Noname manuscript No.
(will be inserted by the editor)

On the impact of the distance between two genes on their interaction curve

Siamak Taati · Enrico Formenti · Jean-Paul Comet · Gilles Bernot

the date of receipt and acceptance should be inserted later

Abstract We analyze a basic building block of gene regulatory networks using a stochastic/geometric model in search of a mathematical backing for the discrete modeling frameworks. We consider a network consisting only of two interacting genes: a source gene and a target gene. The target gene is activated by the proteins encoded by the source gene. The interaction is therefore mediated by activator proteins that travel, like a signal, from the source to the target. We calculate the production curve of the target proteins in response to a constant-rate production of activator proteins. The latter has a sigmoidal shape (like a simple delay line) that is sharper and taller when the two genes are closer to each other. This provides further support for the use of discrete models in the analysis gene regulatory networks. Moreover, it suggests an evolutionary pressure towards making the interacting genes closer to each other to make their interactions more efficient and more reliable.

Keywords gene regulatory networks · stochastic model · Poisson process · Brownian motion

Mathematics Subject Classification (2000) (MSC2010) 92B05 · 92C42 · 60G55 · 60J70

1 Introduction

It is a well-known property in biology that gene interactions are *sigmoidal*. If a gene σ is an activator of a gene τ, and assuming any other interactions being constant, then, when σ gives a signal of activation to τ, the concentration level of the protein produced by τ follows a sigmoidal shape.

It may seem surprising at first glance that the non-linearities reflected by these sigmoids *simplify* the analysis of a gene network model as a whole. This

Université de Nice-Sophia Antipolis,
Laboratoire I3S, 2000 Route des Lucioles, 06903 Sophia Antipolis (France)

is mainly due to the clear *qualitative* distinction that they induce in the set of possible trajectories, facilitating the emergence of a set of clearly distinct stable states and clarifying the high dependency to initial states. Sigmoidal behaviours allow the modeling process to focus on a smaller number of simplified parameters and facilitate the identification activity.

Continuous modeling approaches for gene regulatory networks Tyson et al. (2008); Leloup and Goldbeter (2004) often use Hill functions, for example, to describe this behavior. Discrete modeling approaches even strongly rely on these sigmoidal behaviors because the discrete values representing concentration levels are defined according to thresholds and each threshold is indeed the inflection point of such a sigmoid Glass and Kauffman (1973); Thomas and Kaufman (2001a,b); Bernot et al. (2004).

Although widely observed for a long time in cellular biology, these sigmoidal behaviours became the object of theoretical studies rather recently. For example in biochemistry and biophysics, at the molecular level, in Halford and Marko (2004); Halford (2009); Wunderlich and Mirny (2008) the authors have studied the efficiency of protein binding to specific sites of the DNA strand. The emphasis of these theoretical models is on the recognition of the DNA sites but the models also exhibit a sigmoidal behaviour. In Wunderlich and Mirny (2008) also an interesting distinction is made between 3D diffusion of the proteins and 1D search along DNA.

Also rather recently, biologists have made very interesting observations about the functional organization, folding and evolution of chromosomes. The genes that are co-transcribed under a given stress are preferentially placed at periodic distances along the chromosomes Junier et al. (2010). A chromosome, inside the cell, is dynamically rearranged in such a way that co-working genes are closer to each other when they need to react together, and natural selection seems to favour these periodic distances.

The significance of the spatial arrangement of the interacting genes is, among others, influenced by the fact that the proteins mediating the interactions and playing the role of signals must in a way locate their targets. The mechanism by which a protein locates and binds to its target gene is a subject of some debates, as the 3D diffusion alone is not rapid enough to lead the process (see Wunderlich and Mirny (2008); Halford and Marko (2004)). However, according to Halford (2009), the action of 3D diffusion is predominant before the protein reaches a neighborhood of the target gene ($\sim$50–100 bp).

So, dynamically, the impact of the distance between interacting genes, *independently of any local site recognition phenomena* along DNA, seems to have an important contribution to the biological functions, especially to rapidly start these functions under a stress. The contribution of this article is to precisely study the impact of the *distance* between two interacting genes inside a prokaryote. We relate the quality of the sigmoidal shape of the transient behaviour to the distance between the interacting genes, using an abstract stochastic/geometric model.

The "behaviour" of the genes has been intentionally abstracted (ignoring the two stage translation/transcription production) in such a way that σ could

be also considered as a sort of "source" of transcription factors, or simply a transitory place from which transcription factors come. Similarly, the "place" of a gene intentionally ignores the DNA structure, elasticity or 1D distance along the chromosome, in such a way that we only focus on the actual distance between σ and τ. We prove that considering *the distance between the genes* alone, already induces a sigmoidal interaction. We also show that, the lower the distance, the sharper the sigmoid; thus, a short distance induces a rapid start of the gene function. It may explain by itself why the natural selection favours reconfigurations of the chromosomes that rapidly co-locate co-working genes sharing transcription factors.

2 The Model

We encapsulate the transcription and translation stages into a single step in which new proteins are produced in the vicinity of the genes. We see the genes σ (source) and τ (target) as points in the 3-dimensional Euclidean space that are at distance D from each other. New S proteins are produced at σ according to a Poisson process with rate $\lambda_S > 0$. This amounts to the assumption that the number of proteins produced in disjoint intervals are independent and the probability that a new protein is produced during an infinitesimal time δt is $\lambda_S \delta t$.

Each S protein in our model follows a two-phase process to locate and interact with the target gene τ. The first phase is a simple diffusion; we see it as a Brownian motion in the 3-dimensional space. We denote the diffusion rate by $\beta > 0$. The diffusion of each S protein is assumed to be independent of the production and diffusion of the other S proteins.

The diffusion continues until the protein arrives (if at all) at the "range of interaction" $\mathbf{R}$ of the target τ. We consider the range of interaction $\mathbf{R}$ simply as a sphere with radius $r > 0$ centered at τ. After arriving at $\mathbf{R}$, the process enters its second phase. In this phase, the diffusion is affected by other factors (e.g., the electrostatic forces between the involved molecules, temporary attachments to the DNA strand, etc.) that eventually lead to the production of new T proteins at τ. Since here we are only concerned with the role of the distance D between the two genes, we model the generation of T proteins in this phase simply as a Poisson process: upon its arrival at $\mathbf{R}$, each S protein triggers a Poisson process with rate $\hat{\lambda}_T > 0$ for the production of new T proteins at τ. This T production process continues as long as the initiating S protein has not degraded. The T production processes triggered by different S proteins are assumed to be independent of each other and of the other elements of the model.

Finally, each protein may degrade (become annihilated) according to an exponential decay process, independent of everything else in the model. To simplify the calculations, we assume that S and T proteins both degrade at the same rate $\varepsilon > 0$: the probability that a protein (S or T) degrades during an infinitesimal time δt is $\varepsilon \delta t$.

2.1 Protein Concentrations

The concentration of the proteins of type S or T at a given time can be measured by their actual numbers. We denote the number of S and T proteins at time t by $n_S(t)$ and $n_T(t)$, respectively. These are random variables. We now calculate their expected values. We denote by $p(t_1 - t_0)$, the probability that an S protein produced at time t_0 arrives at the range of interaction of τ before time $t_1 > t_0$, provided it has not degraded during that period. This will be calculated in the next section.

The basic tool used in the calculations of this section is Campbell's theorem (see Kingman (1993)). A Poisson process with rate λ on $\mathbb{R}^+$ can be identified with a random countable set $\xi \subseteq \mathbb{R}^+$. Let $f : \mathbb{R}^+ \to \mathbb{C}$ be an arbitrary measurable function and $I \subseteq \mathbb{R}^+$ a bounded measurable set. Then, Campbell's theorem states that the expected value of the sum over $\xi \cap I$ of f is the same as the integral of λf over I.

Proposition 1

$$\mathbf{E}\left[n_S(t)\right] = \frac{\lambda_S}{\varepsilon}\left(1 - e^{-\varepsilon t}\right) .$$

Proof Denote the set of times $\theta \in \mathbb{R}^+$ in which a new S protein is produced by ξ_S. We write $\xi_S(t)$ for $\xi_S \cap [0, t]$. To calculate $\mathbf{E}\left[n_S(t)\right]$, we condition $n_S(t)$ to ξ_S and use $\mathbf{E}\left[n_S(t)\right] = \mathbf{E}\left[\mathbf{E}\left[n_S(t) \,|\, \xi_S\right]\right]$. Since each S protein may degrade with rate ε, we have

$$\mathbf{E}\left[n_S(t) \,|\, \xi_S\right] = \sum_{\theta \in \xi_S(t)} e^{-\varepsilon(t-\theta)} .$$

Therefore, according to Campbell's theorem we have

$$\mathbf{E}\left[n_S(t)\right] = \int_0^t e^{-\varepsilon(t-\theta)} \lambda_S \mathrm{d}\theta = \frac{\lambda_S}{\varepsilon}\left(1 - e^{-\varepsilon t}\right) . \quad \square$$

Proposition 2

$$\mathbf{E}\left[n_T(t)\right] = \frac{\lambda_S \hat{\lambda}_T}{\varepsilon} \int_0^t \left(e^{-\varepsilon\theta} - e^{-\varepsilon t}\right) p(\theta) \mathrm{d}\theta .$$

Proof For an S protein produced at time s, let $n_T^s(t)$ denote the number of *children* of s at time $t > s$. By the children of s we mean the T proteins produced in the process triggered by s. Clearly, the distribution of $n_T^s(t)$ depends only on $t - s$ and not on t or s separately.

Let us first calculate $\mathbf{E}\left[n_T^0(t)\right]$. Consider a (potential) S protein produced at time 0. Denote by $\delta \in [0, +\infty)$ the time at which this S protein degrades. Let $h \in [0, +\infty]$ represents the time at which this S protein arrives at the range of interaction of τ. If the protein degrades before reaching the target we set $h \triangleq +\infty$. Finally, denote by ξ_T^0 the set of T proteins produced as a result of the interaction of this S protein with the target (i.e., the set of children of

this S protein). In the present model, given h and δ, ξ_T^0 is a Poisson process with rate $\hat{\lambda}_T$ on the interval $[h, \delta]$. We write $\xi_T^0(t)$ for $\xi_T^0 \cap [0, t]$.

Since each T protein annihilates with rate ε, we have

$$\mathbf{E}\left[n_T^0(t) \,\middle|\, h, \delta, \xi_T^0\right] = \sum_{r \in \xi_T^0(t)} e^{-\varepsilon(t-r)} \,.$$

From Campbell's theorem we get

$$\begin{aligned}
\mathbf{E}\left[n_T^0(t) \,\middle|\, h, \delta\right] &= \int_{\min(h,t)}^{\min(\delta,t)} e^{-\varepsilon(t-r)} \hat{\lambda}_T \mathrm{d}r \\
&= \begin{cases} \frac{\hat{\lambda}_T}{\varepsilon}\left(1 - e^{-\varepsilon(t-h)}\right) & \text{if } h \le t < \delta, \\ \frac{\hat{\lambda}_T}{\varepsilon}\left(e^{-\varepsilon(t-\delta)} - e^{-\varepsilon(t-h)}\right) & \text{if } h \le \delta \le t, \\ 0 & \text{if } t < h \text{ or } \delta < h. \end{cases}
\end{aligned}$$

Taking expectation with respect to δ for $h \le t$, we obtain

$$\begin{aligned}
\mathbf{E}\left[n_T^0(t) \,\middle|\, h\right] &= e^{-\varepsilon t}\, \mathbf{E}\left[n_T^0(t) \,\middle|\, h, \{\delta > t\}\right] + \int_h^t \mathbf{E}\left[n_T^0(t) \,\middle|\, h, \{\delta = x\}\right] \varepsilon e^{-\varepsilon x} \mathrm{d}x \\
&= \frac{\hat{\lambda}_T}{\varepsilon} e^{-\varepsilon t}\left(1 - e^{-\varepsilon(t-h)}\right) + \hat{\lambda}_T e^{-\varepsilon t} \int_h^t \left(e^{\varepsilon x} - e^{\varepsilon h}\right) e^{-\varepsilon x} \mathrm{d}x \\
&= \hat{\lambda}_T e^{-\varepsilon t}(t - h) \,.
\end{aligned}$$

Recall that the probability that the S protein hits the target before time h is denoted by $p(h)$. Therefore, resolving the conditioning relative to h, we have

$$\begin{aligned}
\mathbf{E}\left[n_T^0(t)\right] &= \int_0^t \mathbf{E}\left[n_T^0(t) \,\middle|\, h = x\right] \mathrm{d}p(x) \\
&= \int_0^t \hat{\lambda}_T e^{-\varepsilon t}(t - x)\, \mathrm{d}p(x) \\
&= \hat{\lambda}_T e^{-\varepsilon t}\left(t p(t) - \int_0^t x\, \mathrm{d}p(x)\right) \\
&= \hat{\lambda}_T e^{-\varepsilon t} \int_0^t p(x)\, \mathrm{d}x \,.
\end{aligned}$$

Next, let us use $\mathbf{E}\left[n_T^0(t)\right]$ to calculate $\mathbf{E}\left[n_T(t)\right]$. As before, denote the set of time moments $s \in \mathbb{R}^+$ in which a new S protein is produced by ξ_S and write $\xi_S(t)$ for $\xi_S \cap [0, t]$. We have

$$n_T(t) = \sum_{s \in \xi_S(t)} n_T^s(t) \,,$$

and hence

$$\mathbf{E}\left[n_T(t) \,|\, \xi_S(t)\right] = \sum_{s \in \xi_S(t)} \mathbf{E}\left[n_T^s(t)\right] = \sum_{s \in \xi_S(t)} \mathbf{E}\left[n_T^0(t - s)\right] \,.$$

Thus, Campbell's theorem implies that

$$\begin{aligned}\mathbf{E}\left[n_T(t)\right] &= \int_0^t \mathbf{E}\left[n_T^0(t-s)\right]\lambda_S\,\mathrm{d}s = \lambda_S\int_0^t \mathbf{E}\left[n_T^0(r)\right]\mathrm{d}r\\ &= \lambda_S\hat{\lambda}_T\int_0^t e^{-\varepsilon r}\int_0^r p(x)\,\mathrm{d}x\,\mathrm{d}r\\ &= \lambda_S\hat{\lambda}_T\int_0^t p(x)\int_x^t e^{-\varepsilon r}\,\mathrm{d}r\,\mathrm{d}x\\ &= \frac{\lambda_S\hat{\lambda}_T}{\varepsilon}\int_0^t \left(e^{-\varepsilon x}-e^{-\varepsilon t}\right)p(x)\,\mathrm{d}x\ ,\end{aligned}$$

concluding the proof. □

2.2 Mechanism of Diffusion

In this section, we describe the probability distribution $p(t)$ of the time it takes for an S protein produced at σ to reach the range of interaction of τ. As mentioned above, we consider this range of interaction $\mathbf{R}$ as a sphere with radius r centered at τ. In reference to the description given in Halford (2009), r is of order of 50 bp on the target DNA molecule (i.e., ∼20–30 nm), whereas D is of intermolecular scale. We assume that the S protein follows a 3-dimensional Brownian motion (Wiener process) with diffusion rate $\beta > 0$: the total displacement of the protein within an interval $[s,t]$ $(0 \le s < t)$ has a normal distribution with mean 0 and variance $\beta(t-s)$, and the displacements in disjoint intervals are independent (see e.g. Folland (1999) or Krylov (1991)).

For a point $x \in \mathbb{R}^d$ and a compact set $\mathbf{R} \subseteq \mathbb{R}^d$ (not including x) in the d-dimensional Euclidean space, let us denote by $H_d(x,\mathbf{R},t)$ the probability that a standard d-dimensional Brownian motion (i.e., with diffusion rate 1) starting at x and time 0 hits the region $\mathbf{R}$ before time t. Hence, in our model $p(t) = H_3(\sigma, \overline{B_r(\tau)}, \beta t)$, where $\overline{B_r(\tau)}$ is the closed ball with radius r around τ, and the distance between σ and τ is $D > r$.

When $\mathbf{R} = \overline{B_r(y)}$ is a closed ball with radius r whose center y is at distance $D > r$ from x, Yin and Wu (1996) have calculated $H_d(x,\mathbf{R},t)$ in any number of dimensions d as

$$H_d(x,\overline{B_r(y)},t) = \frac{2}{\pi}\left(\frac{r}{D}\right)^{\alpha}\int_0^{\infty}\left(e^{-t\theta^2/2}-1\right)\frac{Q(\alpha,\theta,D,r)}{\theta}\,\mathrm{d}\theta\ , \qquad (1)$$

where $\alpha = d/2 - 1$ and

$$Q(\alpha,\theta,D,r) = \frac{J_\alpha(D\theta)N_\alpha(r\theta) - J_\alpha(r\theta)N_\alpha(D\theta)}{J_\alpha^2(r\theta)+N_\alpha^2(r\theta)} \qquad (2)$$

and J_α and N_α are, respectively, the Bessel functions of the first and second kinds.

It is interesting to observe that formula (1) in dimensions 1 and 3 differ only by a factor of r/D. Namely, Equation (2) can be rewritten as

$$Q(\alpha,\theta,D,r)=\sin(\alpha\pi)\frac{J_\alpha(r\theta)J_{-\alpha}(D\theta)-J_\alpha(D\theta)J_{-\alpha}(r\theta)}{J_\alpha^2(r\theta)J_{-\alpha}^2(r\theta)-2J_\alpha(r\theta)J_{-\alpha}(r\theta)\cos(\alpha\pi)}$$

Note that $Q(\alpha,\theta,D,r)$ is invariant under change of sign of α. In particular, $Q(-1/2,\theta,D,r)=Q(1/2,\theta,D,r)$; that is to say in dimension 1 and 3, the term $Q(\alpha,\theta,D,r)$ is the same. Therefore,

$$H_3(x,\overline{B_r(y)},t)=\frac{r}{D}H_1(x',\overline{B_r(y')},t) \tag{3}$$

for $x,y\in\mathbb{R}^3$ and $x',y'\in\mathbb{R}$ with $|y-x|=|y'-x'|=D$.

In dimension 1, there is a well-known simpler formula

$$H_1(x',\overline{B_r(y')},t)=1-\operatorname{erf}\left(\frac{D-r}{\sqrt{2t}}\right) \tag{4}$$

for the probability distribution of the first time a standard Brownian motion hits an obstacle at distance $D-r>0$ (see e.g. Redner (2001), page 84). Here, $\operatorname{erf}(a)$ refers to the error function:

$$\operatorname{erf}(a)\triangleq\frac{2}{\sqrt{\pi}}\int_0^a e^{-z^2}\mathrm{d}z\ .$$

Combining (3) and (4), we obtain a simple expression

$$H_3(x,\overline{B_r(y)},t)=\frac{r}{D}\left(1-\operatorname{erf}\left(\frac{D-r}{\sqrt{2t}}\right)\right)$$

for the probability that a standard 3-dimensional Brownian motion hits a closed ball with radius $r>0$ and distance $D>r$ before time t.

Proposition 3

$$p(t)=\frac{r}{D}\left(1-\operatorname{erf}\left(\frac{D-r}{\sqrt{2\beta t}}\right)\right)\ .$$

2.3 Analysis

In this section, we verify that the concentration of T proteins in our model is indeed a sigmoidal function of time whose height grows exponentially as the distance between the two genes σ and τ decreases. We also demonstrate numerically that the transition duration of this sigmoidal function grows linearly with the distance between σ and τ. Hence from the current model, the closeness of the interacting genes seem to drastically affect the height and sharpness of the concentration curve of the target proteins. This suggests a strong evolutionary pressure towards making the interacting genes closer to each other.

We call a continuous function $f:(0,+\infty)\to(0,+\infty)$ *sigmoidal* if

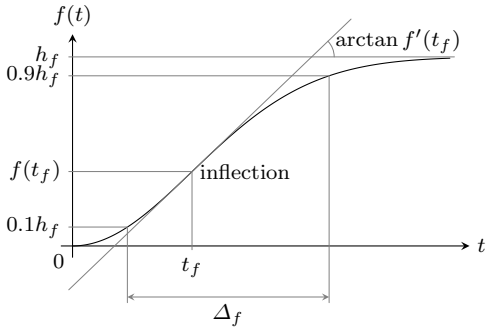

Fig. 1 A generic sigmoidal function with height h_f, steepness $f'(t_f)$, and width Δ_f.

i) its derivatives $f'(t)$ and $f''(t)$ exist and are continuous,
ii) it is monotone,
iii) $\lim_{t\searrow 0} f(t) = \lim_{t\searrow 0} f'(t) = 0$,
iv) $\lim_{t\to+\infty} f(t) < +\infty$, and
v) it has exactly one inflection point $0 < t_f < +\infty$ (at which the curvature changes from upward to downward).

We call the value $h_f \triangleq \lim_{t\to+\infty} f(t)$ the *height* of f and $f'(t_f)$ its *steepness*. We measure the transition duration of f by the value $\Delta_f \triangleq f^{-1}(0.9h_f) - f^{-1}(0.1h_f)$, which we call the *width* of f (see Figure 1). For example, the function $p(t)$ in our model is sigmoidal with height r/D, steepness $\sim 0.23\frac{2\beta r}{(D-r)^2 D}$, and width $\sim 125.9\frac{(D-r)^2}{2\beta}$ (see Figure 2).

The expected number of S proteins $\mathbf{E}[n_S(t)]$ does not have a sigmoidal curve as it has no inflection point on $(0, +\infty)$ (see Figure 3). Yet it is increasing and bounded with $\mathbf{E}[n_S(0)] = 0$, and all its derivatives exist and are continuous. Therefore, we can still define and calculate its height and width: its height is λ_S/ε and its width is $\sim 2.20/\varepsilon$.

Let us verify that the expected number of T proteins $\mathbf{E}[n_T(t)]$ does indeed have a sigmoidal curve (Figure 4(a)).

Proposition 4 *The function $G_T(t) \triangleq \mathbf{E}[n_T(t)]$ is sigmoidal.*

Proof From Proposition 2, we have

$$G_T(t) = \frac{\lambda_S \hat{\lambda}_T}{\varepsilon} \int_0^t \left(e^{-\varepsilon\theta} - e^{-\varepsilon t}\right) p(\theta)\mathrm{d}\theta \,, \tag{5}$$

which is a continuous function of t. Clearly $\lim_{t\searrow 0} G_T(t) = 0$. Since $p(t)$ has continuous derivatives of any degree on $(0, +\infty)$, so does $G_T(t)$. The first derivative of $G_T(t)$ is

$$G'_T(t) = \lambda_S \hat{\lambda}_T e^{-\varepsilon t} \int_0^t p(x)\mathrm{d}x$$

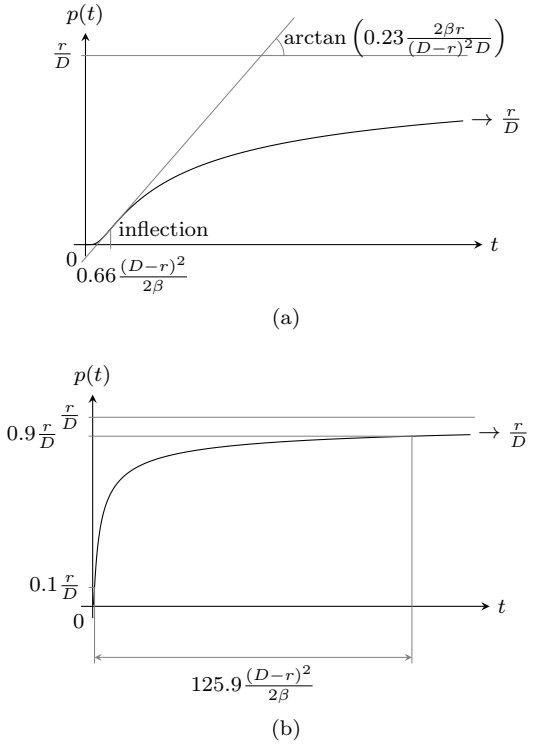

Fig. 2 The probability $p(t)$ in two different time scales.

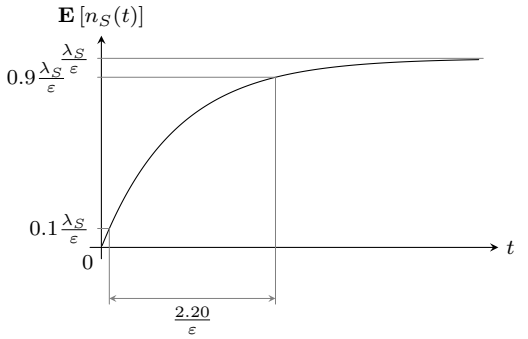

Fig. 3 The expected number of S proteins $\mathbf{E}\left[n_S(t)\right]$.

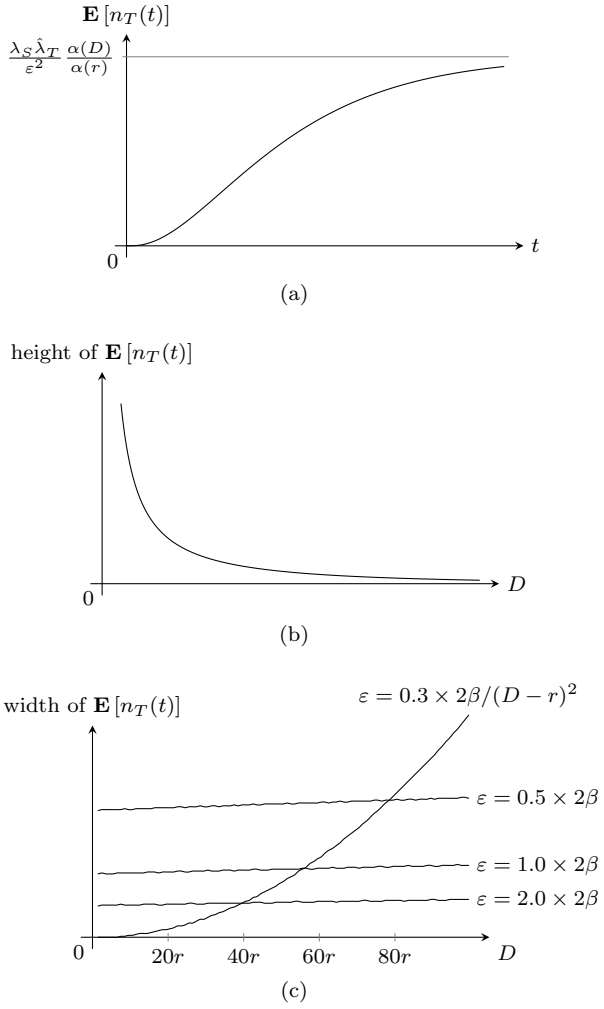

Fig. 4 (a) The expected number of T proteins $\mathbf{E}[n_T(t)]$. (b) The height of $\mathbf{E}[n_T(t)]$ as a function of D. (c) The width of $\mathbf{E}[n_T(t)]$ as a function of D for different values of ε (numerical solution).

Again we have $\lim_{t\searrow 0} G'_T(t) = 0$. Moreover, since $p(t)$ is positive, so is $G'_T(t)$, and therefore $G_T(t)$ is monotonically increasing. Also, since $p(t) < r/D$, we have

$$G'_T(t) \leq \lambda_S \hat{\lambda}_T e^{-\varepsilon t} \int_0^t \frac{r}{D} \mathrm{d}x = \lambda_S \hat{\lambda}_T \frac{r}{D} e^{-\varepsilon t} t \ ,$$

which implies

$$\begin{aligned} \lim_{t\to+\infty} G_T(t) &= \int_0^{+\infty} G'_T(x)\mathrm{d}x \\ &\leq \lambda_S \hat{\lambda}_T \frac{r}{D} \int_0^{+\infty} e^{-\varepsilon x} x \mathrm{d}x \\ &= \frac{\lambda_S \hat{\lambda}_T}{\varepsilon^2} \frac{r}{D} \\ &< +\infty \ . \end{aligned}$$

The first derivative $G'_T(t)$ is everywhere positive and continuous, and converges to 0 as $t \to +\infty$ or $t \searrow 0$. Hence, it has a maximum value. Therefore, in order to show that $G_T(t)$ has a unique inflection point in $(0, +\infty)$, it is sufficient to show that its second derivative $G''_T(t)$ has a unique zero in $(0, +\infty)$. We have

$$G''_T(t) = \lambda_S \hat{\lambda}_T e^{-\varepsilon t} \left(p(t) - \varepsilon \int_0^t p(x)\mathrm{d}x \right) \ . \tag{6}$$

If we substitute $p(t)$ in (6) with its value from Proposition 3 and use the shorthand $c \triangleq \frac{2\beta}{(D-r)^2}$, we get that $G''_T(t) = 0$ if and only if

$$\operatorname{erfc}\left(\frac{1}{\sqrt{ct}}\right) - \varepsilon \int_0^t \operatorname{erfc}\left(\frac{1}{\sqrt{cx}}\right) \mathrm{d}x = 0 \ , \tag{7}$$

where $\operatorname{erfc}(\cdot) \triangleq 1 - \operatorname{erf}(\cdot)$ is the complementary error function. Defining $k \triangleq c/\varepsilon$ and changing the variables using $s \triangleq ct$ and $y \triangleq cx$ we can rewrite (7) as

$$k \operatorname{erfc}\left(\frac{1}{\sqrt{s}}\right) - \int_0^s \operatorname{erfc}\left(\frac{1}{\sqrt{y}}\right) \mathrm{d}y = 0 \ . \tag{8}$$

Note that $s \mapsto s/c$ is one-on-one and onto on $(0, +\infty)$. Hence, $G_T(t)$ has a unique inflection point t in $(0, +\infty)$ if and only if (8) has a unique root s in $(0, +\infty)$. In order to show that for each $\varepsilon, c > 0$ (or equivalently, for each $k > 0$) Equation (8) has a unique root, it is enough to prove the following lemma.

Lemma 1 *The function*

$$u(s) \triangleq \frac{\displaystyle\int_0^s \operatorname{erfc}\left(1/\sqrt{y}\right) \mathrm{d}y}{\operatorname{erfc}\left(1/\sqrt{s}\right)}$$

is one-to-one and onto on $(0, +\infty)$.

A proof of this lemma can be found in Appendix B. □

Let us calculate the height of $\mathbf{E}[n_T(t)]$ as a function of the distance D. Combining Propositions 2 and 3 we have

$$\mathbf{E}[n_T(t)] = \frac{\lambda_S \hat{\lambda}_T}{\varepsilon} \frac{r}{D} \int_0^t \left(e^{-\varepsilon\theta} - e^{-\varepsilon t}\right) \operatorname{erfc}\left(\frac{1}{\sqrt{ct}}\right) \mathrm{d}\theta \,,$$

where $c \triangleq 2\beta/(D-r)^2$. This can be rewritten as

$$\begin{aligned}\mathbf{E}[n_T(t)] &= \frac{\lambda_S \hat{\lambda}_T}{c\varepsilon} \frac{r}{D} \left[\int_0^{ct} \left(e^{-(\varepsilon/c)x} - e^{-(\varepsilon/c)ct}\right) \operatorname{erfc}\left(\frac{1}{\sqrt{x}}\right) \mathrm{d}x\right] \\ &= \frac{\lambda_S \hat{\lambda}_T}{c\varepsilon} \frac{r}{D} \left[F\left(ct, \varepsilon/c\right) - e^{-(\varepsilon/c)ct} G(ct)\right] \,, \end{aligned} \tag{9}$$

where

$$F(s,a) \triangleq \int_0^s e^{-ax} \operatorname{erfc}\left(\frac{1}{\sqrt{x}}\right) \mathrm{d}x \,,$$

and

$$G(s) \triangleq \int_0^s \operatorname{erfc}\left(\frac{1}{\sqrt{x}}\right) \mathrm{d}x \,.$$

Expanding F and G (see Appendix A) it is easy to verify that

$$\lim_{s\to+\infty} e^{-as} G(s) = 0 \,, \qquad \text{and} \qquad \lim_{s\to+\infty} F(s,a) = \frac{1}{a} e^{-2\sqrt{a}} \,.$$

Therefore, we obtain that the height of $\mathbf{E}[n_T(t)]$ is

$$\lim_{t\to+\infty} \mathbf{E}[n_T(t)] = \frac{\lambda_S \hat{\lambda}_T}{\varepsilon^2} \frac{\alpha(D)}{\alpha(r)} \,.$$

where $\alpha(x) = e^{-2\sqrt{\varepsilon/(2\beta)}x}/x$ (see Figure 4(b)).

From (9) we can also see that if we choose $\varepsilon \sim c \sim 1/(D-r)^2$, the value the distance D affects only the scale of the curve $\mathbf{E}[n_T(t)]$ and not its shape. In particular, it follows that with the constraint $\varepsilon \sim 1/(D-r)^2$, the width of $\mathbf{E}[n_T(t)]$ behaves like an increasing quadratic function $\sim(D-r)^2$. For a fixed value of ε, it appears from numerical calculations that the width of $\mathbf{E}[n_T(t)]$ increases linearly with D. See Figure 4(c) for numerical approximations of the width for few choices of the value of ε.

3 Conclusions

The main contribution of this article is to establish an exact formula for the transient behavior of two interacting genes, taking into account their geometric arrangement. We mixed together a probabilistic and a geometric approach, based only on the individual behavior of single molecules and taking into account the Euclidian distance between genes. This provides an alternative to more classical approaches based on concentrations (often using differential equations).

When applied to transcription factors, for example, our analysis shows that the delay of the production of the target gene is considerably shorter when the distance covered by the transcription factors is short. The sharpness of the sigmoidal response of the target gene is abruptly increased for short distances and this result is independent of any "local" molecular recognition phenomenon. When a biological function involving several genes has to be triggered rapidly, it is often observed that the genes share the same transcription factor(s) and that they dynamically co-locate when their transcriptions start Junier et al. (2010). Our result may explain why the natural selection favours such co-localizations.

The modeling approach and the underlying analytical methods used to establish our result can be helpful to address several connected questions at different levels of intracellular phenomena. Among several possible subjects, we plan to develop the following research directions:

- There can be several transcription processes at the same time on a single gene and the number of such simultaneous transcriptions may depend on several limiting factors (e.g. the length of the gene). This phenomenon can be seen as a kind of "pipelining" in the production of RNA. It could be interesting to extend our model to cover such phenomena.
- It has been observed that, in prokaryotes, genes coding for membrane proteins are located near the membrane Jensen and Shapiro (2000), allowing the proteins under construction to anchor into the membrane. Technics similar to the one used in this paper might be adapted to try to explain the localization of such genes.
- A more ambitious extension of our modeling may address the geometry of metabolic pathways. At the metabolic level, it seems that the localization of enzymes plays an important role: several researchers have pointed out the importance of transitory structures in order to optimize the flux of metabolites Thellier et al. (2006); Norris et al. (2007).
- Also, in the present paper, we assume a homogeneous cytoplasm. An interesting question is how to extend our current modeling to consider different compartments inside the cell: fluxes between compartments are crucial in cellular biology.

For all these models an important tricky issue is to properly tune the level of abstraction to smoothly reflect behaviours at the cellular level.

4 Acknowledgements

We wish to thank the anonymous referees for their valuable comments.

References

G. Bernot, J.-P. Comet, A. Richard, and J. Guespin. Application of formal methods to biological regulatory networks: Extending Thomas' asynchronous logical approach with temporal logic. *Journal of Theoretical Biology*, 229(3):339–347, 2004.

G. B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Wiley-Interscience, 2nd edition, 1999.

L. Glass and S.A. Kauffman. The logical analysis of continuous, nonlinear biochemical control networks. *Journal of Theoretical Biology*, 39:103–129, 1973.

Stephen E. Halford. An end to 40 year of mistakes in DNA-protein association kinetics? *Biochemical Society Transactions*, 37:343–348, 2009.

Stephen E. Halford and John F. Marko. How do site-specific DNA-binding proteins find their targets. *Nucleic Acids Research*, 32(10):3040–3052, 2004.

R. B. Jensen and L. Shapiro. Proteins on the move: dynamic protein localization in prokaryotes. *Trends in Cell Biology*, 10(11):483–488, 2000.

I. Junier, O. Martin, and F. Képès. Spatial and topological organization of DNA chains induced by gene co-localization. *PLoS Computational Biology*, 6(2):e1000678, 2010.

J. F. C. Kingman. *Poisson Processes*. Oxford University Press, 1993.

N. V. Krylov. *Introduction to the Theory of Diffusion Processes*. American Mathematical Society, 1991. English Translation.

J.-C. Leloup and A. Goldbeter. Modeling the mammalian circadian clock: sensitivity analysis and multiplicity of oscillatory mechanisms. *J. Theor. Biol.*, 230(4):541–562, 2004.

V. Norris, T. den Blaauwen, A. Cabin-Flaman, R. H. Doi, R. Harshey, L. Janniere, A. Jimenez-Sanchez, D. J. Jin, P. A. Levin, E. Mileykovskaya, A. Minsky, M. Saier Jr., and K. Skarstad. A functional taxonomy of bacterial hyperstructures. *Microbiology and Molecular Biology Reviews*, 71(1):230–253, 2007.

Sidney Redner. *A Guide to First-Passage Processes*. Cambridge University Press, 2001.

M. Thellier, G. Legent, P. Amar, V. Norris, and C. Ripoll. Steady-state kinetic behaviour of functioning-dependent structures. *The FEBS Journal*, 273(18): 4287–4299, 2006.

R. Thomas and M. Kaufman. Multistationarity, the basis of cell differentiation and memory. I. structural conditions of multistationarity and other nontrivial behavior. *Chaos*, 11:170–179, 2001a.

R. Thomas and M. Kaufman. Multistationarity, the basis of cell differentiation and memory. II. logical analysis of regulatory networks in terms of feedback circuits. *Chaos*, 11:180–195, 2001b.

J. Tyson, A. Reka, A. Goldbeter, P. Ruoff, and J. Sible. Biological switches and clocks. *J. of Royal Society Interface*, 5(Suppl 1):S1–S8, 2008.

Zeba Wunderlich and Leonid A. Mirny. Spatial effects on the speed and reliability of protein-DNA search. *Nucleic Acids Research*, 36(11):3570–3578, 2008.

Chuancun Yin and Rong Wu. Some problems on balls and spheres for Brownian motion. *Science in China Series A: Mathematics*, 39(6):572–582, 1996.

A Some Useful Formulas

I. $\operatorname{erf}(x) = 0, \qquad \operatorname{erf}(-x) = -\operatorname{erf}(x), \qquad \lim_{x\to+\infty} \operatorname{erf}(x) = 1.$

II. $\operatorname{erfc}(x) \approx \dfrac{e^{-x^2}}{x\sqrt{\pi}}$ as $x \to +\infty$.

Proof Using l'Hôpital rule. □

III. $\displaystyle\int_0^s \operatorname{erfc}\left(\frac{1}{\sqrt{x}}\right) \mathrm{d}x = (s+2)\operatorname{erfc}\left(\frac{1}{\sqrt{s}}\right) - \frac{2}{\sqrt{\pi}} e^{-1/s}\sqrt{s}$, for $s \geq 0$.

Proof Fubini's theorem $\int_{x=0}^{s}\int_{y=1/\sqrt{x}}^{+\infty} = \int_{y=1/\sqrt{s}}^{+\infty}\int_{x=1/y^2}^{s}$ and integration by parts. □

IV. $\displaystyle\int e^{-a^2x^2-b^2/x^2}\mathrm{d}x = \frac{\sqrt{\pi}}{4a}\left[e^{2ab}\operatorname{erf}\left(ax+\frac{b}{x}\right) + e^{-2ab}\operatorname{erf}\left(ax-\frac{b}{x}\right)\right] + \text{constant}$, for $a, b \geq 0$.

Proof Define $f \triangleq ax+b/x$ and $g \triangleq ax-b/x$. Then $a^2x^2+b^2/x^2 = f^2-2ab = g^2+2ab$ and $(f'+g')/(2a) = 1$. □

V.

$$\int_0^s e^{-ax}\operatorname{erfc}\left(\frac{1}{\sqrt{x}}\right)\mathrm{d}x = \frac{e^{2\sqrt{a}}}{2a}\operatorname{erfc}\left(\frac{1}{\sqrt{s}}+\sqrt{as}\right) + \frac{e^{-2\sqrt{a}}}{2a}\operatorname{erfc}\left(\frac{1}{\sqrt{s}}-\sqrt{as}\right) - \frac{e^{-as}}{a}\operatorname{erfc}\left(\frac{1}{\sqrt{s}}\right),$$

for $a, s \geq 0$.

Proof Integration by parts, change of integration variable to $y = 1/\sqrt{x}$, and using formula IV. □

B Proof of Lemma 1

We verify that the function

$$u(s) \triangleq \frac{\displaystyle\int_0^s \operatorname{erfc}\left(1/\sqrt{y}\right)\mathrm{d}y}{\operatorname{erfc}\left(1/\sqrt{s}\right)}$$

is one-to-one and onto on $(0, +\infty)$. That $u(s)$ is onto follows from the fact that it is continuous with $\lim_{s\searrow 0} u(s) = 0$ (e.g., using l'Hôpital's rule) and $\lim_{s\to+\infty} u(s) = +\infty$ ($\operatorname{erfc}(1/\sqrt{s})$ is positive, increasing and bounded). It remains to show that $u(s)$ is increasing.

Expanding the integral (see Appendix A) and writing $z \triangleq 1/\sqrt{s}$ we get

$$u\left(\frac{1}{z^2}\right) = \frac{\left(\frac{1}{z^2}+2\right)\operatorname{erfc}(z) - \frac{2}{\sqrt{\pi}}\frac{e^{-z^2}}{z}}{\operatorname{erfc}(z)} = 2 + \frac{1}{z^2} - \frac{2}{\sqrt{\pi}}\frac{e^{-z^2}}{z\operatorname{erfc} z}.$$

Note that $z \mapsto 1/z^2$ is onto and decreasing on $(0, +\infty)$. Therefore, we have to show that the function $v(z) \triangleq u(1/z^2)$ is decreasing on $(0, +\infty)$. This happens if and only if the derivative of v is negative for $z > 0$.

Let us use the shorthands $f(z) \triangleq \operatorname{erf}(z)$ and $g(z) \triangleq \left(2/\sqrt{\pi}\right)e^{-z^2}$ so that

$$v(z) = 2 + \frac{1}{z^2} + \frac{g(z)}{z\left(f(z)-1\right)}.$$

We have $f'(z) = g(z)$ and $g'(z) = -2zg(z)$. Differentiating $v(z)$ with respect to z we obtain

$$v'(z) = -\frac{2}{z^3} + \frac{-2z^2 g \cdot (f-1) - g \cdot (f-1) - zg^2}{z^2(f-1)^2} \ ,$$

which is claimed to be negative for $z > 0$. Since $z^3(f-1)^2$ is positive for $z > 0$, it is enough to show that the function

$$w(z) \triangleq -z^3(f-1)^2 v'(z) = 2(f-1)^2 + z(2z^2+1)g \cdot (f-1) + z^2 g^2 \tag{10}$$

is positive on $(0, +\infty)$. Rearranging (10) we have

$$w(z) = 2f^2 + \left[z(2z^2+1)g - 4\right] f + \left[z^2 g^2 - z(2z^2+1)g + 2\right] \ .$$

Recall that $f(z) = \mathrm{erf}(z)$ takes its values on $(0,1)$ for $z > 0$. Therefore, it is sufficient to show that for every $z > 0$, the quadratic function

$$h(x) \triangleq 2x^2 + \left[z(2z^2+1)g - 4\right] x + \left[z^2 g^2 - z(2z^2+1)g + 2\right] \tag{11}$$

is positive for $0 < x < 1$. The discriminant of (11) is

$$\begin{aligned} \Delta &= \left[z(2z^2+1)g - 4\right]^2 - 8\left[z^2 g^2 - z(2z^2+1)g + 2\right] \\ &= z^2 g^2 \cdot \left[(2z^2+1)^2 - 8\right] \ . \end{aligned}$$

If $\Delta < 0$ the function $h(x)$ is strictly positive on $\mathbb{R}$. Otherwise, it has zeros

$$\begin{aligned} x_1 &= \frac{1}{4}\left(-z(2z^2+1)g + 4 - \sqrt{z^2 g^2 \cdot [(2z^2+1)^2 - 8]}\right) \ , \\ x_2 &= \frac{1}{4}\left(-z(2z^2+1)g + 4 + \sqrt{z^2 g^2 \cdot [(2z^2+1)^2 - 8]}\right) \end{aligned}$$

and is positive outside the interval $[x_1, x_2]$. We claim that $x_2 < 0$, that is,

$$z(2z^2+1)g + 4 > \sqrt{z^2 g^2 \cdot [(2z^2+1)^2 - 8]} \ .$$

This becomes clear if we raise both sides to the power 2 and note that

$$8z^2 g^2 + 8z(2z^2+1)g + 16 > 0$$

for every $z > 0$ and $g = \left(2/\sqrt{\pi}\right) e^{-z^2}$.

We conclude that the function $u(s)$ is increasing, and hence one-to-one. □