Methods xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Integrating single-molecule experiments and discrete stochastic models to understand heterogeneous gene transcription dynamics

Brian Munsky [a,b,*], Zachary Fox [b], Gregor Neuert [c,d,*]

[a] Department of Chemical and Biological Engineering, Colorado State University, Fort Collins, CO 80523, USA
[b] School of Biomedical Engineering, Colorado State University, Fort Collins, CO 80523, USA
[c] Department of Molecular Physiology and Biophysics, Vanderbilt University School of Medicine, Nashville, TN 37232, USA
[d] Department of Biomedical Engineering, Vanderbilt University School of Engineering, Nashville, TN 37232, USA

ARTICLE INFO

Article history:
Received 14 March 2015
Accepted 8 June 2015
Available online xxxx

Keywords:
Gene regulation
Single-cell dynamics
Biochemical noise
Chemical master equation
Model identification

ABSTRACT

The production and degradation of RNA transcripts is inherently subject to biological noise that arises from small gene copy numbers in individual cells. As a result, cellular RNA levels can exhibit large fluctuations over time and from one cell to the next. This article presents a range of precise single-molecule experimental techniques, based upon RNA fluorescence *in situ* hybridization, which can be used to measure the fluctuations of RNA at the single-cell level. A class of models for gene activation and deactivation is postulated in order to capture complex stochastic effects of chromatin modifications or transcription factor interactions. A computational tool, known as the finite state projection approach, is introduced to accurately and efficiently analyze these models in order to predict how probability distributions of RNA change over time in response to changing environmental conditions. These single-molecule experiments, discrete stochastic models, and computational analyses are systematically integrated to identify models of gene regulation dynamics. To illustrate the power and generality of our integrated experimental and computational approach, we explore cases that include different models for three different RNA types (sRNA, mRNA and nascent RNA), three different experimental techniques and three different biological species (bacteria, yeast and human cells).

1. Introduction

In recent years, advanced experimental techniques have provided biologists with unprecedented abilities to probe and observe the myriad parts of biological processes. Techniques such as RNA sequencing, super-resolution fluorescent imaging, and flow-cytometry have provided details of individual biological components, even at single-cell and single-molecule resolutions [1–4]. Such detailed observations have largely out-paced our ability to understand, interpret, predict or influence these processes. A key contributor to the disconnect between the availability of high-throughput biological data and quantitative, predictive biological understanding is the extremely complex and often random nature of biological systems. Large numbers of chemical species all interact in complex, non-linear networks to carry out even the most basic biological tasks, such as transcription regulation. Furthermore, inherent in any experimentally observed biological system are several types of "noise", including intrinsic fluctuations in cellular constituents, extrinsic heterogeneities between cells, measurement inaccuracies, and inadvertent environmental fluctuations. When these complex processes and unavoidable noise combine together, the result may make it extremely difficult to match or predict biological phenomena.

Mathematical modeling of biological systems can serve a variety of purposes, such that different models may satisfy different goals. The goal of some models may be to create a comprehensive representation of a biological process by compiling all known understanding of that particular system [5,6]. While such models are qualitative in nature, they can provide a complete picture of how a particular system is currently understood to behave and can be used to test broad qualitative hypotheses. Conversely, the goal of conceptual models may be to capture a small part of larger biological networks, or to reveal physical principles about how an individual subsystem behaves in specific circumstances [7,8]. In this article, we investigate a third goal of modeling: to quantitatively predict how a system will behave under experimental

* Corresponding authors at: Department of Chemical and Biological Engineering, Colorado State University, Fort Collins, CO 80523, USA (B. Munsky). Department of Molecular Physiology and Biophysics, Vanderbilt University School of Medicine, Nashville, TN 37232, USA (G. Neuert).

E-mail addresses: munsky@engr.colostate.edu (B. Munsky), gregor.neuert@vanderbilt.edu (G. Neuert).

http://dx.doi.org/10.1016/j.ymeth.2015.06.009
1046-2023/

conditions. Where comprehensive models may be complex combinations of hundreds of reactions and biophysical parameters, and principle based models may be exceptionally simple, in optimally predicting models, the complexity is dictated by existing quantitative data and predictive goals [9]. Uncertainty from measurement noise combined with highly complex biological systems may lead to poor parameter identification and a resulting loss in predictive power. Two questions naturally arise: how do the challenges presented by biological complexity restrict predictive modeling and in what ways can emerging experimental approaches enable improved predictive understanding?

The first such challenge of model identification is "model sloppiness" – the notion that parameters are often poorly constrained, especially in biological models [10]. For a given amount and type of experimental data, only certain parameter combinations will be well defined, leading to large regions of parameter uncertainty. Moreover, addition of more data of the same type may fail to reduce the parameter uncertainties. This diminishing return from additional data motivates a need for enhanced experiments that complement models or reduced models that complement the available data. In some fortuitous cases, additional data may already exist that has not yet been fully utilized. For example, fitting deterministic models (i.e., sets of ordinary differential equations, or ODEs) to single-cell distributions may partially constrain parameters, but often ignoring cell-to-cell heterogeneity may limit success in model identification [11]. In other words, biochemical noise, the fluctuations inherent to the biological process being measured, may provide information inaccessible when measured with bulk analyses (e.g., PCR or western blot analysis of RNA or protein content) or when modeled by ODE analyses. By reducing parameter uncertainty, it may become possible to constrain more realistic models, and the errors associated with these predictions may be reduced [9,12].

Several approaches have been suggested to utilize biochemical fluctuations to improve parameter estimation for gene regulatory circuits. These approaches have used many different types of experimental data and computational analyses. Several studies have examined regulation at the post-translational level, using fluorescent protein reporters combined with flow cytometry [12–16] or time lapse fluorescence microscopy [17–21]. Others have examined regulation at the level of single mature RNA transcripts [9,22–26] or at the level of active transcription sites [23,27]. Although many studies have focussed on steady state responses [28,29], others have explored how the variability of responses changes over time or from one condition to another [9,30,31]. On the computational modeling side, several studies have used reduced order expressions for parameter moments (i.e., the means and variances) to characterize the variability of the single-cell responses in the presence of intrinsic or extrinsic noise [16,26,32–35]. Other approaches have used kinetic Monte Carlo simulations such as the stochastic simulation algorithm (SSA, [36]) to generate many simulated trajectories to represent the underlying biological system [14,15,19]. Others have used approximate solutions of the infinite dimensional linear equation known as the chemical master equation to directly compare model predictions to measured single-cell distributions [9,12,13,22]. These studies have been applied to natural and synthetic gene regulatory circuits in bacteria [12,13,16,22], yeast [9,19,33,25], and mammalian cells [17,18,26,27,37].

In this article, we will review our approach to fit the full time-varying distributions of a gene regulatory model to single-molecule measurements of RNA at different times and experimental conditions. In the following sections, we will introduce the technique of single-molecule RNA fluorescence *in situ* hybridization (smRNA-FISH [38,39]), which we have used to measure the number and location of RNA molecules in single cells. We will also introduce the computational technique known as the finite state projection (FSP, [40]) algorithm, which can be used to predict the probability distributions of transient gene regulation responses. We will illustrate the use of the smRNA-FISH and FSP approaches to fit models and eventually predict the distributions of RNA in single cells. Finally, we will explore three cases where different models, different FSP analyses and different versions of smRNA-FISH have been combined to explore the temporally changing regulatory characteristics of (i) small RNA in bacteria [22], (ii) serum-activated transcriptional responses in human cells [27], and (iii) osmotic shock response genes in yeast [9].

2. Experimental methods

In order to take advantage of the information contained in single-cell fluctuations, one must measure those fluctuations as precisely as possible. Many recent studies have utilized single-cell measurements of fluorescent protein (FP) markers of gene expression to establish and fit probability distributions at the protein level [12,13,16,19]. One advantage of the FP-based approach is that it allows for the tracking of individual cells over time using time lapse fluorescence microscopy. Alternatively, one can use flow cytometry to measure the FP distributions at specific snapshots in time, which trades the ability to measure temporal correlations within individual cells for an ability to collect statistics of thousands of cells at many points in time. Moreover, the use of FP has a few disadvantages for the analysis of transcriptional responses. First, the use of FP markers requires the genetic manipulation of cells to express a FP marker for each gene of interest. Such modifications could potentially disrupt the natural behavior of gene regulation or the resulting mRNA dynamics. Second, measurement of FP markers in a given cell yields an average fluorescence intensity for each cell, which one must deconvolve from background fluorescence and calibrate against known standards in order to estimate absolute numbers of proteins. Third, the use of FP markers introduces additional dynamics into the process, including processes of translation and fluorescent protein folding and maturation. These processes can add significant delays between the process of transcriptional regulation and the downstream measurable FP signal [31,41]. For fast transcriptional processes, such as stress responses that have time scales on the order of a few minutes, a much faster assay is highly beneficial [9,37].

One such assay that allows for absolute quantification of fast endogenous transcriptional responses is the relatively recent technique of single-molecule fluorescent *in situ* hybridization (smRNA-FISH, [38,39]). Fig. 1A illustrates the basic concept of smRNA-FISH and Fig. 1B–D show three different variants of the approach and images of these approaches applied to human, yeast and bacterial cells. The smRNA-FISH technique was pioneered many years ago using multi-labeled 50 nucleotide long single strand DNA molecules [39] as shown in Fig. 1B (top). About a decade later, this technique was modified to use many single-labeled 20 nucleotide long single strand DNA probes [38] as illustrated in Fig. 1C (top). The advantage of the larger number of smaller probes is to increase the total number of probes on a target RNA while reducing the background fluorescence emitted by unbound probes. To build further on these advances in smRNA-FISH technology, quencher probes as illustrated in Fig. 1D (top) were recently proposed to reduce further the fluorescent signals from unbound probes, reduce background fluorescence and improve single-to-noise ratio [22]. As the background is reduced, smaller "true" signals can be detected, which is particular helpful for the detection of short RNA transcripts. Each of these techniques have been successfully applied to numerous organisms including human-derived cells (Fig. 1B, bottom), yeast (Fig. 1C, bottom)

Please cite this article in press as: B. Munsky et al., Methods (2015), http://dx.doi.org/10.1016/j.ymeth.2015.06.009