2018

Gene Transcription Modeling within a Random and Tethered Environment

Carlos A. Polanco
Bard College

Gene Transcription Modeling within a Random and Tethered Environment

A Senior Project submitted to
The Division of Science, Mathematics, and Computing
of
Bard College

by
Carlos Polanco

Annandale-on-Hudson, New York
May, 2018

Abstract

Gene transcription occurs within an environment highly influenced by random phenomena. Taking into account this constraint and adding a boundary within which genes can diffuse, this project attempts to portray what the co-occurrence of two genes would look like under the conditions of varying binding coefficients. A significant increase in the amount of co-occurrence relevant to proximity denotes a plausible sweet spot for transcription to occur given set conditions for a specific range of binding coefficients.

Contents

List of Figures

Dedication

This project is dedicated to my mother. Te quiero madre bella.

Acknowledgments

This project would not have been made possible without the invaluable assistance of Professors Michael Tibbetts, Karrie-Ann Norton, and Stefan Mendez-Diaz. No matter how busy their days, they always made time for me to listen to my dissatisfaction with my progress. Professor Tibbetts, were it not for your faith and belief in me, I don't know where I'd be in this process. You're the man dude. And for all of the bumps and bruises I've gotten along the way as a result of taking classes with Professors Robert Tynes, Megan Callaghan, Daniel Berthold, Craig Wilder, Japeth Wood, Malik Ndjaye, and every other professor whose names I'm afraid to spell wrong, I appreciate you all. And so we're clear, this academic journey is just beginning.

1
Mathematical Modeling

1.1 Applications of Mathematics in modeling

Mathematical modeling attempts to elucidate many aspects of our world and the dynamic interactions that occur within it. Many scientific, environmental, and engineering fields today rely on modeling to quantitatively and quantitatively explain phenomena arising respectively in each. Its use in biology, information technology and engineering has not only enriched our understanding of the particular environments studied by each field, but also supported life-saving innovations that we use on a day-to-day basis. Just think about how important our ability to forecast weather conditions affecting airplanes flying across the globe or ships traveling across seas is. Quickly, we realize the importance of modeling in our everyday lives.

But for a model to serve various disciplines it must be capable of addressing universal concepts like the conservation of mass, a structures moment of inertia, or the momentum of a fluid [1]. These concepts, however, are influenced by random phenomena affecting their particular environments. Thus, any model attempting to quantify any number of random influences must account for them in their calcula-

tions. And as the environments being modelled get more complicated, the underlying mathematical formulae needed to construct a model of them becomes ever more complex.

In classical mathematical modeling, the main focus has usually been to derive and use both differential and partial differential equations to model natural phenomena and to uncover both the mathematical and numerical methods necessary to compute solutions for these equations. A differential equation is a mathematical equation that defines a relationship between a function (usually representing a physical quantity) and its derivative (the quantitys rate of change). These types of differential equations are usually deterministic; meaning, these equations have discrete solutions given boundaries and initial conditions. Thus, the influence of random phenomena on the subject of analysis is not accounted for in these types of equations.

In order to account for the influence of a stochastic process on an object or a system, we would utilize a specific kind of differential equation that can account for this random influence. A Stochastic Differential Equation (SDE) is a differential equation that accounts for the influence of random phenomena, one whose solution is influenced by a boundary and an initial condition, but is not determined by them. For instance, if we solve an SDE ten different times with the exact same boundary and initial conditions, we would get 10 different answers. However, our answers would likely begin to illustrate a pattern of some sort the more times we solve the equation.

The applications of SDEs run the gamut of disciplines: engineering, finance, physics, and biology to name a few. Take for instance an SDE which has the general form

$$dX_t(\omega) = f_t(X_t(\omega))dt + \sigma_t(X_t(\omega))dW_t(\omega)$$

The deterministic portion of this SDE is called the drift which is expressed by the function f. The subscript t denotes that the function may additionally depend on time t while depending on $X_t(\omega)$. The function σ_t is called the diffusion coefficient. dW_t is the random noise which we call the Brownian Motion. It is the inclusion of this Brownian Motion term which distinguishes ordinary and partial differential equations from SDEs.

Let us take a look at how an SDE is applied within the financial field. Say we would like to model the evolution of a risky asset. Using Hestons model of stochastic volatility, we have that S is the price of a risky asset which evolves according to the equations

$$\frac{dS}{S} = \mu dt + \sqrt[2]{V} \times (\sqrt[2]{1-(\rho)^2}) dW_1 + \rho dW_2$$

$$dV = \kappa \times (\gamma - V)dt + \sigma \times \sqrt[2]{V} dW_2$$

Where ρ,κ, and γ are prescribed values and where $V(t)$ is the instantaneous level of the stocks volatility, dW_1 and dW_2 are differentials of uncorrelated Wiener processes (the random portion of the equation). Since the influence on stock prices depend on unknown variables, this randomness is accounted for through such an SDE. What is important to note is how the stochastic processes governing any inquiry within a specific field can be translated into an equation capable of providing the basis for a predictive model.

To illustrate further, let us take a look at the classic multi-species predator-prey model. Using the Lotka-Volterra two-species model as a basis, we let $F(t)$ and$R(t)$ represent the populations of foxes and rabbits respectively in a given environment. Thus

$$\frac{dR}{dt} = R \times (\alpha - \beta F)$$

$$\frac{dF}{dt} = F \times (\sigma R - \gamma)$$

Where α represents the net birth rate of rabbits when no foxes are present, γ is the natural death rate of foxes and β and σ are the models parameters which control rabbit and foxes interactions. The SDEs resulting from letting a and l becoming random variables are

$$dR = R \times (\alpha - \beta F)dt + \sigma_r R dW_1 \tag{1.1.1}$$

$$dF = F \times (\sigma R - \gamma)dt + \sigma_f F dW_2 \tag{1.1.2}$$

We note, though, that these equations provide a generic solution interpreted as a cyclical process in which the increase of the rabbit population bolsters the food supply for foxes which, in turn, propels an increase in the fox population. Effectively, this increase in fox population will undoubtedly lead to a decrease in the rabbit population which then leads to a decrease in fox population, then an increase in rabbit population...and so on. Essentially, by accounting for a stochastic process which may influence the population growth of rabbits and foxes in our equation permits a more robust projection of both populations. Not surprisingly, such an equation is also necessary when we attempt to model a biological environment like that of a cell. As such, the mathematical formula used to model both a volatile, risky asset and the population of two species can also be used to model biological phenomena as well.

Say we want to model a biological system like that of gene expression within a cell; specifically, let us say we would like to model the process of gene transcription within a cells nucleus. In order to do so, we must utilize mathematical equations capable of accounting for the incredibly volatile cellular environment where gene transcription takes place. Thus, our analysis requires the use of SDEs. Kong et al presents us with the SDE he and his colleagues used to model gene co-localizations in the nucleus that has the form

$$dz(t) = b(\Theta)dt + \sigma_z dB^z(t)$$

Where b is the drift term according to the global activity within the nucleus Θ, σ^2 is the diffusion coefficient of the transcription element z which can take values of x(gene), y(gene), x(transcription factor), and y(transcription factor) and B^z is the random component of the SDE. Thus, modeling stochastic processes occurring in differing fields is made possible by using SDEs.

For a model to produce a reliable projection, however, varying levels of resolution must be accounted for and introduced into various segments of its underlying code. And it was this ability to model complex systems via mathematical equations that inspired the creation of a code which attempts to portray a simplistic model of a very complicated biological process by accounting for random influences at every time step. Accounting for the stochastic influences impacting genes during the process of transcription given the highly volatile environment of a cells nucleus was necessary to depict a more accurate portrait of this cellular event and its possible behavior.

2
Gene Transcription

2.1 Co-occurrence

Interchromosomal interactions have been noted to occur at a frequency that averts a classification of being coincidental. In fact, current evidence suggests interactions between distant regions contribute to gene expression regulation [2]. It was once believed that transcription in eukaryotes conducted by transcription factories highly enriched in the active, hyperphosphorylated forms of RNA polymerase II required this RNA polymerase to produce mRNA by actively travelling and attaching itself onto the promoter regions of genes. Evidence now suggests that these transcription factories occupy distinct regions within a cells nucleus and that gene activation requires the relocation of genes to the transcription factories [2, 3]. Moreover, in the studied eukaryotes of higher species it is well established that individual chromosomes also reside in discrete territories within the three-dimensional tapestry of a cells nucleus [2]. And depending on the type of cell being analyzed, certain chromosomal regions are likelier to loop out in accordance to the specific gene being expressed. Admittedly,the noted molecular organization suggests genomic regions

appear to dynamically relocate to specific subnuclear compartments favoring gene activation [2].

Recognizing the organized tapestry of this nuclear compartment and the tendency of genes to loop out and travel toward a transcription factory during gene activation led me to inquire about how this system functioned. I questioned whether the transcriptional locus was sought out by specific genes as a result of chemical attraction, or whether the unison between gene and factory—as a result of being the least energetically taxing process within the system—was a result of sheer coincidence. This sparked my curiosity. The specific figure/illustration that set off my investigation came from Schoenfelder et als article regarding co-regulation of genes at specific transcription factories. [see Appendix A]

2.2 Existence of Transcription Factories

As I mentioned before, it is well accepted that chromosomes occupy specific spaces within the nucleus, their own little territories if you will. It also seems that transcription occurs in specific sites between chromosomes, sites where transcription factories are believed to be. But how does this figure validate this point?

At this point, it's important to understand what's being projected by Schoenfelder's findings. The figure illustrates the staining of active polymerases by identifying and then tagging the phosphorylated serines of their amino acid chains. Specific genes are also tagged. What the figure demonstrates is an overlap between pairs of genes and activated polymerases (or transcription factories) via the white signal denoted at the top-right corner of each column. Accordingly, pairs of genes were transcribed at the same spot. The other overlapping colors signify the close proximity of genes, alluding to their colocalization within the nucleus. But are genes normally close enough to each other to be transcribed by a polymerase?

Taking into account that transcription seems to occur between chromosomes, and that this figure suggests specific pairs of genes are transcribed in unison, we must ask ourselves whether these genes are on the same chromosome or different ones. If these genes reside on different chromosomes, we could conjecture that an activated polymerase would coincidentally transcribe them as a result of being in the same region of the nucleus. If the transcribed genes were on the same chromosome, then it may well be a coincidence that the chromosome was contorted in such a way that allowed transcription of both genes to be possible. The question we must ask ourselves is whether the incidence of such an occurrence is so high as to deter us from classifying it as a coincidence. This is the main concern underlying the creation of our model via matlab software. We measured the probability of cotranscription between two genes based mainly on Brownian motion and the effect of binding coefficients which were varied by their ability to keep these two genes "stuck."

Returning to the figure in Appendix A, it seems to indicate that genes colocalize in the presence of activated polymerases. In fact, these gene pairs are only found near them. Since genes are normally spatially organized in their specific territories, it would be out-of- place for two genes to be near each other, especially from different chromosomes. Think about it, the genome of mammals is over 3 billion base pairs long. What are the odds of any two genes being together during transcription? Nearly zero. And for colocalization of the same two genes to happen in unison more than once, the probability is almost negligible. This being the case, it is difficult to imagine that an activated polymerase would bind to its target gene then somehow consistently find its other target gene in the vast expanse of a DNA sequence. This would be too energetically taxing for the polymerase and the overall transcriptional system. Additionally, it could only occur if the chromosome somehow contorted itself to facilitate transcription in the case of genes in cis. For colocalization of genes

in trans to occur, chromosomes would have to travel from their specific territories to a site where transcribing both genes would be possible. After all, they reside in specific spaces within the cell.

It is important to note that chromosomes lying in the periphery are least active than ones residing towards the center of the nucleus [4]. Conjecturing that the illustration is a two-dimensional image, if chromosomes are being transcribed in the periphery as illustrated by the figure, then it is likely a result of these chromosomes actively travelling to the transcription sites to get transcribed. If chromosomes must travel, it would be less energetically taxing on the entire system for a polymerase to remain within a given region in order to enable the transcription process. As such, this figure supports the model that chromosomes conform to transcription factories when Replicating. And it's this finding that serves as the basis for this project's model.

Let us not forget how unlikely it is for the same two genes to be transcribed together. Imagine the odds of this occurring 2 or 3 times. This is what we intended to measure by creating a code that accounts for the amount of times that co-occurrence happens within a defined space we labelled our transcription factory.

3

The Co-occurrence Model: Formulation of Matlab Code

3.1 Coding Brownian Motion

In hopes of understanding the possible mechanism influencing the co-occurrence of genes, we formulated a code via matlab. This code accounts for the randomness inherent to the biological system being analyzed. Since randomness is the universe in which most biological processes function, it was important for us to incorporate such a parameter in our model.

Our model tracks the trajectory of two genes: Gene1 and Gene2. And for sake of simplicity, we consider both genes points on a plane. Each point reflects an approximate length of ten base-pairs which we correlate to the size of our target which interacts with a transcription factory at the incept of transcription.

In an attempt to mimic the biology motivating this model, we created two constraints for our genes.One of the constraints we incorporated within our model is the random nature in which these two points (or genes) operate within the plane. We felt it was important and necessary to account for the random environment in which molecules exist within the cell. The influence of chemical attraction and the

impact of fluid dynamics within this nucleic environment avoid precise measurements of how one or the other will affect the movement of chromosomes throughout gene transcription. Inspired by the logic underpinning SDEs, we accounted for the impact of Brownian Motion by having our genes random walk from one point on the graph to the next. We figured that given enough iterations, a pattern would begin to emerge which would allow us to interpret the data in accordance to the observed behavior.

3.2 Coding Restriction: The Tethered Gene

The other constraint we incorporated into our model was a boundary which our points could not go past. The reason for this boundary is to mimic the tethered nature of genes within a nucleus. Since there is a limited space through which genes could diffuse as a result of being part of a chromosome, we felt it important for our points on the plane to remain within a defined space somewhat proportional to the dimensions of chromosomes within the nucleus. Thus, we limited our plane to reflect the extent genes are likely to be limited while diffusing through the nucleus en route to a transcription factory.

3.3 Coding the Transcription Factory

Aside from the boundary, we had to define a space within our plane to serve as our genes transcription factory. This space is somewhat proportional to the scale provided by Kang et al.s measurements of transcription factories and genes. Satisfied with the dimensions of our plane and our transcription factory, we felt it necessary to place both genes equidistant from the transcription factory. We did so because randomizing the location of both points on the plane could have potentially skewed our results and subsequent interpretation of any behaviors noted post-simulation.

By keeping both genes equidistant from the transcriptional locus we could provide a standard from which other interpretations could spawn; that is, we could infer with a greater degree of certainty how these genes are likely to behave if one is closer to the transcriptional mechanism than the other.

4
Model Output: Incidence and Co-occurrence

4.1 Qualifying and Quantifying Transcription

Satisfied with both constraints, the dimension of our transcription factory, and the initial placement of our genes within the plane, we figured that our model should account for two core instances during each iteration of our simulation. We created an array which cataloged the amount of time steps that each gene remained within our defined transcription factory space. And for purposes of our analysis, we consider the incidence (denoted "ind" in our code) as the number of time steps that each particular gene is within our transcription factory.

However, the existence of any gene by itself within our factory space does not constitute transcription. In our model, we define transcription as the existence of both genes simultaneously within our factory. Thus, a transcriptional event is a shared event between both genes which can occur various times within one iteration of our simulation. We coded a rule which classifies a transcriptional event as occurring only if one or both genes were outside of the defined transcription factory space in the previous time step.

To understand this a little better, let us look at both cases that can be considered transcriptional events. Say we looked at our genes over the course of 4 time steps. In time step one, gene1 and gene2 are outside our defined transcription factory space. In time step two, both genes are within the space. This counts as one transcription event. In time step three, gene1 moves outside of our factory space while gene2 remains within it. In time step four, gene1 re-enters the factory space. This reentering of gene1 in the transcription factory while gene2 is still within it constitutes another transcription event. Thus, both genes only transcribe when they are both within the space together after not having been in the space simultaneously in the previous time step. In our code, we label this co-occurrence as "cooc." And since transcriptional events within the nucleus of a cell occur more than once, we felt that cataloging each event according to the aforementioned logic was just and proper for our purposes. So after each event, a co-occurrence array catalogs for the occurrence of each event.

After defining the output of our model, we had to come up with a way to measure how a range of binding coefficients influenced the probability of both genes simultaneously transcribing. To do so, we layered how we defined a non-transcriptional event as the probability of a randomly generated number being less than a set binding coefficient. Since many factors can influence the binding of genes to transcription factories, it was necessary to account for the precarious world in which these nucleic molecules operate. Thus, we accounted for this stochasticity by codifying the statement "if a randomly generated number is less than our set binding coefficient as both genes reside simultaneously within our transcription factory, such an event will not be defined as co-occurrence." This means that transcription is contingent on the probability that a specific binding coefficient will favor simultaneous transcription (co-occurrence). In essence, we can consider the binding coefficient as the "stickiness" of a gene to a transcription factory. Thus, the lower the binding coefficient,

the "stickier" the gene is to the transcription factory. Analyzing this "stickiness" and the range of binding coefficients most likely to favor co-occurrence is the main objective of this project.

5
Discussion

5.1 Interpreting the Data

We ran 100,000 iterations of our simulation per binding coefficient. Each binding coefficient was equally partitioned between a range of 0.025-0.3. We then ran this model 100 times and tallied our outputs into an array. From this array we were able to average out the number of incidences [Appendix B] and co-occurrences [Appendix C] per binding coefficient. We then divided the average number of incidences by the average number of co-occurrences to arrive at the average length of time both genes simultaneously transcribed for per binding coefficient [Appendix D]. Since we ran our model for 10 different binding coefficients, we took all 10 quotients and averaged them out to arrive at an average length of time (in time steps) that both genes simultaneously transcribed for (in code: "sum(mean(R))/10"). So for every 100,000 iterations, the average length of times that transcription occurred for was 434.8057 (or about 435) time steps.

As mentioned, Appendix B illustrates the number of incidences in which both genes resided within the transcription factory. As we would expect, the total number

of incidences decreased as the binding coefficient increased; that is, the less "stickier" the gene became to the transcription factory, the less likely it was to remain attached to it.

In Appendix C we note that the number of co-occurrences peaked when the binding coefficient was at about 0.15. If we correlate this binding coefficient to Appendix B, we also note a sharp increase in the graph at about the same value. This means that while both genes resided within the transcription factory, the binding coefficient was sticky enough to have them both transcribe at the same time.

However, the most significant finding on both graphs occurs when binding coefficients are between 0.21 and 0.27. Note how both graphs illustrate a sharp increase in slope. This means that, given this range of binding coefficients, the ratio of incidences to co-occurrence is lowest; meaning, both genes spent a larger portion of incidences simultaneously transcribing. This is interesting since co-occurrence did not occur at a higher frequency given a lower binding coefficient as we would expect. We also note how there is a sharp decrease in slope between 0.15 and 0.2 binding coefficients on both graphs. This infers that the probability of co-occurrence dropped dramatically given this range of binding coefficients. Interestingly enough, co-occurrence increases the most after this dramatic decrease.

In Appendix D we note how the average length of time genes simultaneously transcribe decreases as the binding coefficient increases. Though this finding is not surprising since we have already established that the smallest ratio between incidences and co-occurrences from Appendix B and C respectively occur at the binding coefficients ranging from 0.21-0.27.

Our model suggests that the greatest amount of transcription relevant to proximity of the transcription factory occurs when the binding coefficients range from about 0.21-0.27. We note this by the sharp increase in slope in Appendix D within

this range of coefficients. Lest we forget, what is most important for our purposes in this project is to find, if one exists, a sweet spot for which a certain binding coefficient value lends itself to a greater probability of co-occurrence. By sheer randomness alone, we can project that the probability is highest for transcription to occur when binding coefficients are restricted to this range (0.21-0.27) given our established parameters.

6
Conclusion

Our goal has been to create a simplified model of a very complex biological system based on constraints that permit reasonable inferences to be made about such a system. Through sheer randomness and restriction of movement, two major influences on the transcription of genes within a cell nucleus, we are able to infer how this system may in fact be impacted by an affinity between two genes promoters. Could it be that they need to be together in order to be able to transcribe? Additionally, is it biologically plausible that both genes would need to be bonded somehow in order to transcribe? And is there a binding coefficient which provides enough stickiness to maintain an efficient system running at optimum capacity? While this model cannot address the former two questions, it provides a basis from which to answer the latter question. Using the random nature in which things occur within the cell as a standard, we note that a pattern emerges throughout our figures. We note that, given a certain binding coefficient range, it is quite possible to determine the amount of times transcription may occur given our constraints. Further studies

will be required to analyze this biological system in accordance to other constraints that may affect transcription.

Bibliography

[1] Alfio Quarteroni, *Mathematical Models in Science and Engineering*, Notices of the AMS **56** (2009), 10–19.

[2] Stefan Schoenfelder et al, *Preferential associations beten co-regulated genes reveal a transcriptional interactome in erythroid cells*, Nature Genetics **42** (2010), 53–62.

[3] Huimin Chen et al, *What have single-molecule studies taught us about gene expression*, Genes and Development **30** (2016), 1796–1810.

[4] Julien Dorier et al, *The role of transcription factories-mediated interchromosomal contacts in the organization of nuclear architecture*, Nucleic Acids Research **38** (2010), 7410–7421.