions of protein numbers before (red) and after duplication (green). B: Corresponding geometric construction, Eqs. 15 and 16. Red line: $h_1(x)$. Green line: $h_1(x)+h_2(x)$. Blue line: $\frac{1}{ab}x+\frac{1}{a}$. Parameters are same as in Fig. 8B.

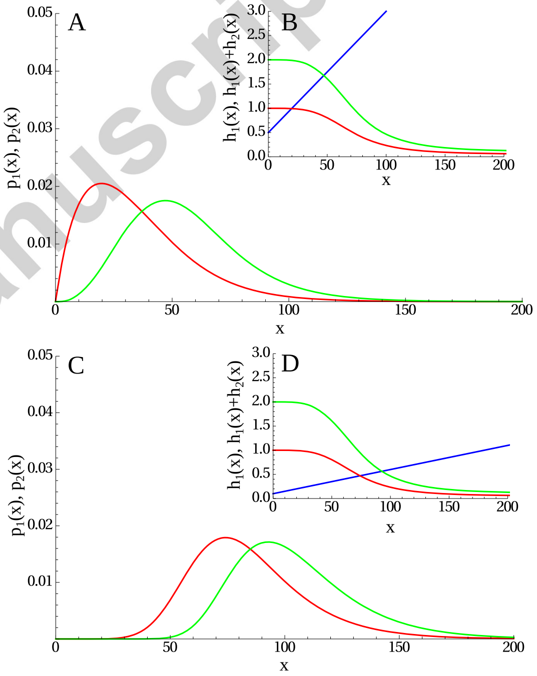

FIG. 14: Relative change in expression after duplication of a negatively self-regulating gene is the smaller, the greater is the maximal burst frequency a (Fig. 8A, green line therein, $K_2 = K_1$). Red and green lines denote quantities before and after duplication, correspondingly. A: Distributions of protein numbers, $a = 2$. B: Corresponding geometric construction. C: Distributions of protein numbers, $a = 10$. D: Corresponding geometric construction. Colors in the geometric constructions are same as in Fig. 13. Parameters are same as in Fig. 8A.

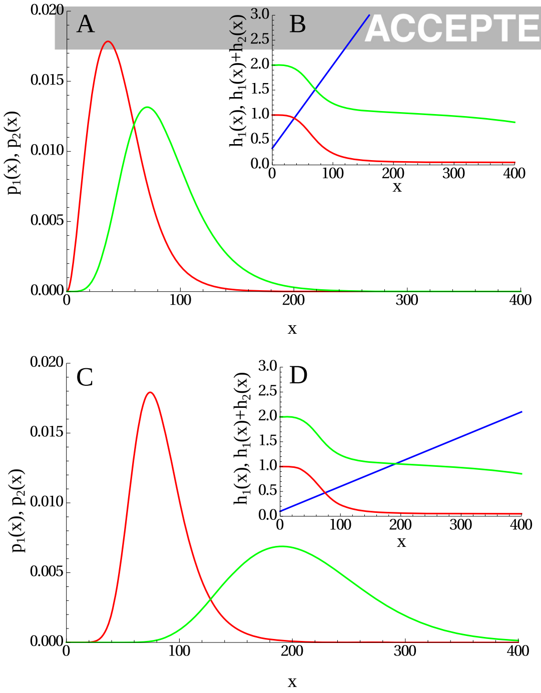

FIG. 15: Relative change in expression after a defective duplication of a negatively self-regulating gene, $K_2 = 8K_1$ (Fig. 8A, magenta line therein). Differently than for the perfect duplication shown in Fig. 14, here the relative change in expression is smaller for small $a = 3$ than for large $a = 10$.

with $s = z + 1$ and $z = x/K$. From (7) and (B2) we get

$$p(x) \;=\; Ax^{-1}e^{-x/b}\prod_{j=1}^{G} x^{a_j}\mathcal{F}_j(x), \qquad \text{(B4)}$$

where A is the normalisation constant, and $\mathcal{F}_j(x)$ reads

$$\mathcal{F}_j(x) \;=\; \frac{\prod_{i_j=2}^{n_j}\exp\left(\frac{a_j(1-\epsilon_j)}{i_j-1}\left(1+\frac{x}{K_j}\right)^{1-i_j}\right)}{\left(1+\frac{x}{K_j}\right)^{a_j(1-\epsilon_j)}}. \text{(B5)}$$

In the case of identical gene copies ($h_j(x) = h(x)$, $a_i = a$), Eq. (B4) can be rewritten as

$$p_G(x) \;=\; Ax^{-1}e^{-x/b}x^{aG}[\mathcal{F}(x)]^G, \qquad \text{(B6)}$$

where

$$\mathcal{F}(x) \;=\; \frac{\prod_{i=2}^{n}\exp\left(\frac{a(1-\epsilon)}{i-1}\left(1+\frac{x}{K}\right)^{1-i}\right)}{\left(1+\frac{x}{K}\right)^{a(1-\epsilon)}}. \qquad \text{(B7)}$$

Note that for $p_G(x)$ given by (B6), identity (9) still holds.

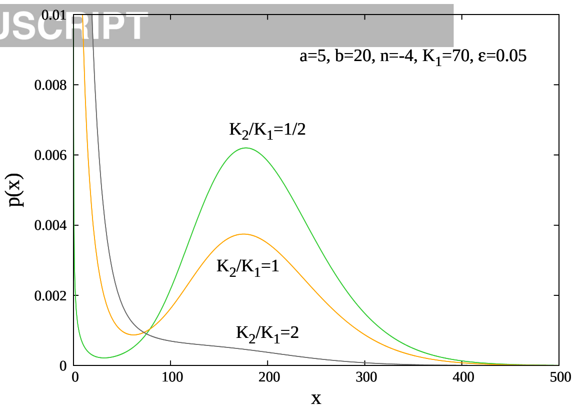

FIG. 16: Probability distribution $p(x)$ for the case of two gene copies and for various values of K_2/K_1: $K_2/K_1 = 1/2$ (green), $K_2/K_1 = 1$ (dark yellow) and for $K_2/K_1 = 2$ (grey); the latter case corresponds to maximum of the Fano factor F as a function of K_2/K_1 for $a = 5$, $b = 20$, $n = -4$, $K_1 = 70$, and $\epsilon_1 = \epsilon_2 = 0.05$, cf. Fig. 9

APPENDIX C: MAXIMA OF F AND η IN FIG. 3C,D DUE TO BIMODALITY OF THE PROTEIN NUMBER DISTRIBUTION

The sharp maxima of both the Fano factor F (Fig. 3 C) and coefficient of variation η (Fig. 3 D) as a function of gene copy number for $K = 700$ correspond to a change in the character of $p_G(x)$ from an apparently unimodal ($G = 7$), through bimodal ($G = 8$), to again apparently unimodal $G = 9$, cf. Fig. 10. Geometric construction (see inset in Fig. 10) reveals that actually all probability distributions shown here are bimodal. However, except for $p_8(x)$, one of their peaks turns out to be much smaller than the other.

APPENDIX D: EXAMPLE PLOTS OF COVARIANCE BETWEEN THE EXPRESSION OF TWO COPIES OF A SELF-REGULATING GENE

In Fig. 11 we show example plots of covariance (Eq. 11) between the expression of two identical gene copies of a negatively self-regulating gene (Fig. 11A) and positively self-regulating gene (Fig. 11B). In some cases the covariance is negative. In the case of positive auto-regulation, abrupt changes in the covariance, from negative to positive, are due to the transition through bimodality regime: If one gene copy produces bimodal distributions of protein numbers and two gene copies have unimodal expression (Fig. 12, red), then the covariance is negative. If one gene copy produces unimodal distributions of protein numbers and the expression of two gene copies is bimodal (Fig. 12, blue), then the covariance is positive.

APPENDIX E: VISUALISATION OF THE CHANGES IN MEAN GENE EXPRESSION AFTER DUPLICATION BY GEOMETRIC CONSTRUCTION

In this Appendix, we show additional figures (Figs. 13-15) for the Subsection III B 2. The geometric construction intuitively visualises the changes in mean gene expression after duplication.

APPENDIX F: MAXIMA OF F AND η IN FIG. 9C,D

The maxima of F and η in Fig. 9C,D roughly correspond to the transition through bimodal distributions. As an example, in Fig. 16 we show probability distributions corresponding to maxima of F and η as a function of relative sensitivity K_2/K_1 of both promoters to auto-regulation as shown in Fig 9 (A) for the case of two non-equivalent copies of a positively self-regulating gene, for $a = 5$. The remaining parameters: $n = -4$, $b = 20$, $K_1 = 70$, $\epsilon_1 = \epsilon_2 = 0.05$. Interestingly, the maximum at $K_2/K_1 = 2$ corresponds to *unimodal* distribution; bimodality is present for $K_2/K_1 = 1$ and $K_2/K_1 = 1/2$.

[1] J. E. Krebs, B. Lewin, E. S. Goldstein, and S. T. Kilpatrick, *Lewin's genes XI* (Jones & Bartlett Publishers, 2013).

[2] W.-H. Li, J. Yang, and X. Gu, Trends in Genetics **21**, 602 (2005).

[3] M. Lynch and J. S. Conery, Science **290**, 1151 (2000).

[4] G. H. Perry, N. J. Dominy, K. G. Claw, A. S. Lee, H. Fiegler, R. Redon, J. Werner, F. A. Villanea, J. L. Mountain, R. Misra, *et al.*, Nature Genetics **39**, 1256 (2007).

[5] R. J. Roper and R. H. Reeves, PLoS Genetics **2**, e50 (2006).

[6] B. Conrad and S. E. Antonarakis, Annu. Rev. Genomics Hum. Genet. **8**, 17 (2007).

[7] J. R. Pollack, T. Sørlie, C. M. Perou, C. A. Rees, S. S. Jeffrey, P. E. Lonning, R. Tibshirani, D. Botstein, A.-L. Børresen-Dale, and P. O. Brown, Proceedings of the National Academy of Sciences **99**, 12963 (2002).

[8] J. T. Kittleson, S. Cheung, and J. C. Anderson, Journal of Biological Engineering **5**, 1 (2011).

[9] D. Volfson, J. Marciniak, W. J. Blake, N. Ostroff, L. S. Tsimring, and J. Hasty, Nature **439**, 861 (2006).

[10] J. Hornos, D. Schultz, G. Innocentini, J. Wang, A. Walczak, J. Onuchic, and P. Wolynes, Physical Review E **72**, 051907 (2005).

[11] N. Friedman, L. Cai, and X. S. Xie, Physical Review Letters **97**, 168302 (2006).

[12] V. Shahrezaei and P. S. Swain, Proceedings of the National Academy of Sciences **105**, 17256 (2008).

[13] A. Ochab-Marcinek and M. Tabaka, Proceedings of the National Academy of Sciences **107**, 22096 (2010).

[14] P. Bokes, J. R. King, A. T. Wood, and M. Loose, Journal of Mathematical Biology **64**, 829 (2012).

[15] T. Aquino, E. Abranches, and A. Nunes, Physical Review E **85**, 061913 (2012).

[16] L. S. Tsimring, Reports on Progress in Physics **77**, 026601 (2014).

[17] A. Ochab-Marcinek and M. Tabaka, Physical Review E **91**, 012704 (2015).

[18] Y. Mileyko, R. I. Joh, and J. S. Weitz, Proceedings of the National Academy of Sciences **105**, 16659 (2008).

[19] A. Loinger and O. Biham, Physical Review Letters **103**, 068104 (2009).

[20] A. J. Stewart, *The construction of transcription factor networks through natural selection*, Ph.D. thesis, UCL (University College London) (2010).

[21] A. J. Stewart, R. M. Seymour, A. Pomiankowski, and M. Reuter, PLoS Computational Biology **9**, e1002992 (2013).

[22] J. Miekisz and P. Szymanska, Bulletin of Mathematical Biology **75**, 317 (2013).

[23] A. L. Van, H. A. Soula, and H. Berry, BMC Systems Biology **8**, 125 (2014).

[24] M. S. Sherman, K. Lorenz, M. H. Lanier, and B. A. Cohen, Cell Systems **1**, 315 (2015).

[25] L. A. Sepúlveda, H. Xu, J. Zhang, M. Wang, and I. Golding, Science **351**, 1218 (2016).

[26] L. Cai, N. Friedman, and X. S. Xie, Nature **440**, 358 (2006).

[27] J. Yu, J. Xiao, X. Ren, K. Lao, and X. S. Xie, Science **311**, 1600 (2006).

[28] J. Stewart-Ornstein, J. S. Weissman, and H. El-Samad, Molecular cell **45**, 483 (2012).

[29] U. Alon, *An introduction to systems biology: design principles of biological circuits* (CRC press, 2006).

[30] A. Crudu, A. Debussche, and O. Radulescu, BMC Systems Biology **3**, 1 (2009).

[31] A. Crudu, A. Debussche, A. Muller, O. Radulescu, *et al.*, The Annals of Applied Probability **22**, 1822 (2012).

[32] R. Yvinec, C. Zhuge, J. Lei, and M. C. Mackey, Journal of Mathematical Biology **68**, 1051 (2014).

[33] N. Popović, C. Marr, and P. S. Swain, Journal of Mathematical Biology **72**, 87 (2016).

[34] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, Science **297**, 1183 (2002).

[35] Y. Taniguchi, P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie, Science **329**, 533 (2010).

[36] M. C. Mackey, M. Tyran-Kamińska, and R. Yvinec, Journal of Theoretical Biology **274**, 84 (2011).

[37] M. Pájaro, A. a. Alonso, and C. Vázquez, Physical Review E **92**, 032712 (2015).

[38] F. Ruf, M.-J. Park, F. Hayot, G. Lin, B. Roysam, Y. Ge, and S. C. Sealfon, Journal of Biological Chemistry **281**,

30967 (2006).

[39] E. Porpiglia, D. Hidalgo, M. Koulnis, A. R. Tzafriri, and M. Socolovsky, PLoS Biology **10**, e1001383 (2012).

[40] Y. Ishihama, T. Schmidt, J. Rappsilber, M. Mann, F. U. Hartl, M. J. Kerner, and D. Frishman, BMC Genomics **9**, 1 (2008).

[41] A. Ishihama, A. Kori, E. Koshio, K. Yamada, H. Maeda, T. Shimada, H. Makinoshima, A. Iwata, and N. Fujita, Journal of Bacteriology **196**, 2718 (2014).

[42] O. Kobiler, A. Rokney, N. Friedman, J. Stavans, A. B. Oppenheim, *et al.*, Proceedings of the National Academy of Sciences of the United States of America **102**, 4470 (2005).

[43] J. S. Weitz, Y. Mileyko, R. I. Joh, and E. O. Voit, Biophysical Journal **95**, 2673 (2008).