Stochastic analysis and inference of a two-state genetic promoter model

Abhyudai Singh, Cesar A. Vargas and Rajesh Karmakar

Abstract—Transcription is the process by which messenger RNA (mRNA) transcripts are synthesized from genes. Measurements in individual living cells reveal fluctuations in mRNA copy numbers over time suggesting that transcription is an intrinsically random process driven by thermal molecular motion of biochemical species. We here use a stochastic model of the transcription process that captures both the extent and timescale of fluctuations in mRNA population counts. In particular, randomness in the transcription process is captured through a two-state model, where the promoter of a gene stochastically switches between an active and inactive state. High levels of transcription occur from the active state, while the inactive state allows for a low basal rate of transcription.

For the two-state model we derive exact analytical formulas for the steady-state mRNA probability distribution and the mRNA auto-correlation function. These results are applied to recent data from the Human Immunodeficiency Virus (HIV) system. Using Akaike Information Criterion (AIC) we select the most likely stochastic model for the transcription process given mRNA histogram data. For the selected model, maximum likelihood estimates of the different kinetic rates associated with the viral promoter are inferred. Analysis reveals that the viral promoter resides mostly in the inactive state and there is a 100-fold difference in the rate of mRNA synthesis from the active and inactive state. In summary, formulas presented here are an important resource for reverse engineering genetic promoters from single-cell mRNA copy number data.

I. INTRODUCTION

Genetically identical cells exposed to the same environment exhibit significant stochastic variability in the levels of mRNAs and protein expressed from a given gene [1]–[4]. Given that such stochastic variability plays important functional roles within cellular processes [5]–[7], much work has focused on understanding the origins of stochasticity in the gene-expression process using both theoretical and experimental techniques. Here, we build systems-level mathematical models of the transcription process for capturing stochasticity in mRNA copy numbers. Particular focus is on model selection and inference using experimentally obtained mRNA copy number distributions.

Perhaps the simplest stochastic model of the transcription process is one where individual mRNA transcripts are created at exponential distributed time intervals (Poisson process) and each mRNA lives for an exponentially distributed time. In this case, the mRNA distribution is predicted to be a Poisson distribution [8]. However, experimentally obtained mRNA histograms show considerable deviations from a Poisson distribution indicating that more complex models are needed to explain the data [9]–[12]. A popular model used for capturing stochasticity in the transcription process is a two-state promoter model where the promoter of a gene randomly fluctuates between an active and inactive state (see Fig. 1; [13]–[17]). For the two-state model we derive an exact analytical expression for the steady-state mRNA probability distribution and auto-correlation function.

The above results are used to investigate the promoter within HIV. Previous results have shown that stochastic expression of viral mRNAs from this promoter during human cell infection drives the virus into latency, a drug resistant state of HIV [18]–[21]. Thus, this is an important system to build models that can capture gene-expression variability and reveal new insight about the promoter. Recent experiments measuring mRNAs expressed from the viral promoter inside individual cells [22] show considerable variation in mRNA copy numbers across a cell population, in spite of the fact that these cells are genetically identical, share a common environment, and have similar shapes/sizes. Combining analytically predicted mRNA distributions with histogram data, different kinetic parameters are estimated for the HIV promoter. Finally, using parameters inferred for the HIV promoter we predict the mRNA auto-correlation function, which could be verified with additional experiments.

The paper is organized as follows. In Section II, we introduce the two-state promoter model and derive an analytical expression for the mRNA probability distribution. Procedure for inferring model parameters from data and its application to HIV is presented in Section III. Auto-correlation function of mRNA copy numbers is derived in Section IV. Finally, our conclusions are presented in Section V.

II. TWO-STATE PROMOTER MODEL

To model the transcription process we consider a two-state promoter architecture where the promoter fluctuates between an active and inactive state with rates $k_{on}$ and $k_{off}$ (see Fig. 1). mRNA transcripts are synthesized from the active state at a rate $k_m$. We assume transcription occurs at a low basal rate $k_b$ from the inactive state. Finally, synthesized mRNAs degrade at a rate $\gamma_m$.

We represent the state of the promoter at time $t$ by $g(t)$, with $g(t) = 1$ ($g(t) = 0$) indicating that the promoter is in the active (inactive) state. Let $m(t)$ denote the mRNA copy number inside the cell at time $t$. Then, the two-state promoter model can be described by a set of discrete events listed in Table I. As in the stochastic formulation of chemical kinetics [23]–[25], events “fire” at exponentially distributed time intervals, and whenever a specific event
Promoter active

$k_{an} \rightarrow k_{off}$

Promoter inactive

$\gamma \rightarrow \emptyset$

Figure 1. Schematic of a two-state model where a promoter stochastically transitions between an active and inactive state. mRNAs are synthesized from the active (inactive) state at a rate $k_{an}$ ($k_{b}$). Since transcription from the inactive state is inefficient, we assume $k_{b} < k_{an}$.

Table I: Frequency and reset maps for different stochastic events in the two-state promoter model

<table>
<thead>
<tr>
<th>Event description</th>
<th>Reset in population count</th>
<th>Propensity function $f(g,m)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Promoter transition to ON state</td>
<td>$g(t) \rightarrow g(t) + 1$</td>
<td>$k_{an}(1 - g(t))$</td>
</tr>
<tr>
<td>Promoter transition to OFF state</td>
<td>$g(t) \rightarrow g(t) - 1$</td>
<td>$k_{off}g(t)$</td>
</tr>
<tr>
<td>mRNA transcription from ON state</td>
<td>$m(t) \rightarrow m(t) + 1$</td>
<td>$k_{on}g(t)$</td>
</tr>
<tr>
<td>mRNA transcription from OFF state</td>
<td>$m(t) \rightarrow m(t) + 1$</td>
<td>$k_{b}(1 - g(t))$</td>
</tr>
<tr>
<td>mRNA degradation</td>
<td>$m(t) \rightarrow m(t) - 1$</td>
<td>$\gamma_{m}m(t)$</td>
</tr>
</tbody>
</table>

occurs the state of the system is updated accordingly to a reset map (see second column in Table I). The third column lists the propensity functions $f(g,m)$, which determine how often an event occurs. More specifically, the probability that an event will occur in the next infinitesimal time interval $(t,t + dt)$ is given by $f(g,m)dt$. In summary, probabilistic events described in Table I define a stochastic model where bursts of mRNAs are produced as the promoter transitions from the inactive to the active state, and then back to the inactive state.

The two-state promoter model described above has been studied in great detail in previous works [9], [14], [15]. However, analysis of this model is mostly restricted to the case where $k_{b} = 0$. Here, we derive an exact analytical expression for the steady-state mRNA probability for the full two-state model, where transcription even occurs when the promoter resides in the inactive state. Let $p_{1}(m,t)$ ($p_{0}(m,t)$) be the probability that at time $t$, the gene is in the active (inactive) state with $m$ number of mRNA molecules. Then, these probabilities evolve according to the following Chemical Master Equation (CME) [24], [25]:

$$
\frac{\partial p_{0}(m,t)}{\partial t} = k_{off} p_{1}(m,t) - k_{on} p_{0}(m,t) + k_{b}[p_{0}(m - 1, t) - p_{0}(m,t)] + \gamma_{m}[m + 1]p_{0}(m + 1,t) - m p_{0}(m,t),
$$

$$
\frac{\partial p_{1}(m,t)}{\partial t} = k_{on} p_{0}(m,t) - k_{off} p_{1}(m,t) + k_{b}[p_{1}(m - 1, t) - p_{1}(m,t)] + \gamma_{m}[m + 1]p_{1}(m + 1,t) - m p_{1}(m,t).
$$

Note that

$$p(m,t) = p_{0}(m,t) + p_{1}(m,t)
$$

is the probability of observing $m$ mRNA transcripts at time $t$. The standard approach for solving the CME is to convert it into a partial differential equation on the probability generating function [26], [27]. Generating functions corresponding to $p_{0}(m,t)$, $p_{1}(m,t)$ and $p(m,t)$ are defined as

$$F_{0}(z,t) := \sum_{m=0}^{\infty} z^{m} p_{0}(m,t),
$$

$$F_{1}(z,t) := \sum_{m=0}^{\infty} z^{m} p_{1}(m,t),
$$

respectively. The CME (1) can be written in terms of the generating functions as follows

$$
\frac{\partial F_{0}(z,t)}{\partial t} = k_{off} F_{1}(z,t) - k_{on} F_{0}(z,t) + k_{b}(z - 1) F_{0}(z,t)
+ \gamma_{m}(1 - z) \frac{d F_{0}(z,t)}{d z},
$$

$$
\frac{\partial F_{1}(z,t)}{\partial t} = k_{on} F_{0}(z,t) - k_{off} F_{1}(z,t) + k_{m}(z - 1) F_{1}(z,t)
+ \gamma_{m}(1 - z) \frac{d F_{1}(z,t)}{d z}.
$$

At steady-state

$$0 = k_{off} F_{1}(z) - k_{on} F_{0}(z) + k_{b}(z - 1) F_{0}(z)
+ \gamma_{m}(1 - z) \frac{d F_{0}(z)}{d z},
$$

$$0 = k_{on} F_{0}(z) - k_{off} F_{1}(z) + k_{m}(z - 1) F_{1}(z)
+ \gamma_{m}(1 - z) \frac{d F_{1}(z)}{d z}.$$
where
\[ \mathcal{F}_0(z) := \lim_{t \to \infty} F_0(z,t), \quad \mathcal{F}_1(z) := \lim_{t \to \infty} F_1(z,t). \] (6)

Adding equations (5a) and (5b) we obtain
\[ k_m \mathcal{F}_1(z) + k_b \mathcal{F}_0(z) = \gamma_m \frac{d\mathcal{F}(z)}{dz}, \quad \mathcal{F}(z) := \lim_{t \to \infty} F(z,t). \] (7)

Solving equations (3b) and (7) yields
\[ \mathcal{F}_1(z) = \frac{\gamma_m}{k_m - k_b} \frac{d\mathcal{F}(z)}{dz} - \frac{k_b \mathcal{F}(z)}{k_m - k_b} \] (8a)
and
\[ \mathcal{F}_0(z) = \frac{k_m \mathcal{F}(z)}{k_m - k_b} - \frac{\gamma_m}{k_m - k_b} \frac{d\mathcal{F}(z)}{dz}. \] (8b)

Using (8), equation (5a) can be written as
\[ (z - 1) \frac{d^2 \mathcal{F}(z)}{dz^2} + (K - (\hat{k}_m + \hat{k}_b) z) \frac{d\mathcal{F}(z)}{dz} + (\hat{k}_b \hat{k}_m z - K) \mathcal{F}(z) = 0, \] (9)
where
\[ K = \hat{k}_on + \hat{k}_off + \hat{k}_m + \hat{k}_b \] (10a)
and
\[ \hat{k}_on = k_{on}/\gamma_m, \hat{k}_off = k_{off}/\gamma_m, \hat{k}_m = k_m/\gamma_m, \hat{k}_b = k_b/\gamma_m \] (11)
are model parameters normalized by the mRNA degradation rate. The exact solution of equation (9) is given by
\[ \mathcal{F}(z) = Ce^{\hat{k}_b z} z_{1F_{1}(\hat{k}_on + \hat{k}_off; (z - 1)(\hat{k}_m - \hat{k}_b))} \] (12)
where \( 1F_1(a;b;z) \) is the confluent hypergeometric function of the first kind and \( C = e^{\hat{k}_b} \) is a normalization constant determined from the condition \( \mathcal{F}(1) = 1 \).

Let \( p(m) \) be the steady-state mRNA probability distribution function, then
\[ p(m) = \frac{d^n \mathcal{F}(z)}{dz^n} \bigg|_{z = 0}. \] (13)

Differentiating equation (12) \( m \) times with respect to \( z \) at \( z = 0 \) and then comparing both sides one obtains
\[ p(m) = e^{-\hat{k}_b} \sum_{i=0}^{m} \binom{m}{i} \frac{\hat{k}_m^{m-i} (\hat{k}_m - \hat{k}_b)^i}{\Gamma(\hat{k}_on+i) \Gamma(\hat{k}_on+\hat{k}_off+i)} \frac{\hat{k}_m^{m-i} (\hat{k}_m - \hat{k}_b)^i}{\Gamma(\hat{k}_on+\hat{k}_off+i) \Gamma(\hat{k}_on+\hat{k}_off+i)} \] (14)
which is characterized by the four normalized model parameters: \( \hat{k}_m, \hat{k}_b, \hat{k}_on, \hat{k}_off \). Plots of \( p(m) \) for different parameter values are shown in Figure 2.

There are two cases under which (14) reduces to a simpler form. Firstly, if \( \hat{k}_b = 0 \) (i.e., no transcription from the inactive state), then
\[ p(m) = \frac{\hat{k}_m^{m} \Gamma(\hat{k}_on+m+1) \Gamma(\hat{k}_on+\hat{k}_off)}{\Gamma(\hat{k}_on) \Gamma(\hat{k}_on+\hat{k}_off+m+1)} \] (15)

Secondly, if \( \hat{k}_b = 0 \) and \( \hat{k}_off \gg 1 \) (i.e., the active promoter state is unstable) then we obtain the following negative binomial distribution
\[ p(m) = \left( 1 + \frac{\hat{k}_m}{\hat{k}_off} \right) \frac{\Gamma(\hat{k}_on+m)}{\Gamma(\hat{k}_on) \Gamma(m+1)} \left( 1 + \frac{\hat{k}_m}{\hat{k}_off} \right)^{-\hat{k}_on} \] (16)
characterized by two parameters \( \hat{k}_on \) and \( \hat{k}_m/\hat{k}_off \) [9].

**III. REVERSE ENGINEERING THE HIV PROMOTER**

Measurements of mRNAs transcribed from the HIV promoter show considerable variation in copy numbers across genetically identical cell populations (see mRNA histogram data in Figure 3 taken from [22]). Given data, our goal is to find the most likely underlying stochastic model. It is important to point out that the mRNA population count data shown in Figure 3 has a Fano factor (variance divided by mean) close to 100, and hence, has a significant deviation from a Poisson distribution, for which Fano factor = 1.

We consider three candidate mRNA distributions that arise from different underlying stochastic models:

1. Distribution associated with the full two-state model characterized by four parameters (Eq. 14).
2. Distribution associated with the two-state model where \( \hat{k}_b = 0 \) and characterized by three parameters (Eq. 15).
3. Negative binomial characterized by two parameters (Eq. 16).

Given measurements \( m_1, m_2, ..., m_N \) of mRNA population counts from \( N \) cells, parameters associated with each distribution can be estimated by maximizing the likelihood.
Table II: Fitting different mRNA probability distributions to data

<table>
<thead>
<tr>
<th>Model description</th>
<th>Probability distribution</th>
<th>Number of parameters $k$</th>
<th>$\Delta$AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-state promoter model</td>
<td>Eq. (14)</td>
<td>4 ($\hat{k}<em>m$, $\hat{k}<em>b$, $\hat{k}</em>{on}$, $\hat{k}</em>{off}$)</td>
<td>0</td>
</tr>
<tr>
<td>Two-state promoter model with $\hat{k}_b=0$</td>
<td>Eq. (15)</td>
<td>3 ($\hat{k}<em>{on}$, $\hat{k}</em>{on}$, $\hat{k}_{off}$)</td>
<td>9</td>
</tr>
<tr>
<td>Two-state promoter model with $\hat{k}<em>b=0$ and $\hat{k}</em>{off} \gg 1$</td>
<td>Eq. (16)</td>
<td>2 ($\hat{k}<em>{on}$/$\hat{k}</em>{off}$, $\hat{k}_{on}$)</td>
<td>7.2</td>
</tr>
</tbody>
</table>

To conclude, model inference using AIC selects the full two-state promoter model as the most likely stochastic model for the HIV promoter. The maximum-likelihood parameter estimates for this model are as follows:

$$\hat{k}_{on} = 0.9, \hat{k}_{off} = 5.75, \hat{k}_m = 759.4, \hat{k}_b = 6.5.$$  \hspace{1cm} (19)

A plot of the mRNA probability distribution (14) for these parameter values together with the histogram data is shown in Figure 3.

IV. mRNA AUTO-CORRELATION TIME

We next focus on the time scale of mRNA fluctuations and derive an analytical formula for the steady-state auto-correlation function of the stochastic process $m(t)$. To determine the auto-correlation function we use the fact that

$$\langle m(t+s) | m(s) \rangle = \langle m(s) \langle m(t+s) | m(s) \rangle | m(s), g(s) \rangle \rangle.$$ \hspace{1cm} (20)

where $\langle \cdot \rangle$ denotes the expected value and

$$\langle m(t+s) | m(s) \rangle$$ \hspace{1cm} (21)

is the expected number of mRNA transcripts at time $t+s$ given $m(s)$ and $g(s)$. To compute the conditional expectation (21) we derive the linear system of differential equations that describe the time evolution of the statistical moments $\langle g(t) \rangle$, $\langle m(t) \rangle$, $\langle g^2(t) \rangle$, $\langle m^2(t) \rangle$ and $\langle g(t)m(t) \rangle$.

For the two-state promoter model described in Table I, the time derivate of the expected value of any differential function $\varphi(g,m)$ is given by

$$\frac{d\langle \varphi(g,m) \rangle}{dt} = \sum_{\text{Events}} \Delta \varphi(g,m) f(g,m),$$ \hspace{1cm} (22)

where $\Delta \varphi(g,m)$ is the change in $\varphi(g,m)$ when an event occurs and $f(g,m)$ is the event propensity function [30], [31]. For appropriate choices of $\varphi(g,m)$ we obtain the following...
\[
R(t) := \frac{(m(t+s)m(s)) - \langle m \rangle^2}{\langle m^2 \rangle - \langle m \rangle^2} = \exp(-\gamma_m t) + \frac{(k_m^2 - k_b^2) \gamma_m (\exp(-k_{on}t - k_{off}t) - \exp(-\gamma_m t))}{(k_{off} + k_{on} - \gamma_m) \left( \frac{(k_{off} + k_{on})(\gamma_m + k_{on})(k_{off} + k_{on})}{k_{off}k_{on}} + k_m^2 - k_b^2 \right)}
\]  

(28a)

moment dynamics:
\[
\begin{align*}
\frac{d\langle g \rangle}{dt} &= k_{on} - (k_{on} + k_{off})\langle g \rangle \\
\frac{d\langle m \rangle}{dt} &= k_m\langle g \rangle + k_b(1 - \langle g \rangle) - \gamma_m\langle m \rangle \\
\frac{d\langle g^2 \rangle}{dt} &= k_{on} + k_{off}\langle g \rangle + k_m\langle g \rangle - 2k_{off}\langle g^2 \rangle - 2k_{on}\langle g \rangle \\
\frac{d\langle m^2 \rangle}{dt} &= k_m\langle g \rangle + k_b(1 - \langle g \rangle) + 2k_b\langle m \rangle + \gamma_m\langle m \rangle + 2(k_m - k_b)\langle gm \rangle - 2\gamma_m\langle m^2 \rangle \\
\frac{d\langle gm \rangle}{dt} &= k_b\langle g \rangle + (k_m - k_b)\langle m^2 \rangle + k_m\langle m \rangle - (\gamma_m + k_{on} + k_{off})\langle gm \rangle.
\end{align*}
\]  

(23a)-(23e)

steady-state analysis of the above linear equations yields
\[
\begin{align*}
\langle g \rangle &= \frac{k_{on} - k_{on} + k_{off}}{k_{on} + k_{off}}, \\
\langle m \rangle &= \frac{k_bk_{off} + k_mk_{on}}{\gamma_m(k_{on} + k_{off})} \\
\langle m^2 \rangle &= \langle m \rangle^2 + \langle m \rangle F \\
F &= 1 + \frac{(k_m - k_b)^2k_{on}k_{off}}{(k_{on}k_m + k_{off}k_b)(k_{on} + k_{off})\gamma_m + k_{on} + k_{off}} \\
\langle gm \rangle &= \langle g \rangle \langle m \rangle + \frac{(k_m - k_b)k_mk_{off}}{(k_{on} + k_{off})^2(\gamma_m + k_{on} + k_{off})},
\end{align*}
\]  

(24a)-(24d)

where \( F \) is the Fano factor (variance divided by mean) of the mRNA population count. Note that when \( k_m = k_b \), i.e., same rate of transcription from both promoter states, then \( F = 1 \), implying Poisson mRNA statistics.

We next use (23) to compute the condition expectation (21). By solving the linear system of differential equations (23a) and (23b) with initial conditions \( g(s) \) and \( m(s) \), respectively, we obtain
\[
\langle m(t+s) | m(s), g(s) \rangle = a_1(t)\langle m \rangle + a_2(t)\langle g \rangle + a_3(t)m(s)
\]  

(25)

for some time-varying coefficients \( a_1(t), \ a_2(t) \) and \( a_3(t) \) that satisfy
\[
\begin{align*}
a_1(0) &= 0, & a_2(0) &= 0, & a_3(0) &= 1, \\
\lim_{t \to \infty} a_1(t) &= 1, & \lim_{t \to \infty} a_2(t) &= 0, & \lim_{t \to \infty} a_3(t) &= 0.
\end{align*}
\]  

(26a)-(26b)

Substituting (25) in (20) yields
\[
\langle m(t+s)m(s) \rangle = a_1(t)\langle m^2 \rangle + a_2(t)\langle gm \rangle + a_3(t)m^2(s)
\]  

(27)

which using values of steady-state moments computed in (23) results in the auto-correlation function \( R(t) \) given by Eq. 28. As expected, when \( k_m = k_b \), mRNA production is a Poisson process and the auto-correlation function reduces to \( R(t) = \exp(-\gamma_m t) \).

Figure 4 plots \( R(t) \) for parameters inferred in Section IV (Eq. 19). Experiments used for quantifying mRNA histograms in Figure 3, report an mRNA half-life of 3 hours [22]. Hence, we assume \( \gamma_m = .23 \) hour\(^{-1} \) in Eq. (28). Figure 4 shows that the mRNA auto-correlation function has shifted to the right compared to \( \exp(-\gamma_m t) \) suggesting that the timescale of mRNA fluctuations is larger than the mRNA half-life. In particular, we obtain the mRNA auto-correlation time \( t_m \) (defined as \( R(t_m) := 0.5 \)) to be approximately 3.7 hours.

Figure 4. mRNA auto-correlation function (dashed line) for the two-state promoter model corresponding to parameters inferred for the HIV promoter in Section IV (Eq. 19). For comparison purposes, \( \exp(-\gamma_m t) \) is also plotted (solid line) where \( \gamma_m \) in the mRNA degradation rate and taken as 0.23 hour\(^{-1} \).

V. Conclusion

The two-state promoter model has been widely used to capture stochasticity in the gene transcription process [8], [9], [16]. We derived exact analytical expressions for the steady-state mRNA probability distribution function (Eq. 14) and auto-correlation function (Eq. 28) for the full two-state promoter model. These results were used to infer kinetic parameters associated with the HIV promoter. Fitting analytically predicted distributions to mRNA histogram data showed that the full two-state promoter model was the most probable stochastic model for the viral promoter. Model inference revealed that
1) Even in the inactive state, the HIV promoter synthesizes transcripts, albeit at low rate ($k_b/\gamma_m = 6.5$). Assuming $\gamma_m = 0.23$ hour$^{-1}$ this corresponds to approximately 1-2 mRNAs being transcribed per hour.

2) In the active state, transcription occurs at a rate $k_m/\gamma_m = 759$, which corresponds to 2-3 mRNAs transcribed per minute.

3) The promoter spends only 15% of its time in the active state.

Using the inferred parameters for the HIV promoter we predicted the mRNA auto-correlation function (Figure 4), which could be verified through additional experiments. In summary, given advances in experimental techniques to measure mRNA levels inside individual cells in real time, analytical results presented here could be useful in inferring genetic promoter models from data.

An implicit assumption in the two-state promoter model is that the time spent in each promoter state is exponentially distributed. However, recent data shows that in some cases time spent in the inactive state is better approximated by a gamma distribution [32]. Future work will extend current analysis to such promoters and develop techniques for discriminating between alternative promoter architectures using measured mRNA population count statistics.

REFERENCES