Published in final edited form as:

J Biotechnol. 2012 April 15; 158(3): 80–90. doi:10.1016/j.jbiotec.2011.08.018.

Single-Cell Behavior and Population Heterogeneity: Solving an Inverse Problem to Compute the Intrinsic Physiological State Functions

Konstantinos Spetsieris and Kyriacos Zygorakis*

Department of Chemical and Biomolecular Engineering, MS-362, Rice University, 6100 Main Street, Houston, TX 77005, USA

Abstract

The dynamics of isogenic cell populations can be described by cell population balance models that account for phenotypic heterogeneity. To utilize the predictive power of these models, however, we must know the rates of single-cell reaction and division and the bivariate partition probability density function. These three intrinsic physiological state (IPS) functions can be obtained by solving an inverse problem that requires knowledge of the phenotypic distributions for the overall cell population, the dividing cell subpopulation and the newborn cell subpopulation.

We present here a robust computational procedure that can accurately estimate the IPS functions for heterogeneous cell populations. A detailed parametric analysis shows how the accuracy of the inverse solution is affected by discretization parameters, the type of non-parametric estimators used, the qualitative characteristics of phenotypic distributions and the unknown partitioning probability density function. The effect of finite sampling and measurement errors on the accuracy of the recovered IPS functions is also assessed. Finally, we apply the procedure to estimate the IPS functions of an *E. coli* population carrying an IPTG-inducible genetic toggle network. This study completes the development of an integrated experimental and computational framework that can become a powerful tool for quantifying single-cell behavior using measurements from heterogeneous cell populations.

Keywords

cell population balance model; inverse problem; population heterogeneity; phenotypic distributions; intrinsic physiological state functions; genetic toggle network

1 Introduction

Current biotechnological applications focus on maximizing the production of bio-products through optimization of the cellular properties, or the phenotype of entire cell populations. For more than 60 years, however, we know that isogenic cell populations exhibit significant and time-dependent variations in a wide range of properties like intracellular content, growth or production rates, transcriptional or lysogenic states and others (Bæk et al., 2003; Chung

© 2011 Elsevier B.V. All rights reserved

*Corresponding Author: kyzy@rice.edu Tel: 1-713-348-5208 FAX: 1-713-348-5478.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

and Stephanopoulos, 1995; Delbrück, 1945; Elowitz et al., 2002; Maloney and Rotman, 1972; Novick and Weiner, 1957; Powell, 1956; Russo-Marie et al., 1993; Spudich and Koshland, 1976; Stocker, 1949).

An approach often used to describe the heterogeneous nature of cell populations involves the cell population balance (CPB) models that were first formulated by Fredrickson and coworkers (Eakman et al., 1966; Fredrickson et al., 1967; Ramkrishna, 2000; Tsuchiya et al., 1966). CPB models describe the dynamics of the distributions of cellular properties (such as size or intracellular content) with partial integro-differential equations that are nonlinearly coupled with ordinary integro-differential equations, describing substrate availability. When the cell state can be adequately described by an extensive single cell property x , the one-dimensional CPB model describing the dynamics of a cell population, cultured in a batch reactor with excess substrate, is given by the following equation:

$$\frac{\partial n(x, t)}{\partial t} + \frac{\partial}{\partial x} [R(x) n(x, t)] + \Gamma(x) n(x, t) + n(x, t) \mu = 2 \int_{x_{\min}}^{x_{\max}} P(x, y) \Gamma(y) n(y, t) dy \quad (1)$$

where $n(x, t)dx$ gives the fraction of cells in the reactor that have a state between x and $x + dx$ at time t , μ is the average specific population growth rate defined by:

$$\mu = \int_{x_{\min}}^{x_{\max}} \Gamma(x) n(x, t) dx \quad (2)$$

and expressing on average the rate at which the cell population is increasing due to cell division, while $x_{\min}$ and $x_{\max}$ are (respectively) the smallest and largest values of the state x .

The three functions, $R(x)$, $\Gamma(x)$ and $P(x, y)$ that appear in the CPB equation (1) are collectively called *intrinsic physiological state* (IPS) functions. The single-cell reaction rate $R(x)$ denotes the rate of change of the physiological state variable x . $\Gamma(x)$ is the single-cell division rate and $\Gamma(x)dt$ represents the probability that a cell with state x at time t will divide in $[t, t + dt]$. In reality, the single-cell reaction and division rates $R(x)$ and $\Gamma(x)$ are averages of rates over cohorts of cells with the same physiological state vector x . However, we will use the term “single-cell” for these rates to distinguish them from the rates obtained by averaging over the entire cell population independently of cellular content. Finally, $P(x, y)$ is the partition probability density function (or PPDF) and expresses the probability that a mother cell with state y at time t , will divide and give birth to two daughter cells with states x and $y - x$, at time $t + dt$. The population balance model (1) is subject to an initial condition:

$$n(x, t=0) = n_0(x) \quad (3)$$

and to the regularity conditions

$$R(x) n(x, t) = 0, \quad \forall x \in \partial\Omega \quad (4)$$

The physical meaning of the latter equation is that the physiological state vector x cannot grow out of the boundaries Ω of the physiological state space $\Omega = [x_{\min}, x_{\max}]$.

Although significant progress has been made over many years in developing numerical algorithms for solving the CPB equation (Godin et al., 1999; Kostova, 1990; Mantzaris et al., 2001a, b, c; Subramanian and Ramkrishna, 1971; Sulsky, 1994), its use for predicting and optimizing cell population behavior has been limited. A major reason for this is that the solution of CPBs requires as inputs the three IPS functions, whose determination presents significant experimental and computational challenges involving the solution of tough

inverse problems (Ramkrishna, 1994; Ramkrishna et al., 1968; Ramkrishna and Schell, 1999).

Collins and Richmond showed in 1962 that under exponential balanced growth conditions (Campbell, 1957; Fredrickson et al., 1967; Ramkrishna et al., 1968), the number density function $n(x, t)$ of a cell population becomes time invariant (Collins and Richmond, 1962). The single-cell reaction rate $R(x)$ can then be computed from the following closed-form expression

$$R(x) = \frac{\mu}{n(x)} \int_0^x (2n_b(y) - n_d(y) - n(y)) dy \quad (5)$$

if we have experimentally determined: (a) the average specific population growth rate μ , (b) the number density function $n(x)$ of the state variable x for the entire cell population, (c) the number density function $n_d(x)$ for the dividing cell subpopulation, and (d) the number density function $n_b(x)$ for the newborn cell subpopulation. The focus of Collins and Richmond's study was the growth of *Bacillus cereus* and they derived equation (5) without considering any mathematical model to describe cell growth.

Ramkrishna and coworkers (1968) re-derived equation (5) using a rigorous approach (Ramkrishna et al., 1968). By applying equation (1) to cell populations cultured in batch reactors at balanced exponential growth conditions and with excess of substrate, they showed that the single-cell division rate $\Gamma(x)$ can be obtained from the following closed-form expression

$$\Gamma(x) = \mu \frac{n_d(x)}{n(x)} \quad (6)$$

and that the partition probability density function $P(x, y)$ satisfies the following equation:

$$n_b(x) = \int_x^{x_{\max}} P(x, y) n_d(y) dy \quad (7)$$

Since $p(x, y)$ is a bivariate probability density, it must also satisfy the normalization condition:

$$\int_0^y P(x, y) dx = 1 \quad (8)$$

equation (5) has also been derived in a systematic way by Harvey *et al.* (Harvey et al., 1967) and the integral equation (7) has been also formulated by Powell (Powell, 1964) and Harvey *et al.* (Harvey et al., 1967).

In a previous study (Spetsieris et al., 2009), we have developed a novel assay based on fluorescence microscopy that can accurately quantify the phenotypic distributions of the overall cell population $n(x)$, of the dividing cell subpopulation $n_d(x)$ and of the newborn cell subpopulation $n_b(x)$. These are the three distributions needed to solve the inverse cell population balance (ICPB) problem defined by equations (5)–(8) in order to determine the three IPS functions required for solving equation (1). This study will present the development of efficient and robust computational methods to solve the inverse (ICPB) problem given the three experimentally determined phenotypic distributions.

2 Solving the Inverse Problem

A large body of work exists in the literature for the solution of inverse problems resulting from population balances describing the dynamic processes like particle nucleation, growth,

breakage and aggregation.(Mahoney et al., 2002; Muralidar and Ramkrishna, 1986; Muralidar and Ramkrishna, 1989; Patruno et al., 2008; Patruno et al., 2009; Sathyagal et al., 1995; Wright and Ramkrishna, 1992). While these studies can provide invaluable guidance, their methods cannot be directly applied to address the challenges of inverse problems resulting from cell population balance models.

Ramkrishna proposed an elegant mathematical framework to determine the IPS functions for the cell population balance problem at self-similar conditions (Ramkrishna, 1994; Ramkrishna and Schell, 1999). His idea about self-similarity is based on scaling transient flow cytometric data with appropriate cell parameters, in order to obtain time-invariant distributions of cell characteristics. The formulation of the mathematical framework requires the postulation of certain functional forms for the IPS functions, which must hold true in order for a cell population to attain self-similarity. Such conditions include power law form expressions for the single-cell division and reactions rates and a certain form for the partition probability density function. This mathematical framework introduces a large number of unknowns whose identification requires transient data for the dividing cell subpopulation, in addition to the transient flow cytometric data for the overall cell population. Moreover, the suitable choice of the scaling parameter is an issue that requires further investigation as Ramkrishna has remarked (Ramkrishna, 1994).

For these reasons, the approach summarized by equations (5)–(8) was the one employed by investigators who sought to gain new insights into the dynamics of heterogeneous cell populations. These efforts, however, frequently relied on limiting assumptions or experimental approaches with significant drawbacks (as we discussed in detail in (Spetsieris et al., 2009)). For example, many investigators were solely interested in determining the single-cell reaction rate for the organism considered (Collins and Richmond, 1962; Koch, 1966; Kromenaker and Srien, 1994; Painter and Marr, 1968; Srien and Dien, 1992; Zusman D. et al., 1971). Other studies postulated functional forms for the number density functions of the dividing and newborn cell subpopulations and for the partitioning frequency at cell division (Harvey et al., 1967; Kempner and Marr, 1979) or assumed that cell division is always symmetric (Anderson et al., 1967; Block et al., 1990). Only in a few cases, the single-cell division rate has been determined (Anderson et al., 1967; Hatzis et al., 1997; Kempner and Marr, 1979; Sweeney et al., 1994). And more importantly, there have not been attempts to recover the bivariate partition probability density function $P(x, y)$ from experimental data obtained for a cell population, by solving the integral equation (7).

This study focuses on developing and testing a robust computational method for solving the inverse problem defined by equations (5)–(8) using experimental data obtained with the assay described in our earlier publication (Spetsieris et al., 2009). Thus, the present study completes the development of an integrated experimental and computational framework that can accurately determine all three IPS functions required to solve the CPB model of equation (1).

2.1 Methodology

The three phenotypic distributions required to solve the inverse problem are determined experimentally from a finite sample of a cell population. Our fluorescence microscopy assay only needs about 3,000 cells to compute the overall phenotypic distribution of a cell population with accuracy comparable to that achievable with flow cytometry (Spetsieris et al., 2009). Such a sample size yields about 300 dividing cells and 600 newborn cells. Moreover, random measurement errors introduce uncertainty in the experimentally obtained cell population data. For these reasons, we will first conduct a systematic parametric study with simulated data to assess the effect of both finite sampling and random errors on the

accuracy of our solution. Figure 1 is a schematic illustration of the steps used in this procedure and discussed in detail in sections 2.1.1, 2.1.2 and 2.1.3.

2.1.1 Finite Sampling—Using known analytical expressions for the three IPS functions and the overall phenotypic distribution $n(x)$, we first generated the true phenotypic distributions $n_d(x)$ and $n_b(x)$ for this population using equations (6) and (7) respectively. To simulate finite sampling, we then generated a finite number of random deviates from these distributions using an acceptance-rejection method and random number generators (Neumann, 1951; Robert and Casella, 2004). Designed to mimic the collection of experimental data with our microscopy assay, this process created three “synthetic” data sets of measurements for the phenotypic property x : $\{X_i, i = 1, 2, \dots, N\}$, $\{X_{dj}, j = 1, 2, \dots, N_d\}$ and $\{X_{bk}, k = 1, 2, \dots, N_b\}$. These three sets correspond, respectively, to the overall cell population, the dividing subpopulation and the newborn subpopulation.

2.1.2 Nonparametric Estimation of the Phenotypic Distributions—Next, we applied nonparametric techniques to the “synthetic” data sets to obtain estimates $\tilde{n}(x)$, $\tilde{n}_d(x)$ and $\tilde{n}_b(x)$ of the true phenotypic distributions $n(x)$, $n_d(x)$ and $n_b(x)$. Nonparametric estimation techniques were used to avoid making any assumptions about the underlying functional forms of the phenotypic distributions. The synthetic data sets can be represented using two nonparametric methods: (a) the estimator of the number density function (NDF), or (b) the estimator of the cumulative distribution function (CDF).

Before comparing the two types of estimation, we need to select the NDF estimator that we will use. We considered two NDF estimators: (a) the simple and widely used histogram estimator, supplemented with the Freedman-Diaconis formula (Freedman and Diaconis, 1981; Izenman, 1991) for optimal selection of the histogram bin size, and (b) the kernel density estimator (Bowman and Azzalini, 1997; Silverman, 1986; Thompson and Tapia, 1990; Wand and Jones, 1995) defined by:

$$\tilde{g}(x) = \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x - X_i}{h}\right) \quad (9)$$

where h is a positive real number, called the bandwidth or the smoothing parameter, and $K(x)$ is a known “kernel” function (usually a normal density with zero mean and variance equal to one). The kernel density estimator is an improvement over the histogram since it obviates the need for a selection of an origin and number of bins. Besides, the kernel density estimator is smoother than the histogram. In order to select the optimal value of the smoothing parameter h , we implemented two fully automatic data-driven methods: the “least squares cross-validation” and the “likelihood cross-validation” (Silverman, 1986; Thompson and Tapia, 1990; Wand and Jones, 1995). To facilitate the comparison between the histogram and the kernel density estimators, we used the histogram in conjunction with an interpolation rule. Three interpolation methods were utilized: (a) nearest neighbor, (b) linear, and (c) spline. Our simulations (see K. Spetsieris, PhD Thesis, Rice University, 2010) revealed faster convergence of the kernel density estimator. A second comparison showed that the CDF estimator was more accurate than the kernel density estimator in representing the same data set of measurements (ibid.). Thus, the CDF estimator became our preferred procedure.

2.1.3 Random Errors—To assess the effect of uncertainty in the experimental data, the “synthetic” data X_i generated by the process presented in section 2.1.1 were modified by adding random errors according to the formula:

$$\widehat{X}_i = X_i + \varepsilon_i \quad (10)$$

where the independent random errors ε_i were drawn from a pre-selected normal distribution with zero mean (see equation (11)) and variance given by equation (12):

$$E(\varepsilon_i) = 0 \quad (11)$$

$$Var(\varepsilon_i) = \left(\frac{1}{SNR} X_i \right)^2 \quad (12)$$

Here, SNR is the signal to noise ratio. The typical SNR obtained with the assay described in an earlier publication (Spetsieris et al., 2009) is about 1,000.

In the final step of the procedure of Figure 1, the inverse problem was solved to obtain approximations $\tilde{R}(x)$, $\tilde{\Gamma}(x)$ and $\tilde{P}(x, y)$ to the original (or true) IPS functions. Several approaches were used for the solution of the inverse problem in order to select the method that minimizes the errors $\|R(x) - \tilde{R}(x)\|$, $\|\Gamma(x) - \tilde{\Gamma}(x)\|$ and $\|P(x, y) - \tilde{P}(x, y)\|$.

2.2 Computing the Single-Cell Reaction and Division Rates

We have used several analytical expressions for $R(x)$ and $\Gamma(x)$ to assess the robustness of our inverse problem solver to recover the original IPS functions from the simulated data. For the sake of brevity, we will only present here the results obtained for the following highly nonlinear true $R(x)$ (Kavousanakis et al., 2009; Kepler and Elston, 2001) and $\Gamma(x)$ (Mantzaris, 2006):

$$R(x) = \frac{a \cdot b + x^2}{b + x^2} - d \cdot x \quad (13)$$

$$\Gamma(x) = \left(\frac{x}{b} \right)^L \quad (14)$$

We computed the single-cell reaction rate $R(x)$ using (a) the integral form of equation (5) and (b) the corresponding differential form given by:

$$\frac{d}{dx} (R(x) n(x)) = \mu [2n_b(x) - n_d(x) - n(x)] \quad (15)$$

To take advantage of the faster convergence of the CDF estimator, we reformulated equation (5) as shown below:

$$R(x) = \frac{\mu}{n(x)} (2c_b(x) - c_d(x) - c(x)) \quad (16)$$

where $c(x)$, $c_d(x)$, $c_b(x)$ are the CDF estimators of, respectively, the overall population, the dividing subpopulation and the newborn subpopulation. Our numerical simulations showed that computing the single-cell reaction rate from the CDF formulation of equation (16) yielded more accurate results than computations using the NDF formulation of equation (15). Figure 2A compares the results obtained for the single-cell reaction rate when $a = 0.4$, $b = 0.001$, $d = 0.005$. This simulation used a finite sample of 3,000 cells with 300 dividing and 600 newborn cells, and a Gaussian overall number density function $n(x)$ with mean 1000 and standard deviation 200. The L_2 error of the NDF method here is ~15%, whereas

the error with the CDF is as low as ~6%. In general, the L_2 error of the CDF formulation for the single-cell reaction rate was in the range of 4–6% for all expressions and parameter sets tested.

Let us now turn to the single-cell division rate $\Gamma(x)$. If we used a CDF formulation to solve equation (6) for the single-cell division rate, we would have lost the benefit of a closed-form solution. For this reason, we have used the kernel density estimator. We again performed numerical simulations with sample sizes of 3,000 cells with 300 dividing cells, the typical sample sizes from our experiments. Figure 2B compares numerical results to the original division rate of equation (14) for $b = 1000$ and $L = 7$, and a Gaussian overall number density function $n(x)$ with mean 1000 and standard deviation 200. The inverse solution is in excellent agreement with the original analytical expression and the corresponding L_2 error is only ~4%. In general, our parametric study showed that the typical sample size of ~300 dividing cells is sufficient for recovering the single-cell division rate $\Gamma(x)$ with acceptable accuracy (4–5%).

Finally, our simulation results have revealed that random measurement errors have negligible effects on the accuracy with which we can recover $R(x)$ and $\Gamma(x)$. For these IPS functions, only the errors introduced by finite sampling affect the accuracy of our solution (K. Spetsieris, PhD Thesis, Rice University, 2010).

2.3 Computing the Partition Probability Density Function

2.3.1 Mathematical Formulation—We now turn our attention to the most interesting and challenging part of the inverse problem: the determination of the bivariate partition probability density function $P(x, y)$ that has no closed-form solution and satisfies the linear integral equation (7). This equation does not belong to the well-studied classes of Fredholm and Volterra integral equations of the first and second kind (Jerri, 1999; Zahrt, 1991).

We have considered two methods to obtain $P(x, y)$. The first one estimated the dividing and newborn phenotypic distributions and recovered $P(x, y)$ directly from equation (7). The second method was based on the following reformulation of equation (7):

$$c_b(x) = \int_{x_{\min}}^x P(z, x_{\max}) c_d(x_{\max}) dz - \int_{x_{\min}}^x \int_x^{x_{\max}} c_d(y) \frac{\partial P(z, y)}{\partial y} dy dz \quad (17)$$

A detailed derivation of equation (17) is presented in the appendix of the supplementary material. Since equation (17) can accept as inputs the more accurate CDF formulations of the phenotypic distributions, we expected that this method would yield more accurate results.

To recover $P(x, y)$ from both equation (7) and (17), we have employed the assumption that it is a homogeneous function of the form

$$P(x, y) = \frac{1}{y} Q\left(\frac{x}{y}\right) \quad (18)$$

where $Q\left(\frac{x}{y}\right)$ is an unknown partitioning function that shows how the daughter-to-mother cell content ratio $f = \frac{x}{y}$ is distributed at cell division. This assumption is supported by theoretical and experimental studies (Harvey et al., 1967; Powell, 1964; Ramkrishna, 1994). In particular, Ramkrishna has shown that a functional form of the form (18) can lead to self-

similar phenotypic distributions (Ramkrishna, 1994), which are observed in the exponential balanced growth regime where the Collins and Richmond approach is valid.

The homogeneity assumption implies that the distribution of the daughter to mother content ratio f is independent of the size of the mother cell y . The benefit of this assumption is that

we only need to solve for the univariate function $Q\left(\frac{x}{y}\right)$ in order to determine the unknown bivariate function $P(x, y)$.

2.3.2 Numerical Solution of the Inverse Problem for the PPDF—To obtain $P(x, y)$, we have discretized the continuous integral equations (7) and (17). We have expressed the

unknown partitioning function $Q\left(\frac{x}{y}\right)$ as a linear combination of m real-univariate basis functions ϕ_j :

$$Q\left(\frac{x}{y}\right) = \sum_{j=1}^m a_j \phi_j\left(\frac{x}{y}\right) \quad (19)$$

We then substituted equations (18) and (19) into (7) and (17) and discretized the resulting equations to obtain the non-square system of linear algebraic equations:

$$\mathbf{G}\mathbf{a} = \mathbf{b} \quad (20)$$

where $\mathbf{G}$ is the $n \times m$ coefficient (or design) matrix, $\mathbf{b}$ is the $n \times 1$ data vector with elements the values of the newborn phenotypic distribution at the n discretization points and $\mathbf{a}$ is the $m \times 1$ vector of the unknown expansion coefficients a_j . To compute the unknown vector $\mathbf{a}$, we transformed the overdetermined ($n > m$) system of linear algebraic equations (20) into the corresponding constrained minimization problem. In order to obtain a meaningful solution for $P(x, y)$, we need to impose: (a) the normalization condition of equation (8) so that $P(x, y)$ is a proper probability density function; (b) the symmetric condition

$$P(x, y) = P(y - x, y) \quad (21)$$

expressing the conservation of the mother's content y upon its partitioning to the two daughter cells during cell division (Fredrickson et al., 1967); and (c) the nonnegativity constraints

$$P(x, y) \geq 0 \quad \text{and} \quad n_b(x) \geq 0 \quad (\text{or } c_b(x) \geq 0) \quad (22)$$

for the probability density function and the phenotypic distribution of the newborn cell subpopulation. We solved the minimization problem by employing a second order Tikhonov regularization (Aster et al., 2005; Tikhonov and Arsenin, 1977) to address the ill-conditioned nature of the inverse problem and to make sure that the recovered solution $P(x, y)$ is a smooth function. The parameter λ^2 controls the amount of regularization used to solve the full minimization problem shown below:

$$\begin{aligned} \min_{\mathbf{a} \in \mathbb{R}^m} & \|\mathbf{G}\mathbf{a} - \mathbf{b}\|_2^2 + \lambda^2 \|\mathbf{D}\mathbf{a}\|_2^2 \\ \text{s.t.} & \\ & \mathbf{c}^T \mathbf{a} = 1 \\ & \mathbf{A}_{\text{sym}} \mathbf{a} = 0 \\ & \mathbf{A}_p \mathbf{a} \geq 0 \\ & \mathbf{G}\mathbf{a} \geq 0 \end{aligned} \quad (23)$$

We have performed a detailed parametric analysis to address the effect of several numerical parameters on the accuracy of the recovered $\tilde{P}(x, y)$ obtained by solving (23) for the true distributions $n(x)$, $n_d(x)$ and $n_b(x)$. In particular, we have looked into how the accuracy of the recovered $\tilde{P}(x, y)$ depends on (a) the type and number of basis functions used in equation (19), (b) the number of points used to discretize the integral equations, and (c) the magnitude of the regularization parameter. In addition, we have investigated the effect of key characteristics of the original phenotypic distributions such as their variance, skewness and number of modes (specifically, the presence of bimodality). Our parametric study established the appropriate choice of basis functions and discretization parameters so that our method can accurately recover $P(x, y)$ for a wide range of qualitatively and

quantitatively different unimodal and bimodal partitioning functions $Q\left(\frac{x}{y}\right)$. Detailed results of the parametric analysis are presented in the supplementary material.

2.3.3 Effect of Finite Sampling—To assess the effect of finite sampling on our solution, we used the NDF- and CDF-based methods to solve again the integral equations (7) and (17) respectively starting this time with the estimated phenotypic distributions $\tilde{n}(x)$, $\tilde{n}_d(x)$ and $\tilde{n}_b(x)$. We performed numerical simulations with both methods varying the number of dividing cells N_d and recorded the corresponding L_2 error between the true $Q(f)$ and the recovered partitioning function $\tilde{Q}(f)$. A typical result is shown in panel A of Figure 3 for the case when a symmetric Beta distribution was used as partitioning function:

$$Q\left(\frac{x}{y}\right) = K\left(\frac{x}{y}\right)^{q-1} \cdot \left(1 - \frac{x}{y}\right)^{q-1} \quad (24)$$

Here, $q = 30$ and we used a Gaussian dividing number density with $CV = 0.2$.

We observed that the L_2 error for both methods gradually declined with increasing sample size N_d . Note also that for approximately 800 dividing cells the true and inverse solutions are practically indistinguishable. Additionally, a 10–15% error in estimating the NDF from a finite sample (with as few as 100 – 200 dividing cells) translates into a relatively small error for the inverse solution ($< 6 - 7\%$). Furthermore, we observed that the error of both methods decreased in a similar fashion with increasing N_d . Contrary to our expectations, however, we did not see any improvement in the accuracy of the inverse solution by using the CDF-based formulation and equation (17).

To understand why the CDF formulation did not perform as expected, we carried out a singular value decomposition on the coefficient matrix G for both methods to obtain the corresponding eigenvalue spectra shown in panel B of Figure 3. Notice that the eigenvalue decayed faster for the CDF-based method. From the theory of inverse problems (Aster et al., 2005; Hansen, 1998; Hofmann, 1986, 1993), we know that fast and abrupt decay of the eigenvalues is a manifestation of a badly ill-conditioned discrete inverse problem. Therefore, our results indicated that the CDF formulation of the inverse problem was more ill-conditioned than the NDF formulation. Although the CDF method had more accurate input data (relative to the NDF method), its more ill-conditioned nature led to a less accurate inverse solution.

In conclusion, the NDF-based formulation is the superior method. It is less complicated to set up, requires less computational time for its solution and offers better accuracy than the more computationally demanding CDF formulation.

2.3.4 Effect of PPDF Characteristics—The previous results established that the NDF-formulation can give acceptable results (i.e. L_2 error smaller than ~5%) for as few as 300 dividing cells. For this reason, a sample size of $N_d = 300$ was used to assess the ability of the NDF formulation to accurately recover unimodal and bimodal $P(x, y)$ functions with widely different characteristics. Figure 4 shows some representative results. In all cases the recovered partitioning functions $\tilde{Q}(f)$ were in very good agreement with the corresponding true functions $Q(f)$. In general, our simulations showed that the error for unimodal partitioning functions was in the range of 4–6%, whereas the error for bimodal partitioning functions could be as high as 11% for sharp distributions with small separation between the two modes (see panel D of Figure 4).

2.3.5 Combined Random and Finite Sampling Errors—Figure 5 presents the combined effect of random and finite sampling errors on the accuracy of the inverse solution for two challenging cases: a unimodal PPDF given by equation (24) with $q = 60$ and a bimodal partitioning function shown below

$$Q(f) = K \left(\exp \left(\frac{-(f - \mu_{part})^2}{2\sigma_{part}^2} \right) + \exp \left(\frac{-(f - 1 + \mu_{part})^2}{2\sigma_{part}^2} \right) \right) \quad (25)$$

with $\mu_{part} = 0.4$ and $\sigma_{part} = 0.035$. With a sample size of 300 dividing cells, we solved the inverse problem for a wide range of values for the signal-to-noise ratio (SNR) and computed the corresponding L_2 error. The L_2 error for the unimodal case remained consistently below ~6% for a wide range of SNR values (Figure 5). Even for the more challenging bimodal case, the L_2 error did not exceed 11% until the SNR values dropped to low levels.

These results are particularly encouraging given that the inverse problem for the PPDF is ill-conditioned and we might expect to see significant adverse effects of small random measurement errors on the accuracy of the solution. We concluded that the chosen numerical method was robust in the presence of random and finite sampling errors with the magnitudes expected for experimental assays.

Application: Computing the IPS functions for populations of *E. coli* with the toggle genetic network

To demonstrate the capabilities and potential of our computational procedure, we applied it to solve the inverse problem for populations of *E. coli* that carried the toggle genetic network equipped with a *gfpmut3* gene functioning as a reporter of intracellular expression levels (Gardner et al., 2000). The required phenotypic distributions $n(x)$, $n_d(x)$ and $n_b(x)$ were nonparametrically estimated using measurements obtained with our fluorescence microscopy assay (Spetsieris et al., 2009). Then, the three unknown IPS functions were determined with the NDF-formulation as presented in the previous section.

Figure 6 presents the estimated densities $n(x)$, $n_d(x)$ and $n_b(x)$ for three concentrations of the chemical inducer IPTG: 2000 μM , 40 μM , and 20 μM . Panels A, C and E of Figure 7 show the single-cell reaction rates as functions of the phenotypic variable (here the total content of Green Fluorescent Protein or GFP) for the three IPTG concentrations. We observe that increasing inducer concentrations expand the dynamic range of the reaction rate values. However, $R(x)$ appears to be linearly correlated with the total GFP content.

Panels B, D and F of Figure 7 show the single-cell division rates $\Gamma(x)$ for the same three inducer concentrations. Here, we notice a clear exponential correlation of the division rate with the phenotype (i.e. GFP content) for all three IPTG concentrations. This implies that

the higher the total GFP content of the cell, the more likely it is to divide. This behavior is realistic which gives credence to our results.

Figure 8 and Figure 9 depict the recovered partition probability density functions for the three inducer concentrations. Two views of the PPD function for 2000 μM are shown in panels A and B of Figure 8, whereas panels A and B of Figure 9 present the PPD functions for 40 μM and 20 μM respectively. The broad three-dimensional shape of these functions is a strong confirmation that unequal cell partitioning is a major source of phenotypic heterogeneity for these isogenic populations. Note also that the range of values of P , x and y changes significantly with the IPTG concentration, even though the surfaces of the three partition probability density functions have qualitatively similar shapes.

The parameters used in this example to solve the inverse PPDF problems for the three IPTG concentrations were: (a) For [IPTG]=2000 μM : $n = 54$, $m = 20$ and $\lambda^2 = 10^{-16}$; (b) For [IPTG]=40 μM : $n = 54$, $m = 25$ and $\lambda^2 = 1.4 \cdot 10^{-25}$; and (c) For [IPTG]=20 μM : $n = 54$, $m = 25$ and $\lambda^2 = 1.4 \cdot 10^{-19}$. Legendre basis functions have been used in all three cases. Additional details for the selection of these parameters can be found in the supplementary material.

4 Summary and Conclusions

This study presented the development of a robust and accurate methodology to compute the three intrinsic physiological state functions of a heterogeneous cell population using the Collins and Richmond approach. We have performed a detailed parametric analysis to determine how the accuracy of the inverse solution is affected by discretization parameters, the type of non-parametric estimators used, the qualitative characteristics of the phenotypic distributions and the nature of the unknown partitioning probability density function. Moreover, we have carefully assessed the effect of finite sampling and measurement errors on accuracy of the recovered IPS functions. This is crucial because both random measurement errors and finite sampling errors are present in experimental data sets for cell populations. Contrary to our expectations, we found that an NDF formulation provided more accurate solutions for $P(x, y)$ than a CDF formulation, even though the latter used more accurate input data. We attributed this finding to the ill-conditioned nature of the CDF formulation. Through detailed numerical simulations, we also demonstrated that our procedure can accurately estimate all three IPS functions in the presence of both aforementioned sources of uncertainty with only a relatively small sample size of ~ 300 dividing cells and an overall population of ~ 3000 cells.

We demonstrated the potential of our method by applying it to estimate the IPS functions of a model system, a heterogeneous *E. coli* cell population carrying the toggle genetic network. The method performed well for a wide range of concentrations of the chemical inducer IPTG. An important observation was that the concentration of IPTG does not affect how each IPS function correlates with the cell phenotype. For instance, the reaction rates for all three IPTG concentrations were found to be linearly correlated with the total GFP content, while the division rates varied in an exponential (or power) fashion with the GFP content for all three inducer concentrations.

The present study completes the development of an integrated experimental and computational framework that can be used to accurately estimate all three IPS functions required to solve CPB models for cell populations. This integrated framework has the potential to become a very useful tool for quantifying single-cell behavior using data collected from highly heterogeneous cell populations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank Prof. J.J. Collins of Boston University for the donation of the plasmid pTAK117 and Dr. Nikolaos Mantzaris for advice provided in the early stages of this work. Financial support by NIH-NIGMS through grant R01 GM071888 and by NSF through grant NSF-BES-QSB 0331324 is also gratefully acknowledged.

5 References

- Anderson EC, Bell GI, Petersen DF, Tobey RA. Cell growth and division IV. Determination of volume growth: Rate and division probability. *Biophysical Journal*. 1967; 9:246–263. [PubMed: 5764232]
- Aster, RC.; Borchers, B.; Thurber, CH. Parameter estimation and inverse problems. Elsevier Academic Press; Amsterdam: 2005.
- Bæk K, Svenningsen S, Eisen H, Sneppen K, Brown S. Single-cell analysis of λ immunity regulation. *Journal of Molecular Biology*. 2003; 334:363–372. [PubMed: 14623180]
- Block DE, Eitzman PD, Wangenstein JD, Srien F. Slit scanning of *Saccharomyces cerevisiae* cells: Quantification of asymmetric cell division and cell cycle progression in asynchronous culture. *Biotechnology Progress*. 1990; 6:504–512. [PubMed: 1366842]
- Bowman, AW.; Azzalini, A. The kernel approach with S-Plus illustrations. Clarendon Press; Oxford: 1997. Applied smoothing techniques for data analysis.
- Campbell A. Synchronization of cell division. *Bacteriological Reviews*. 1957; 21:263–272. [PubMed: 13488884]
- Chung JD, Stephanopoulos G. Studies of transcriptional state heterogeneity in sporulating cultures of *Bacillus subtilis*. *Biotechnology and Bioengineering*. 1995; 47:234–242. [PubMed: 18623397]
- Collins JF, Richmond MH. Rate of growth of *Bacillus cereus* between divisions. *Journal of General Microbiology*. 1962; 28:15–33. [PubMed: 13880594]
- Delbrück M. Effects of specific antisera on the growth of bacterial viruses (bacteriophages). *Journal of Bacteriology*. 1945; 50:137–150.
- Eakman JM, Fredrickson AG, Tsuchiya HM. Statistics and dynamics of microbial cell populations. *Chemical Engineering Progress Symposium Series*. 1966; 62:37–49.
- Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. *Science*. 2002; 297:1183–1186. [PubMed: 12183631]
- Fredrickson AG, Ramkrishna D, Tsuchiya HM. Statistics and dynamics of prokaryotic cell populations. *Mathematical Biosciences*. 1967; 1:327–374.
- Freedman D, Diaconis P. On the histogram as a density estimator: L_2 theory. *Z. Wahrscheinlichkeitstheor. Verw. Geb.* 1981; 57:453–476.
- Gardner TS, Cantor CR, Collins JJ. Construction of a genetic toggle switch in *Escherichia coli*. *Nature*. 2000; 403:339–342. [PubMed: 10659857]
- Godin FB, Cooper DG, Rey AD. Numerical methods for a population balance model of a periodic fermentation process. *AIChE Journal*. 1999; 45:1359–1364.
- Hansen, PC. Numerical aspects of linear inversion. Society for industrial and applied mathematics; Philadelphia: 1998. Rank-deficient and discrete ill-posed problems.
- Harvey RJ, Marr AG, Painter PR. Kinetics of growth of individual cells of *Escherichia coli* and *Azotobacter agilis*. *Journal of Bacteriology*. 1967; 93:605–617. [PubMed: 5335966]
- Hatzis C, Fredrickson AG, Srien F. Cell-cycle analysis in phagotrophic microorganisms from flow cytometric histograms. *Journal of Theoretical Biology*. 1997; 186:131–144.
- Hofmann, B. Regularization for applied-inverse and ill-posed problems. B. G. Teubner; Leipzig: 1986.
- Hofmann B. Regularization of nonlinear problems and the degree of ill-posedness. *Mathematical Research*. 1993; 74:174–188.
- Izenman AJ. Recent developments in nonparametric density estimation. *Journal of the American Statistical Association*. 1991; 86:205–224.

- Jerri, AJ. Introduction to integral equations with applications. 2nd ed.. John Wiley & Sons, Inc.; New York: 1999.
- Kavousanakis ME, Mantzaris NV, Boudouvis AG. A novel free boundary algorithm for the solution of cell population balance models. *Chemical Engineering Science*. 2009; 64:4247–4261.
- Kempner ES, Marr AG. Growth in volume of *Euglena gracilis* during the division cycle. *Journal of Bacteriology*. 1979; 101:561–567. [PubMed: 5413825]
- Kepler TB, Elston TC. Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. *Biophysical Journal*. 2001; 81:3116–3136. [PubMed: 11720979]
- Koch AL. Distribution of cell size in growing cultures of bacteria and the applicability of the Collins - Richmond principle. *Journal of General Microbiology*. 1966; 45:409–417.
- Kostova TV. Numerical solutions to equations modeling nonlinearity interacting age-dependent populations. *Computational and Applied Mathematics*. 1990; 19:95–103.
- Kromenaker SJ, Srien F. Effect of lactic acid on the kinetics of growth and antibody production in a murine hybridoma: secretion patterns during the cell cycle. *Journal of Biotechnology*. 1994; 34:13–34. [PubMed: 7764742]
- Mahoney AW, Doyle FJ, Doraiswami R. Inverse problems in population balances: growth and nucleation from dynamic data. *AIChE Journal*. 2002; 48:981–990.
- Maloney PC, Rotman B. Distribution of suboptimally induced β -D-galactosidase in *Escherichia coli*: the enzyme content of the individual cell. *Journal of Molecular Biology*. 1972; 73:77–91. [PubMed: 4570383]
- Mantzaris NV. Stochastic and deterministic simulations of heterogeneous cell population dynamics. *Journal of Theoretical Biology*. 2006; 241:690–706. [PubMed: 16487980]
- Mantzaris NV, Daoutidis P, Srien F. Numerical solution of multi-variable cell population balance models: I. Finite difference methods. *Computers and Chemical Engineering*. 2001a; 25:1411–1440.
- Mantzaris NV, Daoutidis P, Srien F. Numerical solution of multi-variable cell population balance models: II. Spectral methods. *Computers and Chemical Engineering*. 2001b; 25:1441–1462.
- Mantzaris NV, Daoutidis P, Srien F. Numerical solution of multi-variable cell population balance models: III. Finite element methods. *Computers and Chemical Engineering*. 2001c; 25:1463–1481.
- Muralidar R, Ramkrishna D. An inverse problem in agglomeration kinetics. *Journal of Colloid and Interface Science*. 1986; 112:348–361.
- Muralidar R, Ramkrishna D. Inverse problem of agglomeration kinetics II. Binary clustering coefficients from self-preserving spectra. *Journal of Colloid and Interface Science*. 1989; 131:503–513.
- Neumann, J.v. Various techniques used in connection with random digits. Monte Carlo methods. National Bureau of Standards. 1951; 12:36–38.
- Novick A, Weiner M. Enzyme induction as an all-or-none phenomenon. *Proceedings of the National Academy of Sciences of the United States of America*. 1957; 43:553–566. [PubMed: 16590055]
- Painter PR, Marr AG. Mathematics of microbial populations. *Annual review of microbiology*. 1968; 22:519–548.
- Patruno LE, Dorao CA, Dupuy PM, Svendsen HF, Jakobsen HA. Identification of droplet breakage kernel for population balance modeling. *Chemical Engineering Science*. 2008; 64:638–645.
- Patruno LE, Dorao CA, Svendsen HF, Jakobsen HA. Analysis of breakage kernels for population balance modeling. *Chemical Engineering Science*. 2009; 64:501–508.
- Powell EO. Growth rate and generation time of bacteria, with special reference to continuous culture. *Journal of General Microbiology*. 1956; 15:492–511. [PubMed: 13385433]
- Powell EO. A note on Koch and Schaechter's hypothesis about growth and fission of bacteria. *Journal of General Microbiology*. 1964; 37:231–249. [PubMed: 14247748]
- Ramkrishna D. Toward a self-similar theory of microbial populations. *Biotechnology and Bioengineering*. 1994; 43:138–148. [PubMed: 18615607]
- Ramkrishna, D. Population balances: theory and applications to particulate systems in engineering. Academic Press; San Diego, CA: 2000.

- Ramkrishna D, Fredrickson AG, Tsuchiya HM. On relationships between various distribution functions in balanced unicellular growth. *Bulletin of Mathematical Biophysics*. 1968; 30:319–323.
- Ramkrishna D, Schell J. On self-similar growth. *Journal of Biotechnology*. 1999; 71:255–258.
- Robert, CP.; Casella, G. *Monte Carlo Statistical Methods*. Springer-Verlag; New York: 2004.
- Russo-Marie F, Roederer M, Sager B, Herzenberg LA, Kaiser D. β -Galactosidase activity in single differentiating bacterial cells. *Proceedings of the National Academy of Sciences of the United States of America*. 1993; 90:8194–8198. [PubMed: 8396263]
- Sathyagal AN, Ramkrishna D, Narsimhan G. Solution of inverse problems in population balances - II. Particle break-up. *Computers and Chemical Engineering*. 1995; 19:437–451.
- Silverman, BW. *Density estimation for statistics and data analysis*. 1st ed.. Chapman & Hall; London: 1986.
- Spetsieris K, Zygourakis K, Mantzaris NV. A novel assay based on fluorescence microscopy and image processing for determining phenotypic distributions of rod-shaped bacteria. *Biotechnology and Bioengineering*. 2009; 102:598–615. [PubMed: 18853409]
- Spudich JL, Koshland DE. Non-genetic individuality: chance in the single cell. *Nature*. 1976; 262:467–476. [PubMed: 958399]
- Srienc F, Dien BS. Kinetics of the cell cycle of *Saccharomyces cerevisiae*. *Annals of the New York Academy of Sciences*. 1992; 655:59–71. [PubMed: 1416620]
- Stocker BAD. Measurements of rate mutation of flagellar antigenic phase in *Salmonella typhimurium*. *Journal of Hygiene Cambridge*. 1949; 47:398–413.
- Subramanian G, Ramkrishna D. On the solution of statistical models of cell populations. *Mathematical Biosciences*. 1971; 10:1–23.
- Sulsky D. Numerical solution of structured population balance models II Mass structure. *Journal of Mathematical Biology*. 1994; 32:491–514.
- Sweeney PJ, Srienc F, Fredrickson AG. Measurement of Unequal DNA partitioning in *Tetrahymena pyriformis* using slit-scanning flow cytometry. *Biotechnology Progress*. 1994; 10:19–25.
- Thompson, JR.; Tapia, RA. *Nonparametric function estimation, modeling, and simulation*. Society for Industrial and Applied Mathematics; Philadelphia: 1990.
- Tikhonov, AN.; Arsenin, VY. *Solutions of ill-posed problems*. Winston and Sons; Washington, DC: 1977.
- Tsuchiya HM, Fredrickson AG, Aris R. Dynamics of microbial cell populations. *Advances in Chemical Engineering*. 1966; 6:125–206.
- Wand, MP.; Jones, MC. *Kernel smoothing*. 1st ed.. Chapman & Hall; London: 1995.
- Wright H, Ramkrishna D. Solutions of inverse problems in population balances - I. Aggregation kinetics. *Computers and Chemical Engineering*. 1992; 16:1019–1038.
- Zahrt, JD. *The problem of deconvolution and unfolding*. Society of Industrial and Applied Mathematics; Philadelphia: 1991. A primer on integral equations of the first kind.
- Zusman D, Gottlieb P, E. R. Division cycle of *Myxococcus xanthus*. 3. Kinetics of cell growth and protein synthesis. *Journal of Bacteriology*. 1971; 105:811–819. [PubMed: 4926683]

Highlights

- Cell population balance models and intrinsic physiological state functions (IPSF).
- An inverse problem must be solved to compute IPSF from phenotypic distribution data.
- We develop and test a robust computational method to accurately solve this problem.
- We compute the IPSF of E. coli populations carrying a toggle network to validate it.

Figure 1.

Schematic of the process that uses simulated data to assess the effect of both finite sampling and random errors on the accuracy of the inverse solution.

Figure 2.

Panel A: Effect of finite sampling on single-cell reaction rate. Results obtained with the NDF (dashed-dotted line) and CDF (dotted line) methods are compared to the analytical solution (solid line). Panel B: Comparison of analytical (solid line) and inverse solutions (dashed line) showing the effect of finite sampling on the single-cell division rate.

Figure 3.
 Panel A: Effect of sample size in the accuracy of the partitioning function $Q(f)$. Panel B:
 Eigenvalue spectrum for the coefficient matrix G of both NDF and CDF methods

Figure 4. Effect of PPDF characteristics for $N_d = 300$. Panel A: Symmetric Beta distribution with $q = 10$ and ~4% error, Panel B: Symmetric Beta distribution with $q = 60$ and ~5% error Panel C: Bimodal distribution with $\mu_{part} = 0.3$, $\sigma_{part} = 0.08$ and ~8 % error, Panel D: Bimodal distribution with $\mu_{part} = 0.4$, $\sigma_{part} = 0.035$ and ~10% error.

Figure 5. Effect of combined random and finite sampling errors in the experimental data, on the recovery of the PPDF for a finite sample of 300 dividing cells.

Figure 6. Phenotypic distributions $n(x)$, $n_d(x)$ and $n_b(x)$ obtained via nonparametric estimates from experimental data for *E. coli* populations carrying the toggle genetic network and three different concentrations of the IPTG inducer. Panel A: [IPTG] = 2000 μM , Panel B: [IPTG] = 40 μM , and Panel C: [IPTG] = 20 μM .

Figure 7. Estimated single-cell reaction and division rates for toggle at three IPTG concentrations. Panels A–B: $[IPTG] = 2000 \mu M$, Panel C–D: $[IPTG] = 40 \mu M$, and Panel E–F: $[IPTG] = 20 \mu M$.

A

B

Figure 8. Recovered partition probability density function (PPDF) for a population of *E. coli* carrying the toggle genetic network and a 2000 μM concentration of the chemical inducer IPTG. Panels A and B show the PPDF from different perspectives and with different axes limits.

A

B

Figure 9. Recovered partition probability density function (PPDF) for populations of *E. coli* carrying the toggle genetic network and different concentrations of the chemical inducer IPTG. Panel A: [IPTG] = 40 μ M, Panel B: [IPTG] = 20 μ M.