

Modeling the Activity of Single Genes

10/20/98

Michael Gibson and Eric Mjolsness

Introduction

Motivation - need to understand cis-regulatory logic

The central dogma of molecular biology states that information is stored in DNA, transcribed to messenger RNA (mRNA) and then translated into proteins. The human genome project is expected to determine the complete sequence of all human genes, and the genomes of several other organisms are already completely sequenced. Much of the work in computational biology focuses on understanding what mRNA and proteins will be formed from a given length of DNA, how the proteins will fold, and so on. In this chapter, we consider a different problem, the question of *which* mRNA and proteins are present *at what concentrations* in a given cell at a given time. This will involve understanding transcription and translation, as well as the cellular processes that control those processes. All of these elements fall under the aegis of *gene regulation*.

We may ask: *What genes are expressed in a certain cell at a certain time? How does gene expression differ from cell to cell in a multicellular organism? Which proteins are important in regulating gene expression?* From questions like these, we hope to understand which genes are important for various macroscopic processes. Nearly all of the cells of a multicellular organism contain the same DNA. Yet this same genetic information yields a large number of different cell types. The fundamental difference between a neuron and a liver cell, for example, is which genes are expressed. Thus understanding gene regulation is an important step in understanding development. Furthermore, understanding the usual genes that are expressed in cells may give important clues about various diseases. Some diseases, such as sickle cell anemia and cystic fibrosis, are caused by defects in single, non-regulatory genes; others, such as certain cancers, are caused when the cellular control circuitry malfunctions - an understanding of these diseases will involve pathways of multiple interacting gene products.

We shall concentrate mostly on transcriptional regulation, as that is the mainstay of gene regulation modeling. Translational and post-translational regulation are typically considered as a separate problem. For a given gene, there are two types of regulatory elements - *trans*-regulatory elements, which are diffusible proteins that affect **transcription**, and *cis*-regulatory elements, which are the DNA sites (often upstream of the gene in question) where the *trans*-regulatory elements bind. Much work has been done to understand *cis*-regulatory logic, i.e. how proteins bind to DNA and affect the expression of given genes.

Model characteristics

Rather than advocating a single, definitive model of gene regulation, we will describe a variety of modeling approaches which have different strengths and domains of applicability. There are trade-offs between model precision and computational complexity, for example. A model whose goal is to have tremendous precision may require significant amounts of supercomputer time to simulate, while a model whose goal is to get a simple representation of a system to aid in reasoning may be computationally efficient but less precise. Some of the model characteristics to consider when choosing an approach are:

- **Level of detail.** Depending on the level of detail of the model, certain approximations can be made, whereas others are not valid. For example,
- **Time.** The time scale is absolutely crucial in determining what type of model to use. At very short time scales (seconds or less), the low-level details of binding/unbinding and protein conformational changes have to be modeled. At longer time scales, these can be considered to be in equilibrium, and certain average values can be used (we shall discuss this in more detail later). At very long time scales (e.g. days), processes such as cell division may be very important, which can be ignored at shorter time scales.

- **Number of molecules.** If the number of molecules is very small (i.e. in the 10s or low 100s), stochastic models may need to be used. However, once the number of molecules becomes very large, differential equations become the method of choice.
- **Computational complexity.** The more detail and the longer one wants to model a process for, the higher the complexity, and hence the longer it takes computationally.
- **Available data.** For many systems, there is a lot of high-level qualitative data, but less quantitative, detailed data. This is particularly true of complex eukaryotes. The nature of the data which may be required by a model before it can make predictions is in practice an important property of the model, which may also depend on the power of the data-fitting algorithms available.
- **Predictive ability.** To begin modeling, one must focus on what type of predictions are sought. For simple predictions, simple models are sufficient. For complex predictions, complex models may be needed.

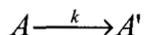
Understanding the biology

Physical chemistry

To understand the mathematical models of gene regulation, let us review some of the main ideas of physical chemistry. In particular, we shall discuss kinetics, equilibrium, free energy and the concept of partition functions.

Kinetics

Consider the chemical equation



which deals with two chemical *species*, A and A'. According to the equation, A is transformed to A' at a rate of k , in other words, the concentration of A, denoted $[A]$, obeys the differential equation

$$\frac{d[A']}{dt} = k[A] = -\frac{d[A]}{dt}.$$

(Notice the signs, because $[A]$ is decreasing as A is converted to A', while $[A']$ is increasing.) Solving these equations, we get

$$[A]_t = [A]_0 e^{-kt} \quad \text{and} \quad [A']_t = [A']_0 + [A]_0 (1 - e^{-kt}).$$

where $[A]_t$ is the concentration of A at time t . For more complicated chemical equations such as



we get differential equations such as

$$\frac{d[AB]}{dt} = k[A][B] \quad \text{or} \quad \frac{d[A_2]}{dt} = k[A]^2.$$

Typically, writing these equations and finding the correct values of the rate constants k are the interesting parts – in most actual gene regulatory systems, the equations become too complicated to solve analytically and so must be solved numerically. For that reason, we shall usually be content to write the equations, rather than provide a detailed solution.

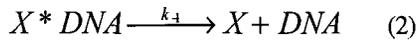
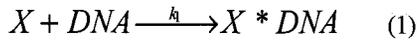
Example (Gene Regulation)

Suppose that protein X is the product of gene X. Assume protein X is synthesized at a constant rate k_1 , i.e. a rate that doesn't depend on the amount of X present. Further, assume protein X is degraded at a rate k_2 proportional to its concentration. Then $[X]$ obeys the differential equation

$$\frac{d[X]}{dt} = k_1 - k_2[X] \quad \text{subject to some initial condition } [X]_0.$$

Equilibrium and Free Energy

Consider a certain DNA binding protein X. Its binding and unbinding to DNA will follow simple and complementary chemical kinetics, i.e.



Where $X * DNA$ means "X bound to DNA." We can write the differential equation that describes the amount of X bound to DNA as a function of time, namely:

$$\frac{d[X * DNA]}{dt} = k_1[X][DNA] - k_{-1}[X * DNA]$$

Note that the right hand side has two terms: the first one is production of new $X * DNA$ due to equation (1), the second is degradation of existing $X * DNA$ due to equation (2). With the additional constraints

$$[X]_{total} = [X]_{free} + [X * DNA]$$

$$[DNA]_{total} = [DNA]_{free} + [X * DNA]$$

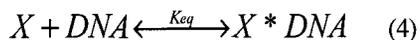
and values of $[X]_{total}$ and $[DNA]_{total}$, we could solve this differential equation as a function of time. Typically, however, we are not interested in these dynamics, which are fast compared to other reactions involved in gene regulation. We assume that on the time scale we're interested in, these two competing processes have reached *equilibrium*, i.e. there are no further *net* changes. Thus,

$$0 = \frac{d[X * DNA]}{dt} = k_1[X][DNA] - k_{-1}[X * DNA]$$

or, equivalently,

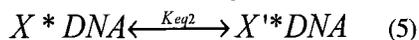
$$\frac{[X * DNA]}{[X][DNA]} = \frac{k_1}{k_{-1}} \equiv K_{eq} \quad (3)$$

where K_{eq} is called the *equilibrium constant* of this reaction. The pair of equations (1) and (2) can be abbreviated as



It is possible to go directly from the chemical equation (4) to the algebraic equation (3) by multiplying all the concentrations of the products (in this case only $[X * DNA]$) together and dividing by the product of the concentrations of the reactants (in this case $[X][DNA]$).

Suppose now, in addition to equation (4), we have the additional equilibrium equation



which could mean that X undergoes a conformational change to X' while bound to DNA. The equilibrium equation in this case is

$$\frac{[X' * DNA]}{[X * DNA]} = K_{eq2} \quad (6)$$

Now notice that

$$\frac{[X' * DNA]}{[X][DNA]} = \frac{[X * DNA]}{[X][DNA]} \frac{[X' * DNA]}{[X * DNA]} = K_{eq} K_{eq2}$$

This is an important property of equilibria: if A and B are in equilibrium and B and C are in equilibrium, then A and C are in equilibrium, and the resulting A-C equilibrium constant is simply the product of the A-B and B-C equilibrium constants.

From thermodynamics in dilute solutions [Hill 1985], it follows that

$$K_{eq} = e^{-\frac{\Delta G}{RT}}$$

where ΔG is the difference in *free energy* between the initial and final states, R is the ideal gas constant and T is the absolute temperature. Typically, values of ΔG are reported instead of the values of K_{eq} .

Partition Functions

We shall introduce the concept of partition functions by means of a detailed biological example. Consider a piece of DNA and a protein P, which can bind at either (or both) of two sites on the DNA (see Figure 1).

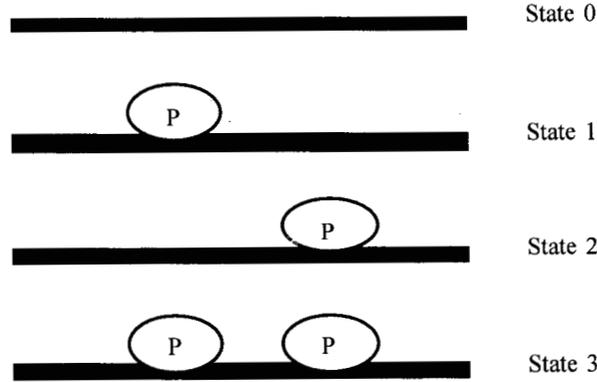


Figure 1 - Binding states of protein P

Let $[DNA]_0$ be the concentration of DNA in State 0, and let $[P]$ be the concentration of free (i.e. unbound) protein P. Then, there are equilibrium constants K_1 , K_2 and K_3 , such that

$$\frac{[DNA]_1}{[P][DNA]_0} = K_1, \quad \frac{[DNA]_2}{[P][DNA]_0} = K_2 \quad \text{and} \quad \frac{[DNA]_3}{[P]^2[DNA]_0} = K_3.$$

Using the fact that

$$[DNA]_{total} = [DNA]_0 + [DNA]_1 + [DNA]_2 + [DNA]_3,$$

we get

$$\begin{aligned} [DNA]_{total} &= [DNA]_0 + K_1[P][DNA]_0 + K_2[P][DNA]_0 + K_3[P]^2[DNA]_0 \\ &= [DNA]_0 Z \end{aligned}$$

where

$$Z \equiv 1 + K_1[P] + K_2[P] + K_3[P]^2$$

is called the *partition function*. It follows that

$$\frac{[DNA]_0}{[DNA]_{total}} = \frac{1}{Z}, \quad \frac{[DNA]_1}{[DNA]_{total}} = \frac{K_1[P]}{Z}, \quad \frac{[DNA]_2}{[DNA]_{total}} = \frac{K_2[P]}{Z} \quad \text{and} \quad \frac{[DNA]_3}{[DNA]_{total}} = \frac{K_3[P]^2}{Z}.$$

The biology

Proteins that regulate gene expression come in many types; some regulate transcription, others translation, still others degradation, RNA splicing, etc. Most work in the mathematical theory of gene regulation focuses on transcription factors (TFs), which are proteins that bind to DNA and control the rate of transcription. The DNA to which TF's bind to control the expression of a particular gene can be called the "promoter", though sometimes that term is reserved just for the TATA box initiation site for transcription. Transcription factors work something like the protein P in Figure 1 above. The binding and unbinding processes can be written out as chemical kinetics or approximated as equilibrium processes, depending on the time scale of the model. However, severe complications ensue when interactions with other transcription factors in a large "transcription complex" become important.

Transcription occurs when RNA Polymerase (RNAP) binds to the DNA, forms a transcriptional complex and moves step by step along the DNA copying it into mRNA (see von Hippel). TFs may affect the rate of binding and/or the rate at which RNAP begins transcribing DNA into mRNA. After the mRNA transcript is made, ribosomes bind to it and begin translation. Like transcription, this is a step-by-step process. Both mRNA and proteins can be degraded once they are created. RNA is degraded by ribonuclease, and proteins are degraded by cellular machinery including proteasomes signalled by ubiquitin tagging and regulated by a variety of more specific enzymes (which may differ from one protein target to another). The rates of transcription and translation vary depending on experimental conditions (Davidson, Watson *et al.*).

Many TFs bind to DNA in a multimeric state, e.g. as homodimers or as heterodimers. It is important to know how many copies of which proteins bind together before the protein is in an active DNA-binding state. Furthermore, the monomer and dimer forms of a protein may be degraded at different rates. Also, there may be extensive cooperativity between binding sites, even in prokaryotes – for example one dimer may bind at one site and interact with a second dimer at a second site. If there were no cooperativity, the binding at the two sites would be independent, so we would have $K_{1,2}=K_1K_2$. Instead, cooperativity would tend to stabilize State (1, 2), so $K_{1,2}>K_1K_2$. Competition is also possible particularly for two different transcription factors binding at nearby sites.

Eukaryotes

Eukaryotic promoters may have large numbers of binding sites occurring in more or less clustered ways. For N binding sites an equilibrium statistical mechanics treatment (possibly oversimplified) will have at least 2^N terms in the partition function, one for each combination of bound and unbound conditions at all binding sites. The most advantageous way to simplify this partition function is not known, because there are many possible interactions between elements of the transcription complex (some of which bind directly to DNA, some bind to each other). In the absence of all such interactions the partition function could be a simple product of N independent two-term factors, or perhaps one such sum for each of a global “active” and “inactive” state.

The “specific” transcription factors are proteins which can bind to DNA and/or interact with one another in poorly understood ways, and it is these protein-protein interactions inside the transcription complex which really cloud the subject of building models for eukaryotic gene expression. A further complication is the “general” transcription factors such as TFIID which assemble at the TATA sequence of eukaryotic transcription complexes, building a subcomplex. Finally signal transduction (e.g. by MAP kinase cascades [Madhani and Fink 1998]) may act on the transcription factor by the phosphorylation of constitutively bound transcription factors, converting a repressive transcription factor into an enhancing one.

Many binding sites occur in spatial and functional clusters such as the 480 base pair *eve* stripe 2 “minimal stripe element” in *Drosophila* [Small et al. 1992], which has five activating binding sites for the *bicoid* (*bcd*) transcription factor, one for *hunchback* (*hb*), and three repressive binding sites for each of *giant* (*gt*) and *Kruppel* (*Kr*). It acts as a “module” which suffices to produce the expression of *eve* in stripe 2 out of its seven stripes in the developing *Drosophila* embryo. Similar modules for stripes 3 and 7 would be less tightly clustered, if they can be properly defined [Small et al. 1996]. These promoter “regions” or “modules” suggest a hierarchical or modular style of modeling the transcription complex and hence single gene expression, such as provided by [Yuh, Bolouri and Davidson] for *Endo16* in sea urchin, or the Hierarchical Cooperative Activation model suggested in Chapter XXX [Mjolsness, this volume]. A different way to think about these binding site interactions is provided by [Gray et al. 1995] who hypothesize three main forms of negative interaction between sites:

- *competitive binding*, in which steric constraints between neighboring binding sites prevent both from being occupied at once,
- *quenching*, in which binding sites within about 50 base pairs of each other can compete, and
- *silencer regions*, promoter regions that shut down the whole promoter when cooperatively activated.

Given these observations, we can see that the biological understanding of eukaryotic cis-acting transcriptional regulation is perhaps ... “embryonic”. We now turn to trans-acting regulation.

Feedback and Gene Circuits

With only the complexity we’ve introduced so far, gene regulatory networks would be complicated, but it would be a relatively straightforward (albeit difficult experimentally) exercise to tease apart the details. The key point we have avoided is *feedback*. Simply stated, *the TFs are themselves subject to regulation*. This leads to interconnected systems that are more difficult to analyze than the feed-forward systems we’ve discussed so far. There are two major kinds of feedback – positive and negative.

Negative feedback is the way a thermostat works: when the room gets too hot, the cooling system kicks in and cools it down; when the room gets too cool, the heater kicks in and warms it up. This leads to stabilization about a fixed point. More complicated negative feedback is also possible, which leads to better control. We shall leave the complete discussion of control systems to the chapters on robustness.

Positive feedback causes amplification and dichotomies. Suppose your thermostat were wired backwards, in the sense that if the room got too hot, the *heater* would turn on. This would make it even hotter, so the heater would turn on even more, etc., and soon your room would be an oven. On the other hand, if your room got too cold, the air conditioner would kick in, and cool it down even more. Thus, positive feedback would amplify the initial conditions – a small hot temperature will lead to maximum heat, a small cold temperature will lead to maximum cooling. This results in two stable final states – a dichotomy of states, as it were – very hot and very cold. In a very hand-waving way, this is how it is possible for cells to pick different fates.

In part B of this section of the book, several models of multiple genes will be presented. In these models, feedback will be key.

Modeling Methods

Let us now look at several different modeling methods. Although this list is by no means complete, it should serve as a good comparison of different types of models and as a jumping-off point for further investigation.

Differential equations when details are known

We first consider the Ackers *et al.* ('82, later extended in Shea and Ackers '85) model of a developmental switch in lambda phage. The model is similar to the equilibrium binding and unbinding model above in the Physical Chemistry section. They consider three proteins – RNAP, repressor and *cro* – which bind to three DNA sites – OR₁, OR₂ and OR₃. In lambda, these sites control equilibrium production of both repressor and *cro*, so that the model will contain feedback.

In principle, each site can be in one of four states – unbound, bound-RNAP, bound-repressor or bound-*cro*. If the sites were completely independent, there would be 4³=64 states. In practice, several of the states are not possible, so their total number is merely 40. For each state, the free energy is measured experimentally. For each *transcriptionally active* state, the rate constants for transcription initiation of repressor and/or *cro* are measured. Using a physical chemistry approach completely analogous to the one above, one may use the experimental data to calculate the probability P_s that the DNA is in a certain state *s* as a function of the concentration of RNAP, repressor and *cro*. In particular

$$P_s = \frac{[DNA]_s}{[DNA]_{total}}$$

Then:

$$\langle rate_{repressor} \rangle = \sum_s rate_{repressor}(s)P_s \quad \text{and} \quad \langle rate_{cro} \rangle = \sum_s rate_{cro}(s)P_s.$$

The sums run over all states *s*, and the terms in the sum can be read “(the rate of repressor transcript initiation given state *s*) times (the probability of state *s*).” Given these terms, they write two differential equations of the same form as one of the equations in the chemical kinetics section, namely

$$\frac{d[repressor]}{dt} = \langle rate_{repressor} \rangle - k_d[repressor],$$

(where k_d is the degradation constant for repressor) and the analogous *cro* equation. Using numerical methods, they solve these equations to get the concentrations of repressor and *cro* as a function of time.

This model requires an enumeration of the states of the system and a selected set of reaction rates for transitions among the states.

Small number of molecules - probabilistic framework

Reinterpretation of Ackers model as probabilistic

It is interesting to note that the Ackers *et al.* model makes correct predictions in the regime where it is used (namely lysogeny maintenance) but, because it is deterministic, would make completely incorrect predictions when applied to a different regime – establishment of lysogeny or of lysis. The key difference is that the former process involves more molecules (200+ molecules of repressor) and occurs on a longer time scale (30-40 minutes): thus it is relatively deterministic. The latter involves relatively few molecules (10-50 of repressor), occurs on a short time scale (the

lysis/lysogeny decision is made within 10-15 minutes) and is probabilistic. Analyzing probabilistic behavior requires a different mode of thinking. See, for example, McQuarrie ('67), McAdams and Arkin ('94) or Van Kampen.

Where does this probabilistic or stochastic behavior come from? As we mentioned in the introduction to the chapter, there are trade-offs to consider in modeling. In the limit of low numbers of molecules and short time steps, the behavior of systems behaves stochastically, while in the limit of high numbers of molecules and long time steps, certain averaging occurs and there is a deterministic outcome. Consider an *E. coli* cell, which acts as the host for lambda. It is a rod shaped bacterium 2µm long with a diameter of 1µm (Watson *et al.*). Thus it has a volume of $\pi r^2 l = \pi/2 \times 10^{-15}$ liters. From the Ackers model, significant differences in binding occur in the range 10^{-9} M to 10^{-7} M. Consider the number of molecules that corresponds to 10^{-8} M.

$$(\pi/2 \times 10^{-15} \text{ liters})(10^{-8} \text{ moles/liter})(6 \times 10^{23} \text{ molecules/mole}) \approx 10 \text{ molecules}$$

In the lambda model, we are dealing with 1 molecule (in the case of DNA) or a few molecules (in the case of mRNA) or even 10s to 100s of molecules (in the case of proteins). Differential equations assume that the concentrations vary continuously, or equivalently, that the fluctuation around the average value of concentration is small relative to the concentration. For the very small number of molecules in the first couple of minutes of lambda infection, those are not good assumptions, thus the need for stochastic models. Later in the chapter, we'll discuss the conditions under which those assumptions *are* met, and differential equations are appropriate.

Binding/unbinding with kinetics (i.e. Markov chain)

The key to the stochastic framework is dealing with probabilities rather than concentrations. For example, in the binding/unbinding example given under the section on Physical Chemistry, we should now consider the *probability* that the single molecule of DNA is in each of the four possible states. In fact, the equilibrium limit is just that – we change the words “fraction of DNA in state *s*” to “probability that DNA is in state *s*.” However, the stochastic version of *kinetics* is more complicated.

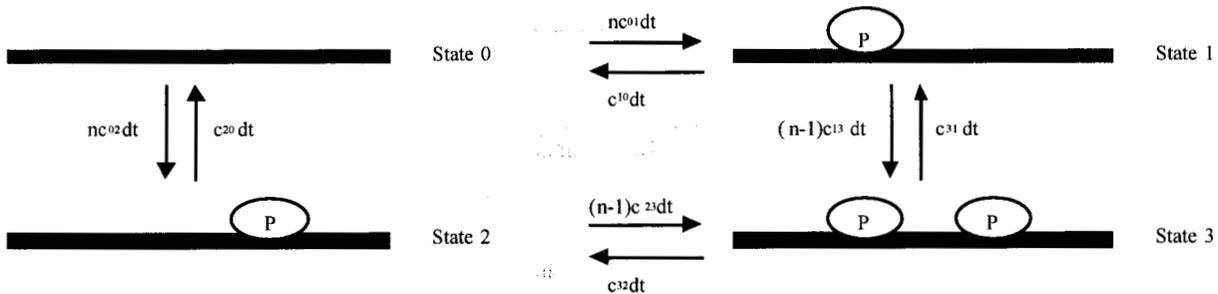


Figure 2 - Kinetics of binding

Consider the more detailed diagram of state transitions for protein/DNA binding in Figure 2. Let $P(0,t)$ be the probability that the single molecule of DNA is in state 0 at time t . Assume there are n molecules of protein P present. Then, the stochastic theory of chemical kinetics assumes that for a small time increment dt ,

$$P(1, t + dt | 0, t) = nc_{01} dt,$$

where c_{01} is a microscopic rate constant. It is related to the macroscopic rate constant k_{01} in a straightforward way. The key assumption is that the probability of a molecule of P binding to the DNA is constant in time. Equivalently, the solution in which the reaction occurs is *well-stirred*, i.e. the mean time to a collision that results in a reaction is large compared to the mean time to a collision that does not. This is assumed to be the case in *E. coli*; it may not be the case in larger cells, so care must be taken to adjust models accordingly. For the system in Figure 2, we can write a system of differential equations

$$\bar{P}(t + \Delta t) \equiv \begin{bmatrix} P(0, t + \Delta t) \\ P(1, t + \Delta t) \\ P(2, t + \Delta t) \\ P(3, t + \Delta t) \end{bmatrix} = A \bar{P}(t)$$

where

$$A = \begin{bmatrix} 1 - nc_{01}\Delta t - nc_{02}\Delta t & c_{10}\Delta t & c_{20}\Delta t & 0 \\ nc_{01}\Delta t & 1 - c_{10}\Delta t - (n-1)c_{13}\Delta t & 0 & c_{31}\Delta t \\ nc_{02}\Delta t & 0 & 1 - c_{20}\Delta t - (n-1)c_{23}\Delta t & c_{32}\Delta t \\ 0 & (n-1)c_{13}\Delta t & (n-1)c_{23}\Delta t & 1 - c_{31}\Delta t - c_{32}\Delta t \end{bmatrix}$$

This leads to the differential equation

$$\frac{d\bar{P}}{dt} = \lim_{\Delta t \rightarrow 0} \frac{\bar{P}(t + \Delta t) - \bar{P}(t)}{\Delta t} = B\bar{P}(t)$$

where

$$B = \begin{bmatrix} -nc_{01} - nc_{02} & c_{10} & c_{20} & 0 \\ nc_{01} & -c_{10} - (n-1)c_{13} & 0 & c_{31} \\ nc_{02} & 0 & -c_{20} - (n-1)c_{23} & c_{32} \\ 0 & (n-1)c_{13} & (n-1)c_{23} & -c_{31} - c_{32} \end{bmatrix}$$

This type of system is called a continuous time *Markov chain*, because the probability of the next transition depends on the current state only, not on the history of states (Feller). Given the number n of molecules of P , we can solve this system of differential equations using standard methods. This is a relatively simple case, in that there is precisely one molecule of DNA. If we were modeling a different process with two substances, each of which could be present in multiple copies, the approach would be similar, but the number of states would grow quickly.

There are efficient algorithms available to simulate chemical reactions in the stochastic framework. See Gillespie ('77) and McAdams and Arkin ('98).

Other processes

For more examples of processes in the stochastic framework, see the next chapter.

Formal basis of the relationship between low-level models and higher

The stochastic methods just mentioned are important for some systems. For others, they can be supplanted with differential equations. Differential equations are an idealization in which the *concentration* of a certain substance varies continuously; stochastic methods model the *number of molecules* of the substance, which varies discretely. In the limit where the number of molecules is large, it is common to perform time- or ensemble-averaging and use differential equations to model the dynamics of the average. This procedure can fail. For example, if the output of a system is stochastic with substantial variance or higher-order moments, a strictly deterministic model may be completely inadequate. There are, however, numerous important cases that have legitimate averaging limits.

Large number of molecules: central limit theorem

From the Central Limit Theorem of probability theory (Feller), we know that the sum of a sequence of independent, identically distributed (i.i.d) random variables with finite mean and variance converges to a normal distribution. Suppose the i.i.d. variables in question are the number of molecules that undergo a certain process. For example, we start out with n molecules of a certain protein (where n is large). Suppose each molecule behaves independently, and may degrade or not degrade according to stochastic kinetics. By the central limit theorem, it is legitimate to average and write this as a differential equation for concentration degradation. The same is true of more complicated reactions, provided that many molecules are involved, and those molecules act independently and identically.

Transcription, translation: limit of a Poisson process

How about transcription and translation? We mentioned a model of those processes, which involved a number of steps along the DNA or mRNA. These are *not* independent, in the sense that the second one can only occur after the first has occurred, etc. However, if we rephrase the question as "how long will it take before we make n steps?" we *may* apply the Central Limit Theorem. In other words, each individual step takes some amount of time, and the times are distributed according to a common distribution (namely an exponential). These times are independent, so the total time will have the sort of average behavior one would expect. This is the reason that differential equations models work well at long time scales, but not so well at very fast time scales.

Longer time scales: equilibrium binding, limit of Markov chain

Many types of Markov chains converge to the equilibrium behavior we assumed in the physical chemistry section. In particular, Markov chains with a finite number of states, where it is possible to get from every state (through a series of transitions) to every other state, will converge to the equilibrium limit. The proof introduces some new concepts to classify the states of a Markov chain, but ultimately it, too, rests on the Central Limit Theorem.

A Counterexample

With all the examples we have given of times where averaging is appropriate, one might be tempted to think that averaging is always valid. Here is a simple example that illustrates some of the problems with averaging. Suppose we have two different species of protein, A and B, which are being produced stochastically at the same rate. We are interested in the question of whether there is more A or more B in a given cell at a given time. Because of the symmetry of the problem, we can use an averaging argument to show that in a *population* of cells, roughly half of the cells will have more A than B and roughly half will have more B than A. Similarly, in a single cell *over a long time*, there will be more A than B roughly half of the time, and more B than A roughly half of the time. However, it is *not* true that in an individual cell, there will typically be nearly equal amount of A and B, with small fluctuations. In fact, there are likely to be arbitrarily large disparities in the number of molecules of A and of B. Thus, when we talk about a single cell over a long time, the amount of time we mean is very long, much longer than might be guessed using intuition alone. See Feller, Chapter 3, for a more detailed discussion of the problem of incorrect averaging based on false intuition.

The Fokker-Planck Equation: An Intermediate Formalism

In between fully averaged formalisms such as deterministic differential equations (see below), and fully stochastic ones such as the use of master equations and a full transition probability matrix, lies a useful level of modeling. This level is described equivalently by either a differential equation with a stochastic noise term added (the Langevin equation) or a deterministic differential equation for the dynamics of a probability distribution of state values (the Fokker-Planck equation). For example, it is sometimes possible to solve for a time-varying mean and a time-varying variance in a Gaussian distribution. The Fokker-Planck formalism and some conditions for its validity are introduced e.g. in [Risken 1989]. It holds some promise for modeling the dynamics of chemical species (including many gene products) where the number of molecules involved is perhaps 10-1000, in between “a few” and “large numbers”.

Differential equations when details are sketchy

When the relevant states and transition rates are unknown or only known to a very limited extent, as is likely for large protein complexes such as subcomplexes of the eukaryotic transcription complex, coarser and more phenomenological alternative models may be required as a matter of practicality. Differential equations intended to describe the dynamics of the time-averaged concentration of gene product are a common choice. It is justifiable whenever it agrees with laboratory experiment, but one might expect it to work best when deviations from average concentration are relatively small as predicted from underlying Markov chain or Fokker-Planck models, where available, or when the typical $1/\sqrt{N}$ noise in a stochastic process is sufficiently small e.g. for $N \geq 100 - 1000$. Whether N is best taken as the number of mRNA's or the number of protein molecules in a cell may depend on nonlinear gene circuit feedback effects including autoregulation. Fortunately we can also appeal to experimental data, rather than further modeling, to settle the question of model applicability.

From quantitative immunofluorescence measurements of *hunchback* and *Kruppel* protein expression levels in *Drosophila* syncytial blastoderms [Kosman et al. 1998], variations in measured fluorescence between nuclei occupying similar positions on the anterior-posterior axis of a single blastoderm would seem to be about 10% of the average value for an “on” signal and 50-100% of the signal average when it is very low, or “off”. These values would seem to be consistent with the requirements of differential equation modeling, and indeed differential equation models have high predictive value in this system.

One example of a phenomenological model for gene regulation is the proposal by [Savageau 1998] to use “Generalized Mass Action” with nonintegral exponents

XXX

as a model of transcription as well as other regulatory processes. Another approach, which has been applied to real gene expression data in several cases as described in Chapter XXX of this volume, is to use analog-valued recurrent Artificial Neural Network (ANN) dynamics with learnable parameters:

XXX

The parameters T , a/t , l , and h XXX have been successfully “trained” from spatio-temporal patterns of gene expression data. Each of these ODE formulations generalizes immediately from one gene to many interacting genes in a feedback circuit.

Two further models can be mentioned as attempts to incorporate promoter-level substructure into gene regulation networks which are otherwise similar to the ANN approach. In Chapter XXX of this volume, “Sigma-Pi units” [PDP XXX] or “higher-order neurons” are introduced to describe promoter-level substructure: regulatory “modules” in sea urchin *Endo16* promoter. In Chapter XXX, partition functions for promoter regulatory regions, dimerization, and competitive binding are proposed as a way to expand a single-node description of selected genes in a regulatory circuit into a subcircuit of partial activation states within a eukaryotic transcription complex.

Other approaches when details are very sketchy

Logical models

There are numerous “Boolean network” models based on Boolean logic [Kauffman 1993, 1969]. The simplest consist of statements of the form “If gene A is expressed now, that will cause gene B to be expressed.” Expression levels of A and B are thus represented as Boolean values – 0 for “not-expressed” and 1 for “expressed.” Interactions are then Boolean functions – for example, “C is expressed if A AND B are expressed.” This kind of reasoning induces a finite state machine (FSM), which consists of all possible n -tuples of values 0 or 1 (where n is the number of proteins considered) and transitions from one state to another, which are consistent with the Boolean functions. This FSM gives a very rough notion of the behavior of the gene regulatory system through time.

One step up the ladder, both in terms of complexity and in terms of predictive value, is the work of Thomas *et al.* ('90 and '95). They have expanded the notion of logical models to include multi-state logic, thresholds and asynchronicity. So instead of 0 and 1, they consider levels 0, 1, 2, 3, etc., which might correspond to “no expression,” “low level expression,” “medium expression” and “high expression.” The interactions between genes now becomes more complex – a typical example might be:

$$C = \begin{cases} 0 & \text{if } A \geq 1 \text{ and } B \leq 2 \\ 1 & \text{otherwise} \end{cases}$$

This type of model aids in enumeration of interactions, and helps in the analysis of steady states. Another key point of the Thomas *et al.* formalism is that genes act independently and asynchronously, so in other words, C might change from 0 to 1, then B change from 3 to 2, but these two are not necessarily at the same time. This is a reflection of the fact that rates of gene production are not uniform across chemical species.

Hybrid logical/differential-equations models

A further modification to the Thomas *et al.* work which moves in the direction of differential equation models is the work of [Mestl *et al.* 1995, 1996]. They have taken a simple differential equation for gene production, namely

$$\frac{d[X]}{dt} = k_1 - k_2[X]$$

and let k_1 and k_2 be functions of the concentrations of [X], [Y], [Z] and any other proteins present in the system. To tie this in with the algebraic (logical) formalism, they assume that the functions are piecewise constant in the discrete ranges of “low,” “medium,” etc. Such functions can specialize to boolean-valued functions. This formalism also

allows one to calculate steady states and more complete trajectories than are possible in the strictly algebraic approach (Boolean networks or their generalization to multiple discrete values). It may require more detailed data to determine the dynamics from rate parameters.

References

General

Physical Chemistry

Cooperativity Theory in Biochemistry: Steady-State and Equilibrium Systems.

T. L. Hill

Springer Series in Molecular Biology, Springer-Verlag, 1985.

See especially pp. 6-8, 35-36, and 79-81.

The Fokker-Planck Equation: Methods of Solution and Applications (Second Edition)

H. Risken

Springer, 1989

General Bio/Development

Developmental Biology

S. F. Gilbert

Sinauer Associates, (1997)

Molecular Biology of the Gene

J. D. Watson, N. H. Hopkins, J. W. Roberts, J. Argetsinger Steitz and A. M. Weiner

The Benjamin/Cummings Publishing Company, Inc. (1987)

Gene Activity in Early Development

E. H. Davidson

Academic Press, Inc. (1986)

Regulation of even-skipped Stripe 2 in the Drosophila Embryo,

S. Small, A. Blair, and M. Levine,

The EMBO Journal 11:11, pp. 4047-4057, 1992.

Regulation of Two Pair-Rule Stripes by a Single Enhancer in the Drosophila Embryo

S. Small, A. Blair, and M. Levine,

Developmental Biology 175:314-324, 1996.

Transcriptional Repression in the Drosophila Embryo

S. Gray, H. Cai, S. Barolo and M. Levine

Phil. Trans. R. Soc. Lond. B 349, pp. 257-262, 1995.

The Riddle of MAP Kinase Signaling Specificity

H. D. Madhani and G. R. Fink

Trends in Genetics 14:4, 1998.

Automated Assay of Gene Expression at Cellular Resolution

D. Kosman, J. Reinitz, and D. H. Sharp

Pacific Symposium on Biocomputing '98

Eds. R. Altman, A. K. Dunker, L. Hunter, and T. E. Klein

World Scientific 1998

Probability and Stochastic Processes

An Introduction to Probability Theory and Its Applications

W. Feller

John Wiley & Sons, Inc. (1966)

Stochastic Processes in Physics and Chemistry

N. G. Van Kampen

North-Holland

Reviews

Simulation of Prokaryotic Genetic Circuits

H. H. McAdams and A. Arkin

Annu. Rev. Biophys. Biomol. Struct. 27 (1998) p199-224

An Integrated Model of the Transcription Complex in Elongation, Termination, and Editing

P. H. von Hippel

Science 281 (31 July 1998) p660-665

Specific modeling techniques

Based on Physical Chemistry

Quantitative Model for Gene Regulation by Lambda Repressor.

G. K. Ackers, A. D. Johnson and M. A. Shea

PNAS USA 79 (1982) p1129-1133

The OR Control System of Bacteriophage Lambda, A Physical-Chemical Model for Gene Regulation.

M. A. Shea and G. K. Ackers

J Mol Biol 181 (1985) p211-230

Exact Stochastic Simulation of Coupled Chemical Reactions

D. T. Gillespie

J Phys Chem 81 #25 (1977) p2340-2361

Stochastic Approach to Chemical Kinetics

D. A. McQuarrie

J Appl Prob 4 (1967) p413-478

Stochastic Mechanisms in Gene Expression

H. H. McAdams and A. Arkin

PNAS USA 94 (February 1997) p814-819

Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected E.coli cells.

A.P. Arkin, J. Ross, H. H. McAdams

Genetics (1998) Accepted.

Phenomenological Models

Rules for the Evolution of Gene Circuitry

M. A. Savageau

Pacific Symposium on Biocomputing '98

Eds. R. Altman, A. K. Dunker, L. Hunter, and T. E. Klein

World Scientific 1998

Logical and Hybrid Models

The Origins of Order

S. A. Kauffman

Oxford University Press, 1993.

See also: J. Theor. Biol. 22, p. 437, 1969.

Biological Feedback

R. Thomas and R. D'Ari

CRC Press (1990)

Dynamical Behaviour of Biological Regulatory Networks - I. Biological Role of Feedback Loops and Practical Use of the Concept of the Loop-Characteristic State

R. Thomas, D. Thieffry and M. Kaufman

Bull Math Biol 57 #2 (1995) p257-276

A Mathematical Framework for Describing and Analysing Gene Regulatory Networks

T. Mestl, E. Plohte and S. W. Omholt

J Theor Biol 176 (1995) p291-300

Chaos in High-Dimensional Neural and Gene Networks

T. Mestl, C. Lemay, L. Glass

Physica D 98, p. 33, 1996.

Genomic Cis-Regulatory Logic: Experimental and Computational Analysis of a Sea Urchin Gene,

C.-H. Yuh, H. Bolouri, and E. H. Davidson,

Science 279:1896-1902, 1998.