Gene Regulatory Networks: The Impact of Distance between Genes

Gilles Bernot, Jean-Paul Comet, Enrico Formenti and Siamak Taati

Abstract— We analyze the basic building block of gene regulation networks using a simple stochastic model. We consider a network consisting only of two interacting genes: an activator (or repressor) gene that produces proteins of type S and a target gene that is activated (or repressed, respectively) by proteins of type S. We identify the role of distance between the two interacting genes by calculating the relative density of those activator proteins that until time t have succeeded in reaching the vicinity of the target gene via an unbiased three-dimensional Brownian motion. The latter quantity seen as a function of time has a sigmoidal shape (like a simple delay line) that is sharper and taller when the two genes are closer to each other. This suggests an evolutionary pressure towards making the interacting genes closer to each other to make their interactions more efficient and more reliable.

I. INTRODUCTION

When modeling gene interactions, threshold phenomena observed in biology [1] constitute one of the bases of the formal semantics. Combined with the additional *in vivo* phenomenon of macromolecule degradation, the interaction curves get a sigmoidal shape (e.g. Hill functions) [2].

Modeling frameworks for gene regulatory networks study trajectories in the concentration space. The majority of those frameworks are based on ordinary differential equations [3], [4] but the sigmoidal shape of elementary interactions also allows one to discretize the state space [5], [6] in such a way that temporal logic techniques can even be applied [7].

These theoretical approaches based on concentrations disregard the spatial structure of cells. Nevertheless, the relative location of interacting genes is a currently "hot" subject in biology. Interacting genes seem to be preferentially co-located (e.g. role of transcription factors [8]).

Our goal is to formally establish the connection between the sigmoidal behavior of the interactions and the distance between the interacting genes, possibly at the price of some simplifying hypotheses.

We consider the following system:

– A source gene σ is at (three-dimensional Euclidean) distance D from its target gene τ. Genes σ and τ do not move during the process.
– New proteins of type S are produced with a constant rate in the vicinity of gene σ.
– The proteins diffuse in the three-dimensional (3D) space, according to non-interacting, independent Brownian motions.
– Each protein may degrade with a constant rate.

The mechanism by which a protein locates and binds to its target gene is a subject of long debates, as biological observations had suggested that the 3D diffusion alone is not fast enough to lead the process (see [9], [10]). However, according to Halford [11], it is not until the protein reaches a certain range of the target gene ($\sim$50–100 bp) that strategies other than the 3D diffusion come into action.

We therefore identify the significance of the distance D by calculating the ratio of those S proteins that until time t have succeeded in reaching this certain range from the target gene τ (which we call the interaction range of τ). Namely, if we denote by $n(t)$ the total number of S proteins at time t, and by $n_1(t)$ the number of those S proteins that have reached the interaction range of τ, we calculate the expectation of the ratio $n_1(t)/n(t)$ to see its dependence on D.

According to our calculations, this value, seen as a function of time, has a sigmoidal curve, which is sharper and taller when the distance S is smaller. This may explain the observation by the biologists that the set of genes that interact to perform a given biological function are geometrically close in the cell, even if not so in the linear sequence of nucleotides.

In the current paper we only present the results along with the ideas of the proofs. Details can be found in [17].

II. 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 space that are at distance D from each other. Starting from time 0, new S proteins are produced at σ according to a Poisson process with rate $\lambda > 0$ (see e.g. [12]). 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 $\lambda \delta t$.

Starting from its birth, each S protein follows a 3-dimensional Brownian motion (Wiener process) with diffusion rate $\beta > 0$: the displacement of the protein from time $s > 0$ to time $t > s$ has a normal distribution with mean 0 and variance $\beta(t-s)$ (see e.g. [13] or [14]). Moreover, each protein may degrade (become annihilated) according to an exponential decay process with rate $\varepsilon > 0$: the probability that a protein degrades during an infinitesimal time δt is $\varepsilon \delta t$.

We assume that the motion of the different S proteins do not interact and are independent of each other and the protein generation process. Similarly, the degradation of the proteins

are independent of each other and the other elements of the model.

We are interested in the ratio of the (survived) S-proteins that have reached the "range of interaction" $\mathbf{R}$ of the target τ before time $t > 0$. For simplicity, we consider the range of interaction $\mathbf{R}$ simply as a sphere with radius $r > 0$ centered at τ.

III. THE DIFFUSION MECHANISM

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.

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, denote $H_d(x, \mathbf{R}, t)$ 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 have calculated $H_d(x, \mathbf{R}, t)$ in any number of dimensions d [15] as an integral involving the Bessel functions. Surprisingly, one can easily see from their formula that the value of $H_d(x, \mathbf{R}, t)$ in dimensions 1 and 3 differ only by a factor of r/D; that is,

$$H_3(x, \overline{B_r(y)}, t) = \frac{r}{D} H_1(x', \overline{B_r(y')}, t) \qquad (1)$$

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

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

for the probability distribution of the first time a standard Brownian motion hits an obstacle at distance $D - r > 0$ (see e.g. [16], page 84). Here, $\mathrm{erf}(a)$ refers to the error function:

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

Combining (1) and (2), we obtain a simple expression

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

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 1:

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

$\square$

The curve of $p(t)$ is shown in Fig. 1. The steady state value of the curve is r/D (i.e., grows if D is decreased) and the time scale of the curve is proportional to $2\beta/(D - r)^2$ (i.e., it varies faster if D is decreased).
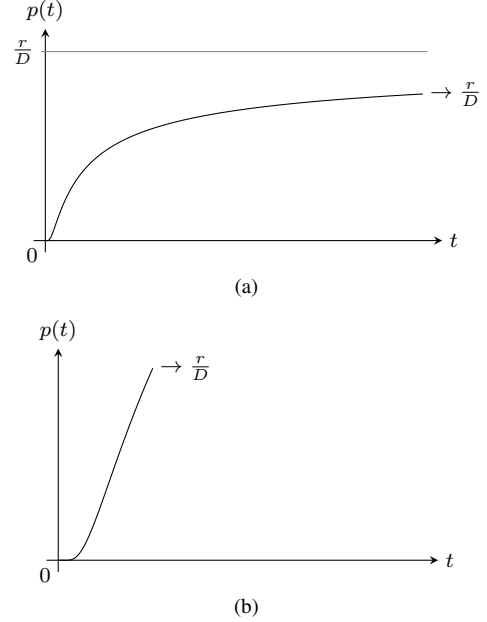

Fig. 1. (a) The probability that an S protein produced at time 0 reaches the range $\mathbf{R}$ before time t. (b) The diagram in (a) focused around $(0, 0)$.

IV. THE SUCCESSFUL PROTEINS

Let $n(t)$ denote the total number of (survived) S proteins at time $t > 0$, and let $n_1(t)$ be the number of those (survived) S proteins that have reached the range of interaction $\mathbf{R}$ before time t. These are random variables. Our aim in this section is to calculate the expected value of the ratio $n_1(t)/n(t)$ and pinpoint its dependence on the distance D between the two genes.

The main tools used in this section are two basic properties of the Poisson processes, namely Campbell's theorem and the so-called Coloring theorem (see [12]). A Poisson process on $\mathbb{R}^+$ with mean measure μ can be identified with a random countable set $\xi \subseteq \mathbb{R}^+$ such that

i) the number of elements of ξ in an interval $I \subseteq \mathbb{R}^+$ has a Poisson distribution with rate $\mu(I)$, and

ii) the number of elements of ξ in disjoint intervals are independent.[1]

Let $f : \mathbb{R}^+ \to \mathbb{C}$ be an arbitrary 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 f over I with respect to μ.

Next, suppose that to each point $s \in \xi$ we associate a random color $c(\theta)$ from a set Γ. The color of each point may depend on its position, but it is essential that the color of different points are chosen independently of each other and of the process itself. Then the Coloring theorem states that the set of points of different colors are independent Poisson processes.

Going back to our model, denote the set of times $\theta \in \mathbb{R}^+$ in which a new S protein is produced by ξ. For brevity, for

---

[1]In our model the mean measure μ is simply λ times the Lebesgue measure.

each $t \geq 0$ let $\xi_t \triangleq \xi \cap [0, t]$.

Proposition 2:

$$\mathbf{E}\left[n(t)\right] = \frac{\lambda}{\varepsilon}\left(1 - e^{-\varepsilon t}\right) \;,$$

$$\mathbf{E}\left[n_1(t)\right] = \int_0^t p(\theta)e^{-\varepsilon\theta}\lambda\mathrm{d}\theta \;.$$

$\square$

To compress the upcoming formulas, let $H(t) \triangleq \mathbf{E}\left[n(t)\right]$ and $G(t) \triangleq \mathbf{E}\left[n_1(t)\right]$. Fig. 2 depicts the curve of $G(t)$. The steady state value of $G(t)$ is $\frac{\lambda}{\varepsilon}\frac{r}{D}e^{-2\sqrt{\varepsilon/(2\beta)}(D-r)}$.
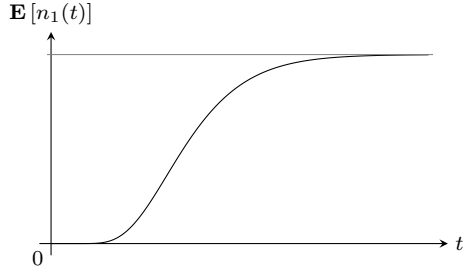

$\mathbf{E}\left[n_1(t)\right]$

Fig. 2. The expected value of the number of S proteins that succeed in reaching the range $\mathbf{R}$ before time t.

Recall that whether an S protein still exists and has reached the target before time $t > 0$ is independent of the other proteins. For $t > 0$, define the random coloring $c_t : \xi_t \to \{0, 1, -\}$ as follows: set $c_t(\theta) \triangleq -$ if the protein produced at θ has degraded before time t; set $c_t(\theta) \triangleq 0$ if the protein produced at time θ has neither degraded nor reached the target until time t; set $c_t(\theta) \triangleq 1$ otherwise. Clearly, $n_1(t) = \left|c_t^{-1}(1)\right|$ and $n(t) = \left|c_t^{-1}(0)\right| + \left|c_t^{-1}(1)\right|$. It follows from the Coloring theorem that the points colored with colors 0, 1 and $-$ form independent Poisson processes. In particular, the random variables $n_1(t)$ and $n(t) - n_1(t)$ are independent Poisson random variables whose expected values are given in Proposition 2.

Proposition 3:

$$\mathbf{Var}\left[n(t)\right] = H(t) \;,$$

$$\mathbf{Var}\left[n_1(t)\right] = G(t) \;.$$

$\square$

If $\mathbf{x}$ and $\mathbf{y}$ are two independent Poisson random variables with rates $\lambda_{\mathbf{x}}$ and $\lambda_{\mathbf{y}}$, then it is straightforward to verify that

$$\mathbf{E}\left[\frac{\mathbf{x}}{\mathbf{x}+\mathbf{y}}\,\middle|\,\mathbf{x}+\mathbf{y} > 0\right] = \frac{\lambda_{\mathbf{x}}}{\lambda_{\mathbf{x}}+\lambda_{\mathbf{y}}} \;.$$

Proposition 4:

$$\mathbf{E}\left[\frac{n_1(t)}{n(t)}\,\middle|\,n(t) > 0\right] = \frac{G(t)}{H(t)} \;,$$

$\square$

See Fig. 3 for the diagram of $\mathbf{E}\left[n_1(t)/n(t)\,|\,n(t) > 0\right]$. The steady state value of this curve is $\frac{r}{D}e^{-2\sqrt{\varepsilon/(2\beta)}(D-r)}$, which is a rapidly decreasing function of D. Moreover, as D increased, the curve is spread in time, making its dynamics slower.
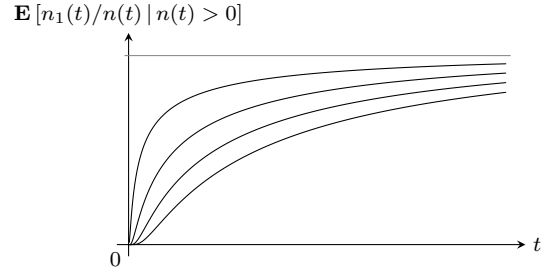

$\mathbf{E}\left[n_1(t)/n(t)\,|\,n(t) > 0\right]$

Fig. 3. The expected value of the ratio of S proteins that succeed in reaching the range $\mathbf{R}$ before time t for different values of D. The functions are scaled so that they have the same steady states values. Decreasing D makes the curve squeeze to the left.

V. CONCLUSIONS

One of our main contributions in this article has been to establish that the simple geometric effect of the distance between two interacting genes is sufficient to induce a sigmoidal behavior of their relative concentration levels. This result is independent of biochemical properties such as the affinities between the corresponding macromolecules. It is only based on the probability that proteins have to meet the set of their interacting genes. Our probabilistic approach provides an alternative to more classical approaches to model gene interactions (such as differential equations) and it has the advantage to precisely identify the impact of the distance between genes on the strength of their interactions. The formulas that we established prove that the interaction between two genes is far more efficient when they are close to each other: it may explain why biologists have observed that the set of genes that interact to perform a given biological function are geometrically neighbors in the cell, even if they are not so in the linear sequence of nucleotides; their positions in the cell have probably been preferentially chosen by the natural selection. Another advantage of our probabilistic and geometrical approach is to take into account such microscopic behaviors.

REFERENCES

[1] J. W. Little, "Threshold effects in gene regulation: When some is not enough," *PNAS*, vol. 102, no. 15, pp. 5310–5311, 2005.
[2] J. Collado-Vides, B. Magasanik, and T. Smith, eds., *Integrative approaches to molecular biology*. MIT Press, 1996.
[3] A. Csikász-Nagy, D. Battogtokh, K. C. Chen, B. Novák, and J. J. Tyson, "Analysis of a generic model of eukaryotic cell-cycle regulation," *Biophysical J.*, vol. 90, no. 12, pp. 4361–4379, 2006.
[4] A. Goldbeter and O. Pourquié, "Modeling the segmentation clock as a network of coupled oscillations in the notch, wnt and fgf signaling pathways," *J. Theoretical Biology*, vol. 252, pp. 574–585, 2008.
[5] E. H. Snoussi, "Qualitative dynamics of a piecewise-linear differential equations: a discrete mapping approach," *Dynamics and stability of Systems*, vol. 4, pp. 189–207, 1989.
[6] E. Snoussi and R. Thomas, "Logical identification of all steady states: the concept of feedback loop characteristic states," *Bull. Math. Biol.*, vol. 55, no. 5, pp. 973–991, 1993.
[7] 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," *J. Theoretical Biology*, vol. 229, no. 3, pp. 339–347, 2004.
[8] F. Képès and C. Vaillant, "Transcription-based solenoidal model of chromosomes," *ComplexUs*, vol. 1, no. 4, pp. 171–180, 2003.

[9] Z. Wunderlich and L. A. Mirny, "Spatial effects on the speed and reliability of protein-DNA search," *Nucleic Acids Research*, vol. 36, no. 11, pp. 3570–3578, 2008.

[10] S. E. Halford and J. F. Marko, "How do site-specific DNA-binding proteins find their targets," *Nucleic Acids Research*, vol. 32, no. 10, pp. 3040–3052, 2004.

[11] S. E. Halford, "An end to 40 year of mistakes in DNA-protein association kinetics?," *Biochemical Society Transactions*, vol. 37, pp. 343–348, 2009.

[12] S. Taati, E. Formenti, J.-P. Comet, and G. Bernot, "On the impact of the distance between two genes on their interaction curve." `HAL:00483286`, 2010.

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

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

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

[16] C. Yin and R. Wu, "Some problems on balls and spheres for Brownian motion," *Science in China Series A: Mathematics*, vol. 39, no. 6, pp. 572–582, 1996.

[17] S. Redner, *A Guide to First-Passage Processes*. Cambridge University Press, 2001.