Kinetics of Assembly of Multiple Transcription Factors

Arvind Thiagarajan
Departments of Biological Engineering and Physics
Massachussetts Institute of Technology
(Dated: May 16, 2011)

The process of transcription is central to biological systems, and for this reason the general transcriptional system driven by a single transcription factor has been well characterized. However, many biological systems are driven by more intricate circuitry, in which the simultaneous presence of multiple transcription factors is required to drive the production of a gene or set of genes. In this paper we examine the general properties of multiple transcription factor binding. We will examine a system in which N binding sites, each for a different transcription factor, are present. Salient properties of the system, such as the time for which all sites are occupied, will be characterized, and a special case of the system, in which all binding events are independent, will be considered. We will consider the limitations of our calculations, and how they might be extended to more accurate theories, for instance the Berg - von Hippel Theory. Finally, we will discuss how our results might be applied to achieve more intelligent circuit design in synthetic systems.

I. INTRODUCTION

The twin fields of systems biology and synthetic biology are at present among the fastest growing subfields of biology, promising to accelerate our understanding of biology and biological engineering to the level at which we understand electrical engineering today. Nonetheless, despite this vast potential, we are still very much in the early stages of research within these fields. In particular, while certain simple motifs have been analyzed thoroughly in natural systems and implemented successfully in synthetic systems, more complicated motifs remain somewhat mysterious, and thus limit our ability to apply synthetic biology more broadly in medical and industrial settings.

One instance of regulation, pervasive in biological systems and well-understood from a mechanistic point of view, is transcriptional regulation. The expression of a gene or set of genes is usually regulated by an operator (promoter) region upstream of the genes to be transcribed. A series of proteins, termed transcription factors, bind to specific sites within these operator regions, in order to regulate transcription. Some of these transcription factors, termed repressors, suppress transcription, while other transcription factors, termed activators, promote transcription. While the specific mechanisms of different activators and repressors may vary, their action usually includes recruiting RNA Polymerase to transcribe the DNA of interest.

Since transcription is needed at a very fundamental level in biology, it is not surprising that the case of a single transcription factor binding to a promoter has been well studied and characterized. However, circumstances rarely permit such a simple system to evolve in nature. The expression of most genes is combinatorially regulated, and as such it would be worthwhile, for a variety of reasons, to examine in some detail the dynamics and equilibrium properties of multiple transcription factor binding. In this paper we first discuss a few examples, both natural and synthetic, that might benefit from such an analysis. We will then begin our analysis. The system under consideration will be that of an operator region in which N binding sites, each for a potentially different transcription factor, immersed in solution with the different transcription factors to which it may bind. We will examine the equilibrium states and their associated probabilities of occupation for this system. We proceed to consider the specific case in which all binding and unbinding events are independent of binding and unbinding events at other sites. For this system, we calculate from first principles the probability distribution and average for the time during which all transcription factors remain bound. We then discuss how this might be affected by dependencies among the different binding sites, as well as how such dependencies might be calculated using the Berg-von Hippel Theory [6]. Finally, we will discuss how the results of our analysis might be used to pursue further research more effectively, whether through improved experimental design or through informed optimization of synthetic systems.

II. MOTIVATING EXAMPLES

There are a great many systems in nature that use combinatorial regulation via multiple transcription factor binding. Indeed, this is not very surprising: most biological systems, whether simple or complex organisms, are composed of subsystems that make decisions, and most decisions, as one might imagine, are neither linear nor tree-like in their structure. Examples abound of this phenomena, so we will name only a few cases explicitly. The CCAAT binding sequence, used ubiquitously in eukaryotic genomes, is in fact typically bound by a heterotrimeric complex that recruits RNA Polymerase [1]. Both interleukin-6 [3] and human complement receptor

2 [5], proteins involved heavily in human immunology, are regulated by multiple transcription factors. Finally, regulation by multiple transcription factors occurs in hormonal signaling pathways as well, as evidenced by the regulation of steroidogenic acute regulatory protein expression [4].

Synthetic systems often require combinatorial regulation as well. For instance, a genetic circuit intended to detect cancer within a cell and induce apoptosis would have to be sensitive to an array of markers, as cancer has no single indicator. Similarly, genetic circuits attempting to treat insulin or any other sufficiently complicated illness will require combinatorial regulation. Thus, it would serve us well to have an understanding of multiple transcription factor binding, so that we may more easily design systems with combinatorial regulation.

III. GENERAL THEORETICAL FRAMEWORK

Let us consider a stretch of DNA, presumably a promoter region, possessing N binding sites. Let us denote by S_i the ith binding site, and let us denote by P_i the corresponding transcription factor. The promoter can obviously be in a number of different possible states, depending on which transcription factors are bound at a given time. Let us encode these states as binary strings of length N, where a 1 or 0 in the ith position indicates that S_i is bound or not bound, respectively. It follows, then, that there is a direct reaction that allows the interconversion of any two states that differ by exactly one bit. Suppose the states are ordered in some way. We associate with each reaction ij interconverting between states i and j (with j having more 1 bits than i) an association constant ${k_{ij}}^+$ and a dissociation constant ${k_{ij}}^-$. These constants are interpreted as follows. Suppose in reaction ij, the mth bit is flipped. Then we have that the mth bit is flipped from 0 to 1 at a rate given by ${k_{ij}}^+[P_m]$, where $[P_m]$ is the concentration of the mth transcription factor. Similarly, the mth bit is flipped from 1 to 0 at a rate given by ${k_{ij}}^-$. This is not a peculiar framework, we are simply imposing the standard reaction kinetics on this system. The relative likelihood of any two given states, and thus the probability of any particular state (subject to the normalization condition that the system must be in some state), is determined by these rate constants and by the concentrations of the transcription factors present. In particular, let $p(i)$ be the probability of being in state i. Multiply each reaction rate by $p(i)$ for the state i that the directed reaction leads away from, and let this be denoted the probability flux of a particular directed reaction. It follows then that the overall probability flux into a given state must be exactly balanced by the overall probability flux out of that state. This relation, which takes the form of a linear system of equations in N variables, can be solved to obtain $p(i)$ for each i. We will now examine a particular example of this framework, one in which the binding events at each site are independent of the state of the promoter.

IV. INDEPENDENT BINDING

Under the constraint of independent binding, a number of the rate constants previously defined become redundant. As such, we shall redifine these rate constants in a manner more suited to this specific system. Let ${k_i}^+$ be the rate constant for the binding of P_i, and let ${k_i}^-$ be the rate constant for the unbinding of P_i. This simplifies a number of properties pertinent to the system. In particular, let $p(i)$ denote for this system the probability that P_i is bound, and let p be the probability that all the transcription factors are bound. We then have that

$$ {k_i}^+[P_i]p(i) = {k_i}^-(1 - p(i)) \tag{1} $$

from which it follows that

$$ p(i) = \frac{{k_i}^-}{{k_i}^+[P_i] + {k_i}^-} \tag{2} $$

$$ p = \prod_{i=1}^{N} \frac{{k_i}^-}{{k_i}^+[P_i] + {k_i}^-} \tag{3} $$

Of course, this probability does not sufficiently characterize the system. The value given merely denotes the fraction of time during which the promoter is bound by P_i. However, the frequency of attachment and detachment is not contained in this value, and in some sense this frequency is actually the most important aspect of the dynamics. After all, if the transcription factor is bound with probability $\frac{1}{2}$, it might spend 3 minutes bound and 3 minutes unbound at a time, or it may spend just 1 millisecond bound and unbound alternately. There is, of course, a minimum time required for the transcription factor complex to recruit RNA Polymerase, and thus it is very important that we be able to calculate the duration for which all the transcription factors stay bound.

A. Time of Dissociation

Suppose that a time $t = 0$ the final transcription factor binds to the promoter (we are not concerned with which transcription factor this is). Now, each site undergoes binding and unbinding events independently of the other sites, and consequently the probability that all sites are still bound at time t is simply the product over all sites i of the probability that site i is still bound. Thus, we would like to determine the probability that site i is still bound at time t, which we will denote as $p_i(t)$. We have that the kinetic rate of unbinding at site i is ${k_i}^-$. It follows, then, that this same rate gives the probability of a decay event occuring in a given second. To find $p_i(t)$, then, we discretize the problem as follows. Suppose

that time occurs in discrete intervals, each containing b seconds. Clearly, then, the probability of not having decayed after n intervals is simply given by $(1-{k_i}^- b)^n = (1-{k_i}^- b)^{\frac{t}{b}}$. Taking the continuum limit of this quantity, i.e. $\lim_{b\to 0}(1-{k_i}^- b)^{\frac{t}{b}}$, gives

$$p_i(t) = e^{-{k_i}^- t} \tag{4}$$

and it follows then that the probability $f(t)$ that all N sites are bound at time t is given by

$$f(t) = \prod_{i=1}^{N} p_i(t)$$

from which it follows that the probability that all N sites are not bound at time t is given by

$$g(t) = 1 - f(t) = 1 - e^{-k_d t}, \text{ where } k_d = \sum_{i=1}^{N} {k_i}^- \tag{5}$$

Now, the probability distribution that we care about, i.e. the probability density for unbinding in the interval $(t, t+dt)$, is given by

$$p(t) = \frac{dg}{dt} = k_d e^{-k_d t} \tag{6}$$

This is the probability distribtution for the actual time at which unbinding occurs. We can now easily calculate the average time required for unbinding as

$$\bar{t} = \int_0^\infty t p(t) dt = \lim_{a\to 0} a g(a) - \int_0^a g(t) dt = \frac{1}{k_d} = \frac{1}{\sum_{i=1}^{N} {k_i}^-}$$

This is actually an interesting result for several reasons. First, it indicates that this system's dissociation time behaves almost exactly like that of a single transcription factor binding promoter, just with a combined rate constant. Second, this result is highly unintuitive. One would likely expect, since the transcription factors all bind independently, that this analysis would simply require a particular approximation, wherein the S_i is bound for average time $\frac{1}{{k_i}^-}$ and unbound for average time $\frac{1}{{k_i}^+[P_i]}$, such that there is no preferred correlation between when two intervals of bound state start for two different binding sites. However, this viewpoint is too simplified to be accurate, even though, ironically, the analysis involved for this approximation is actually more difficult than the analysis conducted here. In actuality, the probability that a given transcription factor stays bound to its corresponding site for a time interval of length t is independent of how long the transcription factor has already been bound, since each interval of time is independent of all intervals of time with which it is disjoint. It is for this reason that we do not consider, in our analysis, how long each transcription was bound before the final transcription factor bound the promoter.

V. COOPERATIVE BINDING

Let us now consider a slightly different arrangement, in which the binding and unbinding events at different sites are not independent, but are instead cooperative. What is meant here by cooperative is that binding of a transcription factor at one site enhances or reduces the rate of transcription factor binding at another site, and/or that unbinding of a transcription factor at one site enhances or reduces the rate of transcription factor unbinding at another site. For both statements, positive cooperativity corresponds to enhanced rates of binding/unbinding, while negative cooperativity corresponds to reduced rates of binding/unbinding.

However, as we noticed in our derivation of the dissociation time in the case of independent binding, only the dissociation constants for each site in the final, completely bound state affect the dissociation time. As such, we could repeat the same derivation here, and we would in fact obtain the same result, namely that the average time spent in one continuous interval in the fully bound state is given by the inverse of the sum of the dissociation constants from the fully bound state for all binding sites. As such, this value is not affected at all by the presence of cooperativity.

It is possible that the dissociation rate constants themselves are particularly low or particularly high for the fully bound state because of this cooperative effect, but this could also be occur in a system with independent binding. Nonetheless, all other things being equal, positive cooperativity leads to low dissociation constants in the fully bound state and thus greater average dissociation time, while negative cooperativity leads to high dissociation constants in the fully bound state and thus lower average dissociation time. Furthermore, the cooperativity changes the distribution of equilibrium probabilities. In particlar, positive cooperativity concentrates the probability density at the two extremes, that is the fully bound and the fully unbound states, and so in this case the fully bound state is not only more probable but also bound more continuously. On the other hand, negative cooperativity concentrates the probability density in the half bound states, and thus in this case the fully bound state is bound less continuously and less often. From this, we can offer a prediction, namely that natural systems with multiple transcription factor binding will often exhibit positive cooperativity, and very rarely, if at all, will they exhibit negative cooperativity.

VI. RATE CONSTANTS AND ALTERNATE THEORIES

It is perhaps obvious that independent binding is an unrealistic assumption, and so a question arises as to what determines the rate constants governing transitions between different states. Is purely positive or purely negative cooperativity a feasible structure? These questions

can be analyzed through mechanistic theories of transcription factor binding. Here we discuss a particularly powerful theory, the Berg von Hippel theory, and discuss how it might be used to predict the actual rate constants to be used for a given system, based solely on the energetics of binding for each transcription factor. The Berg von Hippel theory posits that transcription factors actually bind non-specifically to DNA, search for their binding sites by sliding along the DNA, and eventually fall of the DNA, repeating this process until the binding site is found [6].

This model has actually been shown to be fairly accurate through the use of single molecule experiments. As such, we choose to use it here, to draw some conclusions about the nature of the various rate constants. For the sake of this analysis, let us ignore differences in the binding energies of the different sites. In addition, let us ignore any energetic effects that may arise from structural deformations as a result of transcription factor binding. These two assumptions allows us to consider the geometric aspect of multiple transcription factor binding in isolation.

Now, it seems clear that the first transcription factor to bind, irrespective of the position of it's binding site, will have the greatest association rate constant, since it may slide into its binding site from either side of the DNA molecule. The association rate constant of the second binding event will also not depend to any great degree on the position of the binding site. This rate constant, however, will be approximately half the rate constant for the first binding event, since in this case the transcription factor may only approach from one side, as the first transcription factor, now bound, prevents sliding from the other direction.

Finally, the third transcription factor will have an association rate constant that depends very much on the position of the binding site. In particular, if the position of the binding site is between the two transcription factors that have already been bound, then the transcription factor is forced to bind the DNA directly in between the first two transcription factors, and this is a very unlikely event and the association rate constant is very low. However, if the new binding site is not in between the two sites already bound, then the association rate constant is approximately the same as for the second binding event.

This same logic holds for each subsequent binding event, such that the association rate constant is non-negligible only for sites not lying between already bound sites. Of course, this is only true if the sites are sufficiently close together, and as the distance between binding sites increases, the distinction between these inter-site distances will become more pertinent. Nonetheless, it seems that geometric constraints dictate that the only reasonable way to reach the final, fully bound state is to bind each transcription factor sequentially, either from front to back or from back to front, along the promoter. There seems to be no effect on the dissociation rate constants due to geometric constraints. However, this could change if we were to allow variations in the energetics of these binding and unbinding reactions.

VII. IMPLICATIONS FOR ENGINEERING SYNTHETIC CIRCUITS

There are several ways in which the results of our analyses can be employed to yield more efficient synthetic genetic circuits. First, if at all possible, positive cooperativity should be employed when designing a circuit with a combinatorially regulated promoter element. Second, transcription factors should be chosen with as similar binding energies as possible. While the specific breakdown of contributors to the overall k_d term is not relevant for time of dissociation, it is relevant for the regularity of binding, and so an even distribution of contributors would likely lead to the least erratic transcriptional behavior. Finally, assuming fixed equilibrium constants for each binding site, the dissociation constants should be optimized in such a manner that the average dissociation time is just slightly longer than the time required to recruit RNA Polymerase. For specific systems, a more thorough analysis will be useful as well.

VIII. CONCLUSION

In conclusion, systems with multiple transcription factor binding are prevalent in nature and have great potential in synthetic biology. The equilibrium probabilities of each bound state are dictated in a non-trivial but linear fashion, varying with both association and dissociation rate constants. However, the time of dissociation for the fully bound state depends only one the sum of the dissociation rate constants for this state. Positive cooperativity thus increases this time of dissociation while also making the fully bound state more probable overall. It is unclear whether such coopeartivity is truly present in nature, but the Berg von Hippel Theory can be employed to determine the specific rate constants that characterize a given system. In particular, in the absence of strong energetic biases, this Berg von Hippel Theory suggests that only the sequential binding of transcription factors will successfully lead to the final, fully bound state. Finally, using these results, we determine that genetic circuits utilizing this functionality should employ positive cooperativity, equalize as much as possible the different site binding energies, and optimize overall time of dissociation so as to recruit RNA Polymerase but not spend any extra time bound to the promoter. These analyses can typically be extended for any particular system with known binding energies.

Acknowledgments

The author would like to thank his instructors Professors Mehran Kardar and Leonid Mirny, as well as his teaching assistant Maxim Imakaev, for the help and instruction they have provided throughout this semester. Furthermore, the author would like to thank his classmates, particularly Jonathan Gootenberg, Kenneth Hu, and Lauren McGough, for their willingness to review this manuscript.

[1] McNabb, D. et al. Cloning of yeast HAP5: a novel subunit of a heterotrimeric complex required for CCAAT binding. Genes Development. 9(1) (1995) 47-58

[2] Yeang, C., Jaakkola, T. Modeling the combinatorial functions of multiple transcription factors. MIT CSAIL. http://people.csail.mit.edu/tommi/papers/YeaJaa-recomb05.pdf

[3] Hershko D.D., Robb B.W., Luo G., Hasselgren P.O. Multiple transcription factors regulating the IL-6 gene are activated by cAMP in cultured Caco-2 cells. Am J Physiol Regul Integr Comp Physiol. 2002 Nov;283(5):R1140-8.

[4] Manna P.R., Wang X.J., Stocco D.M. Involvement of multiple transcription factors in the regulation of steroidogenic acute regulatory protein gene expression. Steroids. 2003 Dec;68(14):1125-34.

[5] Vereshchagina, L.A., Tolnay, M., Tsokos, G.C. Multiple Transcription Factors Regulate the Inducible Expression of the Human Complement Receptor 2 Promoter. The Journal of Immunology, 2001, 166: 6156-6163.

[6] Berg O.G., von Hippel P.H. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. J Mol Biol. 1987 Feb 20;193(4):723-50.