Plan of Study

Clayton W. Seitz *May 12, 2022*

Abstract

1 Introduction

The intracellular environment harbors a delicate set of interactions between DNA, RNA, and protein in order to carry out primitive cellular functions and perform signal processing. Recent developments in fluorescent labeling and imaging technologies permit the simultaneous measurement of RNA and protein copy numbers in single cells, allowing us to analyze these interactions in space and time. The scalability of these methods has stimulated a generalization of the notion of a biochemical pathway to the large-scale biochemical network. In the context of gene regulation, this biochemical network perspective has proven quite powerful, as it leverages the power of probabilistic graphical models. Such models allow us to efficiently estimate the conditional dependence structure of high-dimensional joint probability distributions, providing clues towards unknown regulatory interactions or the impact of controlled perturbations on network architecture. Recently, the inference of gene regulatory networks (GRNs) from static experimental data has become a relatively mature class of methods when applied to gene expression data from single-cell RNA-sequencing (Singh, 2018). More recent efforts have gone beyond static datasets, inferring network structure from time-series measurements e.g., oscillations in gene experession during the circadian rhythms of plants, or simulations based on stochastic differential equations (Aalto 2020).

By design, network inference algorithms neglect biochemical kinetics and signed interactions when drawing directed edges between nodes. Inferring precise biochemical kinetics is a much more difficult problem, but remains desirable when perturbations to gene regulation are more nuanced. Gene expression is thought to be inherently stochastic, ultimately governed by an unknown and highly complex coupled system of stochastic differential equations (SDEs). Our inability to assign further detail to network edges stems from the difficulty in determining the form of these differential equations and their parameterization i.e., regression. Nevertheless, Bayesian inference of more detailed kinetic parameters may be possible, under the assumption that transcriptional kinetics can be expressed as relatively simple analytical functions (Burton, 2021) or by using Gaussian processes. Also, a feedback loop between experimental data and Monte Carlo simulations of the chemical master equation (CME) can be estabilished if interaction functions are known.

Of course, there are also a number of other confounding factors, such as the spatial organization of key biomolecules, the biophysical details of transcription factor binding, DNA accessibility, RNA preprocessing, and sample heterogeneity which enter into this hypothetical system of equations. We choose to focus our attention on those factors which can be readily addressed by multiplexed imaging experiments. Challenges associated with sample heterogeneity are closely linked to the curse of dimensionality, which prevents direct inference the joint distribution over biomolecule counts. Sparsity in the data makes it difficult to segregate samples either by discriminative or generative modeling, which may create the illusion that the joint distribution is unimodal - a kind of "modal collapse". In addition, spatial regulation could contribute to non-trivial dynamics. Network inference often neglects the role of the placement of proteins in space and time; however, transcription factors mediate the causual effect between the expression of a source gene and a target gene, making this information valuable in resolving the nature of this interaction. Finally, substantial evidence has surfaced that the biophysical mechanism of TF binding results non-constitutive gene expression; rather, expression occurs in a bursting fashion (Dar 2012; Larrson 2019; Tunnacliffe 2020). Bursting phenomena challenge our assumptions of stationarity and ergodicity of the joint distribution over gene expression and have important implications for parameter inference.

In this study, we use the drug-treatment response of WM989 melanoma cells to treatment with a chemotherapy drug Vemurafenib - a BRAF inhibitor. Statistical treatment has revealed a core set of genes and their logical interactions, which are thought to be of importance during the development of resistance

to chemotherapy treatment *in-vitro* (Shaffer 2017). Development of resistance to treatment has been proposed to a transient state of gene expression which has been attributed to non-genetic expression variability from transcriptional bursting (Schuh 2020). However, the mechanism by which drug treatment alters gene expression irreversibly remains unclear. We expect that this question can begin to be answered by probing alterations in the logical structure of the gene regulatory network alongside Bayesian inference of parameters of the functions governing transcription rates.

1.1 Identifying key genomic regulators in melanoma

1.2 Transcriptional bursting: a source of non-genetic variability

1.3 Spatial properties of gene expression

2 Methods

2.1 Theoretical Methods

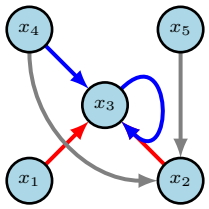
2.1.1 Statistical mechanics of transcription factor binding

2.1.2 The chemical master equation

Suppose we have N distinct chemical species diffusing in a volume Ω. The collection is described by a discrete random variable $\mathbf{X} = \{x_1, ..., x_i\}_{i=1}^{N}$. Suppose molecules can proceed through M distinct chemical reaction channels $\mathbf{R} = \{r_1, ..., r_j\}_{j=1}^{M}$. Perhaps that species x_1 is a dsDNA sequence to be transcribed, x_2 is a transcription factor for x_1, and x_3 is the single stranded RNA. In the following example system, we have five species and three reaction channels

$$\begin{aligned} r_1 &: x_1 + x_2 \rightarrow x_3 \\ r_2 &: x_3 + x_4 \rightarrow 2x_3 \\ r_3 &: x_4 + x_5 \rightarrow x_2 \end{aligned} \qquad \nu = \begin{pmatrix} -1 & -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & -1 & -1 \end{pmatrix}$$

The reaction channels are efficiently described by a stochiometric matrix (columns are species, rows are reactions), as shown on the right. Furthermore, the set of reaction channels $\mathbf{R} = \{r_1, r_2, r_3\}$ has the following graphical representation



Consider then, at $t = 0$ we know the how many of each species are in Ω: $\mathbf{X_0}$ and we would like to predict $\mathbf{X}_t$ at some $t > 0$. Since the evolution of the system is stochastic and the state space is discrete, we have to rely on master equations. Recall the general form of a master equation for a stochastic system where the state space is discrete

$$J_i = \frac{dP_i}{dt} = \sum_j W_{ji}P_j - W_{ij}P_i$$

which, in words, reads that the probability current J_i into the state i is a sum over outgoing current $W_{ij}P_i$ into all other states j and inward current $W_{ji}P_j$ from all other states j. The matrix W is the so-called *transition matrix*. In terms of chemical reactions, a transition between states i and j occurs via a single chemical reaction which occurs with a probability $a_j(\mathbf{x})$ dependent on the current state of the system

$$\frac{dP(\mathbf{x},t)}{dt} = \sum_j a_j(\mathbf{x}-\nu_j)P(\mathbf{x}-\nu_j,t) - a_j(\mathbf{x})P(\mathbf{x},t)$$

The function $a_j(\mathbf{x})$ is generally defined as $a_j(\mathbf{x}) = c_j h_j(\mathbf{x})$ where c_j is a constant factor specific to reaction j while $h_j(\mathbf{x})$ is the number of distinct possible combinations of the reactants available in state $\mathbf{x}$. For example, for r_1 above, we have $h_j(\mathbf{x}) = x_1 x_2$.

2.1.3 The Gillespie algorithm

The chemical master equation is notoriously difficult to solve, and in most cases, we need to simulate from $P(\mathbf{x},t)$ using Monte Carlo methods. Gillespie developed a method for simulating from $P(\mathbf{x},t)$, starting with the following questions: (1) when will the next chemical reaction occur and (2) what kind of reaction will it be? The algorithm proceeds by defining a joint density function $P(\tau,\mu)d\tau$ which describes the probability that the next reaction will be of type μ and occur and a time τ.

2.1.4 Stability analysis of stochastic differential equations

2.1.5 Inferring biochemical networks from data

2.1.6 Learning interaction functions with Gaussian Processes

2.2 Experimental Methods

2.2.1 Multiplexed fluorescence in-situ hybridization (FISH)

2.2.2 Techniques for high-throughput image processing