Isolating the perturbation response of gene regulatory networks in the presence of biological variability and technical noise

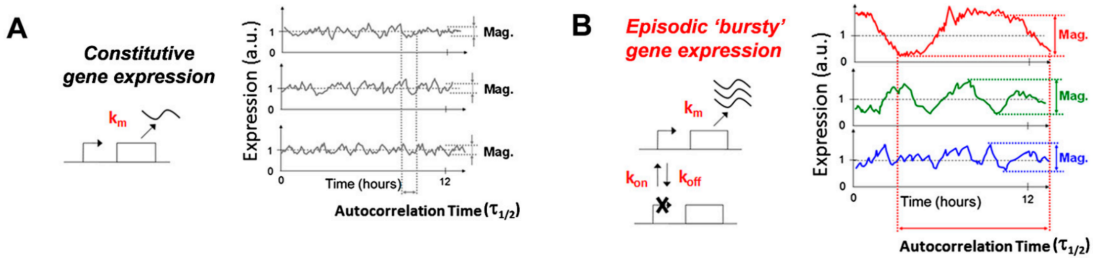
Clayton W. Seitz

May 24, 2022

Outline

Modeling stochastic biochemical reaction networks

Gene expression is stochasic and non-constitutive



- If the biochemical network is known a-priori, we can build parametric dynamical models
- Bayesian inference allows us to fit parametric dynamical models without continuous dynamical trajectories

Stochastic biochemical reaction networks: the repressilator

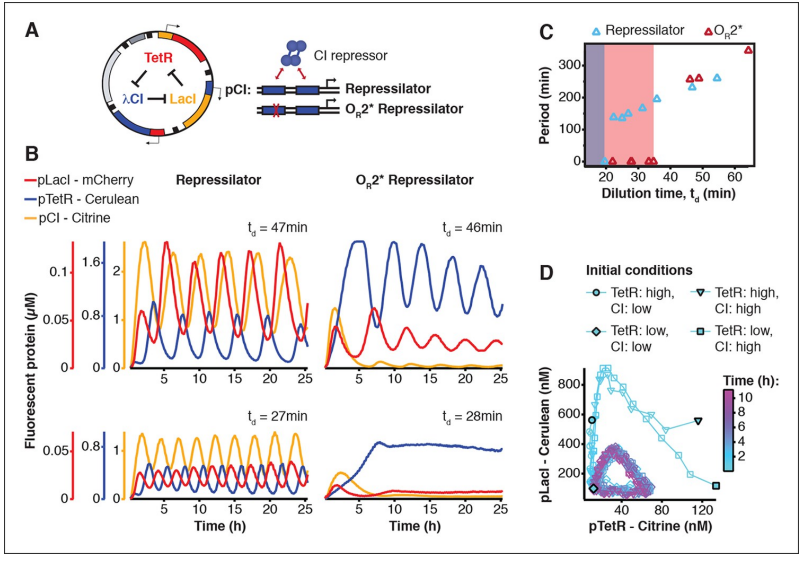


Figure 1: Niederholtmyer et al., eLife 2015

Kolmogorov's forward equation (chemical master equation)

Dynamics on biochemical reaction networks are inherently stochastic and the state space is discrete. We can only write probabilities over the state space

$$\begin{aligned} P(\mathbf{x}_i, t) &= \sum_j T_{ji}(\mathbf{x}_i, t|\mathbf{x}_j, t - \Delta t) P(\mathbf{x}_j, t - \Delta t) \\ &= \sum_k T_k(\mathbf{x}_i, t|\mathbf{x}_i - \nu_k, t - \Delta t) P(\mathbf{x}_i - \nu_k, t - \Delta t) \end{aligned}$$

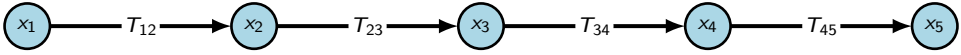
where T_k is the probability of a reaction channel k firing in the interval $(t, t + \Delta t)$.

Taking the limit $\Delta t \to 0$ one can derive the forward Kolmogorov equation or chemical master equation (CME)

$$\frac{dP(\mathbf{x}, t|\mathbf{x}_0)}{dt} = \sum_k T_k(\mathbf{x} - \nu_k) P(\mathbf{x} - \nu_k, t) - T_k(\mathbf{x}) P(\mathbf{x}, t)$$

What about Markov models?

From another point of view, since the dynamics are Markov the state **x** follows the DAG



For MMs, the EM algorithm can be used for MAP estimation, but this requires time-series measurements

Time-series measurements in live cells severely limits the number of species considered simultaneously

More often than not the data we have are *ensemble snapshots*

Bayesian parameter inference using ensemble snapshots

Suppose we have a series of ensemble snapshots of an *in-vitro* population:

$$\mathbf{x} = \{\mathbf{x}_0, ..., \mathbf{x}_t\} \quad \mathbf{y} = \{\mathbf{y}_0, ..., \mathbf{y}_t\}$$

with $\mathbf{x}_t = \{x_1, ..., x_n\}$ and similarly for $\mathbf{y}$. Under perfect measurements $\mathbf{x} = \mathbf{y}$

We would like to use $\mathbf{x}$ to fit a dynamical model $\mathcal{M}(\theta)$. Bayesian inference lets us infer θ from $\mathbf{x}$ while quantifying the uncertainty in our estimate:

$$P(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta) = \pi(\theta)\prod_t f(\mathbf{x}_t|\theta)$$

The likelihood $f(\mathbf{x}_t|\theta)$ is often difficult to define or intractable to compute due to the curse of dimensionality, making even MLE a challenge

Beating the curse of dimensionality for parameter inference

We want to avoid needing to train new networks for every parameter value θ (expensive!) then computing the likelihood of the experimental data. Instead we compute the likelihood of simulated data under set of target variational distributions trained on experimental data. This assumes simulations and experimental data are "exchangeable" when computing the posterior

Variational step: we learn N_t target distributions by training a deep network on the experimental data. In this way we have N_t variational target distributions

ABC step: We sample parameters from our prior $\theta \sim \pi(\theta)$, and produce N Monte Carlo trajectories $\mathbf{x}(t)$. We compute the likelihood of the simulated trajectory with a tolerance ϵ (with a tolerance schedule). This replaces the distance metric in ABC with a variational likelihood

Confounding factors in bursty gene expression

Transcription factors bind to *cis*-regulatory elements of the genes they regulate

Transcription factors allow cells to perform logic operations and integrate information

In some cases, we may have a pool of transcription factors which could be regulate a gene, but we aren't sure which and how

Silencing TFs is time consuming and expensive. Can we infer the logical operation from the data?
Pure induction of transcriptional bursting requires complete control over gene expression by an external perturbation

Finding "purely inducible" genes for which the exact biophysical regulatory mechanism is known is very rare

Even if all factors *are* known, evidence for the transcriptional logic implemented by those factors and their interaction with the DNA must be found in gene expression data

Jointly learning regulators and kinetics

1. Identify a set of potential regulators $\boldsymbol{F}$ using TFBS predictions e.g., using DeepBind
2. A gene coding for (X, Y) is regulated according to

$$
\begin{aligned}
G_{off} + \boldsymbol{F} &\leftrightarrow G_{on} \\
G_{on} &\xrightarrow{k_1} G_{on} + X \\
X &\xrightarrow{\gamma} \emptyset \\
X &\xrightarrow{k_2} Y
\end{aligned}
$$

$\boldsymbol{F}$ are TF conc. - independent observables which determine binding probabilities

Model the first reaction as a Boltzmann machine with weights W where the output node is the probability the promoter is active

Constitutive (Poisson) Transcription + Ising Model

$$\mathcal{H} = -\frac{1}{2}\sum_{i,j} J_{ij} x_i x_j - \sum_j h_j x_j \quad P(\boldsymbol{x}) = \frac{1}{Z}\exp\left(-\beta\mathcal{H}(\boldsymbol{x})\right)$$

We know $p(x_j = 1) = \exp(-\beta H(x_i = 1))$. Then, we define

$$h_j = -\frac{1}{\beta}\log p(x_i = 1) = -\frac{1}{\beta}\log\frac{[x_j]}{K_j + [x_j]}$$

Suppose that there is a single state or set of states $\boldsymbol{x}^*$ for which the promoter is active

$$\lambda = \frac{Z_{on}}{Z_{on} + Z_{off}} \quad \rightarrow \quad P(n|\mu) = \frac{\mu^n}{n!}\exp(-\mu)$$

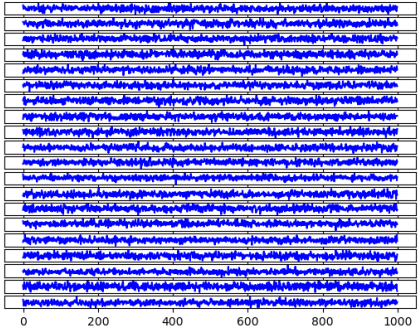
with $\mu = \lambda t$

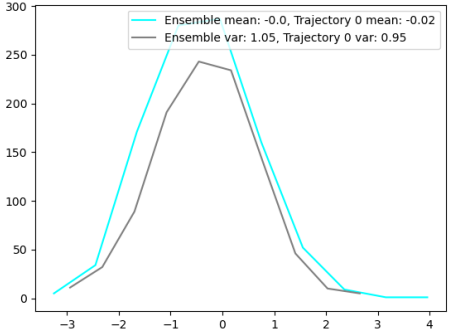
Super-Poissonian Transcription

Unexplained variance suggests in RNA distributions suggests that transcription occurs non-constitutively

Transcription is non-ergodic

Ergodicity of a Markov chain at the heart of MCMC

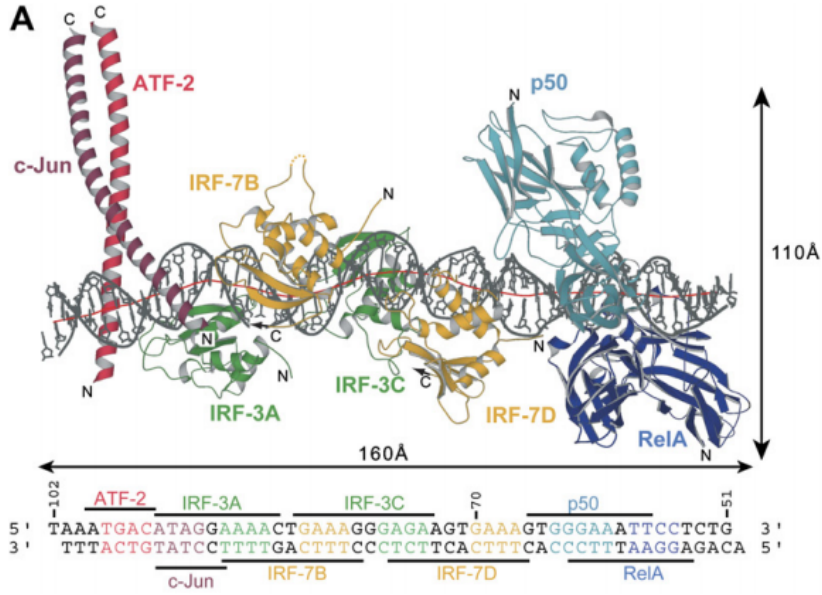




Statistical properties (moments) of an ensemble and a single instance are the same:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} x_i = \int x\ p(x) dx \qquad \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 = \int (x-\mu)^2 p(x) dx$$

The Interferon-β Enhanceosome

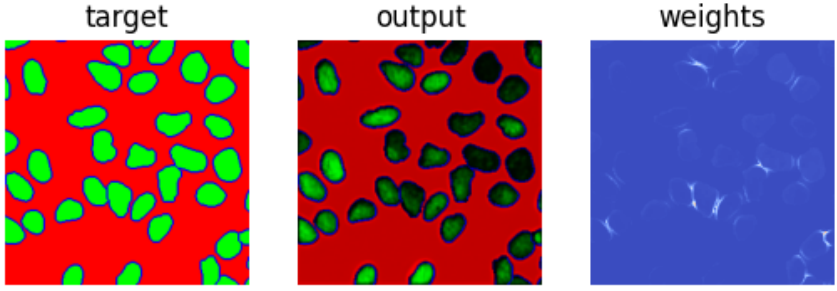


The Interferon-β Enhanceosome

Enhanceosomes are protein complexes which solely regulate gene transcription

Training on BBBC039 U2OS Cells

BBBC039: 200 images, 160 train + 40 validation, 256 x 256 random crop

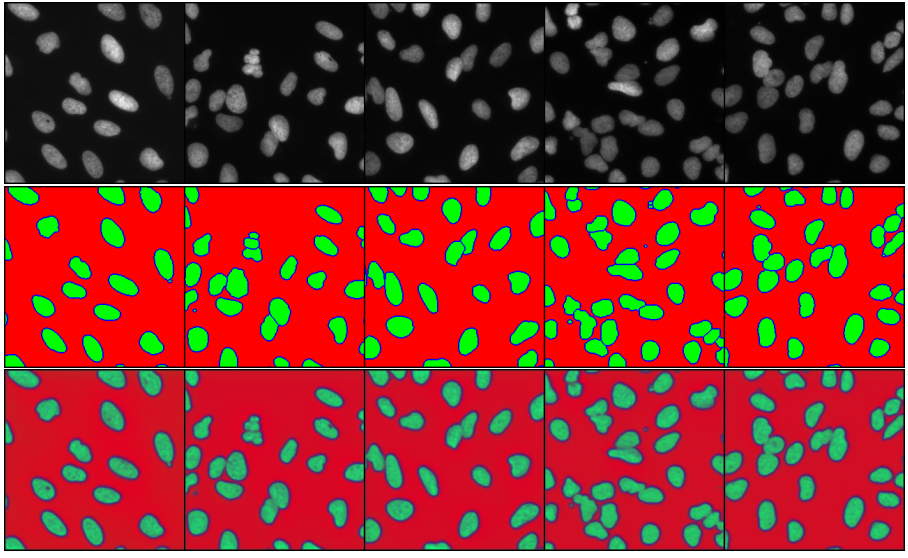


We train a 3-channel semantic segmentation model with **weighted** cross-entropy loss:

$$\mathcal{L} = \sum_{i,j} w_{ij} \log p_{ij}(\tilde{x}) = \sum_{i,j} w_{ij} \log \frac{\exp(-s_{ij}(\tilde{x}))}{\sum_{x \in \chi} \exp(-s_{ij}(\tilde{x}))}$$

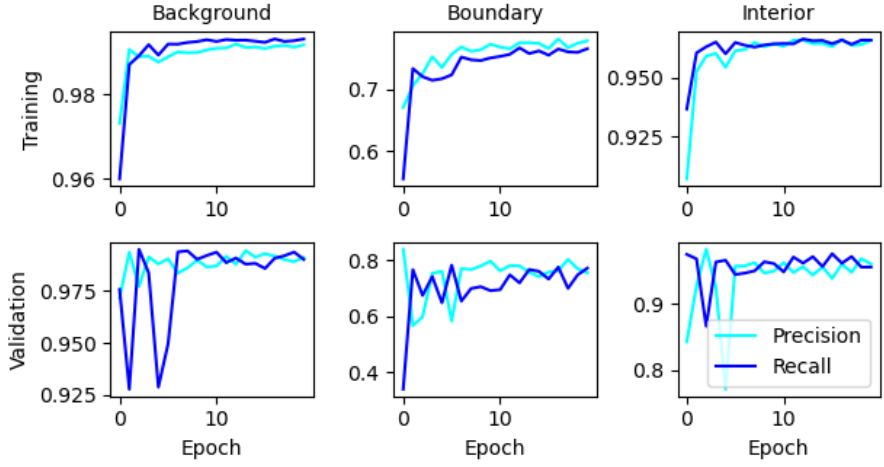
p_{ij} is the probability the model assigns a pixel to the true class $\tilde{x} \in \{\mathrm{a}, \mathrm{b}, \mathrm{c}\}$

Training on BBBC039 U2OS Cells



Training on BBBC039 U2OS Cells

Learning rate $\eta = 0.01$, Batch-size $B = 5$ (32 train iterations, 8 validation)



References I