Deterministic Mathematical Models of Transcription and Translation Processes in Bacteria

Jeffrey D. Varner*

*Robert Frederick Smith School of Chemical and Biomolecular Engineering
Cornell University, Ithaca NY 14853

Summary: Gene expression is at the heart of metabolism. Gene expression consists of a decision, and transcription (TX) and translation (TL), the processes by which information stored in DNA is converted to a working protein through a transient intermediate messenger RNA (mRNA). Toward ultimately manipulating TX/TL, we'll develop a working model of transcription and translation kinetics in prokaryotes, and formulate material balances describing mRNA and protein generation in bacterial systems such as *Escherichia coli*. In so doing, we'll leave many of the biological details out of our working picture, however, as we'll see in future lectures these effective models can still quantitatively describe gene expression.

Student outcomes: At the end of this lecture module, students will be able to:

- O_1 Formulate material balance equations governing mRNA and protein generation
- O_2 Use the simplified mental models of the elementary steps of TX/TL to derive expressions for the kinetic rate of transcription (and translation) in a prokaryote.

Material balance equations for transcription and translation

In *E. coli* (or other cells such as yeast or even human cells) not all genes are continuously transcribed under all conditions. Instead, sets of genes that are required for the cell to grow (or function in some other way) in a specific environmental condition are selectively transcribed [1]. Thus, there is a *choice* made to express certain genes and not others; the cell senses its environment (both internally and externally) and makes a choice to synthesize particular proteins. Given the central importance of gene expression, these has been considerable attention paid to developing mathematical approaches to simulate transcription and translation, and more generally the behavior of gene expression networks [2]. These approaches vary in their biological fidelity, each having specific strengths and weaknesses. Today, we'll start a discussion that is more biophysically motivated, and then later contrast this with other perspectives.

Let's start exploring gene expression by writing material balances around mRNA and protein in a population of *E coli* cells growing in a shake flask at the specific rate μ (units: hr^{-1}). The overall flow of information starts from the coding region of DNA, which is read by the RNA polymerase (non-destructive), to produce a messenger RNA molecule which is then read by a Ribosome (non-destructive) to produce a polypeptide, which in the simplest case becomes a protein. Along the way, both the mRNA and the protein can degrade (at different rates). The control logic in this simple model is confined to a DNA region upstream of the transcriptional start site called the promoter sequence (Fig. 1). Thus, in this simple model if a cell successfully initiates transcription, the protein will get produced; see Hu et al., for an alternative view [3].

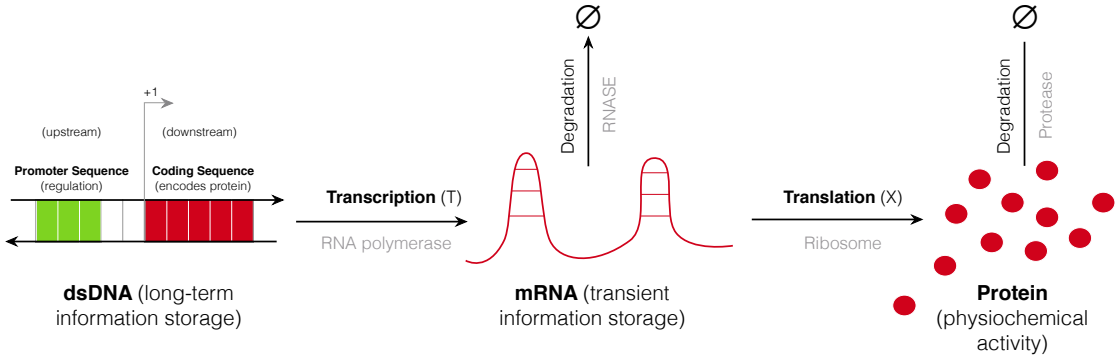

Fig. 1: Schematic of transcriptional (TX) and translation (TL) processes.

Consider gene $\mathcal{G}_i$ (units: nmol/gDW where gDW denotes gram dry weight of cells) which produces protein p_i. The specific material balance equations governing the concentration of mRNA m_i (units: nmol/gDW) transcribed from gene $\mathcal{G}_i$, which is then translated to produce protein p_i (units: nmol/gDW), are given by:

$$\dot{m}_i = r_{X,i}u_i - (\mu + \theta_{m,i})\, m_i + \lambda_i \qquad i = 1, 2, \ldots, N \tag{1}$$

$$\dot{p}_i = r_{L,i}w_i - (\mu + \theta_{p,i})\, p_i \tag{2}$$

where we use the `dot` notation to denote the derivative with respect to time. The term $r_{X,i}u_i$ (units: nmol/gDW-hr) denotes the regulated specific rate of transcription of gene i (production rate of mRNA i), while $r_{L,i}w_i$ (units: nmol/gDW-hr) denotes the specific rate of translation of message i (production rate of protein i). The term λ_i denotes the unregulated rate of transcription (the leak for gene $\mathcal{G}_i$). The $u\left(\dots\right)$ and $w\left(\dots\right)$ terms describe the control logic of the cell for transcription and translation; these terms are dimensionless, and bounded between 0 and 1 e.g., $0 \leq u\left(\dots\right) \leq 1$. There are many ways to formulate these control expressions, and they can be functions of different intracellular (or even extracellular) variables. However, for today let's assume both $u\left(\dots\right)$ = 1.0 and $w\left(\dots\right)$ = 1.0.

The terms in the parenthesis e.g., $(\mu + \theta_{\star,i})$ denote dilution and degradation terms. Degradation rate processes are treated as first-order, where the degradation rate constant $\theta_{\star,i}$ (units: hr^{-1}) governs the rate of lumped non-specific degradation mechanisms (e.g., general RNASE activity). If specific degradation mechanisms are operating in your system of interest, these we'll need to treated separately. Dilution effects come about because of our choice of unit system; we write intracellular concentrations with respect to the total size of the biophase in the shake flask (in this case dry cell mass). This unit system is the prevailing choice for the bioprocess industry as well as the metabolic engineering community (thus, we'll see it many times in subsequent lectures on the analysis and design of metabolic networks).

Where do the dilution terms come from? There are several ways that we can keep track of intracellular species e.g., metabolites, polymerases, proteins, mRNA etc. Suppose we defined intracellular concentration as nmol (or μmol) per unit basis $\mathcal{B}$, where $\mathcal{B}$ is an *abstract* volume basis (like total cell dry weight or total DNA in a flask). To derive the material balance governing the specific intracellular concentration of the j^{th} species x_j ([$\star$mol/$\mathcal{B}$]) we start from the general mole balance and the standard four terms (accumulation, in/out and generation):

$$\dot{n}_{X,acc,j} = \dot{n}_{X,in,j} - \dot{n}_{X,out,j} + \dot{n}_{X,gen,j} \qquad j = 1,2,\dots,\mathcal{M} \tag{3}$$

where $\mathcal{M}$ denotes the number of intracellular species that we want to track, and X denotes a *specific single cell* from our population of cells. There are only two terms that we need to consider for intracellular balances, the accumulation and generation terms; there is no *convective* transport into or from the cellular phase, i.e., cell walls block convective transport of x_j into or from the cell. Thus, the in/out transport terms vanish $\dot{n}_{X,in,j} = \dot{n}_{X,out,j} = 0$ and the intracellular balance around species x_j becomes:

$$\dot{n}_{X,acc,j} = \dot{n}_{X,gen,j} \qquad j = 1,2,\dots,\mathcal{M} \tag{4}$$

We can rewrite the material balance above, assuming some abstract volume $\mathcal{B}$, in a differential-integral form:

$$\frac{d}{dt}\int_{\mathcal{B}} x_j d\mathcal{B} = \int_{\mathcal{B}} \left(\dots\right) d\mathcal{B} \qquad j = 1,2,\dots,\mathcal{M} \tag{5}$$

The first term is the accumulation term, while the right hand side are the generation terms, shown as $\left(\dots\right)$. The integrals with respect to $\mathcal{B}$ can be thought as sums of concentration of species

j, or the chemical reactions in a differential element of the abstract volume $\mathcal{B}$. This calculation is challenging, thus, we almost always involve a simplifying assumption to make the balances represented by Eqn (5) practical, the well mixed assumption.

The well mixed assumption (WMA) implies that the intracellular abundance of species j, denoted as x_j, does not significantly vary between cells i.e., if we sample x_j in many cells, the levels of x_j would be the same to within measurement error. Second, the WMA also implies that x_j does not vary with position *inside* the cells. In reality, the WMA assumption for cells is incorrect on both counts, especially for more advanced organisms; there are many interesting phenomena (both technological and human health related) that occur because of cellular diversity. Moreover, it is well known (especially in eukaryotes) that species abundance varies between the different cellular compartments. However, we'll make this assumption now, and introduce corrections to address some of these shortcomings later. After making the WMA, the integral balance equations reduce to a more recognizable form:

$$\frac{d}{dt}\left(x_i\mathcal{B}\right) = \left(\dots\right)\mathcal{B} \qquad j = 1, 2, \dots, \mathcal{M} \tag{6}$$

Expanding the derivative, and simplifying gives:

$$\dot{x}_j = \left(\dots\right) - x_j\mathcal{B}^{-1}\dot{\mathcal{B}} \qquad j = 1, 2, \dots, \mathcal{M} \tag{7}$$

The term $\mathcal{B}^{-1}\dot{\mathcal{B}}$ on the right hand side of the material balance is an intracellular dilution term; it accounts for the change in the intracellular concentration of species j stemming from changes in the volume basis $\mathcal{B}$. Let's explore a common choice for $\mathcal{B}$, molar specific units.

Molar specific units. In bioprocess engineering and biotechnology, there are several intracellular unit systems which can be used to describe concentration. However, a very common system is to write the concentration with respect to the grams dry weight (gDW) of cells in the culture, for example growing in a shake flash, or bioreactor with working volume of V_R. Thus, intracellular concentration has units of $\star$mol gDW^{-1} which are referred to as *molar specific units*. In this unit system, $\mathcal{B}$ is given by:

$$\mathcal{B} = XV_R \tag{8}$$

where X denotes cellmass concentration (gDW/L) in the culture, and V_R denotes the volume (L) of the culture. Given our concrete choice of concentration basis, we can compute the dilution term in Eqn (7):

$$\mathcal{B}^{-1}\dot{\mathcal{B}} = X^{-1}\dot{X} + V_R^{-1}\dot{V}_R \tag{9}$$

Eqn (9) describes two sources of dilution, first intracellular concentration can dilute because the number of cells (proportional to the gDW of cells) changes, and secondly changes in the the culture volume will also impact concentration. In a batch culture, the volume V_R is constant (excluding volume changes introduced by sampling). Thus, $\dot{V}_R$ = 0 and the dilution term reduces to $\mathcal{B}^{-1}\dot{\mathcal{B}} = X^{-1}\dot{X}$. Using the typical cellmass growth model $\dot{X} = \mu X$, gives a dilution term of $\mathcal{B}^{-1}\dot{\mathcal{B}} = \mu$.

Thus, the dilution of intracellular concentration is proportional to the specific growth rate, leading to intracellular balances of the form:

$$\dot{x}_j = (\dots) - \mu x_j \qquad j = 1, 2, \dots, \mathcal{M} \tag{10}$$

Effective expressions for transcription (and translation) kinetics

To develop expressions for $r_{X,i}$ (or $r_{L,i}$) let's develop a mental model of the elementary steps occurring in transcription (Fig. 1). Our mental model for transcription, based upon the earlier work by McClure [4] and later Bailey [5], consists of a four step elementary reaction scheme:

$$\begin{aligned}
\mathcal{G}_j + R_X &\rightleftharpoons (\mathcal{G}_j : \mathrm{R_X})_C && (11)\\
(\mathcal{G}_j : \mathrm{R_X})_C &\longrightarrow (\mathcal{G}_j : \mathrm{R_X})_O && (12)\\
(\mathcal{G}_j : \mathrm{R_X})_O &\longrightarrow R_X + \mathcal{G}_j && (13)\\
(\mathcal{G}_j : \mathrm{R_X})_O &\longrightarrow m_j + R_X + \mathcal{G}_j && (14)
\end{aligned}$$

where $\mathcal{G}_j$, R_X denote the gene and *free* RNA polymerase (RNAP) concentration, and $(\mathcal{G}_j : \mathrm{R_X})_O$, $(\mathcal{G}_j : \mathrm{R_X})_C$ denote the open and closed complex concentrations, respectively. Let the kinetic rate of transcription be directly proportional to the concentration of the open complex:

$$r_{X,j} = k_{E,j}^{X} (\mathcal{G}_j : \mathrm{R_X})_O$$

where $k_{E,j}^{X}$ is the elongation rate constant for gene j. The key idea behind this derivation is that the RNAP polymerase (or Ribosome) acts as an enzyme. Thus, we might expect that we could use a strategy similar to enzyme kinetics to derive an expression for $r_{X,j}$ (and $r_{L,j}$). The material balances around the closed and open complex for gene j are given by:

$$\begin{aligned}
\frac{d}{dt}(\mathcal{G}_j : \mathrm{R_X})_C &= k_+ (\mathcal{G}_j)(\mathrm{R_X}) - k_- (\mathcal{G}_j : \mathrm{R_X})_C - k_I (\mathcal{G}_j : \mathrm{R_X})_C && (15)\\
\frac{d}{dt}(\mathcal{G}_j : \mathrm{R_X})_O &= k_I (\mathcal{G}_j : \mathrm{R_X})_C - k_A (\mathcal{G}_j : \mathrm{R_X})_O - k_{E,j}^{X} (\mathcal{G}_j : \mathrm{R_X})_O && (16)
\end{aligned}$$

where k_+ (conc^{-1} t^{-1}) and k_- (t^{-1}) denote the on/off rate constant for RNAP at the promoter for gene j, k_I (t^{-1}) denotes the rate constant governing open complex formation and k_A (t^{-1}) denotes the rate constant governing abortive initiation. The total abundance of RNAP, denoted as $R_{X,T}$ is governed by:

$$R_{X,T} = R_X + (\mathcal{G}_j : \mathrm{R_X})_C + (\mathcal{G}_j : \mathrm{R_X})_O \tag{17}$$

At steady state, the abundance of the closed and open complexes can be estimated from the balance equations (where we have neglected the subscript j for simplicity):

$$(\mathcal{G}_j : \mathrm{R_X})_C \simeq \left(\frac{k_+}{k_- + k_I}\right)(\mathcal{G}_j)\,(\mathrm{R_X}) \tag{18}$$

$$(\mathcal{G}_j : \mathrm{R_X})_O \simeq \left(\frac{k_I}{k_A + k_E^X}\right)(\mathcal{G}_j : \mathrm{R_X})_C \tag{19}$$

The ratio of parameters in the open and closed complex expressions have special significance which is apparent from looking at their units. For example, the ratio:

$$K_{X,j}^{-1} \equiv \left(\frac{k_+}{k_- + k_I}\right) \tag{20}$$

is a saturation constant for gene j with units of concentration, while:

$$\tau_{X,j}^{-1} \equiv \left(\frac{k_I}{k_A + k_E^X}\right) \tag{21}$$

is a time constant for gene j comparing the initiation, abortive initiation and elongation constants. We can relate the open complex to the concentration of gene j and *free* RNAP concentration by eliminating the closed complex concentration from the steady state expressions:

$$(\mathcal{G}_j : \mathrm{R_X})_O \simeq (K_{X,j}^{-1})(\tau_{X,j}^{-1})\,(\mathcal{G}_j)\,(\mathrm{R_X}) \tag{22}$$

To estimate the *free* RNAP concentration we can use the total RNAP balance, where we have substituted expressions for the open and closed complex concentrations:

$$R_{X,T} = R_X + (K_{X,j}^{-1})\,(\mathcal{G}_j)\,(\mathrm{R_X}) + (K_{X,j}^{-1})(\tau_{X,j}^{-1})\,(\mathcal{G}_j)\,(\mathrm{R_X}) \tag{23}$$

Starting with Eqn (23), solving for *free* RNAP concentration R_X gives:

$$R_X = \frac{R_{X,T}\,(\tau_{X,j}K_{X,j})}{\tau_{X,j}K_{X,j} + (\tau_{X,j}+1)\,\mathcal{G}_j} \tag{24}$$

Now that we have R_X we can get the open complex concentration in terms of total RNAP:

$$(\mathcal{G}_j : \mathrm{R_X})_O \simeq \frac{R_{X,T}\mathcal{G}_j}{\tau_{X,j}K_{X,j} + (\tau_{X,j}+1)\,\mathcal{G}_j} \tag{25}$$

Lastly, the kinetic rate of transcription is proportional to the open complex concentration which can now be substituted to give:

$$r_{X,j} = k_{E,j}^X R_{X,T}\left(\frac{\mathcal{G}_j}{\tau_{X,j}K_{X,j} + (\tau_{X,j}+1)\,\mathcal{G}_j}\right) \tag{26}$$

In an identical procedure, we can also formulate a model of the translation rate:

$$r_{L,j} = k_{E,j}^{L} R_{L,T} \left(\frac{m_j}{\tau_{L,j} K_{L,j} + (\tau_{L,j} + 1)\, m_j} \right) \tag{27}$$

where m_j denotes the concentration of mRNA j.

Which is limiting, elongation or initiation? Ultimately, this question depends upon the gene of interest. However, we can see some interesting properties of $r_{X,j}$ by considering limiting cases for the value of the time constant $\tau_{X,j}$. Assume the rate constant for abortive initiation k_A is small compared to both k_I and $k_{E,j}^{X}$:

$$\tau_{X,j} \simeq \frac{k_{E,j}^{X}}{k_I} \tag{28}$$

When $\tau_{X,j} \gg 1$ (initiation limited) the kinetic transcription rate becomes:

$$r_{X,j} = \frac{k_{E,j}^{X} R_{X,T}}{\tau_{X,j}} \left(\frac{\mathcal{G}_j}{K_{X,j} + \mathcal{G}_j} \right) \tag{29}$$

while $\tau_{X,j} \ll 1$ (elongation limited) gives:

$$r_{X,j} = k_{E,j}^{X} R_{X,T} \left(\frac{\mathcal{G}_j}{K_{X,j} \tau_{X,j} + \mathcal{G}_j} \right) \tag{30}$$

How do we get values for k_+, k_-, k_I, $k_{E,j}^{X}$ and k_A? Generally speaking, except for $k_{E,j}^{X}$ which we can estimate from first principles, estimating the value of k_+, k_-, k_I and k_A is difficult (especially *in-vivo*). Thus, let's start with $k_{E,j}^{X}$; the elongation rate constant is proportional to the elongation rate of the polymerase e_X (units of nt s^{-1}) multiplied by the length (nt) of the coding region of gene j, or $\mathcal{L}_j$ (the length of DNA the RNAP has to read). However, typically we formulate $k_{E,j}^{X}$ in a slightly different way; first, we compute an average or characteristic elongation rate constant $\left\langle k_E^X \right\rangle$, and then correct this characteristic value by the actual length of gene j:

$$k_{E,j}^{X} = \left\langle k_E^X \right\rangle \left(\frac{\mathcal{L}}{\mathcal{L}_j} \right) \tag{31}$$

where:

$$\left\langle k_E^X \right\rangle = e_X \mathcal{L}^{-1} \tag{32}$$

and $\mathcal{L}$ denotes some characteristic length, e.g., the average length of genes in *E.coli*. For the other parameters, we must estimate them from experimental data.

McClure performed a series of *in vitro* experiments to estimate k_I in transcription and produced a constraint governing permissible values for the remaining transcriptional parameters [4]. In particular, McClure used an abortive initiation assay in which the production of mRNA never completed. Instead, transcription always aborted leaving a stable open complex that could be directly measured. From these measurements, and a mental transcriptional model very similar to

ours, McClure developed the expression:

$$\tau_{obs} = \frac{1}{k_I} + \frac{1}{R_{X,T}}\left(\frac{k_- + k_I}{k_+ k_I}\right) \tag{33}$$

where τ_{obs} is the time required to fully form the open complex (measured) and $R_{X,T}$ denotes the total concentration of RNAP. A value for k_I, and a relationship between the other parameters, can be obtained from the intercept and slope of a $R_{X,T}^{-1}$ versus τ_{obs} plot for a particular promoter of interest.

Summary and conclusions

In this lecture we developed an effective ordinary differential equation (ODE) based model of gene expression for bacteria that used a biophysical, albeit effective, description of transcription and translation kinetics. This model could be used to simulate both steady-state and dynamic gene expression processes if we have a description of the regulation that controls transcription and translation.

References

1. DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278: 680-6.
2. Karlebach G, Shamir R (2008) Modelling and analysis of gene regulatory networks. Nat Rev Mol Cell Biol 9: 770-80.
3. Hu CY, Varner JD, Lucks JB (2015) Generating effective models and parameters for rna genetic circuits. ACS Synth Biol 4: 914-26.
4. McClure WR (1980) Rate-limiting steps in rna chain initiation. Proc Natl Acad Sci U S A 77: 5634-8.
5. Lee SB, Bailey JE (1984) Genetically structured models forlac promoter–operator function in the escherichia coli chromosome and in multicopy plasmids: Lac operator function. Biotechnology and Bioengineering 26: 1372–1382.