How the Number of Alleles Influences Gene Expression

Beata Hat,[1,2] Pawel Paszek,[2] Marek Kimmel,[2,3] Kazimierz Piechor[1] and Tomasz Lipniacki[1,2]

The higher organisms, eukaryotes, are diploid and most of their genes have two homological copies (alleles). However, the number of alleles in a cell is not constant. In the S phase of the cell cycle all the genome is duplicated and then in the G2 phase and mitosis, which together last for several hours, most of the genes have four copies instead of two. Cancer development is, in many cases, associated with a change in allele number. Several genetic diseases are caused by haploinsufficiency: Lack of one of the alleles or its improper functioning. In the paper we consider the stochastic expression of a gene having a variable number of copies. We applied our previously developed method in which the reaction channels are split into slow (connected with change of gene state) and fast (connected with mRNA/protein synthesis/decay), the later being approximated by deterministic reaction rate equations. As a result we represent gene expression as a piecewise deterministic time-continuous Markov process, which is further related with a system of partial differential hyperbolic equations for probability density functions (pdfs) of protein distribution. The stationary pdfs are calculated analytically for haploidal gene or numerically for diploidal and tetraploidal ones. We distinguished nine classes of simultaneous activation of haploid, diploid and tetraploid genes. This allows for analysis of potential consequences of gene duplication or allele loss. We show that when gene activity is autoregulated by a positive feedback, the change in number of gene alleles may have dramatic consequences for its regulation and may not be compensated by the change of efficiency of mRNA synthesis per allele.

KEY WORDS: stochastic gene expression, feedback regulation, diploid genes, haploinsufficiency, piecewise deterministic time-continuous Markov process

---

[1] Institute of Fundamental Technological Research, Swietokrzyska 21, 00-049 Warsaw, Poland; e-mail: tomek@rice.edu and tlipnia@ippt.gov.pl.
[2] Department of Statistics, Rice University, 6100 Main St. MS-138, Houston, TX 77005.
[3] Institute of Automation, Silesian Technical University, 44-100 Gliwice, Poland.

1. INTRODUCTION

Stochasticity in gene expression arises from fluctuation in gene activity, [12] mRNA transcription, protein translation and oligomerization (Refs. 16, 11, 28, 29, recently reviewed in Refs. 9 and 17). Figure 1 illustrates the main steps in gene expression. Control of gene activity is mediated by transcription factors which may bind a specific promoter regions and switch the allele *on* or *off*. When the gene is active, RNA polymerase may bind the gene promoter and initiate mRNA transcription. Next, mRNA becomes edited and exported from the nucleus to the cytoplasm, where the protein translation occurs. Accordingly, a single event of allele activation results (if the activation period is sufficiently long), in a burst of mRNA molecules, which is then translated into an even larger burst of proteins. [12,31,2] Stochasticity in gene expression causes the population of cells to exhibit a large cell-to-cell variability as observed for example by Takasuka *et al.*[27] and Stirland *et al.*[24] for mammalian cells, Raser and O'Shea[22] for budding yeast (*Saccharomyces cerevisiae*) or Elowitz *et al.*[7] for bacteria (*Escherichia coli*).

Prokaryotes are haploid, i.e., most of their genes have only one copy. The higher organisms eukaryotes are diploid, most of their genes have two homologous copies (alleles), which can be independently activated and inactivated. However, the number of gene copies is not constant in cell evolution. In the S phase of cell cycle whole genome is duplicated and then in the G2 phase and mitosis, which together last for several hours, most genes have four copies instead of two. On the contrary, in meiosis 4 daughter haploid cells (gametes) are produced (in two subsequent divisions) out of one diploid cell. The genetic defect of loss of one allele or its transcriptional inactivity can result in haploinsufficiency, which is a hallmark of some diseases. Cancer cells may have gene or chromosomal duplications resulting in a larger number of alleles which may substantially alter cell function and make the disease more aggressive. Transfected cells used in various experiments may have arbitrary and difficult to control number of homologous
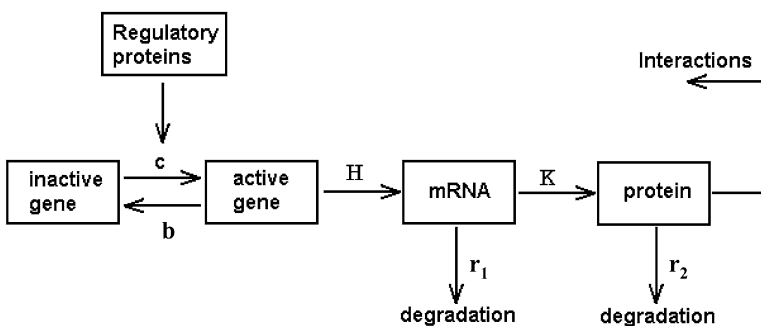

Fig. 1. Simplified schematic diagram of gene expression.

gene copies. In order to deduce the behavior of "normal" (wild-type) cells from experiments performed on transfected cells, one needs to consider N-allelic gene regulation.
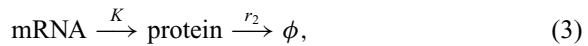
The aim of the current study is to analyze how expression of a single gene depends on the number of its alleles. It is obvious that the number of alleles influences the averaged protein level as well as the variance of the protein distribution. As we will see, when the gene is not regulated by the feedback, both average and variance of the protein distribution are proportional to the number of alleles. However, in the case of feedback there is no general rule relating magnitude and expression noise to the allele number. In many cases, lack of one copy or genome duplication does not influence the health of the organism but, as we will show, there are cases in which allele loss may substantially alter gene expression, and similarly, there are cases in which duplication of a gene may lead to complete deregulation of its expression.

The paper is structured as follows: First, in Sec. 2, we consider single gene regulation without feedback from its own transcriptional activity. The lack of feedback will allow us to take into account three main sources of stochasticity, gene activity, mRNA transcription/decay and protein translation/decay. In this case, we can calculate the mean and variance of the mRNA and protein distribution for arbitrary number of alleles. In Sec. 2.1 we introduce the continuous approximation, in which stochasticity is due solely to flipping of the gene status, [14,15] whereas the stochastic effects associated with mRNA and protein synthesis and decay are neglected. This approximation is used in Sec. 3, where the main results are presented. First we derive an analytical protein distribution in the case of autoregulation of a monoallelic gene (Sec. 3.1) and then the numerical solution for the system of a N-allelic gene (Sec. 3.2). We calculate numerically the mean and the variance of protein distribution in the case of a single gene (with one, two or four homologous copies) regulated by its own protein in negative and positive feedback loop. In the case of positive feedback we discuss three different scenarios of gene activation with respect to the number of alleles. We will show that in the case of diploid gene, allele loss or gene duplication may totally alter gene regulation and the loss of allele may not be fully compensated by the increase of the efficiency of mRNA synthesis of the remaining allele. In Conclusions we discuss potential consequences of our results.

2. PRELIMINARIES

First let us consider a single haploid gene without feedback regulation on its own transcriptional activity. Considered model involves three classes of processes: Gene activation/inactivation, mRNA transcription/decay and protein translation/decay (Fig. 1). It is assumed that the gene can switch to the active state (denoted by A) with a constant propensity $c = c_0$, and return to the inactive state

(denoted by I) with a constant propensity $b = b_0$, due to binding or dissociation of regulatory factors, which are assumed to be present at constant concentrations in cell nucleus. We further assume that mRNA transcript molecules are synthesized at the rate $G(t)H$, where G is a binary variable describing the state of a gene: $G(A) = 1$ and $G(I) = 0$.[12,10,20,15] Protein translation proceeds at a rate $Kx(t)$, where $x(t)$ is the number of mRNA molecules. In addition, mRNA and protein molecules are degraded at rates r_1 and r_2, respectively. The reactions described can be summarized as follows:

$$I \xrightarrow{c} A, \quad I \xleftarrow{b} A, \tag{1}$$

$$A \xrightarrow{GH} \text{mRNA} \xrightarrow{r_1} \phi, \tag{2}$$

$$\text{mRNA} \xrightarrow{K} \text{protein} \xrightarrow{r_2} \phi, \tag{3}$$

where symbol ϕ denotes degradation of gene products. Six parameters c, b, K, H, r_1 and r_2 can vary several orders of magnitude between particular genes and species, which leads to plethora of possible scenarios. There is an upper physiological bound on transcription ($H \simeq 0.1/s$) and translation ($K \simeq 0.5/s$) rates, reached by genes whose products are needed in large quantities within short time. For example, about 100000 particles of an inhibitory protein termed $I\kappa B\alpha$ must be degraded and then resynthesized in about $1/2$ h in order to properly control the early phase of immune response (see Refs. 13, 14, 4 for quantitative discussion of parameters). However, there is a large number of genes, whose products are needed in small quantities, of the order of a few molecules per cell, and it is possible these genes are expressed at much lower rates or their expression is very intermittent. The gene switching propensities c, b are larger for prokaryotes (of order 1/s) for which binding of regulatory factors is typically unstable. On the contrary, in higher organisms (eukaryotes) these rates are expected to be much lower and as a result one may expect that gene activation results in burst of mRNA molecules, translated to even larger burst of proteins. Typically, proteins are more stable than mRNA molecules ($r_1 > r_2$), but again in both cases the degradation half time may vary from minutes to days. In prokaryotes mRNAs are typically unstable with half time of the order of one minute.

The state of the system is decribed by a triple of variables $(x(t), y(t), G(t))$, where $y(t)$ is the number of protein molecules. Their joint discrete distribution can be represented as a pair of the probability mass functions

$$f_{xy} := P[\# \text{ mRNA} = x, \# \text{ protein} = y, G = 0], \tag{4}$$

$$g_{xy} := P[\# \text{ mRNA} = x, \# \text{ protein} = y, G = 1], \tag{5}$$

i.e., probabilities that the number of mRNA molecules (of considered species) is x and the number of protein molecules is y, $x \in \mathbb{N}$, $y \in \mathbb{N}$, in a cell being in the

inactive ($G = 0$) or active ($G = 1$) state, respectively. The marginal distribution $\rho_{xy} := f_{xy} + g_{xy}$ describes the mRNA and protein levels regardless of the gene status.

The time evolution of probabilities (4)–(5) is given by the following system of chemical master equations or Chapman-Kolgomorov equations[30,10]

$$\frac{df_{xy}}{dt} = bg_{xy} - cf_{xy} + G(I)Hf_{x-1,y} + r_1(x+1)f_{x+1,y} - (G(I)H + r_1x)f_{xy}$$
$$+ Kxf_{x,y-1} + r_2(y+1)f_{x,y+1} - (Kx + r_2y)f_{xy}, \tag{6}$$

$$\frac{dg_{xy}}{dt} = -bg_{xy} + cf_{xy} + G(A)Hg_{x-1,y} + r_1(x+1)g_{x+1,y}$$
$$- (G(A)H + r_1x)g_{xy} + Kxg_{x,y-1} + r_2(y+1)g_{x,y+1}$$
$$- (Kx + r_2y)g_{xy}. \tag{7}$$

The above system involves an infinite set of equations. The first two right-hand side terms in Eqs. (6)–(7) correspond to the probability flow due to the regulation of gene activity, further three terms describe the flow of probability due to the synthesis and degradation of mRNA molecules, while the last three terms stand for the synthesis and degradation of protein molecules. Note that, since $G(I) = 0$, the mRNA synthesis terms are absent in Eq. (6). The master equations (6)–(7) provide the time-dependent distribution of the underlying stochastic process and thus their solution is of primary interest. Although these equations (to our knowledge) can not be solved analytically even for $\frac{df_{xy}}{dt} = \frac{dg_{xy}}{dt} = 0$, they can be used to calculate the moments of distributions f_{xy}, g_{xy} or ρ_{xy}. This is accomplished by using the probability generating functions (PGF's) defined as follows:

$$F(z, s) = \sum_{x,y} z^x s^y f_{xy}, \tag{8}$$

$$G(z, s) = \sum_{x,y} z^x s^y g_{xy}. \tag{9}$$

We transform Eqs. (6)–(7) into partial differential equations for the PGF's and use the fact that the factorial moments of the two distributions are equal to the partial derivatives of $F(z, s)$ and $G(z, s)$ calculated at $z = s = 1$. The details of this calculation are found in Ref. 19. The mean $E[y]$ and variance $Var[y] = E[y^2] - (E[y])^2$ of the stationary marginal distribution of the protein $\rho_y = \sum_x \rho_{xy}$ are equal to

$$E[y] = \frac{cKH}{r_1r_2(c+b)}, \tag{10}$$

$$Var[y] = \frac{br_1r_2(r_1 + r_2 + c + b)}{c(r_1 + r_2)(r_1 + c + b)(r_2 + c + b)}(E[y])^2$$

$$+ \frac{K}{(r_1 + r_2)}E[y] + E[y]. \tag{11}$$

The above result, obtained for a haploid gene, can be generalized for a gene having an arbitrary number of copies. Let us mark the mRNAs and proteins resulting from different alleles by different "colors." If the coefficients c and b do not depend on x or y (i.e., there is no feedback) the distributions of differently colored proteins are independent. As a result, the marginal protein distribution is a convolution of these "colored" distributions, the mean of the protein is the sum of means and the variance is the sum of the variances. Thus, for the gene having N alleles we have

$$E_N[y] = NE[y], \tag{12}$$

$$Var_N[y] = NVar[y]. \tag{13}$$

As a result of the above, the Fano Factor defined as $FF[y] = Var[y]/E[y]$ does not depend on the number of gene alleles, while the noise defined as $\eta = \sqrt{Var}/E$ is inversely proportional to $\sqrt{N}$. As for the system without feedback, the mean and the variance calculated per allele are independent of the number of alleles. We will use these measures, viz., $E_0 = E_N/N$ and $Var_0 = Var_N/N$, in the analysis of systems with feedback. As we will see, in this case there is no general rule relating the magnitudes of variance and averages to the number of alleles.

2.1. Continuous Approximation

The aim of this work is to consider feedback systems. Obviously, the feedback causes that the distributions of "colored" proteins (associated with different alleles) are not independent, which makes the analysis more difficult. In order to simplify the problem, we introduce an approximation, well justified for eukaryotes, in which stochasticity related to the synthesis/degradation of mRNA and proteins is neglected (see Refs. 14, 15, 19 for discussion of this approximation). Is short, the approximation is based on the fact that typical numbers of mRNAs (x) and proteins (y) are much larger than the number of gene copies. As a result the fluctuations in gene state give much larger contribution to the protein variance than the mRNA/protein synthesis and degradation processes. Neglecting stochasticity of the mRNA and protein synthesis and degradation we treat x and y as continuous variables and replace relations (2)–(3) by deterministic reaction-rate equations. For monoallelic gene without feedback (i.e., with constant transition coefficients c and b), we obtain the following system

$$I \xrightarrow{c} A, \quad I \xleftarrow{b} A, \tag{14}$$

$$\frac{dx}{dt} = HG(t) - r_1 x, \tag{15}$$

$$\frac{dy}{dt} = Kx - r_2 y. \tag{16}$$

Since $G(t)$ is a binary random variable, equations (15)–(16) generate stochastic trajectories, which can be described as piecewise deterministic time-continuous Markov process. Such processes have been intensively studied in physics. For example, the motion of a charged particle in an electric or magnetic field, randomly changing between two states, can also be considered as a piecewise deterministic time-continuous Markov process. The main difference is that in most physical applications, these transition rates are constant, while here the most interesting case (considered later for auto-regulatory genes) is when the transition coefficients depend on continuous variables $x(t)$ and $y(t)$. A system analogous to our system (15)–(16) in which the transition intensities of the random forcing process $G(t)$ depend on state variables $x(t)$ and $y(t)$ was considered by Basak *et al.*[1]

At each time t, the realizations of mRNA and protein levels $x(t)$ and $y(t)$, and the state of the transcription switch $G(t)$, form a triple of random variables, first two of which are continuous and the third is binary. Therefore, their joint distribution can be described by a pair of probability density functions (pdf) $f(x, y, t)$ and $g(x, y, t)$. The interpretation is that for given time t

$$P[x(t) \in (x, x + \Delta x), \ y(t) \in (y, y + \Delta y), \ G(t) = 0] = f(x, y, t)\Delta x \Delta y,$$

$$P[x(t) \in (x, x + \Delta x), \ y(t) \in (y, y + \Delta y), \ G(t) = 1] = g(x, y, t)\Delta x \Delta y.$$

Now it is possible to write the continuity equations for $f(x, y, t)$ and $g(x, y, t)$ with source terms following from change of gene status (transformation between f and g), Eq. (14):

$$\frac{\partial f}{\partial t} + div\left[\left(\frac{dx}{dt}, \frac{dy}{dt}\right)f\right]_{|G=0} = bg - cf, \tag{17}$$

$$\frac{\partial g}{\partial t} + div\left[\left(\frac{dx}{dt}, \frac{dy}{dt}\right)g\right]_{|G=1} = -bg + cf. \tag{18}$$

Velocity fields $(\frac{dx}{dt}, \frac{dy}{dt})_{|G=0}$ and $(\frac{dx}{dt}, \frac{dy}{dt})_{|G=1}$, transforming f and g respectively, are given by equations (15)–(16). Thus, we have

$$\frac{\partial f}{\partial t} - \frac{\partial}{\partial x}(r_1 x f) + \frac{\partial}{\partial y}[(Kx - r_2 y)f] = bg - cf, \tag{19}$$

$$\frac{\partial g}{\partial t} + \frac{\partial}{\partial x}[(H - r_1 x)g] + \frac{\partial}{\partial y}[(Kx - r_2 y)g] = -bg + cf. \tag{20}$$

Alternatively, the above two equations may be derived from Chapman-Kologomorov equations (6)–(7) via Kramers-Moyal expansion.[30] The first order expansion leads to Eqs. (19)–(20), while the second order expansion results in so the called Fokker-Planck equations with additional diffusion terms. The main advantage of our derivation is that the approximation is made at the level of single cell description, so its accuracy can be validated by comparing simulation of Eqs. (1)–(3) using the Gillespie algorithm to the solutions of (14)–(16).[14] The mathematical analysis of the system (19)–(20) in the case when $b = b(x, y), c = c(x, y)$ was provided recently by Bobrowski.[3]

Similarly as in the discrete case, we define the moment generating functions (MGFs)

$$F(z, s) = \int \int e^{zx+sy} f(x, y) \, dxdy, \tag{21}$$

$$G(z, s) = \int \int e^{zx+sy} g(x, y) \, dxdy. \tag{22}$$

Partial derivatives of first order of $F(z, s)$ and $G(z, s)$ evaluated at $z = s = 0$ give the corresponding moments of $f(x, y, t)$ and $g(x, y, t)$. This allows calculating the mean $E_C[y]$ and variance $Var_C[y]$ of the protein marginal distribution $\rho_y = \int (g(x, y) + f(x, y)) \, dx$.[19]

$$E_C[y] = \frac{cKH}{r_1 r_2 (c + b)}, \tag{23}$$

$$Var_C[y] = \frac{br_1 r_2 (r_1 + r_2 + c + b)}{c(r_1 + r_2)(r_1 + c + b)(r_2 + c + b)} (E_C[y])^2. \tag{24}$$

Let us note that we obtained the same mean as in discrete case (Eqs. (10)) and that the variance differs only in linear terms in $E[y]$: The first missing term $\frac{K}{(r_1+r_2)} E[y]$ corresponds to the transcriptional noise, while the second one, $E[y]$, corresponds to the translational noise. Within frame of the continuous approximation we can easily obtain the answer for a problem posed by Cook *et al.*,[5] viz. how much the variance of the protein distribution changes if one "big" allele is replaced by by two "small" ones (having together the same protein production efficiency). Since the "small" allele produces twice less protein on the average, than the "big" one, then, according to Eq. (24), it has four times smaller protein variance. This implies that the replacement of one "big" allele by two "small" ones would result in twofold reduction of the protein variance. This simple result is due to the continuous approximation, in which the linear terms in the variance (24) are neglected.

System (14)–(16) can be further simplified if we assume that the protein degradation time is much larger than mRNA degradation time, i.e., that $r_1/r_2 \gg 1$. In such case Eq. (15) is much faster than Eq. (16), which allows us to replace Eq. (15) by the equality $x = GH/r_1$. As a result, the system (14)–(16)

is transformed into

$$I \xrightarrow{c} A, \quad I \xleftarrow{b} A, \tag{25}$$

$$\frac{dy}{dt} = \frac{HKG}{r_1} - r_2 y. \tag{26}$$

The above approximation is equivalent to the assumption made by Kepler and Elston, [10] that the protein is synthesized directly from a gene (K-E approximation). Equations for the probability density functions $f(y, t)$, $g(y, t)$, corresponding to the simplified system (25)–(26) read

$$\frac{\partial f}{\partial t} + r_2 \frac{\partial}{\partial y}(-yf) = bg - cf, \tag{27}$$

$$\frac{\partial g}{\partial t} + \frac{\partial}{\partial y}\left(\left(\frac{HK}{r_1} - r_2 y\right)g\right) = -bg + cf. \tag{28}$$

In the same way, we can calculate the mean $E_{KE}[y]$ and variance $Var_{KE}[y]$ of the stationary marginal distribution of protein $\rho(y) = f(y) + g(y)$,

$$E_{KE}[y] = \frac{cHK}{r_1 r_2(c + b)}, \tag{29}$$

$$Var_{KE}[y] = \frac{br_2}{c(r_2 + c + b)}(E_{KE}[y])^2. \tag{30}$$

Let us note that the mean is exactly the same as in previous case and that $Var_{KE}[y]$ can be obtained from $Var_C[y]$ by assuming that $r_1/r_2 \gg 1$.

For $\frac{\partial f}{\partial t} = \frac{\partial g}{\partial t} = 0$, system (27)–(28) can be solved anallytically (see the next section)

$$f(y) = A \, y^{c_r - 1} \left(\frac{HK}{r_1 r_2} - y\right)^{b_r},$$

$$g(y) = A \, y^{c_r} \left(\frac{HK}{r_1 r_2} - y\right)^{b_r - 1} \quad \text{for} \quad y \in \left(0, \frac{HK}{r_1 r_2}\right) \tag{31}$$

and

$$f(y) = g(y) = 0 \quad \text{for} \quad y \notin \left[0, \frac{HK}{r_1 r_2}\right], \tag{32}$$

where $c_r = c/r_2$, $b_r = b/r_2$ and

$$A = \frac{\Gamma(c_r)\Gamma(b_r)}{\Gamma(c_r + b_r)} \left(\frac{r_1 r_2}{HK}\right)^{c_r + b_r} \tag{33}$$

is the constant introduced to normalize the marginal distribution $\rho(y) := f(y) + g(y)$, $\int_0^{\frac{r_1 r_2}{HK}} \rho(y) = 1$,

$$\rho(y) = A \, y^{\, c_r - 1} \left(\frac{HK}{r_1 r_2} - y \right)^{b_r - 1} \frac{HK}{r_1 r_2}. \tag{34}$$

Having the distribution $\rho(y)$ for a monoallelic gene, we can calculate distributions for a gene with an arbitrary number of copies. We use the fact that, when transition coefficients c and b do not depend on the amount of protein y, distributions resulting from different alleles are independent. In such a case distribution of protein $\rho_N(y)$, resulting from N-allelic gene, is given by the Nth order convolution of distributions $\rho(y)$. For example, for diploid gene, we have

$$\rho_2(y) = \int_{s_1}^{s_2} \rho(z) \, \rho(y - z) \, dz, \tag{35}$$

where $s_1 = 0, s_2 = y$ for $0 < y \leq \frac{HK}{r_1 r_2}$ and $s_1 = y - \frac{HK}{r_1 r_2}$, $s_2 = \frac{HK}{r_1 r_2}$ for $\frac{HK}{r_1 r_2} < y < \frac{2HK}{r_1 r_2}$. The tetraploid gene distribution $\rho_4(y)$ can be obtained as a convolution of two $\rho_2(y)$ distributions. Obviously the same formulas are valid for time dependent distributions, however, it is more difficult to calculate analytically $\rho(y, t)$ from the system (27)–(28).

3. RESULTS

In this section we consider the gene regulation with feedback. For simplicity, we restrict our considerations to a single gene with one, two or four alleles whose activity is regulated by its synthesized protein. Accordingly, we assume that the transition coefficients c and b depend on the amount of synthesized protein, i.e., $c = c(y)$ and $b = b(y)$. [10,14,15] We analyze this case within the framework of K-E approximation.

3.1. Autoregulation of a Haploid Gene: Analytical Solution

First, let us consider a monoallelic gene, which can be described by the system (27)–(28), now with $c = c(y)$ and $b = b(y)$. For simplicity we rewrite Eqs. (27)–(28) in a nondimensional form applying units in which $HK = r_2 = 1$,

$$\frac{\partial f}{\partial t} - \frac{\partial}{\partial y}(fy) = b(y)g - c(y)f, \tag{36}$$

$$\frac{\partial g}{\partial t} + \frac{\partial}{\partial y}((1 - y)g) = -b(y)g + c(y)f. \tag{37}$$

The stationary solutions $\frac{\partial f}{\partial t} = \frac{\partial g}{\partial t} = 0$ of the above system can be written in a closed form for $y \in [0, 1]$. Adding equations (36)–(37) side by side, we obtain the

first integral

$$\frac{d}{dy}[-yf + (1-y)g] = 0. \tag{38}$$

This implies

$$-yf + (1-y)g = -f(1) = g(0). \tag{39}$$

Since $f(\cdot)$ and $g(\cdot)$ are, by definition, non-negative, the condition $-f(1) = g(0)$ implies that $f(1) = g(0) = 0$. Thus, from Eqs. (39), we have $g = \frac{y\,f}{(1-y)}$. Inserting this to Eqs. (36) we obtain

$$\frac{yb(y)}{(1-y)}f - c(y)f = -\frac{d}{dy}(yf), \tag{40}$$

or

$$\frac{f'}{f} = \frac{-b(y)}{(1-y)} + \frac{c(y)-1}{y}. \tag{41}$$

Hence

$$f(y) = \exp\left[\int_0^1 \left(\frac{-b(y)}{(1-y)} + \frac{c(y)-1}{y}\right)dy\right], \quad g(y) = \frac{yf(y)}{(1-y)}. \tag{42}$$

For $c(y) = c_0 + c_1 y + c_2 y^2$ and $b(y) = b_0 + b_1 y + b_2 y^2$ the solution can be written explicitly

$$f(y) = A \exp\left[y(b_1 + b_2 + c_1) + \frac{1}{2}y^2(b_2 + c_2)\right] y^{(c_0-1)} (1-y)^{(b_0+b_1+b_2)}, \tag{43}$$

$$g(y) = A \exp\left[y(b_1 + b_2 + c_1) + \frac{1}{2}y^2(b_2 + c_2)\right] y^{c_0} (1-y)^{(b_0+b_1+b_2-1)}, \tag{44}$$

$$\rho(y) = f(y) + g(y) = A \exp\left[y(b_1 + b_2 + c_1) + \frac{1}{2}y^2(b_2 + c_2)\right]$$
$$\times y^{(c_0-1)}(1-y)^{(b_0+b_1+b_2-1)}, \tag{45}$$

where A is chosen so that $\int_0^1 \rho(y) = 1$. Let us note that the solution exists for all $c_0 > 0$, $c_1 \geq 0$, $c_2 \geq 0$, $b_0 \geq 0$, $b_1 \geq 0$, $b_2 \geq 0$, provided that $b_0 + b_1 + b_2 > 0$. Note that conditions $c_0 > 0$ and $b_0 + b_1 + b_2 > 0$ are needed to assure integrability of $f(y)$ and $g(y)$. Physically, $c_0 = 0$ would imply that the gene will switch *off* and will never switch *on* again. Let us note also, that for $c_0 < 1$, $\lim_{y->0^+} f(y) = \infty$ and for $b_0 + b_1 + b_2 < 1$, $\lim_{y->1^-} g(y) = \infty$, thus in this case formally we should write that the solution exists on $(0, 1)$ rather then on $[0, 1]$.

3.2. Autoregulation of N-allelic Gene

An N-allelic gene can exist in one of $N + 1$ states. Discrete function $G_N(t) \in \{0, 1, \ldots, N\}$, where $G_N(t) = i$ corresponds to the state, in which i of N copies are active. Let $f_i(y, t)$, $i \in \mathbb{Z}$, denote probability density functions such that

$$P[y(t) \in (y, y + \Delta y), \quad G_n(t) = i] = f_i(y, t)\Delta y. \tag{46}$$

Obviously $f_i(y, t) = 0$ for $i \notin \{0, 1, \ldots, N\}$. Transition matrix T_N between states f_i results from the transition rates for the single allele

$$I \xrightarrow{c(y)} A, \quad I \xleftarrow{b(y)} A, \tag{47}$$

and reads

$$T_N = \begin{array}{c|c|c|c}
 & f_{i-1} & f_i & f_{i+1} \\
\hline
f_{i-1} & & (N + 1 - i)c(y) & 0 \\
\hline
f_i & ib(y) & & (N - i)c(y) \\
\hline
f_{i+1} & 0 & (i + 1)b(y) & \\
\end{array}, \quad i \in \{0, 1, \ldots, N\}.$$

The time evolution of pdf's f_i is given by the system of $N + 1$ partial differential equations

$$\frac{\partial f_i}{\partial t} + \frac{\partial}{\partial y}((i - y)f_i) = -ib(y)f_i - (N - i)c(y)f_i$$
$$+ (N + 1 - i)c(y)f_{i-1} + (i + 1)b(y)f_{i+1}. \tag{48}$$

The stationary distributions, $\frac{\partial f_i}{\partial t} = 0$, are given by the system of linear ODEs

$$\frac{\partial}{\partial y}((i - y)f_i) = -ib(y)f_i - (N - i)c(y)f_i$$
$$+ (N + 1 - i)c(y)f_{i-1} + (i + 1)b(y)f_{i+1}. \tag{49}$$

Although the system (49) is linear, for $N > 1$ it is not trivial to solve it analytically and even numerically. The support of stationary distributions $f_i(y)$ is the interval $[0, N]$. The reason is that, except for very specific $b(y)$ and $c(y)$, all functions $f_i(y)$ tend either to zero or to infinity at the ends of the interval $[0, N]$. As a result there is no point to start. The method applied in this work was introduced previously.[15] In short, we replace system of $N + 1$ linear ODEs by the system of $M(N + 1)$ linear algebraic equations, where M is the size of the grid, chosen to assure required accuracy. Functions $f_i(y)$ must be further normalized to assure that the marginal distribution $\rho(y) = \sum_i f_i(y)$ satisfies condition $\int_0^N \rho(y)dy = 1$. The algorithm was tested in the case $N = 1$ for which we have analytical solution (42) and proved to be very accurate for $M = 1000$, for which all calculations take

about 1s on an average PC. Having the marginal distribution $\rho(y)$ for $N = 2$ and $N = 4$ we calculate the corresponding mean and variance of the protein.

From now on we concentrate on regulation of a self activating gene. Namely, we assume that the transition rates c and b are

$$c(y) = c_0 + c_2 y^2, \quad b(y) = b_0. \tag{50}$$

Coefficient c_2 measures the strength of self-activation due to binding of protein homodimers. This is a natural assumption since in many cases dimers are much more active and stable than the monomers and the production of homodimers is a quadratic function of the protein concentration.[10] Coefficient b_0 stands for the inactivation due to dissociation of homodimers from gene promoter. The strength of external induction due to binding of some transcription factors is controlled by coefficient c_0. We will thus consider c_2 and b_0 parameters characterizing the gene and for a given pair of coefficients c_2 and b_0 we will analyze the response of the haploid, diploid and tetraploid systems to the varying external stimulation c_0. As we already said, the $c_0 = 0$ leads to singularity in distribution function $\rho(y)$. Numerically, very small c_0 is also a challenge since the distribution concentrates at $y = 0$. To avoid these numerical difficulties we will restrict our numerical investigation to the case $c_0 \geq 0.01$. Biologically this is not a severe restriction since any spontaneous gene activation would assure some minimal value of c_0.

First, let us note that there are three distinctive patterns of N-allelic ($N = 1$, 2, 4) gene activation corresponding to different pairs of (c_2, b_0):

{A}-mode is the trivial mode in which the gene remains Active (i.e. $E_0(y) > 1/2$) for $c_0 \in [0.01, \infty)$. Biologically this implies that the gene can be activated spontaneously and remains active (Fig. 2C).

{B}-mode, in which gene activates for some $c_0 \in [0.01, \infty)$ and distribution $\rho(y)$ is transiently (i.e. for some c_0) Bimodal (Fig. 2D).

{U}-mode, in which the gene activates for some $c_0 \in [0.01, \infty)$ but its distribution $\rho(y)$ remains Unimodal for all $c_0 \in [0.01, \infty)$ (Fig. 2E).

In the {A}-mode the gene is constantly active, the mean per copy $E_0(y)(c_0)$ is close to 1 and it is not sensitive to the magnitude of external stimulation c_0 (Fig. 2A). Variance $Var_0(y)(c_0)$ remains low (Fig. 2B). The {A}-mode is a "perfect choice" for genes whose product is needed in a constant amount. In the {B}-mode the gene activates for relatively low c_0 and the activation pattern is quite interesting. For some c_0 a second maximum (at higher y) appears in distribution $\rho(y)$, (Fig. 2D) (green), then this maximum becomes higher than the first one, Fig. 2D (brown), finally for larger c_0 the distribution becomes unimodal. This pattern of activation implies that in some range of activation parameter c_0, the population of cells splits into two subpopulations having smaller and larger amount of protein. This pattern in some situations could be safer than the unimodal activation {U} shown in Fig. 2E. If the signal is not strong enough only a part of
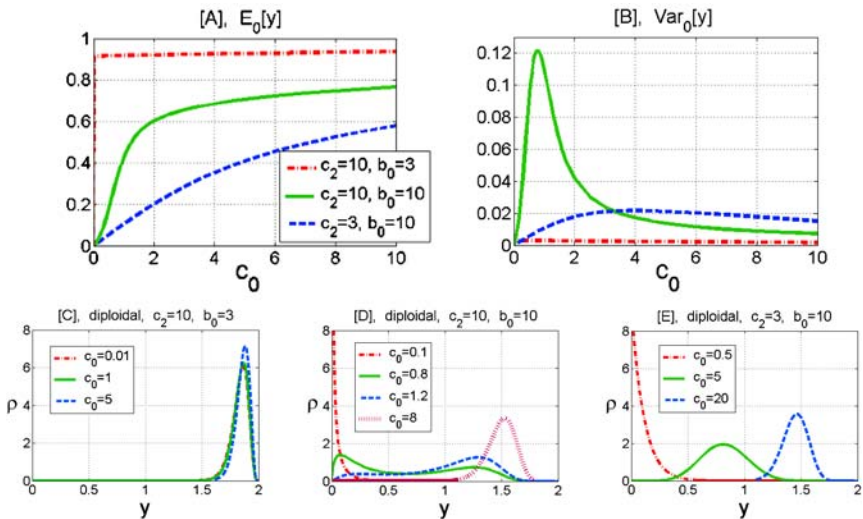
Fig. 2. *Color online:* Three modes of diploid gene activation (positive regulation). Panel A: mean per copy $E_0(y)(c_0)$ and B: variance per copy $Var_0(y)(c_0)$; lines: red (dashed dotted), green (continuous) and blue (dashed) correspond to the {A}, {B} and {U}-modes of activation calculated, respectively, for points (10,3), (10,10) and (3,10) in (c_2, b_0) plane (see Fig. 3B). Panels C, D and E: the protein distributions $\rho(y)$ for the {A} {B} and {U} activation modes.

cells activates, while the rest waits for a "confirmation" by a stronger signal. The {B}-mode of activation takes place for higher c_2/b_0 ratio and thus, the system activates (i.e. $E_0(y) > 1/2$) for lower external stimulation c_0 with much stepper profile of activation (Fig. 2A). Bimodal activation is present also in the variance profile $Var_0(y)(c_0)$, which has a sharp high maximum at c_0 for which $E_0(y)(c_0) \approx 1/2$ (Fig. 2B).

In Fig. 3B we show three regions in (c_2, b_0) plane corresponding to three modes of activation of the diploid gene {A},{B},{U}. The pattern for haploid gene is structurally similar to the diploid one (Fig. 3A), but the {A}-region is much narrower, while the {U}-region is much broader than that for diploid gene. In the case of tetraploid gene the {A}-region is the broadest while the {B} and {U}-regions are relatively narrow (Fig. 3C). In Figs. 3A, 3B and 3C we may also distinguish a small pie-like region for $c_2^2 + b_0^2 \leq 1$. For very small c_2, b_0 and c_0 distribution function $\rho_4(y)$, corresponding to tetraploid gene may have up to five maxima at $y = i$, $0 \leq i \leq 4$, each corresponding to the state in which i out of 4 copies are active. Since this multimodality (caused be very infrequent flipping between gene states) has a different character to the bimodality we are interested in, we will simply exclude region $c_2^2 + b_0^2 \leq 1$ from further considerations.
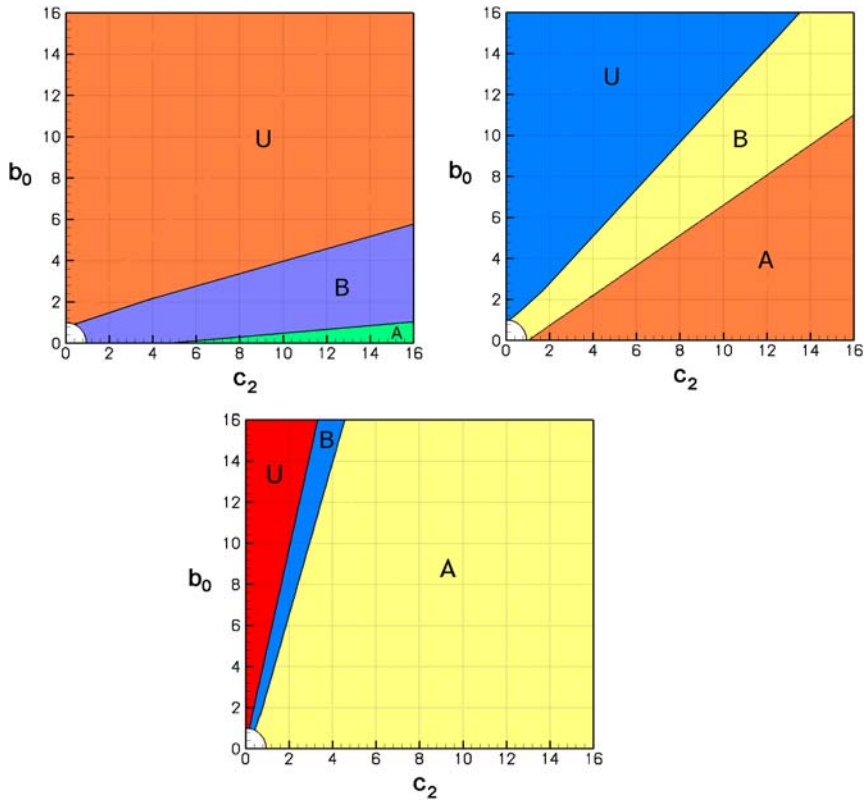
Fig. 3. *Color online:* Three regions in (c_2, b_0) plane corresponding to three modes of gene activation $\{A\},\{B\},\{U\}$. Panel A: the haploid gene, Panel B: the diploid gene and Panel C: the tetraploid gene.

It is interesting that the pattern for diploid gene activation (Fig. 3B) is most robust in the sense that $\{A\}$, $\{B\}$ and $\{U\}$ sectors are of similar width. The haploid genes can be in the $\{A\}$-mode only for very small ratio b_0/c_2, while tetraploid genes are mostly in the $\{A\}$-mode, what makes them insensitive to the magnitude of the external signal c_0.

We can combine these three plots (from Fig. 3) into one (Fig. 4) showing nine possible patterns for simultaneous activation of haploid, diploid and tetraploid genes. This allows an analysis of consequences of haploinsufficiency or gene duplication for diploid genes. According to our notation the sector denoted $\{ABU\}$ gives the range of parameters (c_2, b_0), for which a tetraploid gene is in the $\{A\}$-mode, diploid in the $\{B\}$-mode and haploid in the $\{U\}$-mode. Thus for any pair (c_2, b_0) from the $\{ABU\}$-sector the loss of one allele implies that diploid system would activate in the $\{U\}$-mode instead of the $\{B\}$-mode, while the gene multiplication
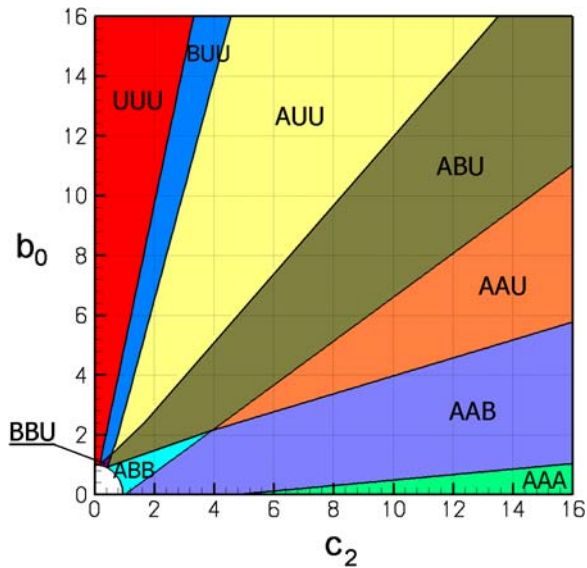
Fig. 4. *Color online:* Nine modes of synergistic gene activation; The notation is such that: {ABU} denotes that the tetraploid gene is in the {A} mode, the diploid gene is in the {B} mode and the haploid gene is in the {U} mode. The activation profiles for three marked points in (c_2, b_0) plane are analyzed in Figs. 5, 6 and 7.

will set it in the {A}-mode. When the diploid system "works" in {AAB} or {AAU} range, then the gene multiplication would not alter the activation mode, but the allele loss would stop the persistent gene activity. Oppositely when the diploid system is in the {AUU} range the loss of an allele would not result in an alternation of the activation mode, but the gene duplication would set it to the {A}-mode.

In Figs. 5, 6 and 7 we analyze three out of nine possible cases: {AAB}, {ABU}, {BUU}. The case {AAB} (Fig. 5) is the one, which resembles the classical picture of haploinsufficiency, see Cook *et al.*[5] The diploid and tetraploid gene remains active, while the haploid gene activates in a bimodal mode, so that its variance profile $Var_0(y)(c_0)$ has a sharp, tall maximum (Figs. 5C and B). It is commonly expected that the loss of one allele decreases the mean but, what is more important, it increases the variability of the gene expression. In this case, if the gene product is needed at a steady level, the loss of one allele may lead to a disease. The case {ABU} (Fig. 6) is quite different; it is the diploid system which transiently exhibits the broadest protein distribution (Fig. 6C). Contrary to the common intuition the variance per copy $Var_0(y)(c_0)$ is in this case larger for a diploid gene than for a haploid one (Fig. 6B). The activation of the haploid gene is much more gradual with a relatively small variance. The tetraploid gene is
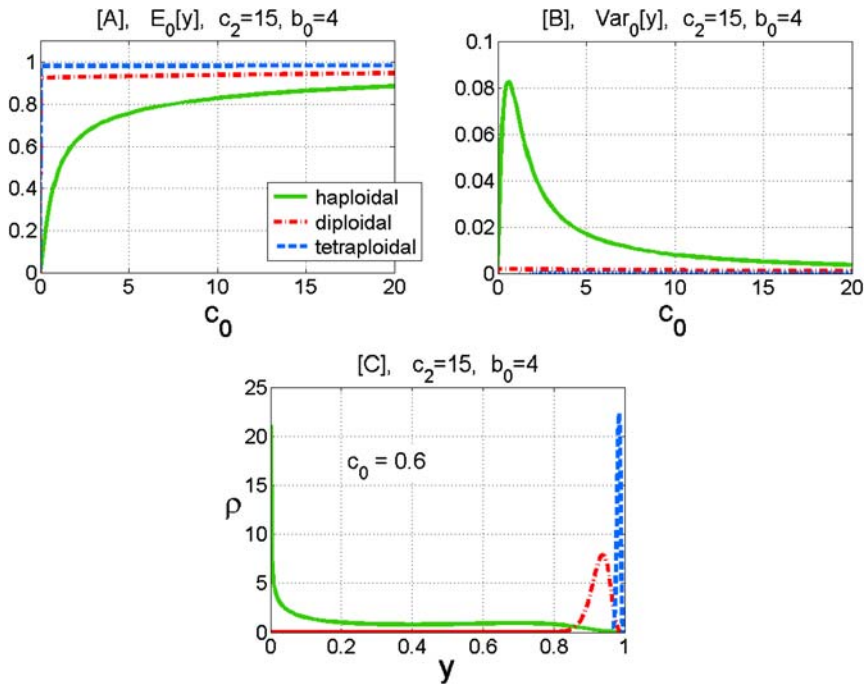
Fig. 5. *Color online:* Mode {AAB}: $c_2 = 15$, $b_0 = 4$; Panels A: mean per copy $E_0(y)(c_0)$ and B: variance per copy $Var_0(y)(c_0)$. Panel C: protein distributions $\rho(y)$ for the haploid, diploid, tetraploid systems for $c_0 = 0.6$ (at which haploid system has $\max[Var_0(y)(c_0)]$).

constantly active with a compact protein distribution. In the case {BUU} (Fig. 7) all three genes activate as stimulation increases. The activation profile (Fig. 7A) is steepest for the tetraploid gene and most gradual for the haploid one. Again, in contrary to the intuition, the variance per copy $Var_0(y)(c_0)$ is the largest for a tetraploid gene, due to its transiently bimodal distribution (Fig. 7B).

3.3. The Missing Allele may not be Compensated by the Doubled Expression of the Remaining One

The loss of one allele can not be fully compensated by the doubled production of the remaining one. In the case without autoregulation, based on the analysis in Sec. 2, we may conclude that the haploid gene has the same mean $E(y)$, but twice larger variance $Var(y)$ than the diploid one. For the negative feedback the mean $E(y)(c_0)$ is slightly higher for the diploid gene. The variance $Var(y)(c_0)$ is about twice higher for haploid gene, and the protein distribution profiles $\rho(y)$ are substantially different for haploid and diploid systems. The difference between
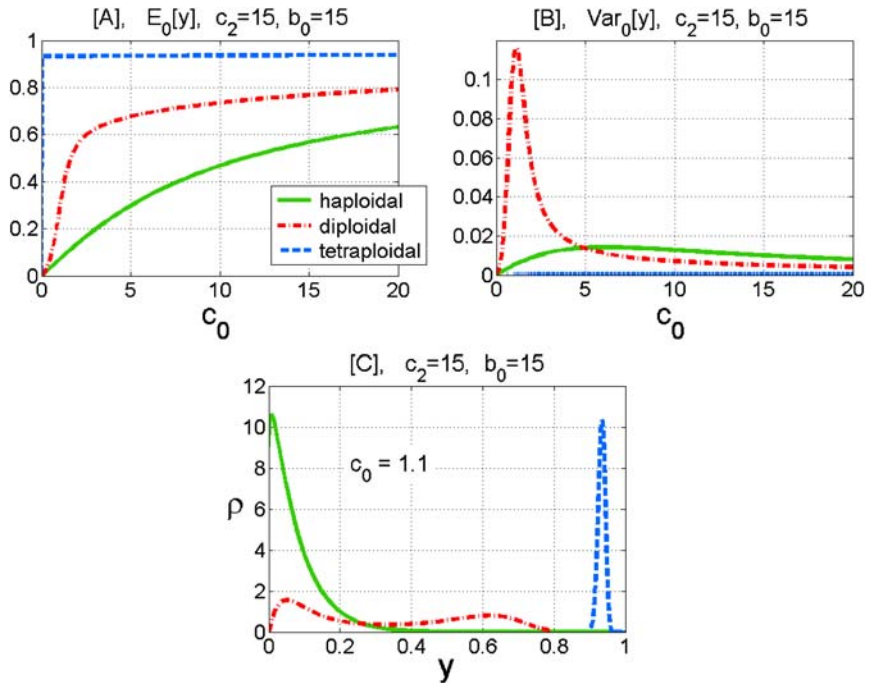
Fig. 6. *Color online:* Mode {ABU}: $c_2 = 15$, $b_0 = 15$; Panel A: mean per copy $E_0(y)(c_0)$ and B: variance per copy $Var_0(y)(c_0)$. Panel C: protein distributions $\rho(y)$ for haploid, diploid, tetraploid systems respectively for $c_0 = 1.1$ (at which diploid system has max[$Var_0(y)(c_0)$]).

the haploid and diploid systems is even more pronounced in the case of positive regulation. As shown in Fig. 8A, the mean $E(y)(c_0)$ is higher for the diploid gene, (which activates at the value of coefficient $c_0 = 0.01$) than for the haploid gene. This is caused by the fact that the case of positive regulation activation of one allele results in protein production which in turn activates the second allele. In the case shown in Fig. 8 the corresponding protein distributions are bimodal when both haploid and diploid genes activate. However, since the diploid gene activates for 10 times lower c_0, the variance of the protein distribution is transiently larger for diploid system than for the haploid one (Fig. 8B). For larger c_0, the haploid system exhibits, as expected, much larger variance than the diploid one.

4. CONCLUSIONS

There are several processes which change the number of gene copies in cell evolution. The most common is genome duplication in the cell cycle. In eukaryotes cell cycle lasts for about 24 h and consists of four phases, G1 in which cell grows,
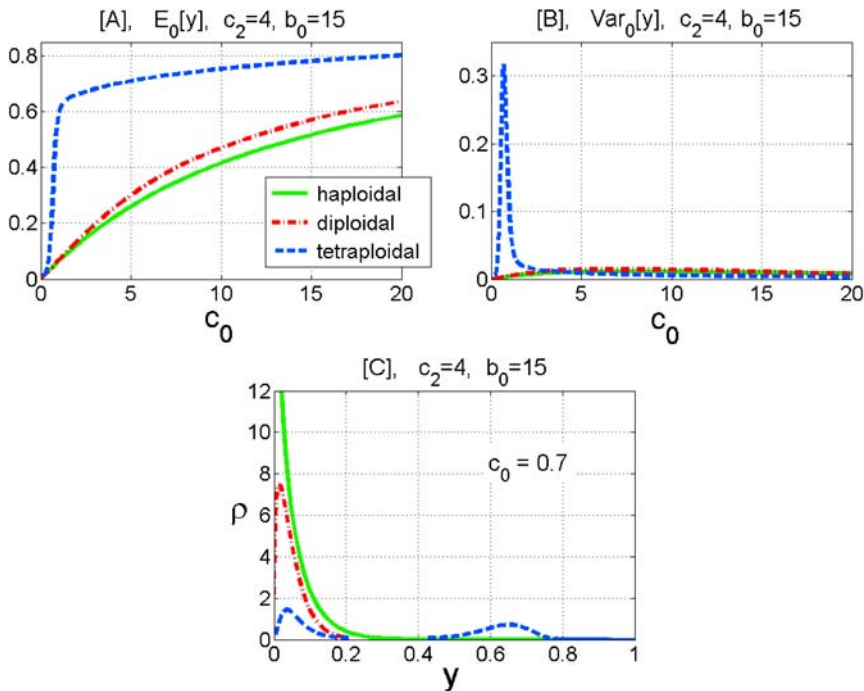
Fig. 7. *Color online*: Mode {BUU}: $c_2 = 4$, $b_0 = 15$; Panels A: mean per copy $E_0(y)(c_0)$ and B: variance per copy $Var_0(y)(c_0)$. Panel C protein distributions $\rho(y)$ for haploid, diploid, tetraploid systems for $c_0 = 0.7$ (at which tetraploid system has max$[Var_0(y)(c_0)]$).

S phase in which the genome is duplicated, G2 in which cell gets ready for the last phase mitosis. In mitosis the duplicated chromosomes are separated and pack into two new nuclei and then cell divides. During the G2 phase and mitosis cell is transiently tetraploid. In contrary in meiosis haploid cells are produced.

Cancer development is in many cases connected with partial genome multiplication, which results in reprogramming of the cancer cells. For example HeLa cells (human ovarian carcinoma) commonly used as a standard in many biological experiments have about 3 times more DNA than normal healthy cells, what implies that in average genes are hexaploid. Also in normal evolution particular genes may be multiplicated in a response for environmental stress, when a given protein is needed in larger quantity.

The aim of this paper was to analyze, via stochastic modeling of gene expression, the potential consequences of the change in gene copy number on its expression. The analysis of this problem was initiated by the seminal work by Cook et al.,[5] on influence of the stochastic gene expression on the haploinsufficiency.
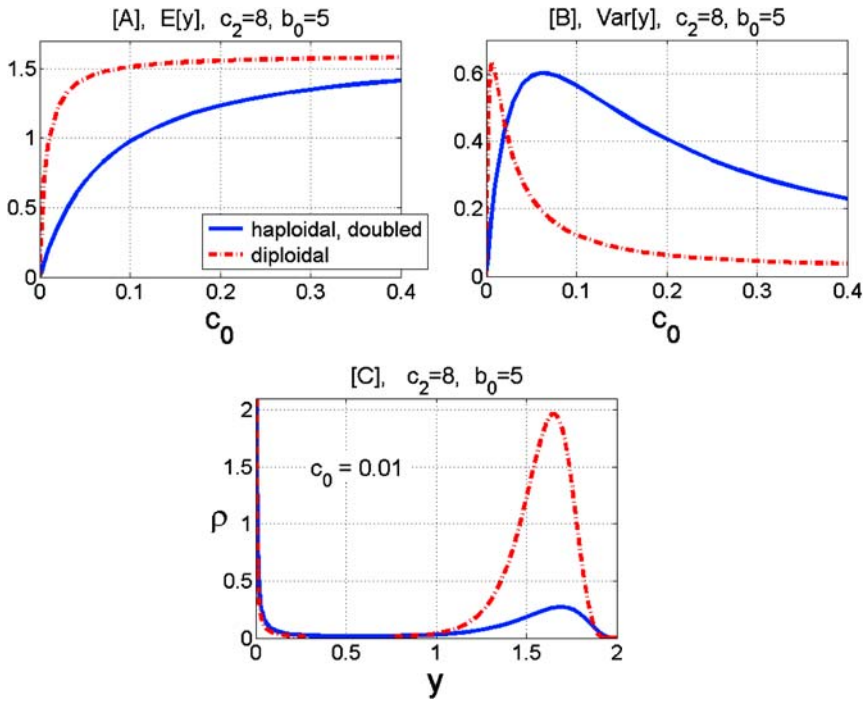
Fig. 8. *Color online:* Regulation of haploid gene with the doubled production and the diploid one (positive feedback); $c_2 = 8$, $b_0 = 5$. Panels A: mean $E(y)(c_0)$ and B: variance $Var(y)(c_0)$; Panel C: the protein distributions $\rho(y)$ for haploid and diploid systems for $c_0 = 0.01$ (at which diploid gene activates).

The recent developments, [10,30,25,26,15] in modeling the stochastic gene expression allowed us to apply a more rigorous treatment of this problem by means of protein probability density functions. We considered expression of haploid, diploid and tetraploid genes, first without any autoregulation, then with positive autoregulation. In the case without the autoregulation the activity of each allele of an N-allelic gene is independent and thus the mean and the variance of the protein distribution are proportional to the number of gene copies provided that the expression from single copy is the same. When protein expression per copy is adjusted, so that the average expression of N-ploid gene equals to that of the haploid one, then (in the continuous approximation in which transcriptional and translational noise is neglected) the variance is inversely proportional to N. This explains the numerical result of Cook *et al.*[5] who found, performing Monte Carlo simulations, that the variance of the protein distribution decreases with the growth of the copy number (provided that averaged expression is constant).

The dependence of the gene expression on the number of alleles is much more complex for systems with feedback. In this case neither the protein level nor variance of protein distribution are proportional to the number of gene copies. In this work we restricted ourself to autoregulatory genes, what surely leaves some interesting phenomena unearthed, but also enables more detail analysis. Despite their simplicity the autoregulatory elements have two features which make them sensitive to the change in the gene copy number. Firstly, the higher protein level implies the stronger feedback, secondly the higher copy number increases the probability of gene activation.

We considered the external induction of haploid, diploid and tetraploid genes with positive feedback to found that in each case there exist three characteristic modes of gene activation; {A}-when gene activity is sustained without external stimulation, {U}-when gene activates at some value of the external stimulation with the unimodal profile of the protein distribution, and {B}-when gene activates via the bimodal distribution. In both {U} and {B} modes gene acts as a switch, in addition in the {B} mode the stochasticity split transiently the population of cells into two subpopulations: one with high other with low protein expression. In some cases this can be the safest strategy for a tissue or a colony of bacteria. Our analysis (Fig. 3) shows that the diploid gene has the most robust pattern of activation in the sense that the {A},{B},{U} sectors in the (c_2, b_0) parameter plane are of the similar width.

The main result of this work is presented in Fig. 4: Considering the simultaneous activation of a haploid, diploid or tetraploid gene we found that there exist nine modes of activation. The analysis of these modes shows the potential consequences of allele loss or gene duplication. When the system is in one of two modes, {AAU} or {AAB} then the allele loss would stop the persistent gene activity and may lead to the disease if the constant level of product is required. As shown in Fig. 6 in the case {AAB} result not only in lower level of protein, but also in much higher variance what implies transient drops of protein level to the very low value. In the case when the autoregulated gene is the transcription factor regulating the transcription of other genes, the variability in its level would propagate further down stream and may lead to a disease. Among over 500 transcription factors encoded in the human genome there are more than 30 that can cause the disease by haploinsufficiency.[23]

Our analysis shows that the gene duplication also may alter its functioning. For example, when the system is in the {AUU} or the {ABU} mode, then the duplication of a diploid gene into tetraploid one would result in a persistent activity. Similarly, when the system is in the {AAB} or the {AAU} mode, duplication of a haploid gene would also result in its persistent activity. In the case when the haploid or diploid gene is "designated" to act as a switch its duplication can lead to a disease. In fact the positive feedback regulation is a frequent element of switch. Based on Fig. 4 we may expect that the temporal genome duplication in

a S phase of cell cycle may result in setting genes regulated in positive feedback into the active state. This implies the different gene expression in G1 (preceding S) and G2 (following S) phases.

Finally, we showed that the loss of a gene copy may not be compensated by higher expression of the remaining one. Even when the haploid gene has a twice higher production per copy than the diploid one it still requires stronger signal to activate. This has an important consequences for modelling: whenever we are interested in the exact description of the diploid gene regulation we should take into account the fact that both gene copies can be activated independently. In the case shown in Fig. 8, a ten times stronger signal is needed to activate the haploid than the diploid gene. Moreover, since activation of the diploid gene proceeds through bimodal distribution it exhibits transiently higher variance. This result is quite counterintuitive, since it is commonly expected that the increased number of gene copies with proportional reduction of transcription rate per copy results in a lower variability.

ACKNOWLEDGMENTS

REFERENCES

1. G. K. Basak, A. Bisi and M. K. Ghosh, Stability of degenerate diffusions with state-dependent switching, *J. Math. Anal. Appl.* **240**:219–248 (1999).
2. W. J. Blake, M. Kaern, C. R. Cantor and J. J. Collins, Noise in eukaryotic gene expression, *Nature* **422**:633–637 (2003).
3. A. Bobrowski, Degenerate convergence of semigroups related to a model of stochastic gene expression. *Semigroup forum* **73**:345–366 (2006).
4. R. Cheong, A. Bergmann, Sh. L. Werner, J. Regal, A. Hoffmann and A. Levchenko, Transient IκB kinase activity mediates temporal NF-κB dynamics in response to wide rage of tumor necrosis factor- doses. *J. Biol. Chem.* **281**:2945–2950 (2006).
5. D. A. Cook, A. N. Gerber and S. J. Tapscott, Modeling stochastic gene expression: Implications for haploinsufficiency, *Proc. Natl. Acad. Sci.* (USA) **95**:15641–15646 (1998).
6. W. D. Cook and B. J. McCaw, Accommodating haploinsufficient tumor suppressor genes in Knudson's model, *Oncogen*e **19**:3434–3438 (2000).
7. M. B. Elowitz, A. J. Levine, E. D. Siggia and P. S. Swain, Stochastic gene expression in a single cell, *Science* **297**:1183–1186 (2002).
8. D. T. Gillespie, Exact stochastic simulations of coupled chemical reactions, *J. Phys. Chem.* **81**:2340–2361 (1977).
9. M. Kaern, C. T. Elston, W. J. Blake and J. J. Collins, Stochasticity in gene expression: From theories to phenotypes, *Nat. Rev. Genet.* **6**:451–464 (2005).
10. T. B. Kepler and T. C. Elston, Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations, *Biophys. J.* **81**:3116–3136 (2001).

11. A. M. Kierzek, J. Zaim and P. Zielenkiewicz, The effect of transcription and translation initiation frequencies on the stochastic fluctuations in prokaryotic gene expression, *J. Biol. Chem.* **276:**8165–8172 (2001).
12. M. S. H. Ko, Stochastic model for gene induction, *J. Theor. Biol.* **153**:181–194 (1991).
13. T. Lipniacki, P. Paszek, A. R. Brasier, B. Luxon and M. Kimmel, Mathematical model of NF-kB regulatory module, *J. Theor. Biol.* **228**:195–215 (2004).
14. T. Lipniacki, P. Paszek, A. R. Brasier, B. Luxon and M. Kimmel, Stochastic regulation in early immune response, *Biophys. J.* **90**:725–742 (2006).
15. T. Lipniacki, P. Paszek, A. Marciniak-Czochra, A. R. Brasier and M. Kimmel, Transcriptional stochasticity in gene expression, *J. Theor. Biol.* **238**:348–367 (2006).
16. H. H. McAdams and A. Arkin, Stochastic mechanisms in gene expression, *Proc. Natl. Acad. Sci.* (USA) **94**:814–819 (1997).
17. J. Paulsson, Models of stochastic gene expression, *Phys. Life Rev.* **2:**157–175 (2005).
18. P. Paszek, T. Lipniacki, A. R. Brasier, B. Tian, D. E. Nowak and M. Kimmel, Stochastic effects of multiple regulators on expression profiles in Eukaryotes, *J. Theor. Biol.* **233**:423–433 (2005).
19. P. Paszek, Modeling stochasticity in gene regulation: Characterization in the terms of the underlying distribution function, *Bull. Math. Biol.* in press (2007) DOI 10.1007/s11538-006-9176-7.
20. J. R. Pirone and T. C. Elston, Fluctuations in transcription factor binding can explain the graded and binary responses observed in inducible gene expression, *J. Theor. Biol.* **226**:111–121 (2004).
21. Ch. V. Rao, D. M. Wolf and A. P. Arkin, Control, exploitation and tolerance of intracellular noise, *Nature* **420**:231–237 (2002).
22. J. M. Raser and E. K. O'Shea, Control of stochasticity in eukaryotic gene expression, *Science* **304**:1811–1814 (2004).
23. J. D. Seidman and Ch. Seidman, Transcription factor haploinsufficiency: When half a loaf is not enough, *J. Clinical Investigations* **109**:451–455 (2002).
24. J. A. Stirland, Z. C. Seymour, S. Windeatt, A. J. Norris, P. Stanley, M. G. Castro, A. S. I. Loudon, M. R. H. White, and J. R. E. Davis, Real-time imaging of gene promoter activity using an adenoviral reporter construct demonstrates transcriptional dynamics in normal anterior pituary cells, *J. Endocrinology* **178**:61–69 (2003).
25. Y. Tao, Intrinsic and external noise in an auto-regulatory genetic network, *J. Theor. Biol.* **229**:147–156 (2004).
26. Y. Tao, Intrinsic noise, gene regulation and steady-state statistics in a two-gene network, *J. Theor. Biol.* **231**:63–568 (2004).
27. N. Takasuka, M. R. H. White, C. D. Wood, W. R. Robertson and J. R. E. Davis, Dynamic changes in prolactin promoter activation in individual living lactotrophic cells, *Endocrinology* **139**:1361–1368 (1998).
28. M. Thattai and A. Oudenaarden, Intrinsic noise in gene regulatory networks, *Proc. Natl. Acad. Sci.* (USA) **98**:8614–8619 (2001).
29. R. Tomioka, H. Kimura, T.J. Kobayashi and K. Aihara, Multivariate analysis of noise in genetic regulatory networks, *J. Theor. Biol.* **229**:501–521 (2004).
30. N. G. van Kampen, Stochastic Processes in Physics and Chemistry (North-Holland, Amsterdam) (1992).
31. M. C. Walters, S. Fiering, J. Eidemiller, W. Magis, M. Groudine and D. I. K. Martin, Enhancers increase the probability but not the level of gene expression, *Proc. Natl. Acad. Sci.* (USA) **92**:7125–7129 (1995).