Inheritance and variability of kinetic gene expression parameters in microbial cells: modeling and inference from lineage tree data

Aline Marguet, Marc Lavielle, Eugenio Cinquemani

▶ **To cite this version:**

Aline Marguet, Marc Lavielle, Eugenio Cinquemani. Inheritance and variability of kinetic gene expression parameters in microbial cells: modeling and inference from lineage tree data. Bioinformatics, 2019, 35 (14), pp.i586-i595. 10.1093/bioinformatics/btz378 . hal-02317115

HAL Id: hal-02317115
https://hal.science/hal-02317115v1

Submitted on 15 Oct 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Bioinformatics, 35, 2019, i586–i595
doi: 10.1093/bioinformatics/btz378
ISMB/ECCB 2019

OXFORD

Inheritance and variability of kinetic gene expression parameters in microbial cells: modeling and inference from lineage tree data

Aline Marguet[1], Marc Lavielle[2] and Eugenio Cinquemani[1,*]

[1]Univ. Grenoble Alpes, Inria, 38000 Grenoble, France and [2]Inria Saclay & Ecole Polytechnique, Palaiseau 91120, France

*To whom correspondence should be addressed.

Abstract

Motivation: Modern experimental technologies enable monitoring of gene expression dynamics in individual cells and quantification of its variability in isogenic microbial populations. Among the sources of this variability is the randomness that affects inheritance of gene expression factors at cell division. Known parental relationships among individually observed cells provide invaluable information for the characterization of this extrinsic source of gene expression noise. Despite this fact, most existing methods to infer stochastic gene expression models from single-cell data dedicate little attention to the reconstruction of mother–daughter inheritance dynamics.

Results: Starting from a transcription and translation model of gene expression, we propose a stochastic model for the evolution of gene expression dynamics in a population of dividing cells. Based on this model, we develop a method for the direct quantification of inheritance and variability of kinetic gene expression parameters from single-cell gene expression and lineage data. We demonstrate that our approach provides unbiased estimates of mother–daughter inheritance parameters, whereas indirect approaches using lineage information only in the post-processing of individual-cell parameters underestimate inheritance. Finally, we show on yeast osmotic shock response data that daughter cell parameters are largely determined by the mother, thus confirming the relevance of our method for the correct assessment of the onset of gene expression variability and the study of the transmission of regulatory factors.

Availability and implementation: Software code is available at https://github.com/almarguet/IdentificationWithARME. Lineage tree data is available upon request.

Contact: eugenio.cinquemani@inria.fr

Supplementary information: Supplementary material is available at *Bioinformatics* online.

1 Introduction

Gene expression variability in isogenic cell populations is known to play a fundamental role in population-level strategies such as bet-hedging, and to explain the existence of certain cellular regulatory patterns (Raj and van Oudenaarden, 2008). Modern experimental technologies allow for the dynamical monitoring of gene expression response in individual microbial cells. Whether in the form of population-snapshot data (Hasenauer *et al.*, 2011) or single-cell gene expression time profiles (Llamosi *et al.*, 2016), this data provides a wealth of information for the quantitative mathematical study of intrinsic and extrinsic gene expression noise.

Among the important sources of variability is gene expression response variability originated at cell division. It is well known that random partitioning of the material among mother and daughter cells contributes significantly to intercellular diversity (Huh and Paulsson, 2011a, b). Several studies have addressed the analysis of how gene expression variability arises along generations based on detailed models of the evolution of cellular constituents over dividing cells (García *et al.*, 2018; Johnston and Jones, 2015; Swain *et al.*, 2002; Thomas, 2017). On the other hand, the inverse problem of reconstructing models of inheritance and variability from single-cell gene expression profiles is extremely challenging and requires an adapted modeling approach.

In particular, lineage information, i.e. known parental relationships among the observed cells, provides invaluable information about inheritance and variability of phenotypic traits at cell division (Ferraro *et al.*, 2016; Hormoz *et al.*, 2016; Taheri-Araghi *et al.*, 2015). Despite this, most mathematical approaches for the

© The Author(s) 2019. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/35/14/i586/5529241 by guest on 15 October 2019

reconstruction of gene expression noise models from single-cell gene expression data treat cells as independent individuals (Hasenauer *et al.*, 2011; Komorowski *et al.*, 2009; Munsky *et al.*, 2009; Neuert *et al.*, 2013; Suter *et al.*, 2011; Waldherr, 2018; Zechner *et al.*, 2012, 2014). Exceptions are few (Feigelman *et al.*, 2016; Kuzmanovska *et al.*, 2017) and are discussed below. Although inheritance and variability at division can still be quantified by post-processing of individual-cell parameter estimates (Llamosi *et al.*, 2016), neglecting parental relationships at a modeling stage is bound to negatively affect reconstruction performance.

In this article, we develop a stochastic model for the evolution of gene expression dynamics along the generations of a cell population, and a method for the direct quantification of variability and inheritance at cell division. Our starting point is mixed-effects (ME) modeling of gene expression. In the ME approach, response variability over different individuals is captured by the variability of the parameters of a structurally identical response model. A population model describes these parameters as random outcomes of a common probability distribution estimated from the data (Dharmarajan *et al.*, 2019; Fröhlich *et al.*, 2018; Llamosi *et al.*, 2016). Crucially, individuals are assumed to be statistically independent. Here, we extend the ME framework by introducing a model that explicitly relates mother and daughter parameters in terms of an autoregressive (AR) process (Ljung, 1999), and formulate estimation of inheritance and variability at division as the identification of the AR process parameters. Then, we develop a direct identification method by extending the SAEM algorithm (Lavielle, 2015) in order to take lineage information and the AR model into the core of the inference procedure. By the nature of our framework, which we call autoregressive ME (ARME), the population distribution of the single-cell parameters also follows naturally.

Next, we apply our method to both *in silico* and *in vivo* experiments. Working *in silico*, we demonstrate the performance of ARME. We benchmark our direct method with the method in Llamosi *et al.* (2016), a state-of-the-art approach among the indirect approaches based on post-processing of individual-cell parameters (Ferraro *et al.*, 2016; Taheri-Araghi *et al.*, 2015). Most importantly, we show that ARME provides unbiased estimates of parameter inheritance from mother to daughter cells, whereas indirect methods systematically underestimate such inheritance. Then, we apply our approach to the *in vivo* measurements of osmotic shock response of Llamosi *et al.* (2016). We show that gene expression response parameters of daughter cells are inherited from mothers to an extent of about 60%, whereas only about 40% of their variability can be attributed to randomness at division. This significant degree of inheritance favors stability of protein concentration levels along a lineage and thus transcriptional memory, a topic of current interest (Ferraro *et al.*, 2016). In addition, the degree of inheritance is found to be roughly the same for all kinetic rates, supporting the conclusion that variability at division uniformly affects the different gene expression regulatory factors.

An approach relevant to ARME is proposed by Kuzmanovska *et al.* (2017), who develop a general Bayesian method for inference of cellular processes from lineage tree data, and demonstrate it on simulated models of different sort. We instead focus on modeling and analysis of inheritance and variability of gene expression kinetic parameters, and apply our methods on real data. Concerning the inference method, we avoid certain approximations used to simplify computation at the price of uncertain accuracy, and require no Bayesian prior on the parameters sought. Despite the theoretical possibility to cast our models into their framework, unfortunately, no software implementation is provided to compare estimation performance. A Bayesian, simulation-based approach is also proposed by Feigelman *et al.* (2016), aimed at model selection among different single-cell regulatory patterns. Different from our framework, a stochastic model for intrinsic gene expression noise is considered along with inheritance of the cellular state at division, whereas parameter variability across different cells is not part of their modeling and estimation methods. In particular, kinetic gene expression parameters are fixed over the entire lineage, which makes their approach inapplicable to our case.

Our work provides effective tools to study the onset of gene expression variability as well as the degree of conservation of parameters and expression levels along generations. Intrinsic noise and parameter fluctuations within the lifespan of a cell are instead very marginally considered here. Although important in general (Swain *et al.*, 2002), their detailed modeling is not crucial for the focus of this work. Provided straightforward generalizations or adaptations, our methods are well suited to the study of many cellular processes for which inheritance and variability at cell division are of concern.

The article is organized as follows. In Section 2, we introduce and discuss the ARME modeling framework. In Section 3, we state the relevant identification problem from lineage tree data and describe our new inference algorithm. In Section 4, we demonstrate the effectiveness of the approach *in silico*, also providing some hints toward experimental design in presence of lineage information. In Section 5, we apply our modeling and inference methods to *in vivo* osmotic shock gene expression data from yeast. Discussion and conclusions are in Section 6.

2 Gene expression modeling over a lineage tree

In this section, we discuss modeling of gene expression dynamics for individual cells that are subject to parental relationships. An illustration of this scenario is given in Figure 1. As a starting point, we rely on the standard approach where the expression dynamics of a given gene of interest are described by the couple of differential equations

$$\begin{cases} \dot{m}(t) &= k_m u(t) - g_m m(t), \\ \dot{p}(t) &= k_p m(t) - g_p p(t), \end{cases} \tag{1}$$

where t denotes time, while $m(t)$ and $p(t)$ denote, respectively, the concentration at time t of messenger RNA and protein molecules of the species encoded by the gene (de Jong, 2002; Llamosi *et al.*, 2016). The first equation describes mRNA transcription at rate $k_m u(t)$, with $u(t)$ the strength of promoter activation at time t, and degradation at rate $g_m m(t)$. The second equation describes translation of protein molecules from the available mRNA molecules at rate $k_p m(t)$, and degradation at rate $g_p p(t)$. Because of cell growth, g_m and g_p account for both biochemical degradation and growth dilution. We assume that $u(t)$ is controlled by a known exogenous stimulus, i.e. it is a known profile. For an individual cell, the above is a viable model as long as intrinsic noise is not dominant for the gene of interest. Stochastic versions of this model should be considered otherwise (Paulsson, 2005). In this work, we rather focus on how individual-cell parameters vary or are conserved across cells.

Rate parameters k_m, g_m, k_p and g_p depend on cell physiology (abundance of ribosomes and polymerase molecules, transcription factors, ...) and may typically differ from cell to cell. In Llamosi *et al.* (2016), a ME modeling approach was shown to be a viable description of this variability. Let $\psi^\nu = (k_m^\nu, g_m^\nu, k_p^\nu, g_p^\nu)$ denote the vector of parameters for the individual cell ν. In the ME approach, for every cell ν, ψ^ν is a random outcome from a common population distribution. Crucially, the different individuals ν, i.e. the different random variables ψ^ν, are assumed to be mutually statistically

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/35/14/i586/5529241 by guest on 15 October 2019

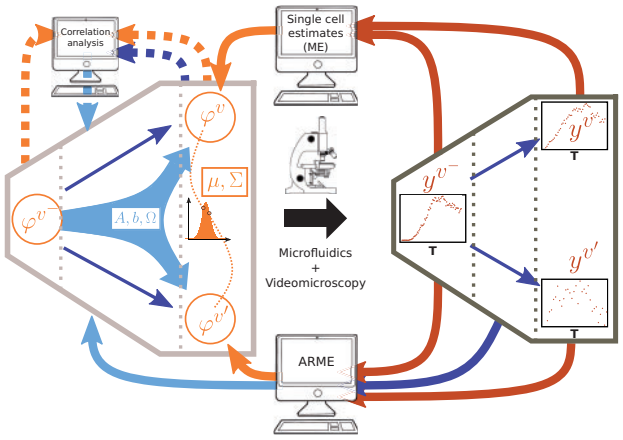

Fig. 1. ARME versus indirect approaches for the estimation of inheritance and variability of gene expression parameters. Left: modeling of single-cell parameters as well as their variability and inheritance across cell division; Right: experimental measurement of gene expression profiles in the same single cells. Orange circles represent cells, straight blue arrows represent the known parental relationships among them (lineage data). The inference problem considered in this article is to reconstruct variability and inheritance dynamics (cyan double-arrow, left) of the single-cell parameters (φ^{ν^-}, φ^{ν}, $\varphi^{\nu'}$; orange, left) from gene expression data (y^{ν^-}, y^{ν} and $y^{\nu'}$; red dots, right) and the known parental relationships. Data processing flow from right to left represents utilization of single-cell data (red arrows) and lineage information (blue arrows) to produce estimates of individual-cell parameters and statistics (orange arrows) as well as of their variability and inheritance dynamics at cell division (cyan arrows). ARME (bottom) is a direct method that, based on explicit modeling of variability and inheritance dynamics, uses single-cell data together with lineage information to estimate the variability and inheritance parameters (A, b, and Ω) at once. Estimates of single-cell parameters and of their statistics (μ and Σ; orange, left) are also obtained as a byproduct. Indirect (e.g. ME based) methods (top), instead, only use individual-cell data to provide estimates of individual-cell parameters and their statistics in a first step. Based on the individual-cell parameter estimates from the first step and lineage information, estimates of inheritance dynamics are produced in a second step

independent. For an osmotic shock-responsive gene in yeast, Llamosi *et al.* (2016) found that statistical independence does not hold for cells in a parental relationship, notably for mother–daughter cell couples. The observed correlation is not surprising, since one expects cell offspring to inherit the physiological state of the parents at least in part. Toward in-depth investigation of this inheritance, we introduce a dedicated statistical framework that is a generalization of ME modeling, as described below and illustrated in Figure 1.

Let us consider $\varphi^{\nu} = \log(\psi^{\nu})$, the log-domain version of the positive rate parameters ψ^{ν} (the reason will be clarified below). Treating $\varphi^{\nu} \in \mathbb{R}^d$ as a column vector (d being the number of individual-cell parameters), we introduce the (first-order) AR model (Ljung, 1999)

$$\varphi^{\nu} = A\varphi^{\nu^-} + (I - A)b + \eta^{\nu}, \tag{2}$$

where ν^- is the direct ancestor of ν, $A \in \mathbb{R}^{d\times d}$, $b \in \mathbb{R}^d$, I denotes the (size-d) identity matrix and η^{ν} is a random variable from a distribution $\mathscr{F}$ independent of ν, with mean zero and size-d covariance matrix Ω. We additionally assume that the random variables η^{ν} are independent across different individuals ν and of φ^{ν^-}. Notice that working with the log-domain parameters φ^{ν} ensures by construction the positivity of the ψ^{ν}. For different values of A, this model expresses the extent to which the offspring parameters φ^{ν} are determined by (inherited from) the parent parameters φ^{ν^-}, or are the result of the randomness brought about by η^{ν}, if not simply of a baseline population value fixed by b. The inheritance of the different entries of φ^{ν^-}, which are different in nature, is duly represented by a diagonal matrix A, whereas a non-diagonal Ω is well suited to capture the onset of statistical dependencies across different entries of φ^{ν}, as e.g. due to global extrinsic regulatory effects. Of course, the model is suited to represent cell division, since two daughters, say ν and ν', may well correspond to the same common parent ν^-, and yet be different as a result of the two independent random quantities η^{ν} and $\eta^{\nu'}$.

Model (2) qualifies $\varphi = (\varphi^{\nu})_{\nu\in V}$ as a stochastic process over V. To further specify the model, we assume that φ^{ν} is in a (weakly) stationary regime. In particular, we assume that mean and covariance of φ^{ν} is the same for all individuals. This assumption is consistent as long as A is Schur-stable (all eigenvalues within the unit circle) (Ljung, 1999). From a biological viewpoint, it represents a form of structural invariance of the system within the time span of interest. In this case, it is easily shown that mean $\mu = \mathbb{E}\varphi^{\nu}$ and covariance matrix $\Sigma = \mathrm{Var}(\varphi^{\nu})$ obey

$$\mu = b, \qquad \Sigma = A\Sigma A^T + \Omega. \tag{3}$$

Therefore, b in (2) fixes the mean of φ^{ν}, whereas Σ depends on A (inheritance matrix) and Ω (covariance matrix of the random component η^{ν}). Moreover, the cross-covariance matrix $\Xi = \mathrm{Cov}(\varphi^{\nu}, \varphi^{\nu^-})$ obeys

$$\Xi = A\Sigma. \tag{4}$$

Thus, normalized by the variance Σ, A plays the role of the (matrix) correlation coefficient between φ^{ν} and φ^{ν^-}. For diagonal A, the closer the diagonal entries to 0 (respectively, to 1), the smaller (resp. the larger) the extent to which daughter cell are determined by mother cell parameters. As a generalization of Equation (4), one finds that the covariance between a given cell and its descendants ℓ generations ahead is given by $A^{\ell}\Sigma$. Thus the model reasonably predicts that two cells are correlated even if none is the daughter of the other. Yet, because of the strict stability of A, correlation fades away along generations. Note that, for the special case $A = 0$ (no inheritance from ν^- to ν), a standard ME model $\varphi^{\nu} = b + \eta^{\nu}$ is recovered, with a fixed term b and random terms η^{ν}, independent across ν, sampled from a common distribution $\mathscr{F}$ with mean zero and covariance matrix $\Sigma = \Omega$. Therefore, our model generalizes ME models by including a variety of possible mother–daughter dependencies ($A \neq 0$).

In summary, the proposed model of gene expression over a population of dividing cells is the combination of model (1) with parameters evolving in accordance with model (2). In a compact form, for any given cell ν, we may rewrite (1) as

$$\dot{x}(t) = F(\varphi^{\nu})x(t) + G(\varphi^{\nu})u(t), \quad x(t_0^{\nu}) = x_0^{\nu}, \tag{5}$$

where the state vector x comprises concentrations m and p, with obvious definition of matrices F and G in terms of parameters φ^{ν}. Vector x_0^{ν} is the initial state of cell ν at its birth time t_0^{ν}. For any $t \geq t_0^{\nu}$, we denote the solution of (5) by $x^{\nu}(t)$. Note that we express all time variables relative to a universal time reference independent of the individual cell. To complete the model, we assume that the daughter cell state at birth is fixed by the mother state at the same time, i.e. $x_0^{\nu} = x^{\nu^-}(t_0^{\nu})$. The resulting model (2), (5) is a description of gene expression over a tree of dividing cells that is stochastic due to the randomness affecting daughter cell parameters at cell division. As such, it can also be interpreted as a model for extrinsic noise (Swain *et al.*, 2002), where kinetic gene expression parameters fluctuate at the time scale of cell division. It naturally accommodates

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/35/14/i586/5529241 by guest on 15 October 2019

several population trees evolving in parallel, as e.g. experimentally observed in microscopy experiments starting form several cells at experimental time 0. By straightforward modifications, the model can be generalized to more complex gene expression dynamics [e.g. inclusion of a protein maturation step in (5)], random inheritance of mRNA and protein concentrations, asymmetric division (e.g. for budding, the mother cell v^- keeps its parameters after generating daughter cell v), and different known inputs affecting different cells (u^v in place of u). Some of these extensions will be used and commented in our application to real data in Section 5. In view of the fact that our approach includes ME modeling as a special case, in the sequel, we will refer to it as ARME modeling of gene expression.

3 Identification from lineage tree data

On the basis of the ARME modeling developed in Section 2, we consider the problem of estimating inheritance dynamics of gene expression over a growing population of cells. Our problem statement stems from but is not limited to videomicroscopy of cells carrying fluorescent reporters, where quantitative individual-cell expression profiles as well as mother–daughter relationships can be established by suitable image processing.

Over an experimental time period $[0, T]$, we assume that gene expression measurements y_j^v are available for individual cells v at cell-dependent time instants t_j^v, with $j = 1, \ldots, n^v$. For every cell, an initial time t_0^v is also available, such that $t_0^v \leq t_j^v$ for all j. Crucially, we assume that parental relationships are available, i.e. a family of pairs of the type (v^-, v) expressing the fact that v has been generated from v^- at time t_0^v. We assume that, for every cell

$$y_j^v = Cx^v(t_j^v) + h\varepsilon_j^v, \qquad j = 1, \ldots, n^v, \tag{6}$$

where $x^v(t_j^v)$ represents the state of cell v at time t_j^v, matrix C selects the components of x^v that are experimentally measured (typically, the protein concentration $p(t)$ or an associated reporter fluorescence), and $h\varepsilon_j^v$ represents random measurement error with standard deviation $h > 0$, where the random variables ε_j^v are assumed of mean zero and unitary variance, independent across j and v and independent of $x^v(t_j^v)$. For true parameters φ^v and initial conditions x_0^v, $x^v(t_j^v)$ is the solution of (5).

Let us denote by V the set of observed cells v, by $Y^v = \{y_j^v : \ j = 1, \ldots, n^v\}$ the measurements for cell $v \in V$, and by $Y = \{Y^v : \ v \in V\}$ the set of all measurements from all cells. Finally, let us denote by $W = \{(v^-, v)\} \subset V \times V$ the set of known mother–daughter relationships. The ARME identification problem that we address is the reconstruction of parameters $\theta = (A, b, \Omega, h)$ from Y and W.

An indirect way to address the problem above is to fit individual-cell parameter values to the data and then, in the light of W, infer parameters θ from the individual-cell parameter estimates $\hat{\varphi}^v$ (Llamosi *et al.*, 2016). In particular, regardless of the inheritance model (2), the (matrix) correlation coefficient A can be defined as the normalized covariance $\mathrm{Cov}(\varphi^v, \varphi^{v^-})\mathrm{Var}(\varphi^v)^{-1}$. Thus, provided a family of individual-cell estimates $\hat{\Phi} = \{\hat{\varphi}^v : \ v \in V\}$ and of mother–daughter pairs W, an estimate of A can be computed as $\hat{\Xi}\hat{\Sigma}^{-1}$, with

$$\hat{\Sigma} = \frac{1}{|V|}\sum_{v \in V}(\hat{e}^v)(\hat{e}^v)^T, \qquad \hat{\Xi} = \frac{1}{|W|}\sum_{(v^-, v) \in W}(\hat{e}^v)(\hat{e}^{v^-})^T, \tag{7}$$

where $\hat{e}^v = \hat{\varphi}^v - \hat{b}$ and $\hat{b}$ is the empirical mean of the individual-cell parameter estimates $\hat{\Phi}$ (e.g. Ljung, 1999; Papoulis, 1991). Analogous empirical estimates can be constructed for the other entries of θ. In turn, individual-cell estimates can be drawn by direct fit of the corresponding cell measurements, or with more advanced methods such as ME identification (Llamosi *et al.*, 2016), as explained shortly. In so doing, however, the inheritance dynamics described by Equation (2) is ignored and the lineage information W is used only in *a posteriori* statistical analysis. The method we develop below instead exploits W and model (2) in conjunction to provide direct estimates of θ from all data Y. It is known from standard ME scenarios that such holistic approaches lead to better estimation performance (Lavielle, 2015). However, standard ME identification assumes independence of individuals and estimates population statistics μ and Σ along with single-cell parameters. From this, estimates of θ can only be computed by the indirect method above. ARME identification instead computes estimates of θ first. From these, in view of Equation (3), estimates for the population parameters μ and Σ follow immediately. For W empty and A fixed to zero, in particular, our method includes ME identification as a special case. We refer to our method as ARME identification and develop it in the next section. A comparison between ARME and indirect (ME) approaches is shown in Figure 1.

3.1 ARME identification

Let $\varphi = (\varphi^v)_{v \in V}$ denotes the collection of all individual-cell parameters. Our approach relies on maximum likelihood (ML) estimation. If $p(Y|W, \theta)$ denotes the probability density of observations Y for putative parameters θ in the light of dependencies W, we define our estimator $\hat{\theta} = (\hat{A}, \hat{b}, \hat{\Omega}, \hat{h})$ of θ as

$$\hat{\theta}(Y, W) = \arg\max_{\theta \in \Theta} L(\theta|Y, W), \qquad L(\theta|Y, W) = \log p(Y|W, \theta),$$

where Θ is a suitable parameter search space. To fully determine the expression of the log-likelihood $L(\theta|Y, W)$, the second-order (mean and variance) description of the random variables η^v and ε^v provided so far does not suffice. To cope with this, from now on, we will fix $\mathscr{F}$ (the distribution of the η^v) to be the multivariate normal $\mathcal{N}(0, \Omega)$. Likewise, we take $\varepsilon^v \sim \mathcal{N}(0, \mathscr{I})$.

Evaluating and maximizing $L(\theta|Y, W)$ over θ is challenging. To achieve this, we rely on the fact that the ARME model of Section 2 is hierarchical. Individual-cell parameters play the special role of hidden variables, i.e. variables whose knowledge would allow one to evaluate the individual-cell likelihoods. A classical algorithm used to seek ML parameter estimates for a model of this type is the expectation-maximization (EM) algorithm. This is an iterative approach where estimates of θ available at iteration k, say $\hat{\theta}_k$, are updated by a two-step procedure. Formally, in our case, these two steps are:

- E-step: compute $Q(\theta, \hat{\theta}_k) := \mathbb{E}_{\varphi|Y, W, \hat{\theta}_k}[\log(p(Y, \varphi|W, \theta))]$;
- M-step: update $\hat{\theta}_{k+1} = \mathrm{argmax}_\theta Q(\theta, \hat{\theta}_k)$.

Notably, the E-step brings to surface and leverages the role of hidden variables φ. However, two main limitations affect this method, possible convergence to local maxima and the typical lack of an expression for the expectation. To address both concerns, Delyon *et al.* (1999) developed a provably convergent randomized version of this method for ME models called stochastic approximation EM (SAEM). Here, we develop a non-trivial extension of SAEM in order to cope with cell-to-cell correlations introduced by the inheritance dynamics model (2).

The rationale of SAEM is to replace the E-step above by the random sampling of the parameters φ in accordance with their currently estimated distribution, which results in intertwining the E-step

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/35/14/i586/5529241 by guest on 15 October 2019

with the M-step along the iterations. The algorithm consists in three steps:

- S-step: simulate φ_{k+1} according to $p(\varphi|Y,W,\hat{\theta}_k)$;
- E-step: compute $Q_{k+1}(\theta)=Q_k(\theta)+\gamma_k(\log(p(Y,\varphi_{k+1}|W,\theta))-Q_k(\theta))$;
- M-step: update $\hat{\theta}_{k+1}=\text{argmax}_\theta Q_{k+1}(\theta)$,

where γ_k is a tunable forgetting factor that trades speed of convergence for exploration of the search space. Typically, the dependence of γ_k on the iteration k is exploited to have a first phase of broad search of the parameter space, followed by a phase that smoothens out the search and stabilizes it around the region of the final optimum. After stabilization of $\hat{\theta}_k$, for k large enough, estimate $\hat{\theta}$ is set equal to $\hat{\theta}_k$. (A discussion about choice of forgetting factor and termination criterion is reported in Supplementary Section S1.5.) The new E-step is simple (Supplementary Section S1.1). A modified version of it also allows for computation of confidence intervals (Supplementary Section 1.4). The M-step can be performed e.g. by numerical optimization (Supplementary Section S1.2). Step S is the most critical. The conditional distribution $p(\varphi|Y,W,\theta)$ is itself unknown. In addition, contrary to ME identification, it cannot be factored out into individual-cell distributions due to cell-to-cell correlation in the model. Our implementation of the S-step, which is of key importance in the ARME framework, is described next.

3.2 Metropolis-Hastings implementation of the S-step

Inspired by Kuhn and Lavielle (2004); Lavielle (2015), we implement the S-step by a Markov Chain Monte Carlo (MCMC) approach based on Metropolis-Hastings (MH) rejection sampling. In what follows, we describe how to get φ_{k+1} for one execution k of the step, and omit k from the notation for simplicity. A Markov chain $(\varphi_j)_{j\in\mathbb{N}}$ is formed by proposing a new candidate $\tilde{\varphi}_{j+1}$ from the current state φ_j of the chain by a random draw from a suitable proposal distribution $q(\varphi_j,\tilde{\varphi}_{j+1})$. Candidate $\tilde{\varphi}_{j+1}$ is accepted as the new chain state with probability

$$\min\left\{1,\frac{p(\tilde{\varphi}_{j+1}|Y,W,\theta)}{p(\varphi_j|Y,W,\theta)}\frac{q(\tilde{\varphi}_{j+1},\varphi_j)}{q(\varphi_j,\tilde{\varphi}_{j+1})}\right\}.$$

If accepted, one sets $\varphi_{j+1}=\tilde{\varphi}_{j+1}$, otherwise one sets $\varphi_{j+1}=\varphi_j$. Convergence of this chain to the distribution sought (i.e. $p(\varphi|Y,W,\theta)$) can be formally proven and practically checked (Lavielle, 2015). For both $\varphi=\tilde{\varphi}_{j+1}$ and $\varphi=\varphi_j$, one may compute factors $p(\varphi|Y,W,\theta)$ above in terms of the likelihood $p(Y|\varphi,W,\theta)$. In turn, the latter can be evaluated easily using (5)–(6) for the given single-cell parameters φ_j. In view of the linearity of (5), this solution can also be implemented explicitly.

The success of this approach depends on the choice of the proposal distribution q for the update of the Markov chain. Ideally, q should be similar to $p(\varphi|Y,W,\theta)$. Importantly, this choice determines the acceptance rate of the sample candidates and thus the efficiency of the procedure. In ARME, contrary to standard ME identification (Kuhn and Lavielle, 2004), the MCMC procedure above cannot be separated out into smaller problems due to cell-to-cell correlation. Yet, due to the high dimension of the cell tree, using a single proposal q for the joint distribution $p(\varphi|Y,W,\theta)$ leads to overly small acceptance rate and thus poor performance.

To address this issue, we implement a hierarchical proposal sampling method that combines a joint population-level proposal with individual-level proposals. Specifically, we consider three proposal distributions: A population-level proposal q_1, a per-generation proposal q_2 and an individual proposal q_3, with expressions

$$\begin{aligned}
q_1(\varphi,\tilde{\varphi}) &= p(\tilde{\varphi}|\theta)\propto e^{-\frac{1}{2}(\tilde{\varphi}_\varnothing-\mu)^T\Sigma^{-1}(\tilde{\varphi}_\varnothing-\mu)}\times \\
&\qquad e^{-\frac{1}{2}\sum_{v\in V}\eta(\tilde{\varphi}^v,\tilde{\varphi}^{v^-})^T\Omega^{-1}\eta(\tilde{\varphi}^v,\tilde{\varphi}^{v^-})},\\
q_2^{(v)}(\varphi^v,\tilde{\varphi}^v) &= p(\tilde{\varphi}^v|\varphi^{v^-},\theta)\propto e^{-\frac{1}{2}\eta(\tilde{\varphi}^v,\varphi^{v^-})^T\Omega^{-1}\eta(\tilde{\varphi}^v,\varphi^{v^-})},\\
q_3^{(v)}(\varphi^v,\tilde{\varphi}^v) &\propto \exp\left(-\frac{(\tilde{\varphi}^v-\varphi^v)^2}{2\sigma^2}\right),
\end{aligned}$$

with $\eta(\tilde{\varphi}^v,\varphi^{v^-})=\tilde{\varphi}^v-(A\varphi^{v^-}+(I-A)b)$, where μ and Σ are fixed by θ via (3). Proposal q_1 is for the joint distribution $p(\varphi|Y,W,\theta)$ and has low acceptance rate. On top of that, iteratively along generations, q_2 is used to make proposals about any individual v of a given generation given the proposed parameters of its ancestor v^-. For the root of the population tree, q_2 is modified into $q_2^{(v)}(\varphi^v,\tilde{\varphi}^v)=p(\tilde{\varphi}^v|\theta)\propto e^{-\frac{1}{2}(\tilde{\varphi}^v-\mu)^T\Sigma^{-1}(\tilde{\varphi}^v-\mu)}$. Finally, separately for every cell v, q_3 allows for local exploration of the cell parameter vector by a random walk in the parameter space which iteratively steps from a current value φ^v to a new random value $\tilde{\varphi}^v$. The standard deviation σ of the step size is chosen adaptively in order to ensure a satisfactory acceptance rate around 0.3 throughout iterations (Lavielle, 2015, Section 9.3). The overall implementation of our MH algorithm results from alternating the usage of these proposal distributions for the generation of the candidate chain samples φ_{j+1}, and propagating changes in the resampled individual parameters to the descendants along the tree. Further technical details are given in Supplementary Section S1.5.2. By the same MCMC approach, single-cell parameter estimates $\hat{\varphi}^v$ can also be obtained (Supplementary Section S1.3).

4 Applications to *in silico* experiments

In this section, we apply our ARME inference method to simulated gene expression data over a lineage tree. We first validate the method in Section 4.1, showing convergence of estimates to the true parameters $\theta=(A,b,\Omega,h)$ as well as the ability to recover individual-cell dynamics. Then, in Section 4.2, we show that our method outperforms existing approaches to estimate mother–daughter relationships. In this analysis, we will consider symmetric division, whereby mother cells generate and are replaced by two newborn daughter cells, each with its own parameters inherited from the mother with additional variability.

4.1 Illustration and validation of estimation approach

In order to test the validity of our ARME identification algorithm, we consider a scenario where gene expression data are collected from individual cells over seven generations subjected to a common perturbation profile u that alternates periods of promoter induction ($u=1$) to periods of lack of induction ($u=0$). In view of later application of the method to the real data from Llamosi *et al.* (2016), both this perturbation profile and the simulated parameters of the model are mostly taken from the same work, where mean values for single-cell parameters were fixed based on literature search and refined based on the data [time units are minutes (min), while the unit for parameters k_m, k_p, g_m and g_p are $(\text{min})^{-1}$].

For identification purposes, in absence of information about the unobserved variable $m(t)$, parameters k_m and k_p of model (1) are indistinguishable from single-cell data, i.e. only their product matters (Llamosi *et al.*, 2016). Accounting for these parameters as separate entities may cause practical issues as well as erroneous interpretation of the results. Without loss of generality, in agreement with existing literature (Llamosi *et al.* (2016) and references therein), we,

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/35/14/i586/5529241 by guest on 15 October 2019

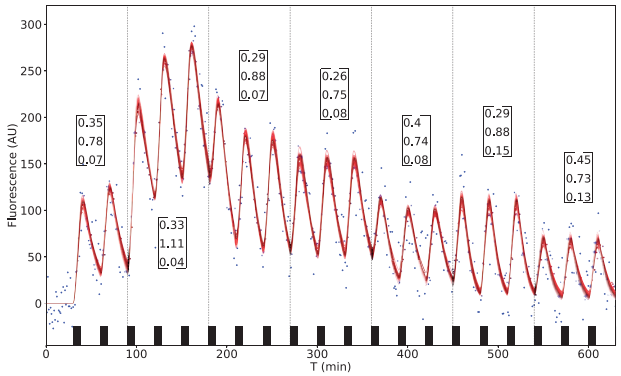

Fig. 2. Single-cell fits of *in silico* gene expression data based on the ARME identification. Results shown are for the seven cells along one branch of one simulated tree spanning seven generations (at every cell division, only one of the two daughter cells is displayed at subsequent times; all branches are statistically similar). Vertical dashed lines indicate cell division times. Black line: true simulated protein profiles; blue dots: noisy protein concentration measurements; and red lines: 30 simulated single-cell trajectories corresponding to single-cell parameters sampled from the posterior $p(\varphi^v|Y,W,\hat{\theta})$, where $\hat{\theta}$ are the parameters of the ARME model identified from data Y. Black bars: promoter activity u. For every cell, true parameter vectors $\psi^v=[g_m,k_p,g_p]$ that generated the data are displayed in square brackets

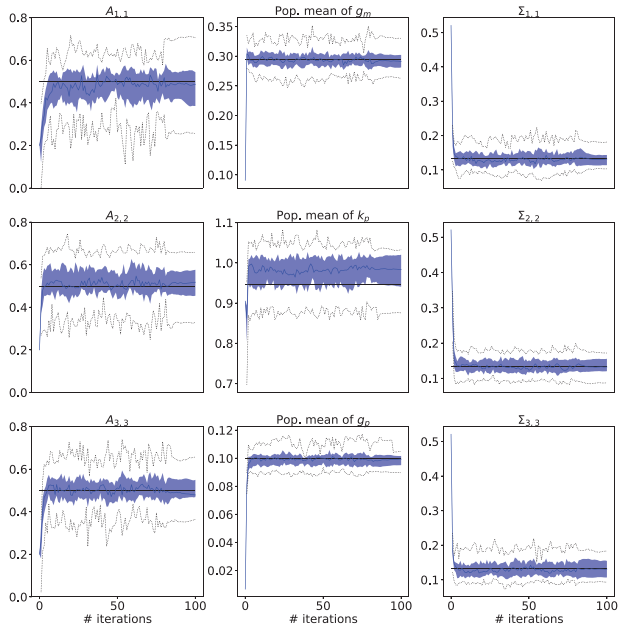

Fig. 3. Iterative ARME identification of parameters $A=\mathrm{Diag}(A_{1,1},A_{2,2},A_{3,3})$, b and $\Sigma=\mathrm{Diag}(\Sigma_{1,1},\Sigma_{2,2},\Sigma_{3,3})$ from the application of the algorithm to 20 simulated datasets Y for 80 search iterations plus 20 stabilizing iterations (100 iterations total). Identification is based on data simulated over seven generations with measurement noise level $h=20$. Horizontal black lines: true parameter values; blue lines: median of the iterative estimation profiles; and shaded blue region and dashed lines: at every iteration, 25% and 75% quantiles of the estimates over the 20 datasets, and extension of corresponding whiskers, as computed for the final parameter estimates in later boxplots (Fig. 4)

therefore, fix k_m to 10 $(\mathrm{min})^{-1}$ in both simulation and identification, focusing our analysis on the reduced parameter vector $\psi^v=(g_m^v,k_p^v,g_p^v)$ and the corresponding size-three parameters A, b and Ω.

To simulate artificial datasets, starting from a single uninduced cell at time 0 (Generation 1), we simulate division of every existing cell into two daughter cells every 90 min over seven generations, thus obtaining a full cell tree. Single-cell parameters are simulated on the basis of model (2), with parameters $A=\mathrm{Diag}(0.5,0.5,0.5)$, $b=[\log(0.294),\log(0.947),\log(0.1)]^T$ and $\Omega=\mathrm{Diag}(0.1,0.1,0.1)$. Gene expression dynamics of every cell are simulated in accordance with model (1). Here, we assume that, besides measurement noise, the reporter protein concentration p coincides with the observed fluorescence intensity. For every cell, we assume that measurements of $p(t)$ are taken every minute. Measurement noise is simulated by adding random Gaussian error with strength $h=20$, corresponding to a standard deviation in order of 10% of the simulated protein concentrations (as observed in real data). This simulation is repeated 20 times, each time with different cell parameters sampled from model (2) and different outcomes of measurement noise. Figure 2 reports an example of the simulated data from one of the 20 datasets.

To assess identification performance, the ARME algorithm of Section 3 is run on every dataset separately, yielding 20 iterative estimation profiles for the unknowns $\theta=(A,b,\Omega,h)$. In our non-optimized implementation in Julia (Bezanson *et al.*, 2017), one estimation run takes about 5 h on an Intel Xeon 3 GHz workstation. Statistics of the estimation process over the 20 datasets are shown in Figure 3. Estimated single-cell dynamics from the model identified on one dataset are shown in Figure 2.

From Figure 3, the first remark is the neat convergence of the iterative procedure around the true parameter values. Unreported results show that the observed convergence is robust to both initial parameter guesses and the randomness of the algorithm. Importantly, the iterative estimation sequences converge on average to the true values used in data generation, i.e. there is no estimation bias (the discrepancy in the estimation of the mean of k_p can be explained by a small sensitivity of the model around its true value and does not exceed the first and third quartiles, see further comments below). The variability of the estimates over the 20 datasets reflects variability in the data due to randomness in parameter inheritance and the realistically large measurement noise (Fig. 2). In general, estimation variability also depends on the richness of the dataset. To verify this, we repeated the same experiment with a higher number of observed cell generations (11 generations, Supplementary Fig. S2) and with a smaller measurement noise strength ($h=10$, Supplementary Fig. S1). Our algorithm converges nicely in all these cases, and the estimation uncertainty is decreased in both cases, as expected (the same Supplementary Figures also show that the discrepancy in the estimation of the mean of k_p observed in Fig. 3 disappears for richer datasets). Additional simulations show that convergence holds for different parameters, notably for a non-diagonal matrix Ω (Supplementary Fig. S3). This case corresponds to a non-diagonal matrix Σ, i.e. a more complex correlation structure among g_m, k_p and g_p. A validation study also shows that the identified model is not overfit and predicts well single-cell parameters of a synthetic validation dataset (Supplementary Section S5). Finally, application of our method to simulated data with various degrees of intrinsic noise shows that inference is robust to small intrinsic noise levels, while estimation uncertainty increases for larger intrinsic noise levels, as expected (Supplementary Section S4).

For the identification of the inheritance matrix A, estimation performance is expected to depend not only on the number of observed cells (i.e. $|V|$, the cardinality of set V) but also on the structure of dependencies W. In particular $|W|$, the number of mother–daughter pairs for which gene expression data is available, plays an important role (unrelated cells do not provide information about A). A preliminary study of this question shows that indeed, for an equal number of cells $|V|$, a larger set of dependencies $|W|$ favors estimation of A.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/35/14/i586/5529241 by guest on 15 October 2019

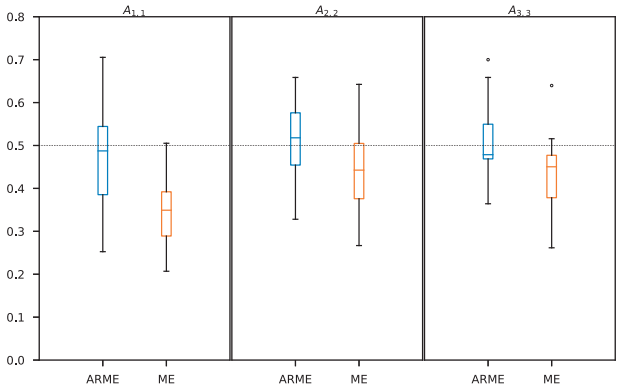

Fig. 4. Statistics of identification of inheritance parameters $A = \text{Diag}(A_{1,1}, A_{2,2}, A_{3,3})$ over 20 simulated datasets. Identification is based on data simulated over 7 generations with measurement noise level $h = 20$. For each parameter, we compare results from ARME identification and from the indirect method based on standard ME identification. Horizontal lines show the parameter values used in simulation

However, it may deteriorate estimation of b and Σ (see Supplementary Section S3 for more details).

4.2 Performance gain over indirect approaches

A key question at the basis of this work is whether full account of parameter inheritance at both modeling and inference level improves reconstruction of statistical mother–daughter dependencies. In this section, we demonstrate that this is the case by comparing ARME identification with the state-of-the-art indirect method in Llamosi *et al.* (2016), where estimates of the inheritance model parameters θ are built on top of ME identification. In a perfectly equivalent manner, we obtain this by running our ARME identification algorithm in the special case where A is fixed to 0, computing individual-cell parameter estimates as described in Supplementary Section S1.3, and then applying the correlation analysis as per Equation (7). To distinguish estimates based on standard ME from estimates based on our ARME approach, in what follows, we append superscript '^{ME}' to the estimates from the former.

We rely on the artificially generated datasets of the previous section. ARME estimates of θ are those of the previous section, whereas indirect ME estimates are obtained for every dataset as explained above. In Figure 4, we show boxplots of estimates $\hat{A}^{ME}$, and analogous boxplots for the ARME estimates $\hat{A}$. The difference is apparent. Estimates $\hat{A}$ are nicely centered around the true values and show little dispersion. On the contrary, despite a rather rich dataset, estimates $\hat{A}^{ME}$ are biased, a signature of poor estimation performance. Bias was also verified by non-parametric hypothesis testing. A sign test applied to $\hat{A}^{ME}_{1,1}$, $\hat{A}^{ME}_{2,2}$ and $\hat{A}^{ME}_{3,3}$ rejected the hypothesis that the estimate is centered around the true value 0.5 for all of them at 0.05 significance (with P-values $< 10^{-5}$, 0.0026 and $< 10^{-5}$, respectively), confirming bias, whereas the same test applied to the ARME estimates $\hat{A}_{1,1}$, $\hat{A}_{2,2}$ and $\hat{A}_{3,3}$ did not reject this hypothesis (P-values 0.50, 0.82 and 0.50). Bias of $\hat{A}^{ME}$ is also reconfirmed on other simulated datasets (Supplementary Figs S4 and S5). Importantly, this bias generally depends on the true values of the unknown parameters under estimation as well as noise strength and observed population size (same Supplementary Figures); therefore, it cannot be easily compensated for. The bias is negative, i.e. traditional approaches systematically underestimate the degree to which parameters are inherited from mother to daughter cells. This is easy

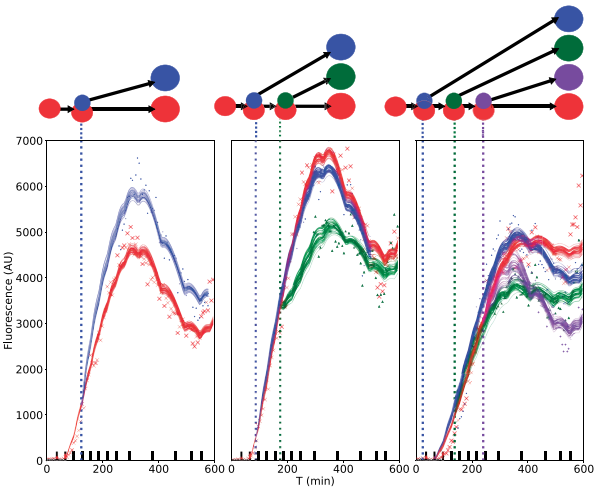

Fig. 5. Illustration of the cell dependencies observed in the yeast experimental data of Llamosi *et al.* (2016) (top) and corresponding single-cell data fits after ARME identification of parameters θ (bottom). In the dataset, 86 cells were monitored, out of which four were discarded after data curation. Observed cell dependencies W result in 15 pairs of one mother generating one daughter cell (top left), 12 triplets of one mother generating two daughter cells (top center) and four quadruples of one mother generating three daughter cells (top right). For each of these three cases, bottom plots provide an example of single-cell fits obtained as in Fig. 2 from ARME identification (30 profiles corresponding to 30 random draws of the individual-cell parameters from the relevant posterior distribution). Vertical lines: daughter cell division times; dots: real fluorescence measurement data; lines: single-cell fits; and black bars: osmotic shock profile u_c. Color coding in bottom plots distinguishes mother from daughter cells as in the top plots

to explain: Methods that do not have a dependency model at the core of the inference approach assume *a priori* independence (no inheritance) of parameters of different cells. In sums, we showed that ARME identification provides unbiased estimates of inheritance and variability of gene expression parameters, whereas indirect methods are affected by an estimation bias that is hard to compensate for.

5 Inheritance of gene expression parameters in yeast osmotic shock response

In this section, we apply our approach to the study of yeast osmotic shock response gene expression data from Llamosi *et al.* (2016). Our study is motivated by the fact that, in Llamosi *et al.* (2016), statistical evidence of correlation between mother and daughter cells was found *a posteriori*, despite the *a priori* modeling hypothesis of independence across cells.

The experiment of Llamosi *et al.* (2016) consists of yeast cells growing in a microfluidic device and subjected to repeated osmotic shocks. A fluorescent reporter protein is expressed in these cells under the control of the promoter of osmosensitive gene STL1, so that new fluorescent reporter molecules are synthesized in response to the shocks. Gene expression response is observed over the experimental time period $[0, 594]$ (min) by videomicroscopy. Fluorescence intensity gene expression measurements are collected for individually tracked cells about every 6 min. Single-cell gene expression data from these recordings are available online (Llamosi *et al.*, 2016). Lineage information were provided to us by the authors for a set of 86 cells, corresponding to the cells observed in a single microfluidic chamber. An illustration of the known parental relationships among these cells is shown in Figure 5. Measurements are shown in Figure 5 on a time axis that also illustrates the delivered osmotic shocks. [In the whole section and figures, florescence measurements

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/35/14/i586/5529241 by guest on 15 October 2019

are in arbitrary units (A.U.), time units is minutes (min), concentrations are in molar units (M) and rate parameters are in $(\text{min})^{-1}$].

The experiments we consider are on budding yeast (*Saccharomyces cerevisiae*). In budding yeast, mother cells generate one newborn daughter cell at a time. Newborn cells are initially smaller than mother cells and do not replace them but rather coexist with them. Because mothers keep most of their material at division, we assume that mothers conserve their own kinetic parameters throughout, while every daughter cell inherits its parameters from the mother with possible variability. This is naturally captured by the modeling framework of Section 2. In accordance with Llamosi *et al.* (2016), an appropriate model to describe fluorescent reporter gene expression response to osmotic shocks in individual cells is an extension of model (1). In this extension, u is the result of a signaling chain that senses exogenous shocks and transduces them into promoter activation via formation and translocation into the nucleus of a transcription factor. Following the characterization of Llamosi *et al.* (2016), we, therefore, rely on the model

$$\begin{cases} \dot{u}(t) &= k_b u_c(t) - g_b u(t), \\ \dot{m}(t) &= k_m u(t) - g_m m(t), \\ \dot{p}(t) &= k_p m(t) - g_p p(t), \end{cases} \tag{8}$$

where m and p are respectively the mRNA and protein concentrations of the reporter species and, up to a known delay, $u_c(t)$ is the commanded (known) microfluidic chamber osmolarity. The first equation models promoter response to shocks with the known parameters $k_b = 0.3968$ and $g_b = 0.9225$. This model is still in the form (5) (with u_c playing the role of u). In addition, the synthesized reporter molecules contribute to the observed fluorescence only after a maturation time $\tau \simeq 30$ (min). Thus, the quantity measured in the experiment is $f(t) = c(g_p) \cdot p(t - \tau)$, where $c(g_p) > 0$ accounts both for the percentage of reporter molecules that mature before degrading (hence the dependence on g_p) and for the conversion of concentration p into corresponding fluorescence intensity. Provided a time shift in the observed data of length τ, this observation model agrees with (6) (dependency of c on g_p is accommodated by the ARME identification algorithm without modifications).

Based on (8) and the inheritance model (2), we ran our ARME identification method to get estimates of the inheritance model parameters θ pertaining the unknown individual-cell quantities g_m, k_p and g_p (in view of the identifiability considerations of Section 4.1, in agreement with Llamosi *et al.* (2016), k_m is fixed to 10, while the remaining parameters k_b and g_b in (8) are known and fixed as above). Results from these estimates are reported in Figure 6. In Figure 5, for various cells ν, predicted single-cell dynamics corresponding to 30 values for φ^ν sampled from the identified model $p(\varphi^\nu|\hat{\theta}, W, Y)$ are compared with the individual-cell measurements. Similar data fits for all cells of the dataset are reported in Supplementary Figures S11–S13.

From Figure 5, it is apparent that the identified model provides an excellent explanation of the data. The *a posteriori* individual-cell simulations agree well with the observations. Variability of these simulations in different cells follows from the estimated variability of φ^ν and matches the stochastic fluctuations in the corresponding single-cell data. The remaining discrepancy between simulations and data is in essential agreement with the estimated measurement noise level $\hat{h} = 427$ (which is similar to the estimate of Llamosi *et al.* (2016) and corresponds to a standard deviation of about 10% of the observed fluorescence levels). An additional validation study confirms that ARME inference does not overfit the data and yields a predictive model (Supplementary Section S5).

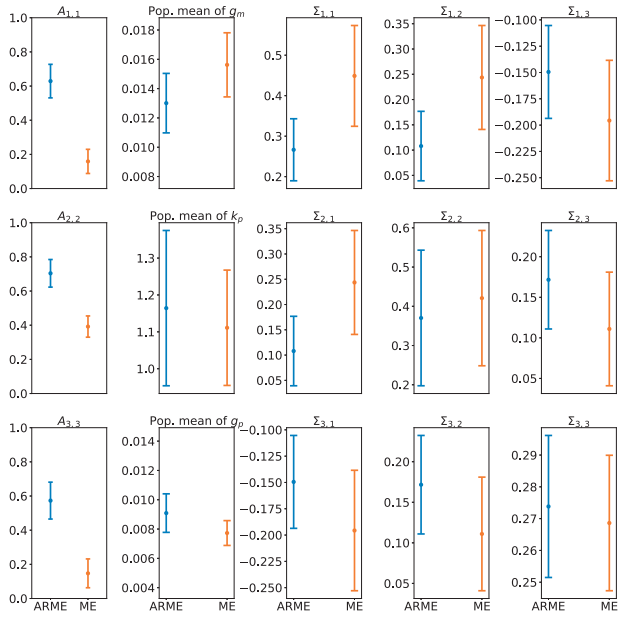

Fig. 6. Results from the identification of an ARME model of yeast osmotic shock response (blue), and comparison with results from a standard ME approach (orange). Plotted are estimates (dots) and 95% confidence intervals (bars). For $\exp(b)$ and Σ, estimates are obtained directly from both ARME and ME identification. For A, estimates are obtained directly for ARME and as described in Section 4.2 for ME. For the computation of confidence intervals see Supplementary Sections S1.4 (ARME) and S1.6 (ME)

ARME estimates of the parameters A, b and Σ drawn from the real data are reported in Figure 6. For comparison, they are shown alongside estimates from the indirect method based on standard ME explained in Section 4.1 and used by Llamosi *et al.* (2016). ARME estimates of inheritance factors $\text{Diag}(A)$ are all around 0.6 (in a scale from 0 to 1). That is, daughter cell parameters are determined by the mother to an extent of about 60%, whereas the remaining 40% follows from the fate inherent in cell division. Because of the unbiasedness of ARME estimates demonstrated in Section 4.2, we interpret this result as a piece of evidence that daughter cell parameters conserve the gene expression kinetics of the mother for the most part.

In particular, our estimates show equal variability of mRNA and protein kinetic parameters at cell division. This may be explained in terms of an equal variability in the partitioning of transcription, translation and degradation regulatory factors. Yet alternative hypotheses, e.g. unmodeled fluctuations of the regulatory processes in the course of a cell lifespan, could support this and deserve further investigation.

Estimates $\hat{A}^{ME}$ based on standard ME, instead, quantify the percentage of inheritance between 20% and 40% depending on the specific parameter. In view of the analysis of Section 4.2, showing that a negative bias affects these estimates, we conclude that correlation analysis studies that do not model inheritance explicitly incur the risk of largely underestimating transcriptional and translational memory. For instance, since correlation between cells ℓ generations apart scales with A^ℓ, estimating $A_{2,2}$ as 0.4 instead of 0.7 reduces the estimated number of generations to achieve a correlation of 10% from 7 to 3.

Estimates of the mean parameter values are similar for ARME and ME identification. This is not surprising since the mean of the ARME model is structurally independent of the presence or absence

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/35/14/i586/5529241 by guest on 15 October 2019

of factor A. Moreover, estimates are biologically reasonable and in essential agreement with those found in Llamosi *et al.* (2016), i.e. $[0.06, 0.81, 0.00645]$ $(\text{min})^{-1}$. On the other hand, the lack of the inheritance factor A in the model for ME identification is reflected into some bias in the estimation of the components of Σ pertaining to g_m (first row and column). Yet, ARME and ME estimates of the correlation structure among the different entries of φ^v (captured by the signs of the off-diagonal elements of Σ) are in agreement. They both predict non-trivial correlations (non-zero off-diagonal elements of Σ), thus reconfirming the presence of mutual correlations observed in Llamosi *et al.* (2016). Overall, it is fair to conclude that ARME and ME estimates of the intercellular parameter variability Σ are similar. Yet, in view of the different estimates of A, ME and ARME provide a different assessment of how this variability is built-up along generations.

6 Discussion and conclusions

In this article, we have addressed reconstruction of gene expression dynamics for a growing population of cell, with focus on the inheritance and variability of transcription and translation parameters at cell division. We have developed an approach for the modeling of parameter inheritance and variability, and a method to identify the model from single-cell quantitative gene expression profiles and information on parental relationships among the observed cells. We have shown that our modeling and identification method, ARME, outperforms indirect methods in recovering inheritance and variability at cell division. In particular, we showed that ARME returns unbiased estimates of inheritance whereas indirect methods systematically underestimate it. We have then applied the method to experimental gene expression data in yeast, showing that daughter cell parameters are determined by the mother to an extent as large as 60%. In comparison, a state-of-the art indirect method assessed this value at $20-40\%$. We concluded that, in yeast as well as other studies, utilization of indirect methods may significantly underestimate population memory for gene expression kinetics. In addition, variability at division was found to affect the different kinetic parameters in a similar manner, hinting that variability is likely associated with aspecific regulatory factors.

Methodologically, our contribution extends ME modeling to the case of tree-structured dependencies among individuals, and provides an original algorithm to reconstruct this type of models that is a significant extension of SAEM (Lavielle, 2015). Developed for and demonstrated on microbial gene expression, it lends itself to a number of applications where from individual-cell data and lineage information are available, for instance, the study of growth of cancerous cell populations. To broaden applicability, a number of extensions are well within reach, notably arbitrary non-linear individual dynamic response. The increased computational burden incurred in the solution of non-linear dynamical modeling, as well as scalability to larger systems (more parameters and states) and larger populations shall then require non-trivial programming efforts. Design and implementation of a suitably general, user-friendly software to the profit of the community is among our work directions.

Our model of transmission of gene expression parameters from mother to daughter cells can be thought of as a description of extrinsic noise at the time scale of cell division. Whereas important, this source of variability does not exhaust all sources of gene expression noise. Although our model was shown to agree well with yeast osmotic shock response single-cell data, in more generality, the proposed model is not sufficient to describe systems where intrinsic noise is the dominant source or variability or the core object of study. In fact, intrinsic noise can be easily included in our framework in terms of stochastic gene expression dynamics. On the other hand, inference of such a model from data becomes more involved, and a non-trivial extension of our identification method is required.

Both the modeling and the inference approaches are statistically well-characterized for the most part, yet non-trivial mathematical questions of practical relevance stand. Whereas the AR model used to describe inheritance dynamics is deeply understood, the properties of the hierarchical model resulting from the combination with ordinary differential equation-type dynamics are much less clear. Falling in the context of piecewise-deterministic systems (Cinquemani *et al.*, 2008), the additional complexity of the tree-like model structure raises analysis and inference questions that do not have a full answer yet. This poses challenges and at the same time great research opportunities. Among the questions that we intend to address analytically are the structural and practical identifiability of the hidden inheritance parameters, and the asymptotics of the measured state dynamics. Along a related line, the study of convergence rates of our ARME identification method as a function, for instance, of population size $|V|$ or number of dependencies $|W|$ will provide us with more guidance toward optimization of experiment design.

The results of the application of our method on experimental data from yeast show that modeling inheritance at division is a fundamental concern to derive reliable estimates of gene expression memory and to understand emergence of variability in a growing population of cells. These results were established based on a simple model of transcription and translation parameters and were demonstrated to be superior to popular data analysis approaches. Of course, these parameters subsume complex biochemical processes. Although this abstraction layer enables reliable analysis of the data and effective interpretation of the results, investigation of specific players and mechanisms behind variability and inheritance requires dedicated biological experimentation. In combination with experiment designs optimized in the light of simulated performance assessment, the proposed approach promises to be an invaluable instrument for in-depth study of memory in gene expression dynamics.

Acknowledgements

The authors thank Gregory Batt, Pascal Hersen and Artemis Llamosi for sharing and providing assistance with lineage data, as well as G.B., Jakob Ruess and Hidde de Jong for insightful discussions and suggestions.

Funding

This work was funded in part by the French national research agency (ANR) via project MEMIP [ANR-16-CE33-0018], and by the Inria IPL CoSy. A.M. acknowledges partial support by the Chaire 'Modélisation Mathématique et Biodiversité' of VEOLIA-Ecole Polytechnique-MnHn-FX.

Conflict of Interest: none declared.

References

Bezanson,J. *et al.* (2017) Julia: a fresh approach to numerical computing. *SIAM Rev.*, **59**, 65–98.

Cinquemani,E. *et al.* (2008) Stochastic dynamics of genetic networks: modelling and parameter identification. *Bioinformatics*, **24**, 2748–2754.

de Jong,H. (2002) Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.*, **9**, 67–103.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/35/14/i586/5529241 by guest on 15 October 2019

Delyon,B. *et al*. (1999) Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist*., **27**, 94–128.

Dharmarajan,L. *et al*. (2019) A simple and flexible computational framework for inferring sources of heterogeneity from single-cell dynamics. *Cell Syst*., **8**, 15–26.

Feigelman,J. *et al*. (2016) Analysis of cell lineage trees by exact bayesian inference identifies negative autoregulation of nanog in mouse embryonic stem cells. *Cell Syst*., **3**, 480–490.e13.

Ferraro,T. *et al*. (2016) Transcriptional memory in the *drosophila* embryo. *Curr. Biol*., **26**, 212–218.

Fröhlich,F. *et al*. (2018) Multi-experiment nonlinear mixed effect modeling of single-cell translation kinetics after transfection. *NPJ Syst. Biol. Appl*., **5**, 1.

García,M.R. *et al*. (2018) Stochastic individual-based modeling of bacterial growth and division using flow cytometry. *Front. Microbiol*., **8**, 2626.

Hasenauer,J. *et al*. (2011) Identification of models of heterogeneous cell populations from population snapshot data. *BMC Bioinformatics*, **12**, 125.

Hormoz,S. *et al*. (2016) Inferring cell-state transition dynamics from lineage trees and endpoint single-cell measurements. *Cell Syst*., **3**, 419–433.e8.

Huh,D. and Paulsson,J. (2011a) Non-genetic heterogeneity from stochastic partitioning at cell division. *Nat. Genet*., **43**, 95–100.

Huh,D. and Paulsson,J. (2011b) Random partitioning of molecules at cell division. *Proc. Natl. Acad. Sci. USA*, **108**, 15004–15009.

Johnston,I.G. and Jones,N.S. (2015) Closed-form stochastic solutions for non-equilibrium dynamics and inheritance of cellular components over many cell divisions. *Proc. Math. Phys. Eng. Sci*., **471**, 20150050.

Komorowski,M. *et al*. (2009) Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC Bioinformatics*, **10**, 343.

Kuhn,E. and Lavielle,M. (2004) Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM Probab. Stat*., **8**, 115–131.

Kuzmanovska,I. *et al*. (2017) Parameter inference for stochastic single-cell dynamics from lineage tree data. *BMC Syst. Biol*., **11**, 52.

Lavielle,M. (2015) *Mixed Effects Models for the Population Approach. Models, Tasks, Methods & Tools. Chapman & Hall/CRC Biostatistics Series*. CRC Press, Boca Raton, FL.

Ljung,L. (1999) *System Identification: Theory for the User*. Prentice Hall, Upper Saddle River, NJ.

Llamosi,A. *et al*. (2016) What population reveals about individual cell identity: single-cell parameter estimation of models of gene expression in yeast. *PLoS Comput. Biol*., **12**, e1004706.

Munsky,B. *et al*. (2009) Listening to the noise: random fluctuations reveal gene network parameters. *Mol. Syst. Biol*., **5**, 318.

Neuert,G. *et al*. (2013) Systematic identification of signal-activated stochastic gene regulation. *Science*, **339**, 584–587.

Papoulis,A. (1991) *Probability, Random Variables, and Stochastic Processes. McGraw-Hill Series in Electrical Engineering*. McGraw-Hill, New York.

Paulsson,J. (2005) Models of stochastic gene expression. *Phys. Life Rev*., **2**, 157–175.

Raj,A. and van Oudenaarden,A. (2008) Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, **135**, 216–226.

Suter,D.M. *et al*. (2011) Mammalian genes are transcribed with widely different bursting kinetics. *Science*, **332**, 472–474.

Swain,P. *et al*. (2002) Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl. Acad. Sci. USA*, **99**, 12795–12800.

Taheri-Araghi,S. *et al*. (2015) Cell-size control and homeostasis in bacteria. *Curr. Biol*., **25**, 385–391.

Thomas,P. (2017) Making sense of snapshot data: ergodic principle for clonal cell populations. *J. Royal Soc. Interface*, **14**, 20170467.

Waldherr,S. (2018) Estimation methods for heterogeneous cell population models in systems biology. *J. Royal Soc. Interface*, **15**, 20180530.

Zechner,C. *et al*. (2012) Moment-based inference predicts bimodality in transient gene expression. *Proc. Natl. Acad. Sci. USA*, **109**, 8340–8345.

Zechner,C. *et al*. (2014) Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. *Nat. Methods*, **11**, 197–202.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/35/14/i586/5529241 by guest on 15 October 2019