Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online
CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 49-26 (2016) 114–119

Validation methods for population models of gene expression dynamics

Andrés M. González-Vargas * **Eugenio Cinquemani** **
Giancarlo Ferrari-Trecate ***,****

* *Departamento de Automática y Electrónica. Universidad Autónoma de Occidente. Cll 25#115-85 Km 2 Vía Cali-Jamundi. Cali, Colombia (e-mail: amgonzalezv@uao.edu.co)*
** *INRIA Grenoble – Rhône-Alpes, Montbonnot, 38334 St.Ismier Cedex, France (e-mail: eugenio.cinquemani@inria.fr)*
*** *Dipartimento di Ingegneria Industriale e dell'Informazione, Università degli Studi di Pavia, via Ferrata 3, 27100 Pavia, Italy (e-mail: giancarlo.ferrari@unipv.it).*
**** *Currently: Automatic Control Laboratory, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland*

Abstract: The advent of experimental techniques for the time-course monitoring of gene expression at the single-cell level has paved the way to the model-based study of gene expression variability within- an across-cells. A number of approaches to the inference of models accounting for variability of gene expression over isogenic cell populations have been developed and applied to real-world scenarios. The development of a systematic approach for the validation of population models is however lagging behind, and accuracy of the models obtained is often assessed on a semi-empirical basis. In this paper we study the problem of validating models of gene network dynamics for cell populations, providing statistical tools for qualitative and quantitative model validation and comparison, and guidelines for their application and interpretation based on a real biological case study.

Keywords: Statistical methods, System Biology, Stochastic modelling, Mixed-Effects modelling, Gene expression

1. INTRODUCTION

Modern experimental techniques for the monitoring of gene expression at the individual cell level provide both evidence of gene expression variability, and quantitative data that can be exploited to describe and analyze variability from a mathematical standpoint (Elowitz et al., 2002; Neuert et al., 2013). Various approaches to the modelling of gene expression variability within and across cells have been developed, along with methods for their inference from experimental data, and applied to the study of real biological systems (Munsky et al., 2009; Zechner et al., 2012; Neuert et al., 2013; Llamosi et al., 2016). Yet, the quality of these models is often difficult to assess, due to the inherent complexity of the models as well as the challenges and costs involved in conducting validation experiments. Model assessment is mostly performed on empirical bases, such as qualitative response shape (Munsky et al., 2009; Zechner et al., 2012), overexpression or knock-out experiments (Cantone et al., 2009), and so on, whereas quantitative predictive capabilities are largely unexplored. The aim of this work is to introduce systematic approaches for the validation of mathematical models of cellular response variability. We are interested in particular in population modelling, i.e. the ability to account for response variability across different cells. Validation methods that will be considered shall emphasize the predictive capabilities of the models, i.e. the ability to correctly anticipate the true system response in new and possibly different experimental conditions. For parametric models, in particular, this rules out approaches based on the analysis of estimated parameter, because parameter inaccuracies are hardly related with predictive capabilities in the common scenario where practical identifiability issues arise (Gutenkunst et al., 2007). For practical utility, methods should be applicable with no further effort by modellers. We will therefore restrict to general validation tools, avoiding to leverage specificities of the different modelling approaches.

We will start by reviewing Mixed-Effects (Lavielle, 2015) and Chemical Master Equation (El Samad et al., 2005) modelling, two somewhat complementary approaches to population modelling that represent well the variety of modelling approaches currently proposed in the literature. We will also summarize the more traditional Mean-Cell modelling, for comparison purposes. Based on simulation of a biological case study, we will infer these models from *in silico* generated data and use them as a running example to introduce and discuss several validation methods derived from the statistical literature. We will illustrate their application for the evaluation of individual models as well as for model comparison, showing that reliable conclusions can be drawn from the ensemble of validation results rather than from the application of a single tool.

2405-8963 © 2016, IFAC (International Federation of Automatic Control) Hosting by Elsevier Ltd. All rights reserved.
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2016.12.112

2. POPULATION MODELS FOR GENE EXPRESSION DYNAMICS

Gene expression dynamics are generally given in terms of a biochemical reaction network operating in a uniform volume, a convenient abstraction of a cell (or a portion of it, e.g. the nucleus). Such a network is then simply characterized by n species, m reaction channels, and a stoichiometry matrix ν with n rows and m columns, each column describing the net change in copy number of the n molecular species over the whole reaction volume when the corresponding reaction takes place. Let $x = [x_1, \ldots, x_n]^T$ denote the amount of molecules of every species. Network dynamics are then fixed by the reaction rates $v(x, \psi)$, an m-dimensional column vector whose entries quantify the velocity at which different reactions take place. As apparent from the notation, $v(x, \psi)$ generally depends on the amount of molecules present in the reaction volume, and on kinetic rate parameters that are typically unknown or only partially known, and need to be determined from experimental data. In more generality, reactions may depend on (possibly time-varying) exogenous variables u affecting rates (e.g. a control signal), in which case we write $v(x, u, \psi)$.

Population models aim at applying this general paradigm to the description of multiple entities (cells) that, despite identical in principle and hence obeying the same model structure, show different responses. Several approaches may be considered, further detailing the meaning of x, ψ and v, as reviewed below.

Mean-Cell (MC) modelling. This approach aims at describing some "typical" behavior of a cell. For a given species abundance x_0 at a time t_0, a deterministic response model for the abundances $x(t)$ at all times t is sought. Under appropriate assumptions on reaction volume and species abundance, allowing in particular to treat $x(t)$ as species concentrations, the entries of $v(x, u, \psi)$ admit the interpretation of (deterministic) number of reaction occurrences per unit time, and are determined by the laws of mass action. In addition, $x(t)$ obeys

$$\dot{x}(t) = \nu v\big(x(t), u(t), \psi\big) \tag{1}$$

with $x(t_0) = x_0$. When confronted with population-average data, x is interpreted as a vector of average concentrations across the cell population, and ψ are considered as typical kinetic parameters. In the context of population modelling, where single-cell profiles are generally different from one another, the solution of (1) is rather interpreted as "mean-cell" dynamics, an oversimplification of the ensemble of single-cell responses. Single-cell measurements are then described as

$$y_i(t) = f\big(t, u(\cdot), x_0, \psi\big) + \mathrm{error}_i$$

where f is determined by the solution of the above ODE for given parameters ψ and initial conditions x_0 under $u(\cdot)$, while error_i accounts for the discrepancy between the mean-cell response f and the response of the ith of N cells, as well as for measurement noise.

Inference of MC models can be addressed by Maximum Likelihood (ML). Suppose that, for every cell $i = 1, \ldots, N$, measurements $\mathcal{Y}_i = \{y_{i,j} = y_i(t_j) : j = 1, \ldots, T_i\}$ are collected at times $\mathcal{T}_i = \{t_{i,j} : j = 1, \ldots, T_i\}$, and denote with $\mathcal{Y}$ the complete dataset. Consider a generic measurement model of the type

$$y_{i,j} = f\big(t_j, u(\cdot), x_0, \psi\big) + h\big(f\big(t_j, u(\cdot), x_0, \psi\big), \epsilon\big)\eta_i(t_j) \tag{2}$$

where errors $\eta_i(t_j) \sim \mathcal{N}(0, 1)$ are mutually independent across i and j, and ϵ are parameters of the noise distribution. Note that h plays the role of error standard deviation, which may be affected in different ways from the current system state. Denoting $\theta = (\psi, \epsilon)$ the set of unknown parameters (possibly including x_0), the ML estimate of θ may be computed by minimizing its negative log-likelihood given $\mathcal{Y}$, i.e., for $f_j(\theta) = f\big(t_j, u(\cdot), x_0, \psi\big)$ and $h_j(\theta) = h\big(f_j(\theta), \epsilon\big)$,

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{N} \sum_{j=1}^{\mathcal{T}_i} \left\{ \frac{1}{2} \left(\frac{y_{i,j} - f_j(\theta)}{h_j(\theta)} \right)^2 + \log h_j(\theta) \right\}.$$

Mixed-Effects (ME) modelling. An alternative approach is to assume that (1) models the individual cell, but different cells may be characterized by different values of ψ. If ψ_i denotes the parameters of the ith cell, one then assumes that

$$y_i(t) = f\big(t, u(\cdot), x_0, \psi_i\big) + \mathrm{error}_i, \quad \text{(individuals model)}$$

where $f\big(t, u(\cdot), x_0, \psi_i\big)$ is the solution of (1) with $\psi = \psi_i$, and error_i accounts for the inaccuracy in modelling single-cell response (and measurement noise). Here x is thought of as concentrations in the relevant cell, and $v(x, u, \psi_i)$ the velocity of reactions in cell i for given intracellular concentrations, while u is common across the population. Mixed-effects modelling enforces the idea of a cell being a variant of a statistically homogeneous population by introducing a common prior on parameters ψ_i,

$$\psi_i = d(a_i, \mu, b_i),\ b_i \sim \mathcal{N}(0, \Omega), \quad \text{(population model).}$$

The entries of the parameter vector μ, common to the whole population, are called fixed-effects. Vectors b_i are mutually independent and contain the random effects, i.e. individual cell discrepancies from the population average. Finally a_i are covariates representing cell-specific known features, if present.

Inference of mixed-effects models from individual data has the primary aim of reconstructing the population properties μ and Ω from the whole dataset $\mathcal{Y}$ of all measurements from all individuals. Consider again a generic measurement model of the form (2), where ϵ is fixed across individuals and ψ is replaced by ψ_i. A statistically powerful approach is provided by Population Likelihood Maximization (PLM). The idea is to leverage all data $\mathcal{Y}$ at once by maximizing with respect to $\Theta = (\mu, \Omega, \epsilon)$ the marginal likelihood $p(\mathcal{Y}|\Theta) = \prod_{i=1}^{N} \int d\psi_i p(\mathcal{Y}_i|\psi_i, \epsilon) p(\psi_i|\mu, \Omega)$, where factorization occurs thanks to the mutual independence of the b_i and of the η_i. By this approach, a single estimate is obtained for all population parameters, including ϵ. From the resulting estimates $\hat{\mu}$ and $\hat{\Omega}$, single-cell parameter estimates $\hat{\psi}_i$ may also be computed, e.g., by maximizing the empirical posterior $p(\psi_i|\mu = \hat{\mu}, \Omega = \hat{\Omega})$. In practice, while all integrands can be written explicitly, no closed form expression exists in general for $p(\mathcal{Y}|\Theta)$. Numerical methods for approximate PLM have been proposed (notably NONMEM (Bauer et al., 2007) and SAEM (Delyon et al., 1999; Bauer et al., 2007)) and are contained in dedicated software packages such as Monolix (Lixoft, 2014).