Math 336 Term Paper

Addressing the Hill Function within a Model of Gene Expression Based on Random Dynamical Systems Reveals Modularity Properties of Gene Regulatory Networks

Differential gene expression has broad implications on topics including but not limited to phenotypic diversity among clonal populations, cellular differentiation and fate, methodologies for biological regulation, tumorigenesis and a vast array of disease dynamics. Unraveling information contained within the genome is thereby critical to understanding cellular dynamics and directing bioengineering efforts. While research is currently being conducted to empirically characterize gene regulatory networks and trends within gene expression, a majority of the genome still remains a mystery. Developing methodologies to characterize and predict gene expression would then serve as an excellent tool to accelerate novel discoveries regarding signaling networks and trends within the genome, as well as to better understand spatiotemporal interactions within the cell. Within A Model of Gene Expression Based on Random Dynamical Systems Reveals Modularity Properties of Gene Regulatory Networks, F. Antoneli, R. C. Ferreira, and M. R. S. Briones introduce a potential methodology to model single-gene systems and network motifs based on random dynamical systems [1]. This paper seeks to address the usage of the Hill Function within this described model of gene expression.

K. M. Warner
6-6-2017

I. Introduction

While the Hill functions were originally constructed as a means to model hemoglobin affinity to oxygen, the functions have since been commonly used to model the equilibrium state of the affinity between *n* ligands binding to one receptor [2]. Beginning in the sixties the Hill function began to be commonly used to model protein binding onto select regions of the genome, as means to accurately predict protein ability to modulate gene expression rates. A key feature of the model is its sigmoidal binding curve, which is nearly universally evidenced in empirically characterized titration curves. Here we attempt to define and explicate the logic behind the Hill functions, and the mechanism by which they enrich F. Antoneli, R. C. Ferreira, and M. R. S. Briones' model of network interactions.

II. Background

I will begin by the providing background on the biological context of the paper. A gene is defined biologically as the contiguous DNA sequence necessary to make a functional protein. As per the central dogma, gene expression may be described as a two-step process. Information contained in the form of a nucleotide sequence in the DNA is first transcribed into an mRNA transcript within of the nucleus. The mRNA transcript is then processed and exported from the nucleus to the cytosol, where the ribosome may bind and translate the mRNA transcript into an amino acid sequence. The sequence may then fold and potentially oligomerize into a functional protein (the product of gene expression). Protein functionality may be orthogonally controlled by availability of coenzymes/cofactors within the cytosol or by ligand binding, depending on the specific conformation of the protein.

All steps within the process are tightly regulated. All of the genome is contained within the nucleus, though mRNA and proteins are free to disburse within the cytosol or among specific organelles. A gene may be turned "on" or "off" depending on its availability to be accessed by the cellular machinery. DNA compaction may be a local or global event depending on specific molecular additions (classified as epigenetic modifications); generally, compaction is analogous to gene silencing ("off") and unwinding of nucleosomes allows for active transcription ("on"). mRNA transcripts are regulated by their stability and accessibility to cellular machinery; the extent to which a gene is transcribed is dictated by the rate at which the mRNA molecule may be translated and conversely by the rate at which the mRNA degrades within the cell.

Proteins, in addition to being regulated by their stability, may also be regulated by ligand binding, degradative tags, substrate availability, and several other mechanisms. Frequently, proteins may have more than one binding site, for which its functionality may be modulated in a variety of mechanisms—allosteric binding, competitive binding, orthogonal modulators, etc. This profound ability ligands within the cytosol to bind and modulate protein activity and or functionality was the original purpose for studying the Hill functions. However, a topic which is even more compelling, is protein ability to modulate and alter the genome via direct binding. Notably, proteins provide most of the molecular machinery of the cell; protein functionality may include regulating gene expression (in the form of a transcription factor) and or altering genetic compaction (in the form epigenetic control). Transcription factors bind specific sequences of

DNA adjacent to coding regions and may accelerate or decelerate transcription by recruiting or blocking machinery necessary for forming a RNA transcript. Epigenetic controls are modifications to DNA which alter gene packaging. Thus, utilizing the Hill functions as a mechanism to model protein binding affinity to select regions of the genome is a far more sophisticated and complex usage of the model. Note that degraded mRNA transcripts and or proteins sequences are nonfunctional.

A single gene system analyzes expression of a single gene and is genetic byproducts. Network dynamics delve into the complex topic of gene to gene interactions, which is often more representative of regulatory networks in the cell. An autoregulated motif is a specific class of interactions in which the gene expression byproduct influences the rate of transcription of the gene from which it was transcribed. A classic example is the Oct4 protein, for which, when transcribed serves as a transcription factor to accelerate its own transcription.

III. Framework for the Hill Function

For the framework of this paper, let us address the scenario for which the transcription factor binding target (which here represents the gene of interest) has *r* binding sites. Affinity may be variable, as will be discussed. The Hill function models a simple scenario. Given *r* transcription factors L which bind to binding site E we have the general equilibria reaction:

$$rL + E \leftrightarrows E_{rL}$$

Let $[X]$ represent the cellular concentration of molecule X. As in accord with the literature, the equilibria constant, K_d will be a constant specific to the gene of interest and transcription factor complex. By definition of an equilibria constant we have also:

$$K_d = \frac{[E_{rL}]}{[L]^r[E]}$$

Furthermore, the model presumes that the total quantity of binding sites exists within a steady state. Thus, the total cellular concentration of a select binding sites, bound or unbound, is constant within the model, that is:

$$[E]_{total} = [E_{rL}] + [E]$$

IV. Hill Function Derivations

The Hill functions described the ratio of bound and unbound binding sites. Let us describe the functions as $H^1(L)$ and $H^2(L)$, respectively [2]. In the single ligand case, we have the **Hill functions**:

$$H^1([L]) = \frac{[E_{rL}]}{[E]_{total}} = \frac{[L]^r}{(K_d)^r + [L]^r} \qquad H^2([L]) = \frac{[E]}{[E]_{total}} = \frac{(K_d)^r}{(K_d)^r + [L]^r}$$

A critical component of both functions is monotonicity [2], which readily models physical phenomena. As the cell becomes saturated with a specific transcription factor, the concentration

of bound sites augments in a sigmoidal fashion—slow as the sites must physically approach and bind the ligand in the correct concentration, then rapidly, as the concentration of ligands exceed available binding sites. On the other hand, the concentration of unbound sites diminishes, which is expected by the steady state assumption. A critical component of the Hill function is that it remains site specific, the binding coefficient will dictate the concentration necessary to saturate the binding site. To demonstrate this binding coefficient dependability, as suggested in [2], take $y = \frac{[L]}{K_d}$. Both Hill functions may then simplify:

$$H^1(y) = \frac{\left(\frac{[L]}{K_d}\right)^r}{(K_d)^r + \left(\frac{[L]}{K_d}\right)^r} = \frac{y^r}{1 + y^r} \qquad H^2(y) = \frac{(K_d)^r}{(K_d)^r + \left(\frac{[L]}{K_d}\right)^r} = \frac{1}{1 + y^r}$$

Let us examine in detail the Hill function which gauges the ratio of ligand bound complexes $(H^1(y))$. Let us consider the variable $k = k_d$, which is fixed depending on the ligand of interest (y). We now have the three-variable function, which may gauge ligand cooperativity to its binding site of interest:

$$H^1(r, k, y) = \frac{y^r}{k + y^r}$$

In this manner, *k* remains the equilibria constant specific to dissociation, *y* indicates the transcription factor interest and *r* functions as the Hill coefficient. Here, let $r \in \mathbb{R}^+$, where $r > 1$ indicates positive cooperativity, and $0 < r < 1$ would indicate negative cooperativity, and $r = 1$ indicate independent binding, in accordance with Michaelis-Menten kinetics [1]. Note that, taking the limit as *r* approaches infinity, we see the cooperativity is highly dependent on whether the concentration of transcription factors exceeds, equals, or is less than the associated dissociation constant:

$$r \to \infty \quad H^1(r, k, y) \to 1 \text{ if } y > 1$$

$$H^1(r, k, y) \to 0 \text{ if } y < 1$$

$$H^1(r, k, y) \to \frac{1}{k+1} \text{ if } y = 1$$

However, considering the limit as *r* approaches zero, we see the concentration of transcription factors will not affect cooperativity, indicating a lack of self-regulation.

$$r \to \infty \quad H^1(r, k, y) \to \frac{1}{k+1}$$

As per [1], the limits as *r* ranges from zero to infinity develop a schematic for depicting the “on” or “off” state of the gene.

V. Mathematical Context

Let us first define the mathematical framework for which this model is approached.

M is a **sigma-field** (sigma-algebra) if it contains the base set and is closed under complements and countable unions [pp.29, 3]. That is, M is a sigma-field if the following are true:

- $\mathbb{R} \in M$
- If $E \in M$, then $E^c \in M$
- If $E_n \in M, n = 1,2, \dots$, then $\bigcup_{n=1}^{\infty} E_n \in M$

Given $B = \bigcap\{M | \ M \ is \ a \ \sigma - field \ containing \ all \ intervals\}$, then B is the sigma-field generated by all intervals. The elements of B are **Borel Sets** [pp.40, 3].

A **probability space** may be represented as $(\mathrm{S}, \mathrm{X}, \mathrm{Q})$, where S is an arbitrary set, X is a sigma-field of subsets of S, and Q is a measure on X such that $Q(S) = 1$. A **random variable** will refer to a measurable function. Hence a random variable is defined as follows: if $(\mathrm{S}, \mathrm{X}, \mathrm{Q})$ is a probability space, then $\mathrm{R}: \mathrm{S} \rightarrow \mathbb{R}$ is a random variable if for all $\mathrm{a} \in \mathbb{R}$, the set $\mathrm{R}^{-1}([\mathrm{a}, \infty)) = \{\omega \in \mathrm{S} | \mathrm{R}(\omega) \geq \mathrm{a}\} \in \mathrm{X}$ [pp. 66, 3]. A **Bernoulli random variable** is a particular class of random variables, such that j if a Bernoulli random variable if $P(j = 1) = p$ and $P(j = 0) = 1 - p$, for some $0 < p < 1$, we denote a $j \sim Ber(p)$.

X, Y are **independent** if the σ-fields generated by them are independent ⇔ for any Borel sets B and C in $\mathbb{R}$, $\mathrm{P}(\mathrm{X}^{-1}(\mathrm{B}) \cap \mathrm{Y}^{-1}(\mathrm{C})) = \mathrm{P}(\mathrm{X}^{-1}(\mathrm{B}))\mathrm{P}(\mathrm{Y}^{-1}(\mathrm{C}))$, where $P_x(B) = P(X^{-1}(B))$ which is the **probability distribution** of the random variable X [pp.68, 3]. A **set of random variables $(\mathbf{X_n})_{\mathbf{n \geq 1}}$ are independent** if for any $\mathrm{k} \in \mathbb{N}$, the variables $\mathrm{X}_1, \dots, \mathrm{X}_\mathrm{k}$ are independent. [pp.244, 3].

A **Random Dynamical System** refers to the triplet $(\mathrm{S}, \Gamma, \mathrm{Q})$ where S is the state space, Γ is a set of maps from S onto itself, and Q is the probability distribution of Γ [pp. 245, 4]. The system has distinct pattern of change:

- The system begins in an initial state $\mathrm{x}_1 \in \mathrm{S}$
- $\gamma_\mathrm{i} \in \Gamma$ is chosen in accord to the distribution dictated in Q
- $\mathrm{x}_2 = \gamma_1(\mathrm{x}_1)$ within the first interval

The sequence of x_n is thereby a **Markov Process**. By definition [pp. 119, 4], on state space S, a sequence of random variables $\{x_0, x_1, \dots, x_n\}$ with values in S is a discrete parameter **Markov Process** if, for each $n \geq 0$, the conditional distribution of x_{n+1}, given $\{x_0, x_1, \dots, x_n\}$, depends only on x_n.

On the other hand, a collection of random variables is considered **IID** (independent and identically distributed) if each random variable $x_i \in \{x_0, x_1, \dots, x_n\}$ is mutually independent and has identical probability distribution.

VI. Introduction to the Single Gene Mathematical Model

Let us introduce the following variables with respect to discrete variable $\mathrm{n} \in \mathbb{N}$, where n indicates a set time point. Let the variable x_n denote the number density of mRNA molecules produced between two successive observations. Let y_n be the number density of proteins produced between two success observations. Let the constant variables $\delta, \gamma, \alpha, \beta$ denote respectively: the mRNA production rate, the mRNA degradation rate, the protein degradation

rate and the protein production rate. Let us introduce the discrete variable I_n that we will use to depict respectively the state of the mRNA molecule or the gene of interest. Let I_n be a state variable with two states, $I_n = \{0 = OFF, 1 = ON\}$. We have then:

$$x_{n+1} = (1 - \gamma)x_n + \delta I_n$$

$$y_{n+1} = (1 - \alpha)y_n + \beta x_n$$

Under this parameter x_0, y_0 are independent random variables independent of the state process I_n. Hence, the process (I_n, x_n, y_n) is Markovian; its state space is $S = \{0,1\}\times\mathbb{N}^2$.

Here we have that, in the state where the gene is considered "off," the number density of mRNA molecules will be solely dependent on the former number density and its rate of degradation:

$$x_{n+1} = (1 - \gamma)x_n$$

And, when the gene is considered "on", an additional factor is considered into the mix:

$$x_{n+1} = (1 - \gamma)x_n + \delta$$

Hence we are now provided a mechanism by which the number density of the mRNA molecules is highly dependent on the state of the gene, as would be predicted biologically. The concept following the number density of proteins follows the same logic, though since the mechanism by which the cellular protein concentration augments is dependent on mRNA, its growth rate is correlated to the precense of mRNA. Here we have derived a mathematical model which is biologically sound.

VII. Hill Function Compatibility

Theorem B: The classical rate equations of kinetic theory (as described in Section VI) coupled through Hill input functions can be readily obtained from the stochastic model, by taking the averages of the state variables in the internal dynamics of each single gene sub-system. [1].

Hence the Hill function may be a useful mechanism by which gene expression may be modelled among a variety of network dynamics.

VIII. Conclusion

Within the context of a clonal population (for which the genomic information contained is identical), alternate gene expression is responsible for profound phenotypic differences among select individuals. Only recently was the full human genome fully sequenced, yet incredible discoveries in gene regulation, gene expression and gene to gene crosstalk continue to be published. Empirical characterization may often be monotonous and labor intensive. Currently our ability to gather precise spatiotemporal data regarding gene regulation is limited due to the rapid rate of cellular reactions, and the limits of current experimental methodologies. Mathematical models may serve as a strong predictive force for novel discoveries in genomics and gene signaling networks, and hopefully accelerate and enhance our understanding of the mysteries of the cell, with profound implications in bioengineering, disease dynamics and pharmacology. Hence, fine-tuning mathematical models to accurately represent biological

phenomena is critical for accelerating discoveries in this rapidly growing field. Antoneli, Ferreira and Briones provide a novel and insightful mechanism by which we many model gene to gene interactions in eukaryotic and prokaryotic cell models based on the theory of random dynamical systems, which is thereby enriched by its use of the Hill function. Their insight may pave the future for computational models of gene expression, with the potential to rapidly enhance our understanding of molecular biology.

Bibliography

[1] **Antoneli, Fernando, Renata C. Ferreira, and Marcelo R. S. Briones. "A Model of Gene Expression Based on Random Dynamical Systems Reveals Modularity Properties of Gene Regulatory Networks." Mathematical Biosciences 276 (2016): 82-100. ArXiv. June 2016. Web. May 2017**

[2]Santillán, M. "On the Use of the Hill Functions in Mathematical Models of Gene Regulatory Networks." *Mathematical Modelling of Natural Phenomena* 3.2 (2008): 85-97. Arxiv. November 2008.

3Capinski, Marek. Kopp, Ekkehard. (2005). Measure, Integral and Probability, 2nd edition. Springer.

4 Bhattacharya, Rabi, and Mukul Majumdar. *Random Dynamical Systems Theory and Applications*. Cambridge: Cambridge UP, 2007. Print.