A Unified Probabilistic Modeling Framework for Eukaryotic Transcription Based on Nascent RNA Sequencing Data

Adam Siepel

Simons Center for Quantitative Biology
Cold Spring Harbor Laboratory
Cold Spring Harbor, NY, USA

Abstract

Nascent RNA sequencing protocols, such as PRO-seq and NET-seq, are now widely used in the study of eukaryotic transcription, and these experimental techniques have given rise to a variety of statistical and machine-learning methods for data analysis. These computational methods, however, are generally designed to address specialized signal-processing or prediction tasks, rather than directly describing the dynamics of RNA polymerases as they move along the DNA template. Here, I introduce a general probabilistic model that describes the kinetics of transcription initiation, elongation, pause release, and termination, as well as the generation of sequencing read counts. I show that this generative model enables estimation of separate pause-release rates, termination rates, and the initiation/elongation rate ratio up to a proportionality constant. Furthermore, if applied to time-course data in a nonequilibrium setting, the model can be used to estimate elongation rates. This model leads naturally to likelihood ratio tests for differences between genes, conditions, or species in various rates of interest. If read counts are assumed to be Poisson-distributed, convenient, closed-form solutions are available for both parameter estimates and likelihood-ratio-test statistics. Straightforward extensions of the model accommodate variability in the pause site and steric hindrance of initiation by paused polymerases. Additional extensions address Bayesian inference under the Poisson model and a generalized linear model that can be used to discover genomic features associated with rates of elongation. Finally, I address technicalities concerning estimation of library size, normalization and sequencing replicates. Altogether, this modeling framework enables a unified treatment of many common tasks in the analysis of nascent RNA sequencing data.

Introduction

In recent years, several experimental protocols have emerged for measuring newly produced RNAs on a genome-wide scale [1–5]. These *nascent RNA sequencing* methods can be thought of as adaptations of traditional nuclear run-on assays that exploit modern massively-parallel sequencing technologies to simultaneously map the positions of engaged RNA polymerases across an entire genome. They have proven useful in a wide variety of applications, including measurement of transcription levels independent of RNA decay [1], measurement of rates of elongation [6], mapping of transcription start sites [7], identification of active enhancers [8,9], and estimation of relative RNA half-lives [10]. These measurements have led, in turn, to a number of new insights into the biology of transcription, such as the widespread incidence of promoter-proximal pausing and divergent transcription [1, 2], the remarkably similar architecture of transcription initiation at promoters and enhancers [7], pause-release as a crucial rate-limiting step in transcription [11], and distinct waves of transcriptional responses to the induction of transcription [9].

A variety of computational and statistical methods have been used to analyze these data, ranging from relatively simple count- and ratio-based tests to hidden Markov models and discriminative machine learning

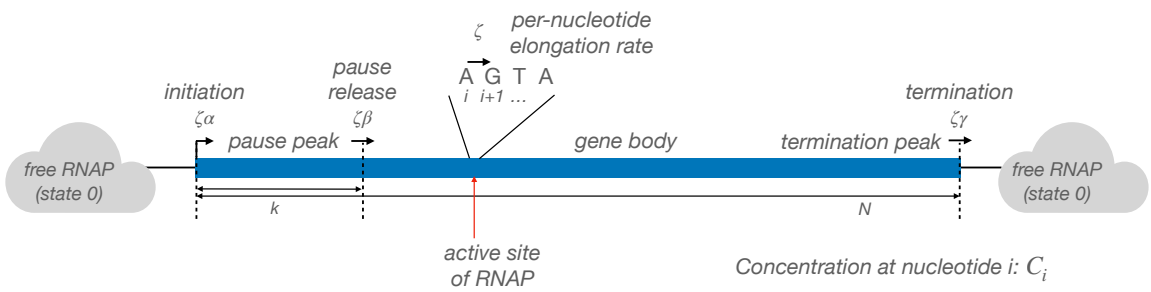

Figure 1: General kinetic model for transcription initiation, elongation, pause release, and termination along a transcription unit.

approaches (e.g., [6, 8, 9, 12, 13]). Some of these methods are sophisticated, powerful, and widely used. For the most part, however, they have focused fairly narrowly on specific prediction tasks (such as elongation rate estimation or TSS prediction), rather than on more generally modeling the biophysical processes underlying transcription. As a result, with a few partial exceptions [14, 15], the existing computational methods for nascent RNA sequence data, or closely related data types, generally do not permit direct estimation of biophysical parameters of interest, such as rates of initiation or termination. In addition, owing to their heuristic description of the underlying processes, they can be difficult to adapt for nuances in the process, such as promoter-proximal pausing or variation in elongation rate along the gene body.

In this article, I introduce a unified framework that combines a simple kinetic model of transcription with a generative model for nascent RNA sequence data, and permits direct inference of kinetic parameters from sequence data. Separate layers in the model account for stochasticity in the process and noise in the sequencing data, and extension allows for uncertainty in the location of the pause site. I show that this model can be applied to a number of problems of interest, including estimation of initiation and elongation rates. It also leads directly to powerful statistical tests for differences in rates of interest, which naturally consider both the information in the read count data and the underlying physical process. In its simplest form, the modeling framework ignores collisions between RNA polymerase molecules, but I show that it can be extended to allow for certain classes of collisions. In addition, the model has natural extensions to Bayesian inference and to a generalized linear model that accommodates covariates of elongation rate. Finally, I show how to address technical issues concerning estimation of library size, normalization, and sequencing replicates within this modeling framework. I conclude with a discussion of some current limitations and possible extensions of the current framework.

General Kinetic Model for Transcriptional Initiation, Elongation, Pause-Release, and Termination

I begin with a general kinetic model for the process by which RNA polymerase (RNAP) enzymes initiate transcription at the transcription start site (TSS) of a transcription unit (TU) and move along the DNA template, synthesizing an elongating nascent RNA molecule. This model is defined by a series of states corresponding to each nucleotide position of RNAP as it moves along the template (Figure 1). In particular, state i represents the positioning of the active site of RNAP at nucleotide i of the TU and corresponds to a nascent RNA of length $i - 1$. I assume that each RNAP eventually traverses the entire DNA template, borrowing from evidence suggesting that premature termination occurs at relatively low rates across most transcription units [16] (but see Discussion).

I will distinguish between two segments of a TU of length N: (1) the first k nucleotides, known as the *pause peak*, where RNAP tends to accumulate owing to promoter-proximal pausing (typically $k \approx 50$) [1];

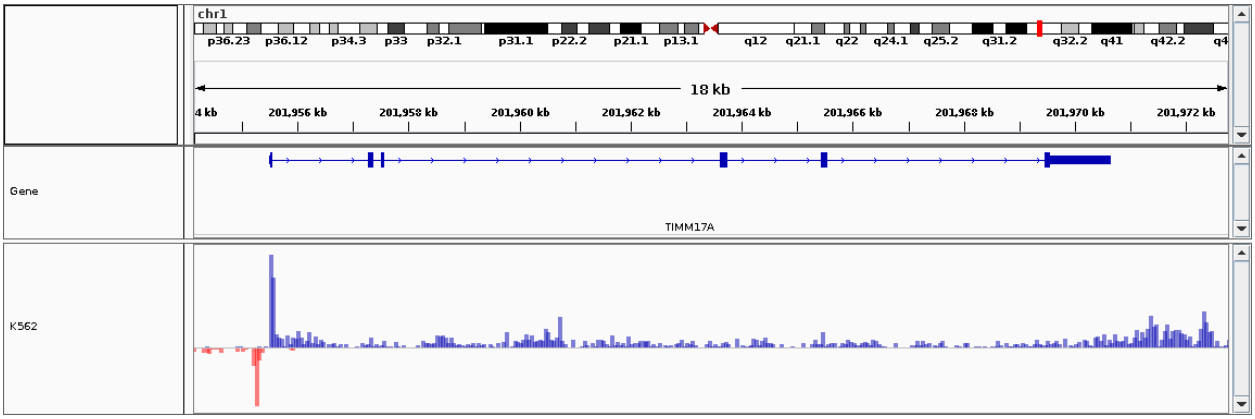

Figure 2: Example of real nascent RNA sequencing data in the region of the human *TIMM17A* gene. The data were obtained by applying PRO-seq to K562 cells, as described in ref. [9]. The pause peak and termination peak are clearly visible. Notice, however, that the read counts are quite noisy along the gene body. In addition, transcription in this case seems to have run past the annotated transcription termination site, a phenomonen that occurs frequently. As a result, the methods described in this paper would likely require some adjustment of annotated gene coordinates based on the raw data itself (e.g., see [13]).

and (2) the subsequent $N-k$ nucleotides, where RNAP tends to be relatively unimpeded, which is typically referred to as the *gene body*. I will also allow for a *termination peak* near the end of the TU, owing to delays from termination of transcription (see Figure 2). The model has $N+1$ states, corresponding to the N nucleotide positions of the TU plus an additional state (labeled 0) that abstractly represents *free RNAP*, that is, RNAP that is not currently engaged in transcription and is available for a new initiation.

I assume, as with most existing protocols, that the data summarize a relatively large population of cells, and denote by C_i the concentration of cells in the sampled population having RNAP positioned at nucleotide i. As will be seen below, the sequencing reads from a PRO-seq experiment that map to position i at their $3'$ ends should have a depth roughly proportional to C_i. The situation is similar for other protocols.

The kinetic model is defined by four rates: an initiation rate α, a pause-release rate β, a termination rate γ, and a constant per-nucleotide elongation rate ζ. For mathematical convenience, I assume that the initiation, pause-release, and termination steps are coupled with single-nucleotide elongation steps and occur at rates $\zeta\alpha$, $\zeta\beta$, and $\zeta\gamma$, respectively. As a result, as long as ζ is the same across nucleotides, it can be considered a scaling factor that applies equally to all steps in the process. (Later I will consider a generalization that allows this rate to vary across nucleotides.) I will also assume, for now, that the RNAPs are relatively sparse along each DNA template, and ignore the effects of collisions between them, although this simplification, too, will be revisited.

With these assumptions, the concentrations C_i, for $i \in \{0, \ldots, N\}$, are governed by the following

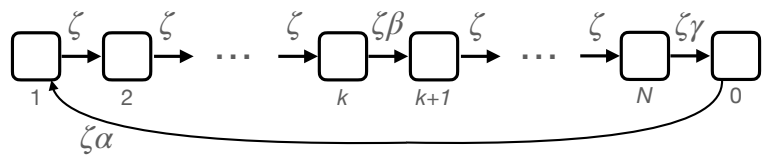

Figure 3: State-transition diagram for continuous-time Markov model for a single RNAP molecule moving along the DNA template.

system of $N+1$ differential equations,

$$\begin{aligned}
\frac{d}{dt}C_1(t) &= \zeta\alpha C_0(t) - \zeta C_1(t) \\
\frac{d}{dt}C_2(t) &= \zeta C_1(t) - \zeta C_2(t) \\
&\vdots \\
\frac{d}{dt}C_k(t) &= \zeta C_{k-1}(t) - \zeta\beta C_k(t) \\
\frac{d}{dt}C_{k+1}(t) &= \zeta\beta C_k(t) - \zeta C_{k+1}(t) \\
&\vdots \\
\frac{d}{dt}C_N(t) &= \zeta C_{N-1}(t) - \zeta\gamma C_N(t) \\
\frac{d}{dt}C_0(t) &= \zeta\gamma C_N(t) - \zeta\alpha C_0(t).
\end{aligned}$$

I will consider both the time evolution of the C_i values and their values at steady-state.

Continuous-time Markov Model

If instead of thinking in terms of a population of cells, one considers the probability distribution over states (nucleotide-specific positions) for a given RNAP, the same model can be represented using the convenient and flexible formalism of continuous-time Markov models. The continuous-time Markov version of the model consists of $N+1$ states and transition rates exactly analogous to the ones above (Figure 3). In this case, however, each state i corresponds to a binary random variable Z_i, indicating whether a given RNAP is ($Z_i = 1$) or is not ($Z_i = 0$) at a particular position at a particular time t. The assumption that collisions are rare allows the stochastic process followed by each RNAP to be considered independent of all of the others.

It is easy to see that this model is identical to the previous one up to a normalization term. Given a sufficiently large collection of independent RNAP molecules, the overall concentration $C_i(t)$ is expected to be proportional to the probability of occupancy $P(Z_i \,|\, t, \alpha, \beta, \gamma, \zeta)$ for any individual RNAP. Furthermore, for $P(Z_i \,|\, t, \alpha, \beta, \gamma, \zeta)$ to remain a proper probability distribution, the constant of proportionality must be given by $\mathcal{Z}(t) = \sum_{i=0}^{N} C_i(t)$.

This Markov chain is defined by an $(N+1) \times (N+1)$ infinitesimal generator matrix, $\mathbf{Q} = \zeta\mathbf{R}$, such

that,

$$\mathbf{R} = \begin{pmatrix} -\alpha & \alpha & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ & \vdots & & & & & \vdots & & & & & \vdots & & & & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -\beta & \beta & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ & \vdots & & & & & \vdots & & & & & \vdots & & & & \vdots \\ \gamma & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & -\gamma \end{pmatrix}.$$

The element at row i and column j of $\mathbf{Q}$ indicates the instantaneous rate at which transitions occur from state i to state j. By convention, the values along the diagonal are set such that the rows sum to zero. Because the states must be visited in a sequence, this matrix has a particularly simple form: it has nonzero terms only on the main diagonal, the diagonal immediately above it, and the single element at bottom left (which allows RNAPs to "wrap around" from the last state to the first one).

The probability of transitioning from any state i to any other state j over a period of time $t \geq 0$ can be computed in the usual way as,

$$\mathbf{P}(t, \alpha, \beta, \gamma, \zeta) = e^{\mathbf{Q}t} = e^{\mathbf{R}\zeta t}, \tag{1}$$

where the element at row i and column j of the matrix $\mathbf{P}()$ indicates the conditional probability of occupancy of state j at time t given occupancy of state i at time 0, and the matrix exponential is given by the Taylor expansion $e^{\mathbf{R}\zeta t} = \mathbf{I} + \mathbf{R}\zeta t + \frac{(\mathbf{R}\zeta t)^2}{2!} + \frac{(\mathbf{R}\zeta t)^3}{3!} + \ldots$ [17]. If transcription is initiated at time 0 (meaning that the initial state is state 0), then the distribution over states at a later time t is simply given by row $i = 0$ of $\mathbf{P}()$. In particular, the element at the jth column of row $i = 0$ indicates the quantity I have denoted $P(Z_j \,|\, t, \alpha, \beta, \gamma, \zeta)$.

Notice that the elongation rate ζ and the time t are nonidentifiable under these assumptions; one of these quantities must be known to estimate the other.

As discussed further below, it will sometimes be desirable to allow for uncertainty in k, according to some prior distribution $P(k)$. In this case, the transition matrix can simply be defined as a mixture

$$\mathbf{P}(t, \alpha, \beta, \gamma, \zeta) = \sum_k P(k) e^{\mathbf{R}_k \zeta t}, \tag{2}$$

where each $\mathbf{R}_k$ reflects a different choice of k.

Generative Model for Sequence Data

In the analysis of nascent RNA sequencing data, we have the additional complication that the positions of RNAP molecules are not directly observed, but instead are indirectly indicated by aligned sequence reads. The read counts reflect a sampling process that probes only a relatively small subset of cells in the starting material. Moreover, this process can be biased by factors such as base composition or RNA secondary structure that can lead to subtle differences in the efficiency of various steps in library preparation, including DNA fragmentation, purification, and PCR amplification.

I will address this problem by adding a second layer to the model that describes the probabilistic process by which sequencing read counts are generated conditional on an underlying "true" density of RNAP at each nucleotide. In this way, I can obtain a full generative model for the observed sequence data that is defined by the parameters of both the kinetic model described above and these conditional distributions for sequencing reads, enabling inference of both sets of parameters from data.

Let μ_i be the expected read depth at position i. I will assume that μ_i is proportional to $P(Z_i | t, \alpha, \beta, \gamma, \zeta)$ (as computed from the continuous-time Markov chain), with a proportionality constant that depend on the

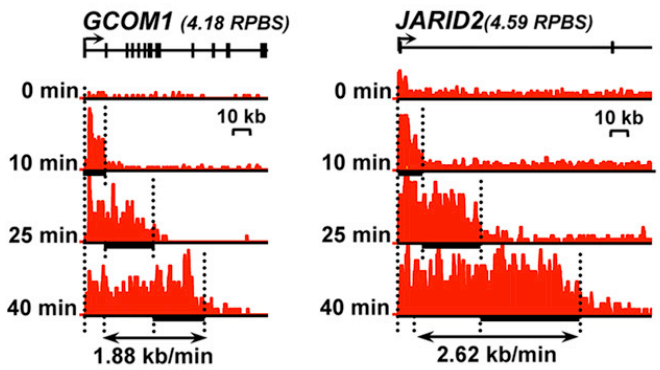

Figure 4: Examples of nascent RNA sequencing data collected in a time-course and used to estimate rates of elongation. Data were collected by GRO-seq at 0, 10, 25, and 40 minutes in MCF-7 cells [6]. Results are shown for the *GCOM1* (left) and *JARID2* (right) genes. In both cases, the left edges of the *x*-axes correspond to the annotated TSS. The RNAP "wave" can be seen to move rightward as time advances. Figure adapted from reference [6].

sequencing depth, that is, with $\mu_i = \lambda P(Z_i \,|\, t, \alpha, \beta, \gamma, \zeta)$, where λ is a measure of sequencing depth. In addition, I will abstractly assume a general generating distribution, ψ, for read counts given μ_i and any other relevant parameters, which will be cumulatively denoted θ. Let the data be denoted $\mathbf{X} = (X_1, \ldots, X_N)$, where X_i representes the number of sequencing reads having their $3'$ end aligned to position i. Assuming independence of sequencing reads, the likelihood for the data in terms of the time t and the free parameters of the model can be defined as,

$$P(\mathbf{X} \,|\, t, \alpha, \beta, \gamma, \zeta, \theta) = \prod_{i=1}^{N} \psi(X_i \,|\, \mu_i, \theta), \tag{3}$$

where $\mu_i = \lambda P(Z_i \,|\, t, \alpha, \beta, \gamma, \zeta)$ and $P(Z_i \,|\, t, \alpha, \beta, \gamma, \zeta)$ is obtained from the matrix $\mathbf{P}(t) = e^{\mathbf{Q}t}$, as described above. As will be seen below, λ can generally be preestimated from global properties of a sequencing data set. Notice that typically there will be a data point corresponding to each state of the model, with the exception of state 0, which is unobservable.

Nonequilibrium Case and Application to Time-Course Data

Suppose nascent RNA sequencing data have been collected in a time course, with read counts $\mathbf{X}^{(t)}$ for $t \in \{t_1, t_2, \ldots, t_M\}$. Typically $t_1 = 0$ and represents the case prior to induction of transcription (Figure 4) [6]. Assuming independence of the data for the time points, a joint likelihood for all data can be defined as,

$$P(\mathbf{X}^{(t_1)}, \ldots, \mathbf{X}^{(t_M)} \,|\, \alpha, \beta, \gamma, \zeta, \theta) = \prod_{j=1}^{M} \prod_{i=1}^{N} \psi(X_i^{(j)} \,|\, \mu_i^{(j)}, \theta), \tag{4}$$

where $\mu_i^{(j)} = \lambda^{(j)} P(Z_i \,|\, t_j, \alpha, \beta, \gamma, \zeta)$, analogous to the case above. Notice that I allow here for a time-point-specific sequencing depth, $\lambda^{(j)}$.

In this case, α, β, γ, and ζ can all be estimated from the data, at least in principle, although the information about some of them may be weak. However, there is a practical problem relating to polymerase

positions that are impossible under the model (for example, when $t = 0$ and transcription has not yet commenced) but for which the corresponding read count X_i is nonzero simply owing to the background noise of the assay. This problem can be addressed by defining $\psi(X_i \mid \mu_i, \theta)$ by a suitable background distribution whenever $\mu_i = 0$ (or perhaps whenever $\mu_i < T$, where T is a background-derived threshold). In this way, it should be possible to estimate an elongation rate ζ for each transcription unit directly from time-course data by numerical optimization of equation 4. This approach would provide an alternative to current hidden Markov model-based inference methods and would have the advantage of allowing estimation of α and β as well. The termination rate γ may be more difficult to estimate, but it may be possible to do so with a sufficiently long time course.

As observed above, ζ and t cannot be separately estimated from data, because they appear in the likelihood only as the product ζt. However, in a time-course setting, t is known and thus ζ can be estimated.

Stationary Distribution and Inference at Steady-State

This Markov chain is ergodic and, therefore, will eventually reach a steady-state equilbrium defined by a unique stationary distribution over states. This stationary distribution, denoted $\boldsymbol{\pi}$, is invariant to ζ and can be found by solving the equation $\boldsymbol{\pi}(\mathbf{R} + \mathbf{I}) = \boldsymbol{\pi}$, or equivalently, $\boldsymbol{\pi}\mathbf{R} = \mathbf{0}$. Owing to the simple structure of $\mathbf{R}$, the solution for $\boldsymbol{\pi}$ can easily be determined by substitution.

Because state 0 is simply an abstraction to allow for the recirculation of RNAP and cannot be measured, we will ignore its stationary probability and instead describe the stationary distribution conditional on RNAP occupancy along the DNA template. This conditional stationary distribution is given by $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$ such that,

$$\pi_i = \frac{1}{\mathcal{Z}} \cdot \begin{cases} \frac{\alpha}{\beta} & i = k \\ \frac{\alpha}{\gamma} & i = N \\ \alpha & i \in \{1, \ldots, N-1\}, i \neq k, \end{cases} \tag{5}$$

with normalization constant $\mathcal{Z} = \alpha \left(N - 2 + \frac{1}{\beta} + \frac{1}{\gamma} \right)$.

This distribution has an intuitive interpretation. First, note that it will typically be the case that $\beta, \gamma < 1$, owing to slowdowns in elongation from pausing and termination. By contrast, RNAP should proceed unimpeded in the gene body, and therefore its density at steady state should reflect the initiation rate. Accordingly, π_i is proportional to α in the gene body, but elevated by factors of $\frac{1}{\beta}$ and $\frac{1}{\gamma}$ at the pause peak and termination peak, respectively (Figure 5). The quantity $\frac{1}{\beta}$ is sometimes estimated from nascent RNA sequencing data and described as the *pause index*. By analogy, the quantity $\frac{1}{\gamma}$ can be described as a *termination index*.

Notice that I have chosen to write the elements of $\boldsymbol{\pi}$ such that α appears in each numerator, despite that these α terms would cancel with the denominator and hence could be omitted. The reason for this choice will become clearer later in the article when multiple genes are considered, each with its own version of α. The inclusion of α in the stationary distribution will allow these values to be compared across genes.

Of course, when comparing different genes, it is possible that their elongation rates also differ. When examining nascent RNA-sequencing data at steady-state, the read depth in each TU j will be proportional to $\frac{\alpha_j}{\zeta_j}$, where α_j and ζ_j are, respectively, the initiation and elongation rates specific to that TU. Furthermore, the observed counts will be also be proportional to the sequencing depth, λ. Because these three parameters are not identifiable at steady state, I will represent them by the compound parameter $\chi = \frac{\lambda\alpha}{\zeta}$. Thus, the stationary distribution can be written (Figure 5),

$$\pi_i = \frac{1}{\mathcal{Z}} \cdot \begin{cases} \frac{\lambda\alpha}{\beta\zeta} = \frac{\chi}{\beta} & i = k \\ \frac{\lambda\alpha}{\gamma\zeta} = \frac{\chi}{\gamma} & i = N \\ \frac{\lambda\alpha}{\zeta} = \chi & i \in \{1, \ldots, N-1\}, i \neq k, \end{cases} \tag{6}$$

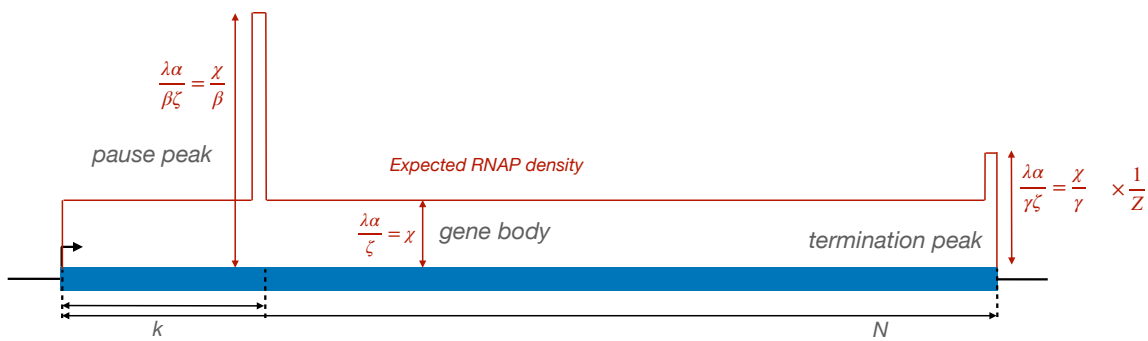

Figure 5: Expected density of RNAP at steady-state under the continuous-time Markov model.

with $\mathcal{Z} = \chi\left(N - 2 + \frac{1}{\beta} + \frac{1}{\gamma}\right)$.

This stationary distribution can be used to define a likelihood function for a system in steady state, described by a single data set (as opposed to a time course), denoted $\mathbf{X} = (X_1, \ldots, X_N)$. In this case, let the expected read depth at each site i be $\mu_i \propto \pi_i$, so that the likelihood of the data is given by (cf. equation 3),

$$\begin{aligned} P(\mathbf{X}|\chi, \beta, \gamma, \theta) &= \prod_{i=1}^{N} \psi(X_i \,|\, \mu_i, \theta) \\ &= \left[\prod_{i=1}^{N} \psi\left(X_i \,|\, \chi, \theta\right)\right] \times \frac{\psi\left(X_k \,|\, \chi/\beta, \theta\right)}{\psi\left(X_k \,|\, \chi, \theta\right)} \times \frac{\psi\left(X_N \,|\, \chi/\gamma, \theta\right)}{\psi\left(X_N \,|\, \chi, \theta\right)}, \end{aligned} \tag{7}$$

where the normalization constant for $\boldsymbol{\pi}$ is assumed to be absorbed by χ.

As it turns out, the termination peak is typically difficult to characterize with real nascent RNA sequencing data, owing to transcriptional run-on, poorly characterized $3'$ ends of genes, and other factors. Therefore, from this point on, I will omit the termination component of the model and the corresponding parameter γ. With this simplification, the likelihood function can be written,

$$P(\mathbf{X}|\chi, \beta, \theta) = \left[\prod_{i=1}^{N} \psi\left(X_i \,|\, \chi, \theta\right)\right] \times \frac{\psi\left(X_k \,|\, \chi/\beta, \theta\right)}{\psi\left(X_k \,|\, \chi, \theta\right)}. \tag{8}$$

The Poisson Case

In some applications, the generating distribution ψ may need to be complex to capture realistic patterns in sequencing reads, but in others, it may be sufficient to assume a simple form for this distribution. The assumption of a Poisson distribution at each nucleotide for ψ leads to particularly straightforward closed-form estimators in many applications of interest.

With a Poisson generating distribution, the steady-state likelihood becomes simply,

$$\begin{aligned} P(\mathbf{X}|\chi,\beta) &= \left[\prod_{i=1}^{N} \text{Pois}\left(X_i \mid \chi\right)\right] \times \frac{\text{Pois}\left(X_k \mid \chi/\beta\right)}{\text{Pois}\left(X_k \mid \chi\right)} \\ &= \left[\prod_{i=1}^{N} \frac{\chi^{X_i} e^{-\chi}}{X_i!}\right] \times \beta^{-X_k} e^{-\chi\left(\frac{1}{\beta}-1\right)} \\ &= \frac{1}{\mathcal{Z}} \chi^{s} \beta^{-X_k} \exp\left[-\chi\left(N + \frac{1}{\beta} - 1\right)\right], \end{aligned} \tag{9}$$

where $s = \sum_{i=1}^{N} X_i$ is a sufficient statistic equal to the sum of all read counts and $\mathcal{Z} = \prod_{i=1}^{N} X_i!$ is a normalization term that does not depend on the model parameters and can be ignored during optimization. The log likelihood is given by,

$$\ell(\mathbf{X};\chi,\beta) = s \log \chi - X_k \log \beta - \chi\left(N + \frac{1}{\beta} - 1\right) - \log \mathcal{Z}. \tag{10}$$

The maximum-likelihood estimators for χ and β have simple closed-form solutions:

$$\hat{\chi} = \frac{s - X_k}{N-1} \tag{11}$$

$$\hat{\beta} = \frac{s - X_k}{X_k(N-1)}. \tag{12}$$

Thus, in this case, the estimator for χ is simply the average read depth, excluding the pause peak, and the estimator for β is the ratio of that same average read depth to the read depth in the pause peak. Notably, these are estimators that have been widely used in the analysis of nascent RNA sequencing data, with more heuristic justifications.

In practice, it is often best to avoid the complex signal in the pause region in estimating χ and instead estimate it from a downstream portion of the gene body. In this case, one can define $s' = \sum_{i=j}^{j+M-1} X_i$, where M is the length of the interval considered, and estimate χ and β as,

$$\begin{aligned} \hat{\chi} &= \frac{s'}{M} \\ \hat{\beta} &= \frac{s'}{X_k M}. \end{aligned} \tag{13}$$

In most cases below, I will make use of these simpler, more robust estimators for χ and β.

While the nucleotide-specific counts in real data sets tend to be overdispersed, the simplicity and speed of the Poisson model may still make it an attractive option in many analyses. For example, as will be seen below, it may be desirable to perform statistical tests based on the Poisson model and then adjust *p*-values using an empirical calibration.

Allowing for Variation in the Pause Site

In practice, the pause site tends to vary across cells, leading to a broad pause peak in real nascent RNA sequencing data. This problem can be addressed by allowing the location of the pause peak, k, to vary between some k_{min} and some k_{max} according to an appropriate distribution, and then assuming each read count X_k within this range reflects a mixture of cells that do and do not have their pause site at position k. More specifically, I will assume that, for $k \in \{k_{\text{min}}, \ldots, k_{\text{max}}\}$, f_k represents the fraction of cells with pause

site k, and that the read count X_k derives from a mixture of one Poisson distribution with rate $\chi/\beta \cdot f_k$ and a second Poisson distribution with rate $\chi \cdot (1 - f_k)$. If we denote by Y_k the (unknown) portion of the read count that derives from the first process, then the likelihood function can be expressed as (cf. equation 9),

$$P(\mathbf{X} \mid \chi, \beta) = \left[\prod_{i=1}^{k_{\text{min}}-1} \text{Pois}\left(X_i \mid \chi\right) \right] \times \left[\prod_{k=k_{\text{min}}}^{k_{\text{max}}} \sum_{Y_k=0}^{X_k} \text{Pois}\left(Y_k \mid \chi/\beta \cdot f_k\right) \text{Pois}\left(X_k - Y_k \mid \chi \cdot (1 - f_k)\right) \right] \times \left[\prod_{i=k_{\text{max}}+1}^{N} \text{Pois}\left(X_i \mid \chi\right) \right]. \tag{14}$$

This likelihood function is straightforward to maximize by expectation maximization. The complete-data log likelihood function, with known values of Y_k, can be expressed in terms of compact sufficient statistics (analogous to equation 10) as,

$$\ell_c(\mathbf{X}, \mathbf{Y}\,;\, \chi, \beta) = s \log \chi - t \log \beta - \chi \left(N + \frac{1}{\beta} - 1 \right) + \sum_{k=k_{\text{min}}}^{k_{\text{max}}} Y_k \log f_k + (X_k - Y_k) \log(1 - f_k) - \log Z \tag{15}$$

where $s = \sum_i X_i$ and $t = \sum_k Y_k$. For simplicity, assume that χ is pre-estimated for a portion of the gene body downstream of k_{max} using equation 13. For the moment, assume also that the values of f_k are fixed in a preprocessing step. Then β can be estimated as,

$$\hat{\beta} = \frac{\hat{\chi}}{\langle t \rangle}, \tag{16}$$

where $\langle t \rangle$ denotes the posterior expected value of t, which can be computed on each iteration by summing the posterior expected values $\langle Y_k \rangle$. These values, in turn, can be computed by observing that, because X_k is the sum of two Poisson-distributed variables, $Y_k \mid X_k$ is binomially distributed with probability,

$$p_k = \frac{f_k \chi/\beta}{f_k \chi/\beta + (1 - f_k)\chi} = \frac{1}{1 - \beta + \beta/f_k}. \tag{17}$$

Therefore, $\langle Y_k \rangle = X_k \cdot p_k$, and

$$\langle t \rangle = \sum_k X_k \cdot p_k = \sum_k \frac{X_k}{1 - \beta + \beta/f_k}. \tag{18}$$

Thus, an EM algorithm can be implemented by iteratively applying equations 16 and 18, in turn, until convergence.

However, it may also be of interest to define the f_k values in terms of a parametric distribution and estimate the parameters of that distribution also by EM. In this way, the distribution over pause sites can be estimated from the data at each TU of interest. A natural choice is a Gaussian distribution with mean μ and variance σ^2,

$$f_k = \frac{1}{\mathcal{Z}} \times \begin{cases} \frac{1}{\sigma} e^{-\frac{1}{2}\left(\frac{k-\mu}{\sigma}\right)^2} & k \in \{k_{\text{min}}, \ldots, k_{\text{max}}\} \\ 0 & \text{otherwise,} \end{cases} \tag{19}$$

where the explicit normalization constant $\mathcal{Z}$ is needed because the distribution is applied to a bounded interval and is defined at integer values only.

In this case, it can be shown that the EM updates for μ and σ^2 are simply,

$$\hat{\mu} = \frac{u - z}{t - w}, \qquad \hat{\sigma}^2 = \frac{v - r}{t - w} - \hat{\mu}^2, \tag{20}$$

where,

$$t = \sum_k \langle Y_k \rangle, \qquad w = \sum_k \frac{f_k}{1 - f_k}(X_k - \langle Y_k \rangle),$$

$$u = \sum_k \langle Y_k \rangle k, \qquad z = \sum_k \frac{f_k}{1 - f_k}(X_k - \langle Y_k \rangle)k,$$

$$v = \sum_k \langle Y_k \rangle k^2, \qquad r = \sum_k \frac{f_k}{1 - f_k}(X_k - \langle Y_k \rangle)k^2.$$

Notice that all of these quantities are easily computed from the $\langle Y_k \rangle$ values together with f_k and X_k.

Allowing for Collisions

As noted, the continuous-time Markov model described in this article makes the important simplification of ignoring collisions between polymerase molecules by modeling the movement of each RNAP independently of the others. In reality, an RNAP molecule can never pass another RNAP molecule along the DNA template, and will be blocked from forward progress if it catches up to its predecessor. The approximation should be adequate when the density of RNAPs per TU is low, but it will begin to fail when either the average density becomes high or when the elongation rate in certain local regions (such as the pause site) becomes sufficiently low that polymerases begin to back up.

Extension to the ℓ-TASEP

It turns out that the dynamics of "biopolymerization"—that is, the assembly of a protein or nucleic acid polymer by polymerases moving along a common nucleic acid template—has been investigated for decades (e.g., [18]). The relevant stochastic process is known in the literature as the *totally asymmetric simple exclusion process* (TASEP) [19]. In the case of heterogeneous elongation rates and a polymerase molecule that has an extended size ℓ—meaning that the centers of successive molecules can never be less than ℓ nucleotides apart—inference under this model is thought to be analytically intractable and is typically approached using Monte Carlo methods.

Interestingly, however, Erdmann-Pham et al. have recently given an approximate closed-form solution for the steady state density of this version of the model (the inhomogenous ℓ-TASEP) that considers pairs of interacting molecules [20]. Their solution was initially developed for the case of protein synthesis (i.e., with ribosomes moving along an RNA template), but in a subsequent work, they also applied it to the case of RNA transcription [21].

It appears that the solution of Erdmann-Pham et al. could be adapted in a fairly straightforward manner for the steady-state cases above. Essentially, one could discard the continuous-time Markov model and use the equations from ref. [20] in place of those given in equation 6. For example, this alternative stationary distribution could simply be substituted into equation 8 or equation 9 to obtain a new likelihood function. Statistical inference in this framework could be accomplished by numerical optimization. Notably, the generality of Erdmann-Pham et al.'s solution would allow it to be used also in the case where the elongation rate is different at each nucleotide position, as discussed later in the article.

Still, this solution to the ℓ-TASEP may be more complex than needed here. Instead, I will consider a simple extension of the Markov model introduced above that may be adequate for most analyses of eukaryotic transcription.

The special case of steric hindrance of initiation

In eukaryotic cells, it appears that, by far, the most pronounced reduction in elongation rate—and hence, the highest probability of collision—occurs at the promoter-proximal pause site. Indeed, based on current estimates of average rates of initiation and elongation, RNAP molecules should generally be fairly sparse along the gene body. For example, with an average elongation rate of 2.3 kb/min and an average initiation rate of 2.7 RNAP molecules per cell per minute [22], RNAPs would have an average spacing of 850 bp in the gene body. The rate of collision, of course, depends on other factors, such as the variance across cells in both initiation and elongation rate, as well as the variability over time in these rates (e.g., if transcription occurs in "bursts" collisions will be more likely), but it seems likely that, at least in many cases, collisions along the gene body will be infrequent enough that they can be ignored without too much cost.

Upstream of the pause site, however, collisions will be many times more likely. In addition, there is now considerable evidence that RNAP molecules frequently pile up behind the pause site to such an extent that they begin to block new RNAP molecules from initiating transcription. Because the pause site generally occurs ∼50 bp downstream of the TSS and the "footprint" of an engaged RNAP is ∼30–40 bp on the DNA template, the geometry seems to permit only about one, perhaps sometimes two, RNAP molecules to be held in the paused position before they begin to interfere with new initiation events [23]. This phenomenon could potentially lead to an effective decrease in the initiation rate owing to pausing—what Gressel et al. [22] have called the "pause-initiation limit."

It turns out that the specific problem of a decrease in initiation owing to steric hindrance from paused RNAP molecules can be accommodated fairly easily in the framework introduced here, at least in the steady-state setting. If it is true that this is the dominant manner in which collisions impact nascent RNA sequencing data for eukaryotes, then the Markov model augmented to allow for steric hindrance may be adequate to describe many aspects of the complex interplay between initiation and pausing.

To define a simple model for steric hindrance, I will first decompose the rate of initiation, α, into a *potential* rate of initation in the absence of occlusion of the initiation site, ω, and a *conditional* probability that initiating events that would otherwise occur are blocked by an existing RNAP molecule, which I will denote ϕ. Thus, $\alpha = \omega(1-\phi)$. In turn, ϕ can be reasonably assumed to be equal to the probability that an RNAP molecule occupies any of the first ℓ nucleotide positions of the DNA template of interest in any given cell, where a "landing pad" of ℓ unoccupied nucleotides is needed for initiation. Assuming further that it is possible to obtain a reasonably good estimate of the average number of RNAP molecules per TU per cell in the sample of interest, denoted R, then, at steady state, $\phi = 1 - \exp\left(-R\sum_{i=1}^{\ell}\pi_i^{(k)}\right)$, where $\pi_i^{(k)}$ represents the stationary probability that an RNAP molecule occupies position i given that it is present on the DNA template and given a particular pause site k. Now, for $i \in \{1,\ldots,\ell\}$, $\ell < N$ (c.f. equation 6),

$$\pi_i^{(k)} = \frac{1}{\mathcal{Z}}\begin{cases}\frac{\alpha}{\beta\zeta} & i = k\\ \frac{\alpha}{\zeta} & \text{otherwise,}\end{cases} \tag{21}$$

where $\mathcal{Z} = \frac{\alpha}{\zeta}(N - 1 + \frac{1}{\beta})$ (omitting the termination state). In addition, k can be assumed to follow the

distribution defined by f_k (e.g., equation 19). Thus,

$$\begin{aligned}
\phi &= 1 - \exp\left\{ -R \sum_{i=1}^{\ell} \sum_{k=k_{\min}}^{\ell} \pi_i^{(k)} f_k \right\} \\
&= 1 - \exp\left\{ -\frac{R}{\mathcal{Z}} \left[\ell \cdot \frac{\alpha}{\zeta} + \sum_{k=k_{\min}}^{k_{\max}} f_k \left(\frac{\alpha}{\beta\zeta} - \frac{\alpha}{\zeta} \right) \right] \right\} \\
&= 1 - \exp\left\{ -\frac{R\alpha}{\mathcal{Z}\zeta} \left[\ell + \rho \left(\frac{1}{\beta} - 1 \right) \right] \right\} \\
&= 1 - \exp\left\{ -\frac{R}{N - 1 + \frac{1}{\beta}} \left[\ell + \rho \left(\frac{1}{\beta} - 1 \right) \right] \right\}, \qquad (22)
\end{aligned}$$

where $\rho = \sum_{k=k_{\min}}^{\ell} f_k$ is the fraction of cells having a pause site $k \leq \ell$ and I assume $\ell \leq k_{\max}$.

Equation 22 shows that the probability of occlusion, ϕ, increases with R, ℓ, and ρ, and will tend to decrease with β. Notice also that α, β, λ, and all derived quantities can be estimated in pre-processing under steady-state assumptions, and ϕ can be computed afterward, which then allows ω to be computed.

Allowing for Overdispersion

In general, read counts in nascent RNA sequencing data tend to be overdispersed (with variance exceeding the mean), just as in RNA-seq data. A relatively straightforward way to allow for overdispersion is to assume a mixture of Poisson rates within each region while maintaining the assumption of independence across nucleotide sites. In this way, some sites are allowed to accumulate reads at higher-than-average rates, and others at lower-than-average rates. If the mixture over rates is assumed to be Gamma-distributed, then this assumption simply implies that the site-specific Poisson distributions for read counts are replaced with negative binomial distributions.

In particular, if the read depth at each nucleotide i is assumed to be scaled by a random variable ρ_i, which is Gamma-distributed with shape parameter A and rate parameter A (so that $E[\rho_i] = 1$), and the assumption of independent Poisson read counts is maintained as in equation 9, then the likelihood function becomes,

$$\begin{aligned}
&P(\mathbf{X} \mid \chi, \beta, A) \\
&\quad = \left[\prod_{i=1}^{N} \int \text{Gamma}\left(\rho_i \mid A, A\right) \text{ Pois}\left(X_i \mid \chi\rho_i\right) d\rho_i \right] \frac{\int \text{Gamma}\left(\rho_k \mid A, A\right) \text{ Pois}\left(X_k \mid \chi\rho_k/\beta\right) d\rho_k}{\int \text{Gamma}\left(\rho_k \mid A, A\right) \text{ Pois}\left(X_k \mid \chi\rho_k\right) d\rho_k} \\
&\quad = \left[\prod_{i=1}^{N} \text{NB}\left(X_i \mid A, \frac{A}{A+\chi} \right) \right] \frac{\text{NB}\left(X_i \mid A, \frac{A}{A+\chi/\beta} \right)}{\text{NB}\left(X_i \mid A, \frac{A}{A+\chi} \right)} \\
&\quad = \left(\frac{\chi}{A+\chi} \right)^{s} \left(\frac{A}{A+\chi} \right)^{NA} \left(\frac{A+\chi}{A+\chi/\beta} \right)^{X_k+A} \beta^{-X_k} \prod_{i=1}^{N} \binom{X_i + A - 1}{A-1} \\
&\quad = \chi^s A^{NA} \left(A+\chi\right)^{X_k+A-s-NA} \left(A+\chi/\beta\right)^{-X_k-A} \beta^{-X_k} \prod_{i=1}^{N} \binom{X_i + A - 1}{A-1}, \qquad (23)
\end{aligned}$$

where $\text{NB}(X \mid r, p)$ indicates that X has a negative binomial distribution with parameters r and p, that is, such that,

$$p(X = x \mid r, p) = \binom{x + r - 1}{r - 1} (1-p)^x p^r. \qquad (24)$$

Notice that A must be a positive integer for this likelihood function to be well-defined.

In general, there are no convenient closed-form expressions for the maximum likelihood estimates of the parameters in this case, and they would need to be estimated numerically or by expectation maximization. For this reason, I will focus on the Poisson model in many of the sections below (but see Discussion).

Likelihood Ratio Tests for Differences Between Transcription Units, Conditions, or Species

These likelihood functions lead naturally to a class of likelihood ratio tests (LRTs) for comparing aspects of the transcriptional dynamics of different TUs, or of the same TU under different conditions. For example, one might test whether two genes have different initiation rates or different pause-release rates. In each case, it might be desirable to allow some parameters to vary freely, while testing for evidence in the data of a difference in another parameter. For example, one might test for a difference in the pause-release rate allowing for differences in the initiation rate, or a difference in the initiation rate allowing for differences in the pause-release rate. I will focus on tests based on steady-state data, but one could equally well conduct a test for a difference in elongation rate based on time courses representing different conditions.

These LRTs all have the same general form. Consider read counts captured at steady-state for two TUs, $\mathbf{X}^{(1)} = (X_1^{(1)}, \ldots, X_{N_1}^{(1)})$ and $\mathbf{X}^{(2)} = (X_1^{(1)}, \ldots, X_{N_2}^{(2)})$. As noted, $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ could represent the same TU under different conditions, different TUs under the same conditions, or even different TUs under different conditions. In a comparative genomic analysis, they may represent orthologous TUs in different species (say, human and chimpanzee). Let the free parameters for the two genes be denoted $\Theta_1 = (\alpha_1, \beta_1, \theta_1)$ and $\Theta_2 = (\alpha_2, \beta_2, \theta_2)$, respectively (ignoring the ζ parameters for now, and focusing on the steady-state), where θ_1 and θ_2 represent any additional parameters of interest. Finally, let the log likelihood of a particular data set be denoted $\ell(X; \Theta)$, and let the maximized value of that likelihood be denoted $\hat{\ell}(X; \Theta) = \max_\Theta \ell(X; \Theta)$.

In general, one wishes to compare a null hypothesis H_0 that the two data sets can be jointly described by (at least some) shared parameters against an alternative hypothesis H_A that they require different parameters. The parameter can be partitioned into three mutually exclusive and exhaustive categories:

- *Fully tied* parameters (Θ_{T}). These parameters are shared by both TUs under the null and alternative hypotheses.
- *Fully free* parameters (Θ_{F1} and Θ_{F2}). These parameters vary freely in the two data sets under both the null and alternative hypotheses.
- *Tested* parameters (Θ_* or Θ_{*1} and Θ_{*2}). These parameters vary freely under the alternative hypothesis but are tied under the null hypothesis.

With these assumptions, the test statistic for an LRT for a difference in the tested parameters can be defined generally as,

$$T = \hat{\ell}(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}; \Theta_{\text{T}}, \Theta_{\text{F1}}, \Theta_{\text{F2}}, \Theta_{*1}, \Theta_{*2}) - \hat{\ell}(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}; \Theta_{\text{T}}, \Theta_{\text{F1}}, \Theta_{\text{F2}}, \Theta_*). \tag{25}$$

Here, each term represents the joint likelihood of the two data sets. Generally, $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ will be assumed to be conditionally independent given the parameters, so that their joint likelihood is a product of their individual likelihoods; however, I present them in this more general form to emphasize the need to estimate tied parameters jointly. In the usual manner, the first term represents the maximized likelihood under the alternative hypothesis and the second term represents the maximized likelihood under the null hypothesis. If the tested parameters, Θ_{*1} and Θ_{*2}, are completely unconstrainted under the alternative hypothesis, T represents a two-sided test. However, they can be constrained, if desired, to achieve a one-sided test (for example, allowing only $\alpha_2 \geq \alpha_1$ to test for an increased initiation rate in the second TU).

To make this framework concrete, let us consider some specific examples of tests. First, suppose one wishes to test for a difference in initiation rate between two TUs from the same data set, assuming equal elongation rates. In this case, $\Theta_{\text{T}} = \{\lambda, \zeta\}$, $\Theta_{\text{F1}} = \{\beta_1\}$, $\Theta_{\text{F2}} = \{\beta_2\}$, and $\Theta_* = \{\alpha\}$.

Next, suppose one wishes to test for a difference in the pause-release rate between two TUs from the same data set. In this case, $\Theta_{\text{T}} = \{\lambda\}$, $\Theta_{\text{F1}} = \{\alpha_1, \zeta_1\}$, $\Theta_{\text{F2}} = \{\alpha_2, \zeta_2\}$, and $\Theta_* = \{\beta\}$.

Finally, suppose one wishes to test for a difference in initiation rate in the same TU under two conditions, represented by separately sequenced nascent RNA sequencing data sets. Here, $\Theta_{\text{T}} = \{\zeta\}$, $\Theta_{\text{F1}} = \{\beta_1, \lambda_1\}$, $\Theta_{\text{F2}} = \{\beta_2, \lambda_2\}$, and $\Theta_* = \{\alpha\}$.

In general, for two-sided tests, the asympotic distribution of $2T$ under the null hypothesis will be a chi-square distribution with $|\Theta_*|$ degrees of freedom, enabling convenient calculation of nominal *p*-values. In many cases, however, it may be preferable to use empirical *p*-values instead, for example, by computing T for a collection of TUs believed to be representative of the null hypothesis, and using this empirical distribution in place of the null.

Poisson-based Likelihood Ratio Tests

In the case of the Poisson generating distribution at steady state, it is often possible to find closed-form solutions for the test statistic T. For example, in the test for differences in initiation rate ($H_0 : \alpha_1 = \alpha_2$ vs. $H_A : \alpha_1 \neq \alpha_2$), T can be expressed as,

$$\begin{aligned} T = s_1' &\left[\log\left(\frac{s_1'}{M_1}\right) - \log\left(\frac{\lambda_1/\zeta_1(s_1' + s_2')}{\lambda_1 M_1/\zeta_1 + \lambda_2 M_2/\zeta_2}\right)\right] \\ &+ s_2' \left[\log\left(\frac{s_2'}{M_2}\right) - \log\left(\frac{\lambda_2/\zeta_2(s_1' + s_2')}{\lambda_1 M_1/\zeta_1 + \lambda_2 M_2/\zeta_2}\right)\right], \end{aligned} \tag{26}$$

where the simplified estimation method in equation 13 is assumed, and where s_i' denotes the sum of read counts and M_i denotes the length of the gene body in TU i. Thus, the LRT for a difference in α depends only on the data in the gene body, not on the pause peak.

It is sometimes convenient to re-express T in terms of weights τ_1 and τ_2, such that,

$$T = s_1' \log \frac{s_1'}{\tau_1 \left(s_1' + s_2'\right)} + s_2' \log \frac{s_2'}{\tau_2 \left(s_1' + s_2'\right)}, \tag{27}$$

where,

$$\tau_1 = \frac{\lambda_1 M_1/\zeta_1}{\lambda_1 M_1/\zeta_1 + \lambda_2 M_2/\zeta_2}, \qquad \tau_2 = 1 - \tau_1 = \frac{\lambda_2 M_2/\zeta_2}{\lambda_1 M_1/\zeta_1 + \lambda_2 M_2 \zeta_2}. \tag{28}$$

When $\lambda_1 = \lambda_2$, as when comparing the read counts for two different TUs from the same data set, when $M_1 = M_2$, as can typically be ensured by construction, and assuming $\zeta_1 = \zeta_2$, the weights are equal ($\tau_1 = \tau_2 = \frac{1}{2}$) and T reduces to an expression that depends only on s_1' and s_2':

$$T = s_1' \log \frac{s_1'}{\left(s_1' + s_2'\right)/2} + s_2' \log \frac{s_2'}{\left(s_1' + s_2'\right)/2}. \tag{29}$$

In this way, T can be seen to be a measure of discordance in read depth, normalized for differences in library size, gene-body length, and elongation rate. Notice in particular that, when the (normalized) average read depths are the same in the two gene bodies (e.g., when $s_1' = s_2'$ in equation 29), then $T = 0$ and the null hypothesis cannot be rejected. T grows larger as these average read depths become more different from one another. Importantly, however, T depends on the raw read counts, not only on their ratios. Thus, T tends to be small when the s_i' values are small, and to grow larger as they increase, reflecting greater confidence in the rejection of the null hypothesis with more data.

The test statistic for a difference in pause-release rates β_1 and β_2, with a fixed pause site k but allowing for a difference in initiation rates, is messier but can still be expressed in closed form as,

$$\begin{aligned} T = {} & s_1' \log s_1' + s_2' \log s_2' + X_k^{(1)} \log X_k^{(1)} + X_k^{(2)} \log X_k^{(2)} \\ & - \left(s_1' + s_2'\right) \log \left(s_1' + s_2'\right) - \left(s_1' + X_k^{(1)}\right) \log \left(s_1' + X_k^{(1)}\right) \\ & - \left(s_2' + X_k^{(2)}\right) \log \left(s_2' + X_k^{(2)}\right) - \left(X_k^{(1)} + X_k^{(2)}\right) \log \left(X_k^{(1)} + X_k^{(2)}\right) \\ & + \left(s_1' + s_2' + X_k^{(1)} + X_k^{(2)}\right) \log \left(s_1' + s_2' + X_k^{(1)} + X_k^{(2)}\right), \end{aligned} \tag{30}$$

where, for simplicity, I focus on the case of shared $M_1 = M_2 = M$ and further assume $k_1 = k_2 = k$ (this test is independent of λ_1, λ_2, ζ_1, and ζ_2). Here, it can be shown easily that $T = 0$ if the ratios of read counts at the pause peak and the gene body are equal, i.e., if $X_k^{(1)}/s_1' = X_k^{(2)}/s_2'$, and that T grows larger as these ratios become more divergent, but in a manner that depends not only on the ratios but also on the absolute counts. Notably, this test can be shown to be a special case of a *G*-test and asymptotically equivalent to both a chi-squared test and Fisher's exact test based on the four counts, $s_1', s_2', X_k^{(1)}$, and $X_k^{(2)}$.

The test for a difference in β becomes more complicated when allowing for uncertainty in the pause site, but this situation be accommodated relatively easily by taking advantage of the EM framework and precomputed posterior expected values of $t^{(1)}$ and $t^{(2)}$, which can be used in place of $X_k^{(1)}$ and $X_k^{(2)}$, respectively. In particular, the test statistic is exactly the same as in equation 30 but with these posterior expected values in place of observed ones,

$$\begin{aligned} T = {} & s_1' \log s_1' + s_2' \log s_2' + \langle t^{(1)} \rangle \log \langle t^{(1)} \rangle + \langle t^{(2)} \rangle \log \langle t^{(2)} \rangle \\ & - \left(s_1' + s_2'\right) \log \left(s_1' + s_2'\right) - \left(s_1' + \langle t^{(1)} \rangle\right) \log \left(s_1' + \langle t^{(1)} \rangle\right) \\ & - \left(s_2' + \langle t^{(2)} \rangle\right) \log \left(s_2' + \langle t^{(2)} \rangle\right) - \left(\langle t^{(1)} \rangle + \langle t^{(2)} \rangle\right) \log \left(\langle t^{(1)} \rangle + \langle t^{(2)} \rangle\right) \\ & + \left(s_1' + s_2' + \langle t^{(1)} \rangle + \langle t^{(2)} \rangle\right) \log \left(s_1' + s_2' + \langle t^{(1)} \rangle + \langle t^{(2)} \rangle\right), \end{aligned} \tag{31}$$

where the quantities of the form $\langle t^{(j)} \rangle$ are computed iteratively as in equation 18. Strictly speaking, however, the $\langle t^{(j)} \rangle$ values that appear in combination with other parameters, which reflect H_0 (lines 2–4 of equation 31) should be calculated using the version of the model where β is tied, and the $\langle t^{(j)} \rangle$ values that appear alone, which reflect H_A (line 1), should be calculated using the version of the model with free parameters β_1 and β_2. The EM algorithm for the tied parameters is similar to the one for free parameters (equations 16–18) but the update for the shared β on each iteration is,

$$\hat{\beta} = \frac{s_1' + s_2'}{M \left(\langle t^{(1)} \rangle + \langle t^{(2)} \rangle\right)}. \tag{32}$$

The iterative calculation of $\langle t^{(1)} \rangle$ and $\langle t^{(2)} \rangle$ is as specified in equation 18 but using the shared value of β.

Finally, the LRT for a difference in initiation rate can be extended to allow for steric hindrance. By explicitly allowing for interactions between pausing and initiation, it is possible to test whether an observed difference in initiation rates is beyond what can be explained by steric hindrance alone. In other words, this more conservative test excludes the possibility that what appears to be a difference in initiation rates is simply a consequence of different effects of the pause-initiation limit.

In this case, we assume $\alpha_1 = w_1(1 - \phi_1)$ and $\alpha_2 = w_2(1 - \phi_2)$ such that $\omega_1 = \omega_2$ under the null hypothesis and $\omega_1 \neq \omega_2$ under the alternative hypothesis. As above, β_1 and β_2 are free parameters. As it

turns out, the test statistic T is equivalent to the one given in equation 29, but with the weights τ_1 and τ_2 redefined to incorporate the terms $(1-\phi_1)$ and $(1-\phi_2)$:

$$T = s_1' \log \frac{s_1'}{\tau_1 \left(s_1' + s_2'\right)} + s_2' \log \frac{s_2'}{\tau_2 \left(s_1' + s_2'\right)}, \tag{33}$$

where,

$$\begin{aligned} \tau_1 &= \frac{\lambda_1 M_1/\zeta_1 \left(1-\phi_1\right)}{\lambda_1 M_1/\zeta_1 \left(1-\phi_1\right) + \lambda_2 M_2/\zeta_2 \left(1-\phi_2\right)}, \\ \tau_2 = 1 - \tau_1 &= \frac{\lambda_2 M_2/\zeta_2 \left(1-\phi_2\right)}{\lambda_1 M_1/\zeta_1 \left(1-\phi_1\right) + \lambda_2 M_2/\zeta_2 \left(1-\phi_2\right)}. \end{aligned} \tag{34}$$

In the case where k is unknown, the test can be executed by first fitting the model by EM under the null hypothesis, estimating ϕ_1 and ϕ_2 based on the MLEs for α, β_1, and β_2, and then substituting these values into equations 33 and 34. The EM algorithm in this case of a shared ω value has updates,

$$\begin{aligned} \omega &= \frac{s_1' + s_2'}{\lambda_1 M_1/\zeta_1 \left(1-\phi_1\right) + \lambda_2 M_2/\zeta_2 \left(1-\phi_2\right)}, \\ \beta_1 &= \frac{\tau_1 \left(s_1' + s_2'\right)}{\langle t^{(1)} \rangle}, \\ \beta_2 &= \frac{\tau_2 \left(s_1' + s_2'\right)}{\langle t^{(2)} \rangle}, \end{aligned} \tag{35}$$

where on each iteration ϕ_1 and ϕ_2 are updated using equation 22, based on the previous values of ω, β_1, and β_2, and these new ϕ_1 and ϕ_2 values are used to compute new values of τ_1 and τ_2 according to equation 34.

Bayesian Inference Under the Poisson Model

Another advantage of the Poisson model at steady-state is that it leads to relatively straightforward Bayesian inference of the key model parameters, χ, β, and γ. To simplify the mathematics, I will introduce a change of variables from β to $\nu = \frac{1}{\beta}$ and assume Gamma priors for χ and ν:

$$\begin{aligned} \chi &\sim \text{Gamma}(a, b), \\ \nu &\sim \text{Gamma}(A, B), \end{aligned} \tag{36}$$

where the first parameter of each Gamma distribution represents its shape and the second parameter represents its rate (inverse scale).

With these assumptions, and assuming k is known, the joint posterior distribution of the parameters is given, up to a normalization constant, by:

$$\begin{aligned} P(\chi, \nu \,|\, \mathbf{X}) &\propto P(\chi \,|\, a, b)\, P(\nu \,|\, A, B)\, P(\mathbf{X} \,|\, \chi, \nu) \\ &\propto \chi^{a-1} e^{-b\chi} \nu^{A-1} e^{-B\nu} \chi^{s} \nu^{X_k} \exp\left[-\chi \left(N + \nu - 1\right)\right] \\ &= \chi^{a+s-1} \nu^{A+X_k-1} \exp\left[-B\nu - \chi \left(b + N + \nu - 1\right)\right]. \end{aligned} \tag{37}$$

Because χ is entangled with ν in the last term of equation 37, the marginal posterior distributions for the individual parameters do not have standard forms. However, one can sample from the joint posterior distribution by rejection sampling or sampling-importance resampling using the unnormalized density in equation 37.

In addition, it can be shown that the conditional distribution of each parameter given the others reduces to a standard Gamma distribution:

$$P(\chi \,|\, \mathbf{X}, \nu) = \text{Gamma}(a + s, b + N + \nu - 1) \tag{38}$$

$$P(\nu \,|\, \mathbf{X}, \chi) = \text{Gamma}(A + X_k, B + \chi). \tag{39}$$

As a result, it is straightforward to implement a Gibbs sampling algorithm to sample from the joint posterior, from which marginal posterior distributions can easily be approximated. Using this framework, it would be possible to calculate Bayes factors for model comparison, analogous to the LRTs introduced above.

In the case of the steric-hindrance model, the posterior distributions of ω and ϕ can be obtained from those for χ and $\beta = \frac{1}{\nu}$ in post-processing, using equation 22.

Generalization to Site-Specific Elongation Rates

The model can easily be generalized to allow for a different elongation rate ζ_i at each nucleotide position i, instead of a constant rate across all nucleotides. In this case, the elongation rate no longer serves as a simple scaling factor for the infinitesimal generator matrix $\mathbf{Q}$, and there is potentially information in the data about site-specific elongation rates, even at steady-state.

In this case, I will ignore the effects of both pausing end termination, and focus on the gene body, where the elongation signal is easiest to interpret. Thus, with site-specific elongation rates ζ_i, the conditional stationary distribution for $i \in \{1, \ldots, N\}$ becomes simply (cf. equation 6):

$$\pi_i = \frac{1}{\mathcal{Z}} \frac{\alpha}{\zeta_i}, \tag{40}$$

with normalization constant, $\mathcal{Z} = \alpha \sum_{i=1}^{N} \frac{1}{\zeta_i}$, where I now assume that all positions $i \in \{1, \ldots, N\}$ fall in the gene body. Notice that I have maintained the parameter α here, despite that it appears only as a compound parameter together with the site-specific elongation rates and hence is nonidentifiable. The reason will become clear in the next section.

A Generalized Linear Model for Discovering Features Associated with Elongation

In practice, nascent RNA sequencing data tends to be quite noisy at individual nucleotides, and it is not likely to be useful for direct inference of each ζ_i value. However, it may be possible to learn something about the site-specific elongation process by pooling information across many sites and many TUs.

In particular, I will define a generalized linear model to make use of a collection of generic features at each transcribed nucleotide across the genome. These could include any property known or suspected to influence elongation rates, including, for example, the degree of chromatin accessibility in a cell type of interest (assayed by DNase-seq or ATAC-seq), proximity to a splice site, presence of nucleosomes (MNase-seq), presence of stem-loops or other secondary structures of RNAs, or local G+C composition. Suppose there are D such features, summarized by a vector $\mathbf{Y}_i$ at each site i. Let ζ_i be an exponentiated linear function of these features,

$$\zeta_i = e^{\boldsymbol{\kappa} \cdot \mathbf{Y}_i}, \tag{41}$$

where $\boldsymbol{\kappa}$ is a vector of D real-valued coefficients shared across all sites, and the function e^x ensures that ζ_i takes a positive value. It may be useful to define the first element of $\mathbf{Y}_i$ to be a constant of 1 to accommodate an intercept for the linear function at the corresponding position in $\boldsymbol{\kappa}$ (implying $D - 1$ true features).

In order to separate global aspects of the elongation process from TU-specific (and highly variable) rates of initiation, I will maintain a separate instance of the parameter α for each TU, but share the coefficient

vector κ across all TUs. Thus, for M (independent) TUs indexed by j, the joint log likelihood function under the Poisson model is (cf. equations 9 & 10),

$$\begin{aligned}\ell(\mathbf{X};\boldsymbol{\alpha},\boldsymbol{\kappa},\lambda,\mathbf{Y}) &= \sum_{j=1}^{M}\sum_{i=1}^{N_j} X_{j,i}\log\left(\frac{\lambda\alpha_j}{\zeta_{j,i}}\right) - \frac{\lambda\alpha_j}{\zeta_{j,i}} - \log\mathcal{Z} \\ &= \sum_{j=1}^{M}\sum_{i=1}^{N_j} X_{j,i}\log\left(\lambda\alpha_j\right) - X_{j,i}\left(\boldsymbol{\kappa}\cdot\mathbf{Y}_{j,i}\right) - \lambda\alpha_j e^{-\boldsymbol{\kappa}\cdot\mathbf{Y}_{j,i}} - \log\mathcal{Z}, \end{aligned} \tag{42}$$

where α_j is the initiation rate and N_j is the total length for TU j; and $X_{j,i}$ is the read-count, $\mathbf{Y}_{j,i}$ is the feature vector, and $\zeta_{j,i}$ is the elongation rate at the ith nucleotide in TU j;

Fitting the GLM to data

The log likelihood can be written in terms of simple summaries of the data, as follows,

$$\ell(\mathbf{X};\boldsymbol{\alpha},\boldsymbol{\kappa},\lambda,\mathbf{Y}) = \sum_{j=1}^{M}\left[s'_j\log\left(\lambda\alpha_j\right) - \boldsymbol{\kappa}\cdot\mathbf{T}_j - \lambda\alpha_j U_j\right] - \log\mathcal{Z}, \tag{43}$$

where s'_j is the now-familiar sum-of-read-counts for the gene body, but restricted here to the jth TU, and $\mathbf{T}_j$ and U_j are analogously defined as,

$$\mathbf{T}_j = \sum_{i=1}^{N_j} X_{j,i}\mathbf{Y}_{j,i}, \tag{44}$$

$$U_j = \sum_{i=1}^{N_j} e^{-\boldsymbol{\kappa}\cdot\mathbf{Y}_{j,i}}. \tag{45}$$

Notice that U_j is not strictly a sufficient statistic because it depends on the parameter $\boldsymbol{\kappa}$ as well as on the data.

This joint log likelihood can be maximized by gradient ascent. The partial derivative with respect to the nth component of $\boldsymbol{\kappa}$ is given by,

$$\frac{\partial}{\partial\kappa_n}\ell(\mathbf{X};\boldsymbol{\alpha},\boldsymbol{\kappa},\lambda,\mathbf{Y}) = \sum_{j=1}^{M}\lambda\alpha_j\mathbf{V}_{j,n} - \mathbf{T}_{j,n}, \tag{46}$$

where the final subscript n indicates the nth element of a vector, and $\mathbf{V}_j$ is defined as,

$$\mathbf{V}_j = \sum_{i=1}^{N_j} e^{-\boldsymbol{\kappa}\cdot\mathbf{Y}_{j,i}}\mathbf{Y}_{j,i}. \tag{47}$$

For a given value of $\boldsymbol{\kappa}$, the maximum for α_j can be determined analytically as,

$$\hat{\alpha}_j = \frac{s'_j}{\lambda U_j}. \tag{48}$$

Thus, a gradient ascent algorithm can be implemented to iteratively improve estimates of $\boldsymbol{\kappa}$ and, on each iteration, to fully optimize α_j conditional on the previous parameter values. Importantly, the sufficient statistics s'_j and $\mathbf{T}_j$ need only be computed once, in pre-processing, although U_j and $\mathbf{V}_j$ must be recomputed on each iteration of the algorithm.

Notably, these results are closely related to those for standard Poisson regression. As in that case, the log likelihood function is guaranteed to be convex.

Filters for genomic features

In applications to real data, the genomic features will generally require some preprocessing, because they will have different noise properties and levels of resolution along the genome. A general and flexible way to convert them to a common representation that is matched to the nascent RNA sequencing data is to apply a separate smoothing filter to each feature.

In particular, let a filter $F^{(n)}_{r,\sigma,\delta}$ for each feature $n \in \{1, \ldots, D\}$ be defined by three parameters: a *radius* of application r, a *smoothing bandwidth* σ, and an *offset* δ (which allows for a translation along the genome sequence in the relationship between the covariate and the data X_i). Each (scalar) covariate $Y^{(n)}_i$ can then be replaced with a filtered version,

$$\bar{Y}^{(n)}_i = \frac{1}{\mathcal{Z}} \sum_{k=-r}^{+r} Y^{(n)}_{i+k+\delta} F^{(n)}_{r,\sigma,\delta}(k), \qquad \text{where } \mathcal{Z} = \sum_{k=-r}^{+r} F^{(n)}_{r,\sigma,\delta}(k). \tag{49}$$

The filter $F^{(n)}_{r,\sigma,\delta}(k)$ can take a variety of forms. For example,

- Gaussian filter: $F^{(n)}_{r,\sigma,\delta}(k) = \frac{1}{\sigma} e^{-\frac{1}{2}(k/\sigma)^2}$
- Laplace filter: $F^{(n)}_{r,\sigma,\delta}(k) = \frac{1}{2\sigma} e^{-|k|/\sigma}$
- Negative exponential filter: $F^{(n)}_{r,\sigma,\delta=-r}(k) = \frac{1}{\sigma} e^{k/\sigma}$
- Positive exponential filter: $F^{(n)}_{r,\sigma,\delta=r}(k) = \frac{1}{\sigma} e^{-k/\sigma}$
- Sine filter: $F^{(n)}_{r,\sigma,\delta}(k) = \sin\left(\frac{(k+r)\pi}{r}\right)$.
- Cosine filter: $F^{(n)}_{r,\sigma,\delta}(k) = \cos\left(\frac{(k+r)\pi}{r}\right)$.
- Averaging filter: $F^{(n)}_{r,\sigma,\delta}(k) = 1$. The *identity* filter $\bar{Y}^{(n)}_i = Y^{(n)}_i$ can be obtained by the special case, $F^{(n)}_{r=0,\sigma,\delta=0}(k)$.
- Generalized filter: $F^{(n)}_{r,\sigma,\delta}(k) = \lambda_{k+r}$, which is defined by a vector of nonnegative scale factors, $\boldsymbol{\lambda} = \{\lambda_0, ..., \lambda_{2r}\}$.

Notably, these filters can also be *composed* to define a richer, more general family of filters. For example, a composition filter $G(k)_{r_1,r_2,\sigma_1,\sigma_2,\delta_1,\delta_2} = F^{(1)}_{r_1,\sigma_1,\delta_1}(k) \circ F^{(2)}_{r_2,\sigma_2,\delta_2}(k)$ can be defined such that,

$$\bar{Y}^{(n)}_i = \frac{1}{\mathcal{Z}} \sum_{k=-r_1}^{+r_1} Y^{(n)}_{i+k+\delta_1} F^{(1)}_{r_1,\sigma_1,\delta_1}(k) + \sum_{k=-r_2}^{+r_2} Y^{(n)}_{i+k+\delta_2} F^{(2)}_{r_2,\sigma_2,\delta_2}(k). \tag{50}$$

In principle, the parameters that define the filters could be estimated from the data along with the parameters of the GLM, but this approach is likely to be unwieldy and computationally expensive. Instead, it will most likely be preferable to determine them in a preprocessing step by examining each of the covariates in turn. For example, it may be possible to arrive at reasonable choices of a filter and its parameters by examining metaplots of nascent RNA sequencing readcounts centered on features of interest, such as ChIP-seq or ATAC-seq peak calls, or splice sites called from RNA-seq data.

Technicalities

Estimation of λ and interpretation of rate parameters

As noted, the parameter for sequencing depth, λ is confounded with those for TU-specific initiation and elongation rates, α_j and ζ_j at steady-state, in the sense that all three parameters appear only in the relevant likelihood functions as the product $\chi_j = \frac{\lambda\alpha_j}{\zeta_j}$. Because these parameters all contribute linearly to increases in the expected read counts, they cannot be independently estimated without some exogenous calibration, which in most cases will not exist.

The most pragmatic path forward seems to be to estimate λ in a preprocessing step, hold it fixed throughout the analysis, and interpret estimates of α_j/ζ_j relative to each other and to the meaning of λ. Notably, β_j must be interpreted in the same relative manner, since it is effectively tied to α_j and ζ_j through the structure of the continuous-time Markov model. In some cases, it may be reasonable to assume ζ_j is constant across TUs (as when comparing orthologs between closely related species), in which case relative values of α_j can be obtained. Overall, it is worth emphasizing that the general framework presented in this article should be regarded as a means for estimating *relative*, rather than absolute, rates of initiation, pause-release, and, potentially, elongation. Nevertheless, these rates can effectively be compared with each other, across TUs, and—with appropriate care, as discussed below—across experiments.

Two possible ways to estimate λ are to make use of (1) the average read counts across all TUs, or (2) the read counts at one or more designated "reference" TUs. In either case, it would be sensible to make use of the gene bodies only of confidently annotated and likely expressed TUs. If the set of such TUs is denoted V (where V is a large, inclusive set in the first case or a smaller, more restricted one in the second case), then λ can be estimated in preprocessing simply as,

$$\hat{\lambda} = \frac{\sum_{j \in V} s'_j}{\sum_{j \in V} M_j}. \tag{51}$$

(Notice that this formula will implicitly weight TUs by length; if desired, an unweighted average of averages could be used instead.) After estimation of $\hat{\lambda}$, an initiation/elongation rate ratio can be determined for each TU j as,

$$\frac{\alpha_j}{\zeta_j} = \frac{\hat{\chi}_j}{\hat{\lambda}}. \tag{52}$$

The choice of V will determine the meaning of these α_j/ζ_j ratios, and in turn, the meaning of the β_j parameters. If V is taken to include all (expressed) TUs, then α_j/ζ_j will be expressed relative to the average behavior across the genome, with $\alpha_j/\zeta_j = 1$ indicating an average initiation/elongation rate ratio, $\alpha_j/\zeta_j < 1$ indicating a lower-than-average such ratio, and $\alpha_j/\zeta_j > 1$ indicating a higher-than-average such ratio. If instead, V is a set of unusually highly or lowly expressed genes, then α_j/ζ_j must be interpreted accordingly.

It is worth noting that if one were to have access to a set V of TUs with known absolute values of α_j and ζ_j, then these TUs could serve as a calibration for the α_j/ζ_j ratios, allowing them to be be assigned absolute, rather than relative, values.

Batch effects and normalization

A major practical problem in the analysis of all transcriptomic data, including nascent RNA sequencing data, is bulk differences in the distributions of values from separate experiments that are not biologically meaningful but instead reflect differences in sample or library preparation, sequencing, or some other technical aspect of data collection. In our case, such "batch effects" could be particularly problematic in tests

of differences in initiation or pause-release rates across different conditions or species. If they are ignored, they may produce many false-positive predictions of differences.

In many cases, batch effects can be effectively addressed in a preprocessing step that eliminates systematic differences across many TUs that are likely to reflect technical artifacts, leaving differences more likely to be driven by the biological processes under study. I will focus here on quantile normalization [24], although the discussion will apply to other normalization methods as well, such as ones based on principal component analysis.

Assume that quantile normalization is applied to the gene bodies of all TUs in each data set under study, such that the original value of s'_j is replaced by a new normalized value s''_j. For now, assume that a proportional correction is made to the read counts across the TUs (including in the pause and termination peaks), so that each count $X_{j,i}$ is replaced by $X'_{j,i} = \rho_j X_{j,i}$ where $\rho_j = s''_j / s'_j$.

One possible strategy would be to apply the methods described in this paper directly to the normalized data set, without change. An issue with this approach is that the normalized read counts, in general, will no longer have integer values, and for some applications, they would either need to be rounded or a continuous distribution would need to be used in place of the Poisson (for example, a Gamma distribution).

An alternative strategy, which may be preferred in some cases, is to absorb the TU-specific rescaling induced by normalization into the λ constant. In particular, let each TU have its own read-depth parameter λ_j and set $\lambda_j = \hat{\lambda}/\rho_j$, where $\hat{\lambda}$ is the original (global) estimate of λ. In this way, the analysis can still make use of the original raw data $\mathbf{X}$, rather than the normalized data $\mathbf{X}'$, but the inferred rate parameter will effectively reflect the normalization transformation.

In tests of differences in the pause-release rate β, a second normalization may be needed. There are some indications that the degree to which RNAP accumulates in the promoter region may differ from one experiment to another, perhaps owing to differences in the concentration of factors that contribute to pause release. As a result, the data may display batch effects in pause peaks beyond those that can be explained by differences in the gene bodies. This problem could be addressed by a second round of normalization applied to the X_k values after having normalized for differences in s', as described above. As in the first case, this normalization could also be accommodated by altering the λ values per TU, but it would require the introduction of a second λ_j that applied only to the pause peaks.

Replicates

In many applications, it is desirable to perform multiple replicates (biological or technical) of each individual experiment, to mitigate the noisiness of the sequence data and increase confidence in parameter estimates or hypothesis tests. Replicates can easily be accommodated in the model by working with a joint likelihood function that assumes independence across replicates but shares parameters that represent an underlying biological "truth" expected to be the same in each replicate. Typically, the key rate parameters—α, β, and ζ—will be shared, whereas technical parameters—such as λ—will be separate for each replicate.

With R replicates per time point, the likelihood function for the nonequilibrium case (cf. equation 4) generalizes to,

$$P(\mathbf{X}^{(t_1)}, \ldots, \mathbf{X}^{(t_M)} \,|\, \alpha, \beta, \zeta, \theta) = \prod_{j=1}^{M} \prod_{r=1}^{R} \prod_{i=1}^{N} \psi(X_{i,r}^{(j)} \,|\, \mu_i^{(j)}, \theta), \tag{53}$$

where $X_{i,r}^{(j)}$ represents the read count for nucleotide i and time-point j as measured in replicate r and $\mu_i^{(j)} = \lambda_r^{(j)} P(Z_i \,|\, t_j, \alpha, \beta, \zeta)$ represents a shared "truth" across replicates for site i and time-point j that is separately normalized for each replicate r using read-depth parameter $\lambda_r^{(j)}$.

In the case of the Poisson model at steady-state, the likelihood function reduces to (cf. equation 9),

$$P(\mathbf{X}|\alpha,\beta,\boldsymbol{\lambda}) = \prod_{r=1}^{R} \frac{1}{\mathcal{Z}} \left(\frac{\lambda_r \alpha}{\zeta}\right)^{s_r} \beta^{-X_{k,r}} \exp\left[-\frac{\lambda_r \alpha}{\zeta}\left(N + \frac{1}{\beta} - 1\right)\right]$$
$$= \frac{1}{\mathcal{Z}'} \left(\frac{\alpha}{\zeta}\right)^{S} \beta^{-T} \exp\left[-\frac{\lambda' \alpha}{\zeta}\left(N + \frac{1}{\beta} - 1\right)\right], \tag{54}$$

where $S = \sum_r s_r$, $T = \sum_r X_{k,r}$, $\lambda' = \sum_r \lambda_r$, and where we absorb the terms that depend only on λ_r in $\mathcal{Z}'$, assuming that the λ_r values will be pre-estimated. Thus, the likelihood function can be expressed in terms of sufficient statistics that are simply sums over replicates of the original sufficient statistics. Notice that this function would be identical if the read counts were simply pooled across replicates. Predictably, it leads to maximum-likelihood estimators of,

$$\hat{\chi} = \frac{S - T}{(N-1)}, \tag{55}$$

$$\hat{\beta} = \frac{S - T}{T(N-1)}, \tag{56}$$

where, in this case, $\chi = \frac{\lambda' \alpha}{\zeta}$.

The likelihood ratio tests and generalized linear model under the Poisson model can similarly be expressed in terms of these new aggregate sufficient statistics.

Discussion

In this article, I have outlined a new probabilistic modeling framework for the analysis of nascent RNA sequencing data. This framework allows for a unified treatment of many problems of interest in the study of eukaryotic transcription, including estimation of the rates of initiation, elongation, and pause-release. It leads naturally to likelihood ratio tests for differences in rates, and it has extensions to Bayesian inference and a generalized linear model for evaluating covariates of elongation rate. It can also be adapted to capture the phenomenon of steric hindrance of initiation from paused polymerases at steady state. In addition, several issues of practical importance in the analysis of real data—such as accomodating differences in library size, batch effects, and experimental replicates—can be readily addressed in this framework.

A number of machine-learning and statistical methods already exist for analyzing nascent RNA sequencing data [6, 8, 12, 13], so it is reasonable to ask what is gained by addressing these problems using the generative probabilistic model described here. In my view, this generative model has several key advantages. First, the model does allow estimation of some quantities that have been out of reach of most previous methods, such as the relative rates of pause-release and initiation/elongation. These quantities have been indirectly characterized by more heuristic methods—as when pausing is quantified using the pause index—but it has been difficult to know precisely how to interpret these measures. Second, the model can be fitted to the raw read counts in a completely unsupervised manner, with no need for labeled training data. This property is important because, while high-quality, independent training data is available for some applications (e.g., [8]), no such data is available for others, including estimation of pause-release or elongation rates, or tests for differences in rates. Third, because the model allows several distinct problems to be addressed in a unified, coherent manner, it allows various parameter estimates and hypothesis tests to be interpreted jointly, without the concern that separate estimates might reflect different biases or different modeling assumptions—as when, say, applying one machine-learning method to estimate elongation rates and a second, completely separate method to characterize promoter-proximal pausing or correlates of elongation rate. Finally, by directly describing the actual process by which RNAP moves along the DNA template, the

model potentially leads to new insights into this process, beyond simply providing predictions of quantities of interest, as a discriminative machine-learning method would. Altogether, the generative model both opens up new applications and provides a new perspective on the fundamental process of eukaryotic transcription.

The model does require a number of simplifying assumptions, some of which may not hold in all settings. For example, it assumes that RNAPs that initiate transcription gradually make their way all the way along the DNA template to the termination site, with negligible rates of premature termination. The frequency with which premature termination occurs is debated in the literature, with some studies suggesting it occurs at low rates across most genes [16] and others indicating higher rates [25, 26]. My hope is that this phenomenon occurs at low enough rates across most of the TU (with the possible exception of locations near the promoter) that it will have at most a minor impact on the rate estimators and hypothesis tests described here, but it could result in appreciable biases. The challenge in this setting is that, if premature termination were added to the model, its rate would be confounded with the initiation and elongation rates. I do not see a way of accommodating it in this framework except by conditioning on independent estimates of its rate.

As noted throughout the article, another major assumption is that each RNAP molecule moves independently along the DNA template, without interference from other molecules. This "no collisions" assumption should be reasonable in the regime where the density of RNAPs per TU per cell is low, as it appears to be on average. However, this assumption is likely violated frequently upstream of the pause peak, especially when the intitiation rate is high and/or the pause release rate is low. I have shown that the special case of collisions upstream of the pause site can be addressed at steady-state, within the framework of the Markov model, by taking advantage of the limited space for RNAP molecules in this region and assuming that the primary consequence of collisions is an effective decrease in the initiation rate. That is, I assume, to a first approximation, that either an RNAP is or is not present upstream of the pause site, and if it is, it will block initiation by new RNAPs. I have argued that, if indeed most collisions occur in this small local region, this adapted Markov model may be adequate for many purposes. But it is still unclear how frequently collisions occur downstream of the pause site, and more work will be needed to explore this case. In particular, high levels of temporal variation in the initiation rate (transcription "bursts", e.g., [27]) could conceivably produce many more collisions than average rates would suggest. I have argued that the steady-state solution for the ℓ-TASEP recently introduced by Erdmann-Pham et al. [20] could be used in place of the stationary probabilities from the Markov model to accommodate collisions in a more general way. Still, as I understand it, this model considers pairwise interactions between RNAP molecules but ignores higher-order interactions, so Monte Carlo methods or other approximations would presumably be needed for a fully general treatment of the issue.

The assumption of a Poisson distribution for read counts leads to a great deal of mathematical simplicity, and I have therefore relied heavily on this version of the model. I see three major points in favor of the Poisson model. First, it leads to closed-form expressions that are easy to interpet, and help to develop intuition for how the data drive parameter estimates and hypothesis tests. Second, it provides a convenient and rapid means for performing many computations of interest, without the need for complicated, slow, and error-prone software for numerical optimization. Third, it is likely that, even when the Poisson assumption is unrealistic (as when read counts are clearly overdispersed), it will nevertheless have limited impact on many of the estimators or tests of interest, particularly if care is taken in post-processing and interpretation. For example, in LRTs, one Poisson model is compared with another, and if log likelihood ratios are interpreted with respect to an empirical null distribution, the effects of overdispersion should be considerably mitigated. On the other hand, if it does prove necessary to discard the Poisson model, the negative binomial model is still relatively convenient to work with, despite requiring numerical optimization. Notably, the fully general framework would potentially allow the use of distributions for read counts that allowed not only for overdispersion but also for autocorrelation along the genome sequence, such as a Gaussian process model or hidden Markov model. More work will be needed to determine whether the use of more complex and computationally intensive generating distributions is justified.

It is worth noting that the analysis of real nascent RNA sequencing data requires addressing a number of practical issues that have been glossed over in this article. For example, the model addresses uncertainty in the pause site in a reasonably general way, but it ignores uncertainty in the TSS, which can lead to a "smear" of read counts at the $5'$ end of TUs and/or to a mixture arising from alternative TSSs at considerable distance from one another. We have recently introduced a program, called DENR [28], that attempts to deconvolve the contributions of multiple overlapping pre-RNA isoforms, and it could be used to preprocess the data. However, even with DENR, some filtering of the data is typically still necessary to ensure a clean signal. I have also largely avoided the issue of the termination site in this article, because we find that it is quite difficult to pinpoint precisely in real data, owing to transcriptional run-on and variability across cells. For many purposes it may be adequate to omit the signal near the $3'$ end of the TU, as I have assumed, but in others it may be desirable to model this portion of the gene. In general, different inference problems will require different strategies for preprocessing and filtering.

Overall, I believe that the combination of a continuous-time Markov model and a generalized generating distribution for read counts, as outined in this article, provides a flexible and powerful modeling framework for nascent RNA sequencing data. The version of the model described here does make use of a number of simplifying assumptions, some of which may need to be relaxed over time. Nevertheless, the basic framework described here should, if nothing else, be useful as a starting point for further model and algorithm development.

Acknowledgments

I thank Charles Danko, Yifei Huang, Yixin Zhao, Lingjie Liu, and Noah Dukler for helpful discussions, and Yixin Zhao for help with figure preparation.

References

[1] Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science. 2008;322:1845–1848.

[2] Churchman LS, Weissman JS. Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature. 2011;469(7330):368–373.

[3] Kwak H, Fuda NJ, Core LJ, Lis JT. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science. 2013;339(6122):950–953.

[4] Mayer A, di Iulio J, Maleri S, Eser U, Vierstra J, Reynolds A, et al. Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. Cell. 2015;161(3):541–554.

[5] Mahat DB, Kwak H, Booth GT, Jonkers IH, Danko CG, Patel RK, et al. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nat Protoc. 2016;11(8):1455–1476.

[6] Danko CG, Hah N, Luo X, Martins AL, Core L, Lis JT, et al. Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. Mol Cell. 2013;50(2):212–222.

[7] Core LJ, Martins AL, Danko CG, Waters C, Siepel A, Lis JT. Analysis of nascent RNA identifies a unified architecture of transcription initiation regions at mammalian promoters and enhancers. Nat Genet. 2014;46(12):1311–1320.

[8] Danko CG, Hyland SL, Core LJ, Martins AL, Waters CT, Lee HW, et al. Identification of active transcriptional regulatory elements from GRO-seq data. Nat Methods. 2015;12(5):433–438.

[9] Dukler N, Booth GT, Huang YF, Tippens N, Waters CT, Danko CG, et al. Nascent RNA sequencing reveals a dynamic global transcriptional response at genes and enhancers to the natural medicinal compound celastrol. Genome Res. 2017;27(11):1816–1829.

[10] Blumberg A, Zhao Y, Huang YF, Dukler N, Rice EJ, Krumholz K, et al. Characterizing RNA stability genome-wide through combined analysis of PRO-seq and RNA-seq data. BioRxiv. 2019; p. 690644.

[11] Jonkers I, Kwak H, Lis JT. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. Elife. 2014;3:e02407.

[12] Azofeifa JG, Allen MA, Lladser ME, Dowell RD. An Annotation Agnostic Algorithm for Detecting Nascent RNA Transcripts in GRO-Seq. IEEE/ACM Trans Comput Biol Bioinform. 2017;14(5):1070–1081.

[13] Anderson WD, Duarte FM, Civelek M, Guertin MJ. Defining data-driven primary transcript annotations with primaryTranscriptAnnotation in R. Bioinformatics. 2020;36(9):2926–2928.

[14] wa Maina C, Honkela A, Matarese F, Grote K, Stunnenberg HG, Reid G, et al. Inference of RNA polymerase II transcription dynamics from chromatin immunoprecipitation time course data. PLoS Comput Biol. 2014;10(5):e1003598.

[15] Azofeifa JG, Dowell RD. A generative model for the behavior of RNA polymerase. Bioinformatics. 2017;33(2):227–234.

[16] Elrod ND, Henriques T, Huang KL, Tatomer DC, Wilusz JE, Wagner EJ, et al. The Integrator Complex Attenuates Promoter-Proximal Transcription at Protein-Coding Genes. Mol Cell. 2019;76(5):738–752.

[17] Karlin S, Taylor HM. A First Course in Stochastic Processes. 2nd ed. Academic Press; 1975.

[18] MacDonald CT, Gibbs JH, Pipkin AC. Kinetics of biopolymerization on nucleic acid templates. Biopolymers: Original Research on Biomolecules. 1968;6(1):1–25.

[19] Zia R, Dong J, Schmittmann B. Modeling translation in protein synthesis with TASEP: A tutorial and recent developments. Journal of Statistical Physics. 2011;144(2):405–428.

[20] Erdmann-Pham DD, Dao Duc K, Song YS. The Key Parameters that Govern Translation Efficiency. Cell Syst. 2020;10(2):183–192.

[21] Fischer J, Song YS, Yosef N, di Iulio J, Churchman LS, Choder M. The yeast exoribonuclease Xrn1 and associated factors modulate RNA polymerase II processivity in $5'$ and $3'$ gene regions. J Biol Chem. 2020;295(33):11435–11454.

[22] Gressel S, Schwalb B, Decker TM, Qin W, Leonhardt H, Eick D, et al. CDK9-dependent RNA polymerase II pausing controls transcription initiation. Elife. 2017;6.

[23] Ehrensberger AH, Kelly GP, Svejstrup JQ. Mechanistic interpretation of promoter-proximal peaks and RNAPII density maps. Cell. 2013;154(4):713–715.

[24] Amaratunga D, Cabrera J. Analysis of Data From Viral DNA Microchips. Journal of the American Statistical Association. 2001;96(456):1161–1170.

[25] Kamieniarz-Gdula K, Proudfoot NJ. Transcriptional Control by Premature Termination: A Forgotten Mechanism. Trends Genet. 2019;35(8):553–564.

[26] Krebs AR, Imanci D, Hoerner L, Gaidatzis D, Burger L, Schubeler D. Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. Mol Cell. 2017;67(3):411–422.

[27] Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S. Stochastic mRNA synthesis in mammalian cells. PLoS Biol. 2006;4(10):e309.

[28] Zhao Y, Dukler N, Barshad G, Toneyan S, Danko CG, Siepel A. Deconvolution of Expression for Nascent RNA sequencing data (DENR) highlights pre-RNA isoform diversity in human cells. Bioinformatics. 2021;.