Plan of Study

Clayton W. Seitz *May 5, 2022*

Abstract

1 Introduction

The intracellular environment harbors a delicate set of interactions between DNA, RNA, and protein in order to carry out cellular functions and perform signal processing. Modern multi-omics technologies have fostered a generalization of the notion of a biochemical pathway to the large-scale biochemical network or circuit, in an attempt to understand the structure of these interactions across different cellular states. In the context of gene regulation, the biochemical network perspective has proven quite powerful, as it leverages the power of probabilistic graphical models. Graphical models allow us to efficiently summarize the conditional dependence structure of high-dimensional joint probability distribution over protein and RNA copy numbers. Such an approach simultaneously permits clues towards unknown regulatory interactions or the impact of controlled perturbations on the dynamics of biochemical reaction networks. 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). Interactions between different genes are represented as a directed graph in which nodes represent genes and directed edges indicate a causual relationship between a source gene and destination gene.

We have identified several shortcomings of this approach in need of development. First, there exists biological varability and technical noise in our datasets, introducing heterogeneity which is difficult to detect. This can be thought of as a byproduct of the curse of dimensionality, in which we are unable to directly infer the joint distribution over biomolecule counts. The inability to segregate samples either by discriminative or generative modeling may create the illusion that the joint distribution is unimodal - a kind of "modal collapse". Furthermore, by design, any network inference algorithm neglects biochemical kinetics when drawing directed edges between nodes. Inferring precise biochemical kinetics is a much more difficult problem, but remains desirable. Gene expression is thought to be inherently stochastic, 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 our inability to determine the form of these differential equations and their parameterization, or even its stationary statistics. 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, and RNA preprocessing, which enter into this hypothetical system of equations. We choose to focus our attention on spatial organization, as this information can be readily measured in multiplexed RNA-imaging experiments. Spatial regulation could contribute to non-Hill dynamics. Network inference is frequently carried out on RNA-seq datasets or multiplexed RNA-FISH while neglecting the role of the placement of proteins in space and time. 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 genes are not expressed constitutively; rather, gene expression occurs in a bursting fashion (Dar 2012; Larrson 2019; Tunnacliffe 2020) which has been suggested to occur due to switching of a gene promoter between inactivate and activate transcriptional states. Bursting phenomena challenge our assumptions of stationarity and ergodicity of the joint distribution over gene expression, posing important theoretical questions.

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 Gillespie algorithm

2.1.3 Stability analysis of stochastic differential equations

2.1.4 Inferring biochemical networks from data

2.1.5 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