Selecting stochastic models of gene activity underlying transcriptional bursting in single mammalian cells

Benjamin Zoller, Nacho Molina, David Suter, Felix Naef

Computational System Biology Lab, Ecole Polytechnique Fédérale de Lausanne

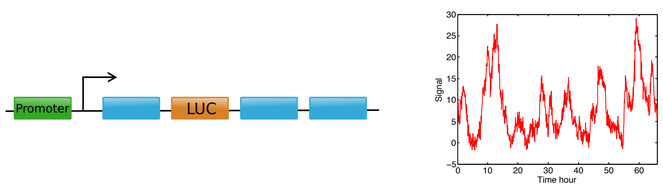
benjamin.zoller@epfl.ch

Abstract

The intrinsic stochasticity in the dynamics of mRNA and protein expression has important consequences on gene regulation and on non-genetic cell-to-cell variability. Recently experimental work in prokaryotes and eukaryotes using single cell resolution time lapse imaging has enabled a quantitative analysis and modeling of the stochastic processes underlying observed fluctuations. Gene transcription was found to occur mainly during short and intense periods referred to as transcriptional bursts, interspersed by silent periods. The fine transcriptional kinetics of endogenous genes in mammalian cells has recently been measured (Suter et al. 2011) by live imaging at high temporal resolution of short-lived luciferase reporters. Here, we further develop the probabilistic framework to model these recordings based on three-layered Hidden Markov Models that describe the three main processes of gene expression: gene activation, transcription and translation. We propose models with different number of sequential gene states describing the activation and inactivation events. To study those, we developed and tested several approximations to efficiently compute the transition probabilities and the likelihood of the models. We select the optimal model using Markov-Chain Monte Carlo (MCMC) sampling, which provide new insights about the number of gene activity processes and their characteristic timescale leading to transcriptional bursts.

1. Real-time measurement of transcription

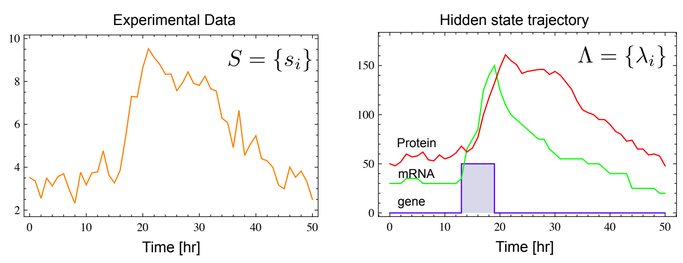
1.1 Experimental setup



1. *Constructs with a modified luciferase reporter are made*
2. *A lentovirus is used to insert the construct into the genome of NIH 3T3 cells*
3. *Special modifications are carried out to ensure short half-lives of both mRNA and protein (30'-40' and 20' respectively)*
4. *Images are captured every 5 minutes at single-cell resolution with a high-sensitive camera*
5. *Recordings of endogenous promoter, Bmal promoter and synthetic promoter are obtained*
6. *The camera is calibrated :* $P(s|p) \sim \mathcal{N}(s; \alpha p, \sigma_b + \beta p)$

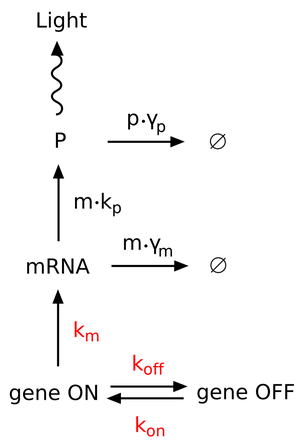
2. Stochastic Modeling of Gene Expression

2.1 HMM framework



1. *The system being observed is assumed to be a Markov process with unobserved states* $\lambda_i = (p_i, m_i, g_i)$*, where* p_i*,* m_i *and* g_i *are the protein copy number, the mRNA copy number and the gene activity at each time point.*
2. *Given a model* Θ*, the states of the system* Λ *are infered from the signal time trace* S*.*
   - *Forward algorithm*
     → *Likelihood* $\mathcal{L}(S|\Theta) = \sum_{\text{all paths } \Lambda} P(s_1|\lambda_1)P(\lambda_1)\prod_{i=2}^{T} P(s_i|\lambda_i)P(\lambda_i|\lambda_{i-1})$
   - *Forward and Backward algorithm*
     → *Posterior decoding* $P(\Lambda|S\Theta) = \frac{P(\Lambda|\Theta)}{P(S|\Theta)}$
   - *Viterbi algorithm*
     → *Maximum Path* Λ_{max}
3. *To compute those quantities one needs the transition probabilities* $P(\lambda|\lambda')$ *which can be computed by solving the master equations of the model* Θ*.*

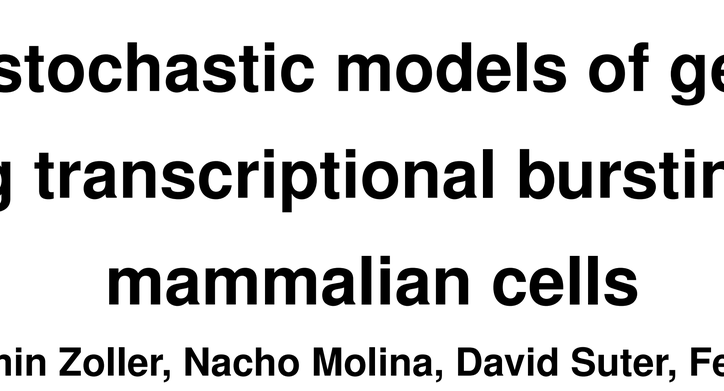
2.2 The Telegraph Model



1. *Transcriptional bursts can be modelized by a gene stochastically switching between a state of transcriptional activity (ON-state) and inactivity (OFF-state).*
2. *This is the simplest model accounting for the 3 main processes : translation, transcription and gene activity.*
3. *All the parameters are assumed to be constant, the rates* k_p, γ_p, γ_m *are measured whereas* k_m, k_{on}, k_{off} *are inferred.*

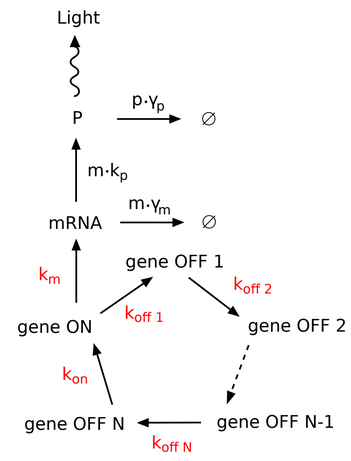
2.3 Refractory Period

The telegraph model predicts exponentially distributed ON-time and OFF-time and yet the data show a refractory period in the OFF-time distribution (Suter et al. 2011). We need at least 2 inactive states to account for this refractory period.



3. Extension : multi sequential OFF-states Model

3.1 Minimal Model for refractory period



1. *Simplest extension : gene activity modelized by one ON-states and* N *sequential OFF-states.*
2. *Minimal Model accounting for refractory period :* $N \geq 2$*.*
3. *Questions : How many off-states do we need to best explain the data ? What are the typical timescale of these OFF-states ? What are the biological mechanisms behind this refractory period ?*

3.2 Propagators

Assuming $\gamma_m \Delta t \ll 1$*, one can factorize the propagator which simplify considerably the problem :*

$$P(pmg|p'm'g') \simeq P(p|p'm')P(mg|m'g') \quad \textit{if} \quad \gamma_m \Delta t \ll 1$$

3.2.1 Protein transition

The transition probabilities of the protein $P(p|p'm')$ *are described by a simple birth and death process :*

$$\frac{d}{dt}P(p,t) = -(k_p + p\gamma_p)P(p,t) + k_p P(p-1,t) + (p+1)\gamma_p P(p+1,t)$$

The exact solution of this equation is given by :

$$P(p|p'm') = \sum_{q=0}^{p'} \binom{p'}{q} \mathcal{P}(p-q; \frac{k_p m'}{\gamma_p}(1-e^{-\gamma_p t}))(e^{-\gamma_p t})^q (1-e^{-\gamma_p t})^{p'-q}$$

The mean and the variance of the distribution :

$$\mu_p(m',t) = \frac{k_p m'}{\gamma_p}\left(1-e^{-\gamma_p t}\right) + p' e^{-\gamma_p t}$$

$$\sigma_p^2(m',t) = \frac{k_p m'}{\gamma_p}\left(1-e^{-\gamma_p t}\right) + p' e^{-\gamma_p t}\left(1-e^{-\gamma_p t}\right)$$

For large amount of proteins, the linear noise approximation ca be used if $\sigma_p \ll p$ *:*

$$P(p|p'm') = \frac{1}{\sqrt{2\pi(\mu_p(m',t) - p'e^{-2\gamma_p t})}} \exp\left(-\frac{(p-\mu_p(m',t))^2}{2(\mu_p(m',t) - p'e^{-2\gamma_p t})}\right)$$

3.2.2 mRNA and the gene state transition

The transition probabilities of the mRNA and the gene $P(mg|m'g')$ *are obtained by solving the following master equation :*

$$\frac{d}{dt}P(m,g,t) = -k_m\delta_{g,on}P(m,g,t) - m\gamma_m P(m,g,t) + k_m\delta_{g,on}P(m-1,g,t)$$
$$+(m+1)\gamma_m P(m+1,g,t) + k_g P(m,g-1,t) - k_{g+1}P(m,g,t)$$

The solution of this master equation can be obtained by applying the exponential on the state rate matrix $\mathbf{M}$ *:*

$$P(mg|m'g') = \langle mg| \exp(\mathbf{M}\Delta t)|m'g'\rangle$$

4. Model Selection : MCMC Sampling

4.1 Reversible Jump MCMC

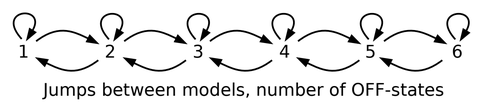
The model selection is based on the posterior distribution :

$$P(m|S) = \frac{\int \mathcal{L}(S|m,\theta'_m)P(\theta'_m)P(m)d\theta'_m}{\sum_{m'}\int \mathcal{L}(S|m',\theta'_m)P(\theta'_m)P(m')d\theta'_m}$$

Instead of computing this quantity, one can sample the targeted distribution $P(m,\theta_m|S)$ *with the Reversible Jump MCMC algorithm.*

A move set has to be defined to change the dimensionality of the sampling space according to the model (the jump), for instance :

1. *split randomly a given component* θ_k *such that* $\theta_k = \theta'_k + \theta'_{k+1}$
2. *merge two components* θ_k *and* θ_{k+1} *such that* $\theta_k + \theta_{k+1} = \theta'_k$
3. *stay in the same model, update the components according to the proposal distribution*



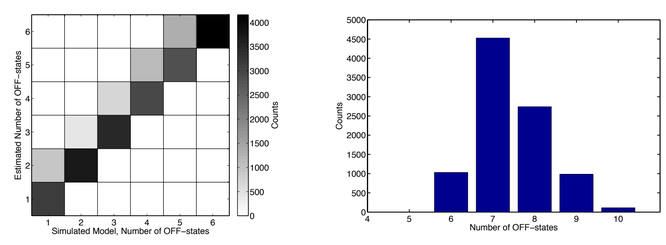
This algorithm is very similar to the Metropolis-Hastings, except that the sampling space $\mathcal{X} = \bigcup_{m\in\mathcal{M}}(\{m\} \times \theta_m)$ *involves spaces of different dimensions.*

The iterative protocol :

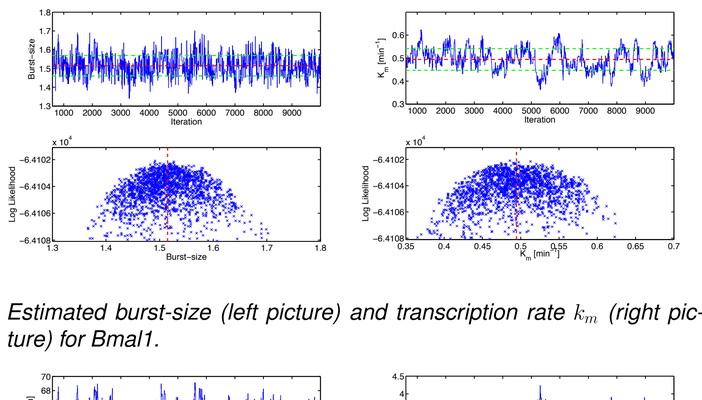
1. *Choose a move for the chain randomly*
2. *Compute the likelihood for the proposed parameters* n*,* θ_n
3. *Sample* u *from a uniform distribution* $\mathcal{U}_{[0,1]}$
4. *If* $u < \mathcal{A}_{m\to n}$*, then accept the move and update the chain*

$$\mathcal{A}_{m\to n} = \min\left\{1, \frac{\mathcal{L}(S|n,\theta_n)P(n)P(\theta_n)Q(m|n)Q_{n\to m}}{\mathcal{L}(S|m,\theta_m)P(m)P(\theta_m)Q(n|m)Q_{m\to n}}\mathcal{J}_{m\to n}\right\}$$

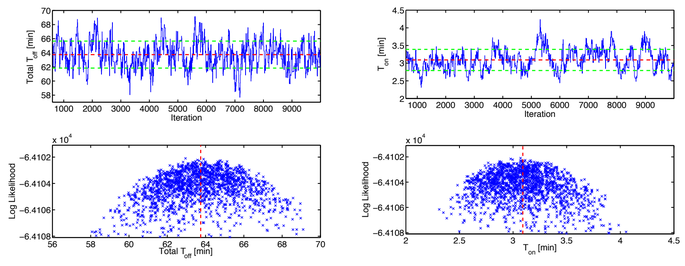
4.2 Results



Estimated number of OFF-states for simulated data with Gillespie algorithm (left picture). Estimated number of OFF-states for Bmal1 (right picture).



Estimated burst-size (left picture) and transcription rate k_m *(right picture) for Bmal1.*



Estimation of the total OFF-time $\tau_{off} = k_{on}^{-1} + \sum_{i=2}^{N} k_{off,i}^{-1}$ *(left picture) and the ON-time* $\tau_{on} = k_{off,1}^{-1}$ *(right picture) for Bmal1.*

5. Conclusion

1. *We have presented a minimal extension of the telegraph model to account for the refractory period.*
2. *We are able to estimate the kinetic parameters and the number of OFF-states applying an MCMC approach.*
3. *The presence of this refractory period may reflect the requirement of several processes before elongation : chromatin remodeling, transcription factors binding, polymerase recruitment, etc.*
4. *The main advantage of this framework is the great flexibility it provides. The gene activity part of the model can be easily changed to account for transcription factors binding or other mechanisms.*
5. *There are still open questions : Is the infered number of OFF-states common between different genes ? What are the biological mechanisms accountable for this refractory period ?*