Quasi-Newton Stochastic Optimization Algorithm for Parameter Estimation of a Stochastic Model of the Budding Yeast Cell Cycle

Minghan Chen, Brandon D. Amos, Layne T. Watson, John J. Tyson, Yang Cao,
Clifford A. Shaffer, Michael W. Trosset, Cihan Oguz, Gisella Kakoti

Abstract—Parameter estimation in discrete or continuous deterministic cell cycle models is challenging for several reasons, including the nature of what can be observed, and the accuracy and quantity of those observations. The challenge is even greater for stochastic models, where the number of simulations and amount of empirical data must be even larger to obtain statistically valid parameter estimates. This work describes a new quasi-Newton algorithm class QNSTOP for stochastic optimization problems. QNSTOP directly uses the random objective function value samples rather than creating ensemble statistics. QNSTOP is used here to directly match empirical and simulated joint probability distributions rather than matching summary statistics. Results are given for a current state-of-the-art stochastic cell cycle model of budding yeast, whose predictions match well some summary statistics and one-dimensional distributions from empirical data, but do not match well the empirical joint distributions. The nature of the mismatch provides insight into the weakness in the stochastic model.

Index Terms—Budding yeast, cell cycle model, model parameter estimation, quasi-Newton algorithm, stochastic optimization problem.

1 Introduction

A fundamental challenge of molecular systems biology is to build accurate dynamical models of the molecular mechanisms underlying various aspects of cell physiology, e.g., cellular chemotaxis or the regulation of cell growth and division. Typically, these models are expressed in terms of differential equations, i.e., the models are 'deterministic', and their validity is assessed by comparison of model simulations to the observed (average) properties of large populations of cells responding to various experimental conditions. In recent years, however, cell biologists are increasingly able to measure the behavior and molecular constitution of single cells as they go about their business in space and time. As might be expected, the specific behavior of any given cell may be quite different than the average behavior of a population of cells, reflecting molecular variability between cells. Regardless of the source(s) of such variability, which may be at the levels of DNA, mRNA, protein, and/or signaling molecules, deterministic models of the behavior must be converted into realistic stochastic models to deal with the variability of responses from one cell to another.

An important and difficult aspect of any modeling project is estimation of the kinetic constants ('model parameters') that appear in any dynamical model (deterministic or stochastic) of a molecular regulatory process. The parameters (e.g., rates of gene expression, rate constants for mRNA and protein degradation, rates of association and dissociation of molecular complexes, etc.) are estimated by comparison of model simulations to relevant experimental measurements of molecular turnover in cells. For deterministic models the problem is difficult enough, because any reasonably complete model will have dozens of molecular species and many dozens of undetermined parameters, but the available experimental data is often quite extensive, and there exist powerful algorithms for fitting deterministic simulations to experimental data points. For stochastic models the problem is considerably more difficult, because a stochastic model adds many more parameters to the deterministic model on which it is based, and the specific sorts of data required to estimate these 'stochastic' parameters is often difficult to obtain experimentally. Furthermore, stochastic simulations generate statistical distributions of observables, and these distributions must be compared to experimentally observed distributions, and the parameters estimated by optimization of an objective function that is a random variable. Algorithms for such stochastic optimization problems are still being developed and assessed.

This paper presents results on optimization of the parameters in a stochastic model of cell cycle regulation in budding yeast. Section 2 describes the model briefly. Section 3 states the mathematical optimization problem precisely. Section 4 outlines a new quasi-Newton algorithm (QNSTOP) for stochastic optimization, and Section 5 presents the results of using QNSTOP to fit the stochastic cell cycle model to observed distributions of cell cycle observables (mass at birth, cell cycle time, duration of G_1 phase). Section 6 discusses some biological implications of the results, and conclusions are drawn in Section 7.

2 Stochastic Cell Cycle Model

The cell cycle model used in this paper was developed originally by Teeraphan Laomettachit and is described in full in his Ph.D. thesis [4]. The deterministic version of the model uses a set of nonlinear differential algebraic equations (DAEs) to track the temporal evolution of 26 variables (proteins governing progression through the budding yeast cell cycle). These equations involve 126 parameters (kinetic constants) that are estimated by fitting simulations of the model to the observed phenotypes of 119 budding yeast strains.

The initial determination of the 'best' parameter values was done 'by hand' as follows. Starting with an initial 'basal' parameter vector X_{basal}, simulate the sequence of cell cycle events in 'wild-type' cells growing in glucose and in galactose, making sure that the cells are viable under both conditions. Then simulate the phenotypes of 117 mutant strains of budding yeast growing in either glucose or galactose. Each mutant strain is characterized by a set of genetic changes (e.g., gene A is knocked out and gene B is overexpressed two-fold). The strain is simulated by appropriate changes to the basal parameter vector (e.g., the rate constant for synthesis of protein A from gene A is set to zero, and the rate constant for synthesis of protein B from gene B is set to twice the basal value). Each mutant strain has an observed phenotype: viable or inviable; if viable then some observed birth size relative to wild-type cells; if inviable then stuck at some particular stage of the cell cycle. The simulated phenotype of each mutant strain is compared to the observed phenotype, and the basal parameter vector is scored accordingly. Then the basal parameter vector is modified, the simulations are repeated and rescored, and the process is repeated until no further improvement seems to be possible.

Surprisingly, despite the immensity of the parameter space, a good modeler can make significant improvements to the basal parameter vector by hand in a few weeks, and the process is necessary (from the modeler's point of view) in order to understand the vagaries of the model with respect to the experimental data. In the process, the modeler often makes slight 'tweaks' to the underlying molecular model (the DAEs) in order to get better agreement between the model and the mutant phenotypes.

Once the deterministic model (from [4]) was fitted as well as possible to the data set (the phenotypes of 110 of 119 strains correctly simulated), it was converted to a stochastic model in order to explore the observed variability of cell cycle progression among single cells (wild-type and mutant strains). The conversion was made in two steps. First, the dimensionless variables of the DAE model (call them $z_i(t)$, $i = 1, \ldots, 26$) had to be converted into numbers of molecules of species i per cell, $c_i \; z_i(t)$, where c_i is the 'characteristic concentration' of species i. Then each of the differential equations of the system of DAEs was converted into a stochastic differential equation of the Langevin type by adding two random noise terms to the right hand side. The first noise term had the usual form of a birth-death process for the protein species, and the second term was designed specifically to model the effects of mRNA fluctuations on noisy protein expression; see Laomettachit's thesis [4]. These two steps introduced 52 new 'stochastic' parameters into the model: 22 characteristic concentrations, and 30 parameters describing the coupling between mRNA expression and protein synthesis.

Laomettachit estimated these stochastic parameters by hand, as well. From experimental estimates of the average numbers of protein molecules per cell for each cell cycle gene, he could estimate the 22 characteristic concentrations. From reasonable guesses about mRNA dynamics in budding yeast cells, he could estimate the 30 other parameters. These estimates gave quite acceptable agreement with the limited amount of statistical data at his disposal for the distributions of cell cycle observables in populations of wild-type cells.

The Laomettachit model of the budding yeast cell cycle was further examined by Oguz et al. [5], who explored the utility of differential evolution (DE) as a tool for characterizing the parameter space of the model. These authors started from an intermediate stage of Laomettachit's search (a basal parameter vector that accounted correctly for the phenotypes of only 72 of 119 strains). They found that DE could quickly improve the score (i.e., the number of phenotypes correctly simulated) of the basal parameter vector, but could not improve on the score that Laomettachit achieved by hand. That is to say, 92.5% (110/119) seems to be about the best fit that Laomettachit's deterministic model can achieve. In a later publication, Oguz et al. [6] applied DE to the stochastic version of Laomettachit's model. They held the 126 deterministic parameters fixed at the values determined by the mutant phenotypes, and they estimated the 52 stochastic parameters by DE. The objective function in this case was constructed by comparing simulated values and observed values for the means and variances of certain cell-cycle observables: total cycle time and duration of G_1 phase of the cell cycle, for mother cells and daughter cells. The purpose of this paper was not so much to estimate the stochastic parameters of the model as to use the parametrized model to study the synchronization of cell division in budding yeast populations by external perturbations; see [6] for details.

3 The Mathematical Problem

As explained in the previous section, stochastic models of the cell cycle are necessary to explain the observed variability in cell cycle progression among individual cells. Estimating the parameters in a stochastic cell cycle model is challenging, both mathematically and empirically. Obtaining accurate and useful data from individual cells is difficult, and very little such data exists in the literature. Regardless of what criterion is minimized to estimate the model parameters, the mathematical problem is a stochastic optimization problem, meaning that the objective function $\theta(x)$ itself is a random variable. To further complicate matters, the random noise in the objective function is not additive, i.e., the objective function is not of the form (deterministic $\theta(x)$) + (random noise). The randomness is buried deep in the simulation model, and has no simple representation at the output level of the simulation model.

For a real colony of cells and a simulated colony, several properties (e.g., mass at birth m_B and duration of G_1 phase T_{G_1}) can be observed. It is common practice to compute statistics (e.g., mean, variance) of these observables and then to estimate the simulation model parameters by minimizing the difference (measured somehow) between the empirical colony's statistics and the simulated colony's statistics. For example, both Laomettachit [4] and Oguz et al. [6] approximated the scatter plot of the two-dimensional joint distribution of m_B versus T_{G_1} by a dogleg (continuous piecewise linear function with two line segments), and then estimated stochastic model parameters by matching the slopes of the line segments in the two (empirical and model predicted) doglegs. Matching these statistics is certainly a *necessary* condition for the correctness of the model, but such summary statistics do not capture all the available information. What one really wants to do, for example, is match the empirical colony's distribution of m_B with the simulated colony's distribution of m_B. Even better, match the distributions for all the observables simultaneously, or even match the joint distributions. The proposal here is to do exactly that—for both mother and daughter budding yeast cells, match the joint distributions of the pair (mass at birth, duration of of G_1 phase) from the empirical and simulated cell colonies.

Postponing until later the details of obtaining (approximations of) these distributions, let $p(i)$ and $q(i)$ denote the probability mass (after discretization of the probability density) functions of the empirical and of the simulated colony's observable, respectively. There are several standard, well-justified ways to compare distributions. From an information theoretic perspective comes the Kullback-Leibler divergence

$$d_{\mathrm{KL}}(p,q) = \sum_i p(i) \log_2 \left(\frac{p(i)}{q(i)} \right),$$

which is nonnegative and zero if and only if $p = q$, but is not a metric. Another criterion from statistics is the Hellinger distance

$$d_H(p,q) = \left(\sum_i \left(\sqrt{p(i)} - \sqrt{q(i)} \right)^2 \right)^{1/2},$$

which is a metric. Depending on how discretization (one- or two-dimensional histograms) is done, some of the simulation probabilities $q(i)$ might be zero (a histogram interval or box has no points in it), which makes d_{KL} infinite. d_H is better behaved in such cases. Both d_{KL} and d_H were tried for this work, but only results for d_H are reported.

Let $X \in \mathbb{R}^n$ be the vector of parameters to be estimated in the stochastic cell cycle model. Let $p(i)$ and $q(i)$ be the probability mass functions of the observable (e.g., m_B or the pair (m_B, T_{G_1})) from the empirical cell colony and from the simulated cell colony, respectively. $p(i)$ is constant, but $q(i)$ is a random variable determined by a stochastic simulation. The objective function is the random variable

$$f(X) = d_H(p,q),$$

and the stochastic optimization problem to be solved is

$$\min_{L \leq X \leq U} f(X),$$

where $[L, U]$ is a box in $\mathbb{R}^n$ defining the feasible set (allowable values for the model parameters X).

The approach taken here, aptly described as simulation-based parameter estimation, has a long history in statistics, which is discussed, with historical references, in Castle's Ph.D. thesis [2]. The original version of the algorithm QNSTOP summarized in the next section is due to Castle [2]; based on experience with applications, and considerations of numerical stability and computational efficiency, the original version of QNSTOP has evolved to that of Amos et al. [1], the version used here.

4 Quasi-Newton Algorithm for Stochastic Optimization

QNSTOP is a class of quasi-Newton methods developed for stochastic optimization that can also be used for deterministic global optimization. Complete mathematical details, convergence theory, and programming implementation details are in [1]. The essential steps are outlined here. In iteration k, QNSTOP computes the gradient vector $\hat{g}_k$ and Hessian matrix $\hat{H}_k$ of a quadratic model

$$\begin{aligned}\widehat{m}_k(X - X_k) = \hat{f}_k + \hat{g}_k^T (X - X_k) \\ + \frac{1}{2} (X - X_k)^T \hat{H}_k (X - X_k)\end{aligned}$$

of the objective function f centered at X_k, where $\hat{f}_k$ is generally not $f(X_k)$. The next iterate is

$$X_{k+1} = \left(X_k - \left[\hat{H}_k + \mu_k W_k \right]^{-1} \hat{g}_k \right)_{\Theta},$$

where μ_k is the Lagrange multiplier of a trust region subproblem, W_k is a symmetric, positive definite scaling matrix, and $(\cdot)_\Theta$ denotes projection onto the feasible set $\Theta = [L, U]$.

To estimate the gradient, QNSTOP uses an ellipsoidal design region centered at the current iterate $X_k \in \mathbb{R}^n$. Let

$$\begin{aligned}W_\gamma = \big\{ W \in \mathbb{R}^{n \times n} : W = W^T,\ \det(W) = 1, \\ \gamma^{-1} I_n \preceq W \preceq \gamma I_n \big\}\end{aligned}$$

for some $\gamma \geq 1$ where I_n is the $n \times n$ identity matrix. The elements of the set W_γ are valid scaling matrices that control the shape of the ellipsoidal design regions with eccentricity constrained by γ. Let the ellipsoidal design regions, with radius τ_k, be given by

$$E_k(\tau_k) = \left\{ X \in \mathbb{R}^n : (X - X_k)^T W_k (X - X_k) \leq \tau_k^2 \right\}$$

where $W_k \in W_\gamma$.

In each iteration, QNSTOP chooses a set of N uniformly sampled design sites $\{X_{k1}, \ldots, X_{kN}\} \subset E_k(\tau_k) \cap \Theta$. Let $Y_k = (y_{k1}, \ldots, y_{kN})^T$ denote the N-vector of responses modeled by the linear model $y_{ki} = \hat{f}_k + X_{ki}^T \hat{g}_k + \epsilon_{ki}$ where ϵ_{ki} accounts for lack of fit. $\hat{g}_k$ is then the least squares estimate of the linear model gradient.

Depending on the context, QNSTOP either constrains the Hessian matrix update to satisfy

$$-\eta I_n \preceq \hat{H}_k - \hat{H}_{k-1} \preceq \eta I_n$$

for some $\eta \geq 0$, using a variation of the SR1 (symmetric, rank one) quasi-Newton update, or uses the unconstrained BFGS quasi-Newton update

$$\hat{H}_k = \hat{H}_{k-1} - \frac{\hat{H}_{k-1} s_k s_k^T \hat{H}_{k-1}}{s_k^T \hat{H}_{k-1} s_k} + \frac{\nu_k \, \nu_k^T}{\nu_k^T s_k},$$

where $s_k = X_k - X_{k-1}$, $\nu_k = \hat{g}_k - \hat{g}_{k-1}$.

QNSTOP utilizes an ellipsoidal trust region concentric with the design region for controlling step length. In one usage mode, the trust region ellipsoid radius ρ_k is taken equal to the design ellipsoid radius τ_k, and the optimization problem

$$\min_{X \in E_k(\rho_k)} \hat{g}_k^T (X - X_k) + \frac{1}{2} (X - X_k)^T \hat{H}_k (X - X_k)$$

is solved for X_{k+1} and μ_k related by

$$X_{k+1} = X(\mu_k) = X_k - \left[\hat{H}_k + \mu_k W_k \right]^{-1} \hat{g}_k.$$

In another usage mode, μ_{k-1} is directly updated to μ_k, giving $X_{k+1} = X(\mu_k)$ as above. If necessary, X_{k+1} is projected back into the feasible set Θ.

Finally, the experimental design region $E_k(\tau_k)$ is updated to approximate a confidence set by updating the scaling matrix W_k. The updated scaling matrix is given by

$$W_{k+1} = \left(\hat{H}_k + \mu_k W_k \right)^T V_k^{-1} \left(\hat{H}_k + \mu_k W_k \right),$$

where V_k is the covariance matrix of $\nabla \widehat{m}_k(X_{k+1} - X_k)$. For numerical stability, W_{k+1} is constrained (by modifying its eigenvalues) to satisfy the constraints $\gamma^{-1} I_n \preceq W_{k+1} \preceq \gamma I_n$ and $\det(W_{k+1}) = 1$, so $W_\gamma \ni W_{k+1}$.

Algorithm summary: It is generally desirable to run QNSTOP from multiple start points, and the algorithm described below is repeated for each start point.

Step 0 (initialization): Given a function evaluation budget $\tilde{B}$ per start point and operating mode (choices of quasi-Newton update, ways of updating the ellipsoidal design region radii τ_k and ellipsoidal trust region radii ρ_k, etc.), set values for $\tau_0 > 0$, $\gamma \geq 1$, $\eta \geq 0$, N, X_0, $k := 0$, $W_0 := \hat{H}_0 := I_n$.

Step 1 (regression experiment): Depending on the usage mode, compute the design ellipsoid radius τ_k. Uniformly sample $\{X_{k1}, \ldots, X_{kN}\} \subset E_k(\tau_k) \cap \Theta$. Observe the response vector $Y_k = (y_{k1}, \ldots, y_{kN})^T$. Compute $\hat{g}_k$ by linear regression.

Step 2 (secant update): If $k > 0$, compute the model Hessian matrix $\hat{H}_k$ using either the BFGS or SR1 variant update, depending on the usage mode.

Step 3 (update iterate): Compute μ_k depending on the usage mode, solve $[\hat{H}_k + \mu_k W_k]s_k = -\hat{g}_k$ for the step s_k, and compute $X_{k+1} = \big(X_k + s_k\big)_\Theta$.

Step 4 (update subsequent design ellipsoid): Compute a new scaling matrix $W_{k+1} \in W_\gamma$.

Step 5: If $(k+2)(N+1)+1 < \tilde{B}$ then increment k by 1 and go to **Step 1**. Otherwise, the algorithm terminates. (f is also observed at each ellipsoid center X_k.)

The Fortran 2003 parallel (OpenMP) subroutine QNSTOPP from [1] is used here, and the nondefault values for all inputs to QNSTOPP are reported with the results later.

5 Numerical Results and Discussion

The budding yeast stochastic cell cycle model in [6], called 'Laomettachit's stochastic model' here, has 52 parameters that are exclusive to the stochastic aspects of the model, of which some are chosen to be equal to others, leaving 44 independent variables (parameters) to be determined by some mathematical procedure (here, solving a stochastic optimization problem). The parameter names follow a pattern: the species $\star$ is denoted by an index 1, 2, . . ., 10, referring to species Cln3, Bck2, Cln2, CKI, Clb5, Clb2, Swi5, Cdc20, Pds1, and POLO, respectively. The parameters $k_{tr\star}$, $k_{dm\star}$, $m_{min\star}$ for species $\star$ are, respectively, translation rate, mRNA degradation rate, and minimum number of mRNA molecules. This accounts for 30 parameters. The remaining 22 parameters c_x, where x is the species name, are the characteristic concentrations of the above ten species and 12 other species: Whi5, SBF, Cdh1, APCP, Clb14, Net1, PPX, Esp1, Cdc15, Tem1, MEN, Mcm1. These characteristic concentrations are introduced to convert the dimensionless concentrations of the species in the deterministic version of Laomettachit's model into numbers of molecules per cell for each species in the stochastic version of the model. Since some of the species in the model bind with each other to form stoichiometric complexes, the characteristic concentrations of such binding partners must be identical. Therefore, as in Oguz et al. [6], making the eight assignments $c_{\mathrm{SBF}} \equiv c_{\mathrm{Whi5}}$, $c_{\mathrm{Clb2}} \equiv c_{\mathrm{Clb5}} \equiv c_{\mathrm{CKI}}$, $c_{\mathrm{APCP}} \equiv c_{\mathrm{Cdc20}}$, $c_{\mathrm{Net1}} \equiv c_{\mathrm{Cdc14}}$, $c_{\mathrm{Esp1}} \equiv c_{\mathrm{Pds1}}$, and $c_{\mathrm{MEN}} \equiv c_{\mathrm{Tem1}} \equiv c_{\mathrm{Cdc15}}$ leaves 44 independent parameters defining the vector X.

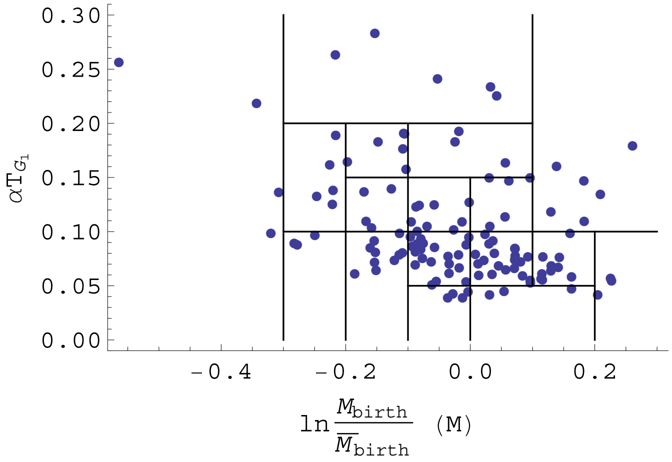

Fig. 1. Discretization for empirical correlations of mass at birth and scaled duration of G_1 phase of mother cells. The x-axis is ln (individual mass/mean mass), where the mean is of all mother cells.

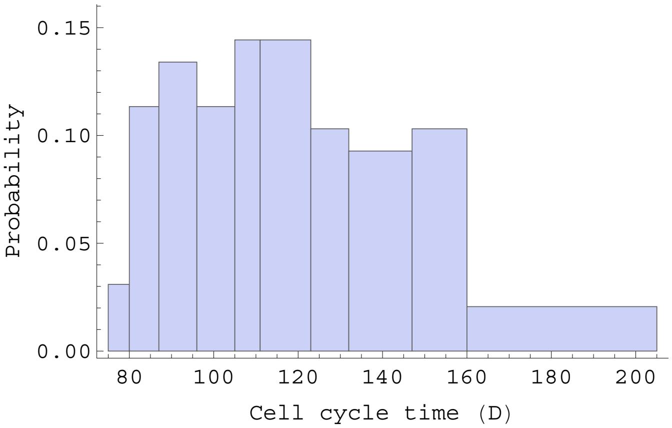

Fig. 2. Discretization for empirical daughter cell cycle time.

Table 1 lists the 52 stochastic parameters in Laomettachit's model. The nominal vector X_0 defines the search box $[L, U]$, where the bounding interval for the ith component $(X)_i$ of X is $\big[(1/\varphi_i)(X_0)_i, \varphi_i(X_0)_i\big]$ and each factor φ_i is either 2 or 5. The 'best [6] vector' is the best estimate of the parameter vector found by [6] using differential evolution.

The empirical data from Di Talia et al. [3] includes mass at birth, duration of G_1 phase, and cell cycle time of both mother and daughter budding yeast cells. Using the Hellinger distance to measure the difference between the empirical data distribution and the simulated data distribution requires approximating the continuous distributions by (one- or two-dimensional) histograms. For example, Figure 1 shows the histogram box boundaries for the joint distribution of the (scaled) pair (mass at birth, duration of G_1 phase) for mother cells. The strategy is to define rectangles (or intervals in one dimension) that roughly evenly divide the empirical data points and such that every rectangle (or interval) contains some data points. The 122 data points yield 17 bins (divided by black lines in Figure 1). The particular discretization has no effect on the optimization algorithm. Figure 2 shows the one-dimensional histogram for the empirical data of daughter cell cycle times. Here the 97 data points are divided into 10 bins. Given the sparsity and

TABLE 1
List of parameters in stochastic cell cycle model.

parameter	nominal value	best [6] value	$[L, U]$
k_{tr1}	0.22	0.3870	$[0.044, 1.1]$
k_{dm1}	0.7	2.9459	$[0.14, 3.5]$
m_{min1}	1.0	5.0	$[0.2, 5.0]$
k_{tr2}	0.22	0.6166	$[0.044, 1.1]$
k_{dm2}	0.7	0.6033	$[0.14, 3.5]$
m_{min2}	4.0	17.0	$[0.8, 20.0]$
k_{tr3}	0.22	0.0761	$[0.044, 1.1]$
k_{dm3}	0.7	2.9502	$[0.14, 3.5]$
m_{min3}	1.0	2.0	$[0.2, 5.0]$
k_{tr4}	0.22	0.2024	$[0.044, 1.1]$
k_{dm4}	0.7	1.4652	$[0.14, 3.5]$
m_{min4}	4.0	1.0	$[0.8, 20.0]$
k_{tr5}	0.22	0.6878	$[0.044, 1.1]$
k_{dm5}	0.7	0.1975	$[0.14, 3.5]$
m_{min5}	4.0	8.0	$[0.8, 20.0]$
k_{tr6}	0.22	0.6974	$[0.044, 1.1]$
k_{dm6}	0.7	1.6668	$[0.14, 3.5]$
m_{min6}	4.0	15.0	$[0.8, 20.0]$
k_{tr7}	0.22	0.8867	$[0.044, 1.1]$
k_{dm7}	0.7	2.4182	$[0.14, 3.5]$
m_{min7}	4.0	16.0	$[0.8, 20.0]$
k_{tr8}	0.22	0.7344	$[0.044, 1.1]$
k_{dm8}	0.7	3.4411	$[0.14, 3.5]$
m_{min8}	4.0	6.0	$[0.8, 20.0]$
k_{tr9}	0.22	0.6737	$[0.044, 1.1]$
k_{dm9}	0.7	1.2706	$[0.14, 3.5]$
m_{min9}	4.0	9.0	$[0.8, 20.0]$
k_{tr10}	0.22	0.4258	$[0.044, 1.1]$
k_{dm10}	0.7	0.1469	$[0.14, 3.5]$
m_{min10}	4.0	5.0	$[0.8, 20.0]$
c_{Cln3}	10.0	19.0957	$[5.0, 20.0]$
c_{Bck2}	10.0	16.3317	$[5.0, 20.0]$
c_{Whi5}	22.0	21.8688	$[11.0, 44.0]$
c_{SBF}	22.0	21.8688	$[11.0, 44.0]$
c_{Cln2}	45.0	84.2260	$[22.5, 90.0]$
c_{CKI}	80.0	101.9969	$[40.0, 160.0]$
c_{Clb5}	80.0	101.9969	$[40.0, 160.0]$
c_{Clb2}	80.0	101.9969	$[40.0, 160.0]$
c_{Swi5}	57.5	50.4561	$[28.75, 115.0]$
c_{Cdc20}	100.0	93.1338	$[50.0, 200.0]$
c_{Cdh1}	100.0	59.4664	$[50.0, 200.0]$
c_{APCP}	100.0	93.1338	$[50.0, 200.0]$
c_{Cdc14}	14.0	20.2049	$[7.0, 28.0]$
c_{Net1}	14.0	20.2049	$[7.0, 28.0]$
c_{PPX}	100.0	81.0649	$[50.0, 200.0]$
c_{Pds1}	3.3	2.3993	$[1.65, 6.6]$
c_{Esp1}	3.3	2.3993	$[1.65, 6.6]$
c_{Cdc15}	8.0	8.7958	$[4.0, 16.0]$
c_{Tem1}	8.0	8.7958	$[4.0, 16.0]$
c_{MEN}	8.0	8.7958	$[4.0, 16.0]$
c_{POLO}	100.0	155.2614	$[50.0, 200.0]$
c_{Mcm1}	100.0	183.1687	$[50.0, 200.0]$

accuracy of the data, and the stated goal for how to discretize the continuous distributions, the result is a histogram shape as in Figure 2.

Altogether there are eight distributions being matched (and eight Hellinger distances $d_{H,i}$, $i = 1, \ldots, 8$): joint pair (mass at birth, duration of G$_1$ phase) for mothers (17 boxes), joint pair (mass at birth, duration of G$_1$ phase) for daughters (18 boxes), mass at birth for mothers (12 intervals), mass at birth for daughters (10 intervals), G$_1$ duration for mothers (9 intervals), G$_1$ duration for daughters (11 intervals), cell cycle time for mothers (10 intervals), and cell cycle time for daughters (10 intervals). There are thus a total of 97 discrete probabilities being matched (one for each bin/box/interval) using 44 degrees of freedom (the independent stochastic cell cycle model parameters X), which is a well-posed problem. The objective function is

$$f(X) = \sum_{i=1}^{8} d_{H,i}(p, q),$$

where p, q were described earlier. Trying different weights on the $d_{H,i}$ in the sum had little effect on the final results, and hence results for different weights are not reported here.

Nondefault values for the input arguments to the computer code QNSTOPP are described next. MODE is ‘G’ for global optimization, ‘S’ for stochastic optimization; N is the number of design ellipsoid sample points (from the statistical rule of thumb that at least $1.5n$ data points are needed to estimate n parameters); XI is the initial start point; $[L, U]$ is the feasible box; TAU is the initial design ellipsoid radius τ; GAIN, relevant only for MODE ‘G’, defines the decay factor such that the design ellipsoid radius at iteration k is $\tau_k = \mathrm{GAIN}/(\mathrm{GAIN} + k - 1) \cdot$ TAU.

Using MODE = ‘G’ (global optimization), TAU $= 20$ (5% of the search box diameter), GAIN $= 10$, $N = 64$, Figure 3 shows the iteration histories starting from three points chosen from the box $[L, U]$ in Table 1 (lower box corner, upper box corner, and the best value in Oguz’s model). The Hellinger distance starting from the upper corner point shows a clear descent from ≈ 2.9 to ≈ 1.75 in 13 iterations. Starting from the best point in [6], the Hellinger distance decreases from ≈ 2.5 to ≈ 1.75 in 15 iterations and then oscillates around that value. The same oscillation happens starting from the lower corner point, which suggests that ≈ 1.75 is the best objective function value, and that every point in the box near the corner L has about the same objective function value ≈ 1.8. Dozens of other different start points in the box $[L, U]$ produced similar best function values (QNSTOP can automatically generate a Latin hypercube design of start points including a given start point XI). Note that the best objective function values (≈ 1.75) are not particularly small in the (summed) Hellinger distance measure, meaning that the empirical data is not being matched especially well, although

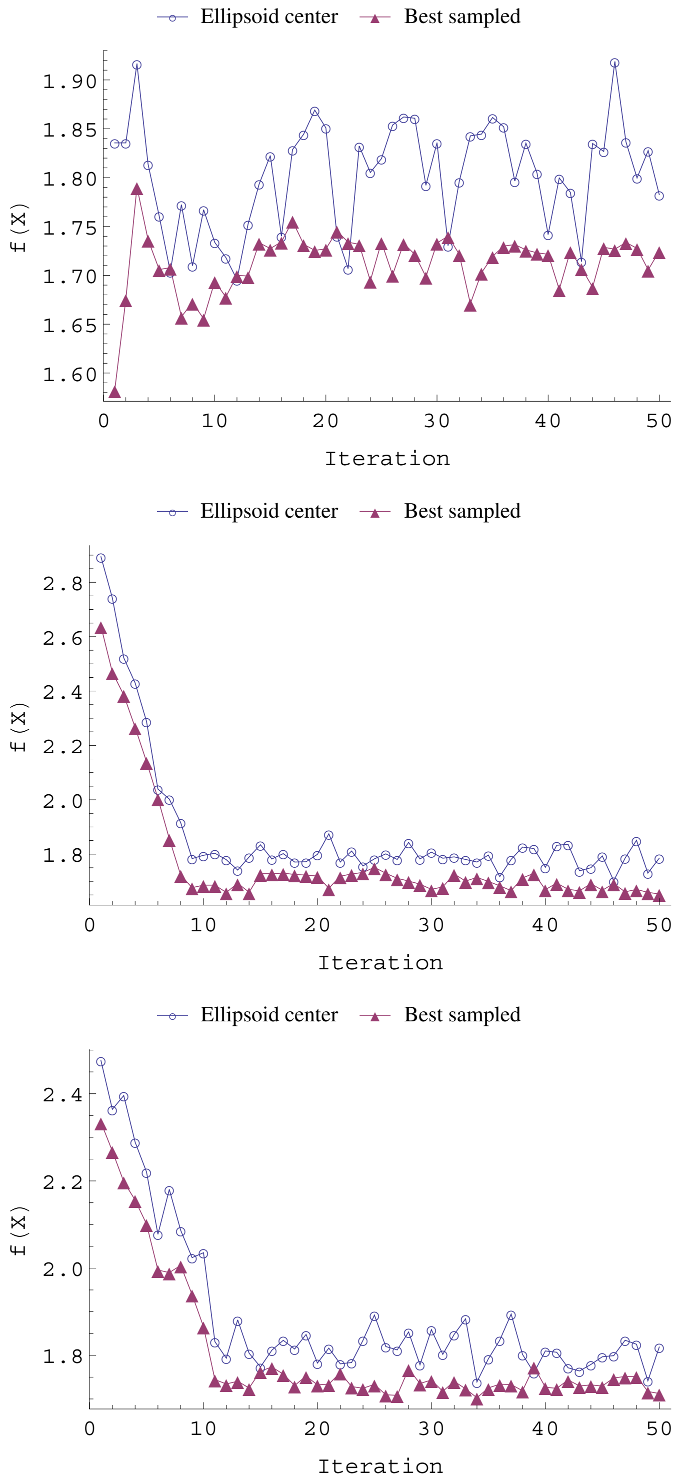

Fig. 3. Execution trace of QNSTOP for three start points from Table 1, lower box corner (top), upper box corner (middle), best value in [6] (bottom). The x-axis shows the iteration number, and the y-axis is the objective function value. For each iteration, the ellipsoid center (circles) and best sampled (triangles) objective function values are reported.

the average Hellinger distance of $\approx 1.75/8 = 0.21875$ is not bad.

To demonstrate that the stochastic parameters in Laomettachit's stochastic cell cycle model are not entirely arbitrary, and that QNSTOP can make progress on stochastic optimization problems, consider an enlarged search box $[(1/2)L, 2U]$ with the start point XI taken as the upper bound corner of this box. This start point is far away from the best point in [6] and has a much larger objective function value. The initial design ellipsoid radius TAU is also changed to 5% of the diameter of the new box, and GAIN $= 10$. A larger value for GAIN causes the ellipsoid radii to decrease more slowly, which is advantageous when starting far away from the optimum point. The execution trace in Figure 4 drops rapidly to near 1.9 in less than 10 iterations, and stays around that value, apparently a local minimum. Figure 5 shows the execution trace of QNSTOP from an even worse starting point (upper bound corner) in the much larger box $[(1/4)L, 4U]$, with the initial TAU adjusted as for Figure 4, and GAIN $= 10$. The plot shows a downward trend and drops sharply around 20 iterations to get near ≈ 2.1, apparently another local minimum.

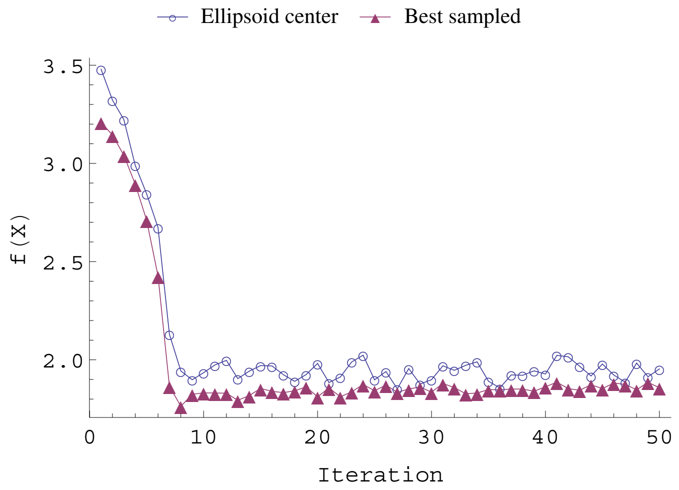

Fig. 4. Execution trace of QNSTOP starting at the upper corner of the larger box $[(1/2)L, 2U]$. The x-axis shows the iteration number, and the y-axis is the objective function value. For each iteration, the ellipsoid center (circles) and best sampled (triangles) objective function values are reported.

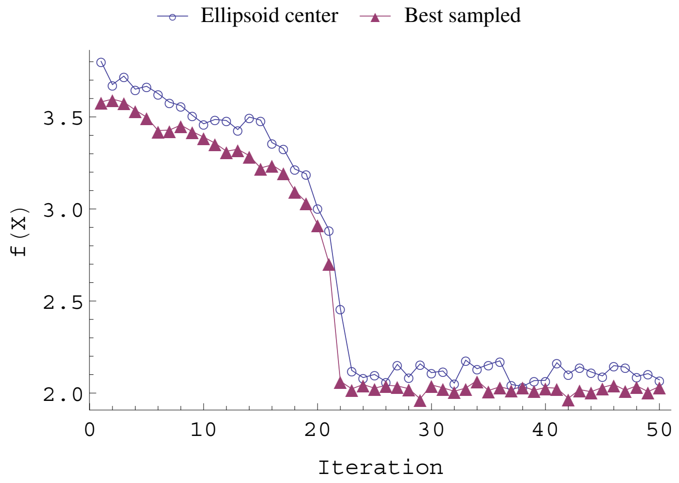

Fig. 5. Execution trace of QNSTOP starting at the upper corner of the larger box $[(1/4)L, 4U]$. The x-axis shows the iteration number, and the y-axis is the objective function value. For each iteration, the ellipsoid center (circles) and best sampled (triangles) objective function values are reported.

TABLE 2
Individual Hellinger distances between empirical distributions and simulated distributions using the best point from Table 1 and the best point found by QNSTOP.

	Table 1	QNSTOP
$d_{H,1}$	0.57	0.44
$d_{H,2}$	0.37	0.22
$d_{H,3}$	0.16	0.09
$d_{H,4}$	0.19	0.12
$d_{H,5}$	0.45	0.31
$d_{H,6}$	0.37	0.22
$d_{H,7}$	0.19	0.10
$d_{H,8}$	0.18	0.15
$f(X)$	2.48	1.65

TABLE 3
Best parameter vector found by QNSTOP.

parameter	value	parameter	value
k_{tr1}	0.6470	k_{tr2}	0.4938
k_{dm1}	0.8598	k_{dm2}	1.4749
m_{min1}	0.2085	m_{min2}	9.0806
k_{tr3}	0.4768	k_{tr4}	0.6377
k_{dm3}	2.1048	k_{dm4}	1.4175
m_{min3}	3.3014	m_{min4}	12.2215
k_{tr5}	0.5411	k_{tr6}	0.4676
k_{dm5}	1.9824	k_{dm6}	1.5821
m_{min5}	8.7150	m_{min6}	10.7990
k_{tr7}	0.5430	k_{tr8}	0.59856
k_{dm7}	1.3543	k_{dm8}	1.6878
m_{min7}	9.7407	m_{min8}	12.3070
k_{tr9}	0.5941	k_{tr10}	0.5638
k_{dm9}	2.0224	k_{dm10}	1.8554
m_{min9}	12.2770	m_{min10}	8.7718
c_{Cln3}	11.0180	c_{Bck2}	13.0380
c_{Whi5}	25.612	c_{Cln2}	59.8760
c_{CKI}	94.6380	c_{Swi5}	67.5470
c_{Cdc20}	123.9300	c_{Cdh1}	121.8000
c_{Cdc14}	20.1910	c_{PPX}	110.2900
c_{Pds1}	3.9074	c_{Cdc15}	11.0270
c_{POLO}	126.2300	c_{Mcm1}	125.5000

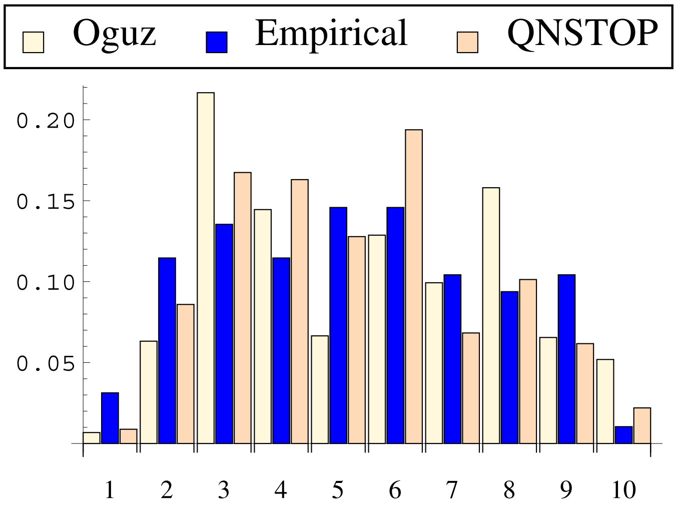

Fig. 6. Comparison of histograms of the cell cycle time for daughter cells, from the simulation using the best point from Table 1 (Oguz), from the empirical data, and from the simulation using the best point found by QNSTOP.

As QNSTOP iterates, the design ellipsoid (in which samples are taken to build a quadratic model of the objective function) radius τ_k decreases. Figure 3, showing the objective function value at the ellipsoid center and at the best sampled point inside that ellipsoid, thus gives a good indication of the variability of the stochastic objective function values within that ellipsoid. Observe that this variability shows little change with respect to the iteration number, meaning that the inherent simulation variance for a fixed parameter vector is roughly comparable to the variance within the (small) design ellipsoid.

Table 2 shows the individual Hellinger distances $d_{H,i}(p,q)$ comprising the objective function $f(X)$, and that the best point (from all runs) found by QNSTOP is considerably better than that found by differential evolution in [6]. For completeness, Table 3 reports that best point X found by QNSTOP. In summary, QNSTOP performs well on this stochastic budding yeast cell cycle model, quickly finding the best known Hellinger distance even from a poor starting point, and significantly improving the result from differential evolution in [6]. From very distant starting points, QNSTOP converges to a (not globally optimal) local minimum point, which is not unexpected behavior.

6 Implications for the Cell Cycle Model

Mathematically, Table 2 shows how well the distributions of the various cell cycle observables (mass at birth, etc.) are being captured by Laomettachit's stochastic cell cycle model. The smallest Hellinger distances are associated with the distributions of birth masses for mother and daughter cells, $d_{H,3}$ and $d_{H,4}$, and with the cycle time distributions for mother and daughter cells, $d_{H,7}$ and $d_{H,8}$. The histograms of daughter cell cycle times (Figure 6) show how good the fit is between the model and the data in this particular case. The major discrepancies are in the tails of the distribution. In contrast, the distribution of G_1 durations for mother cells is not a good match: $d_{H,5} = 0.31$ in Table 2, and the histograms in Figure 7 show clearly that the model overestimates the time spent by mother cells in G_1 phase of the cell cycle. This discrepancy points to a 'structural' problem of the model: the 'G_1-stabilizing' proteins in the model (Cdh1 and CKI) seem to be too active in mother cells, delaying the exit of mother cells from G_1 into S phase. On the other hand, the time spent by daughter cells in G_1 phase is not nearly so discrepant, $d_{H,6} = 0.22$ in Table 2, suggesting that the structural problem is related to some subtle difference between mother cells and daughter cells, which has escaped modelers' attention so far.

The other data that are poorly matched by the model are the joint distributions of (mass at birth, duration of G_1 phase) for mother and daughter cells. The Hellinger distances from QNSTOP are $d_{H,1} = 0.44$ and $d_{H,2} = 0.22$, respectively, which are clear improvements over the best point from Table 1; nonetheless, the Hellinger distances are hard to interpret. Figures 8 and 9 contain histograms of these joint distributions from the empirical data, from the simulation using the best point from Table 1, and from the simulation using the best

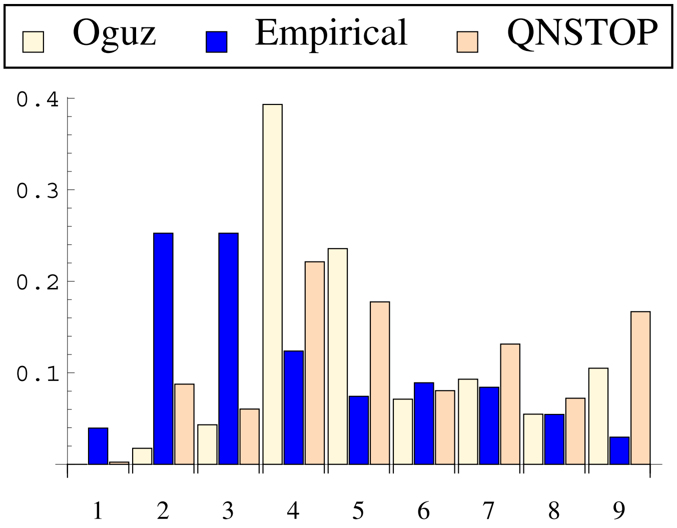

Fig. 7. Comparison of histograms of G_1 duration for mother cells, from the simulation using the best point from Table 1 (Oguz), from the empirical data, and from the simulation using the best point found by QNSTOP.

point found from QNSTOP. Each simulation produces about 1,000 data points, compared to about 100 empirical data points. From these histograms it is evident that the major discrepancies between the model and the empirical joint distributions are in one specific region of the joint distribution: the region where T_{G_1} is large ($T_{G_1} > 0.2\alpha = 26$ min) and m_B is not too much different from the mean mass of mother cells at birth ($-0.3 < \ln(m_B/\overline{m_B}) < 0.1$). In this case, the model is clearly overestimating the number of cells (both mothers and daughters) that spend a long time in G_1 phase, which is complementary to the 'structural' problem noted above. The model underestimates the number of cells with short G_1 durations and overestimates the number of cells with long G_1 durations.

7 Conclusions

As observed in the Introduction, to understand fully the molecular basis of many aspects of cell physiology requires the construction of detailed mathematical models that take into account the intricate interactions among the genes, mRNAs, and proteins involved in regulating each process. Deterministic models, expressed as sets of nonlinear differential equations describing the temporal and spatial interactions of these molecules, are appropriate for understanding the average behavior of large populations of cells. On the other hand, to get at the statistical variability of how individual cells behave requires stochastic models that accurately describe cell-to-cell variability. Stochastic differential equations (SDEs) are often used for this purpose.

In either case—deterministic or stochastic models—the modeler is faced with a daunting task of estimating dozens of parameters (rate constants) by fitting model simulations to experimental observations. The parameter estimation problem is difficult enough for a deterministic model, because of the high dimension of the parameter space of any reasonably complete, molecular-level model of some aspect of cell physiology, and because of the general paucity of accurate and pertinent experimental data. For stochastic models, parameter estimation is more difficult indeed because one must compare statistical distributions (computed and observed) and vary the parameter values to optimize the fit. The computations are more expensive (typically hundreds or thousands of replica simulations to approximate the probability distribution function), and relevant

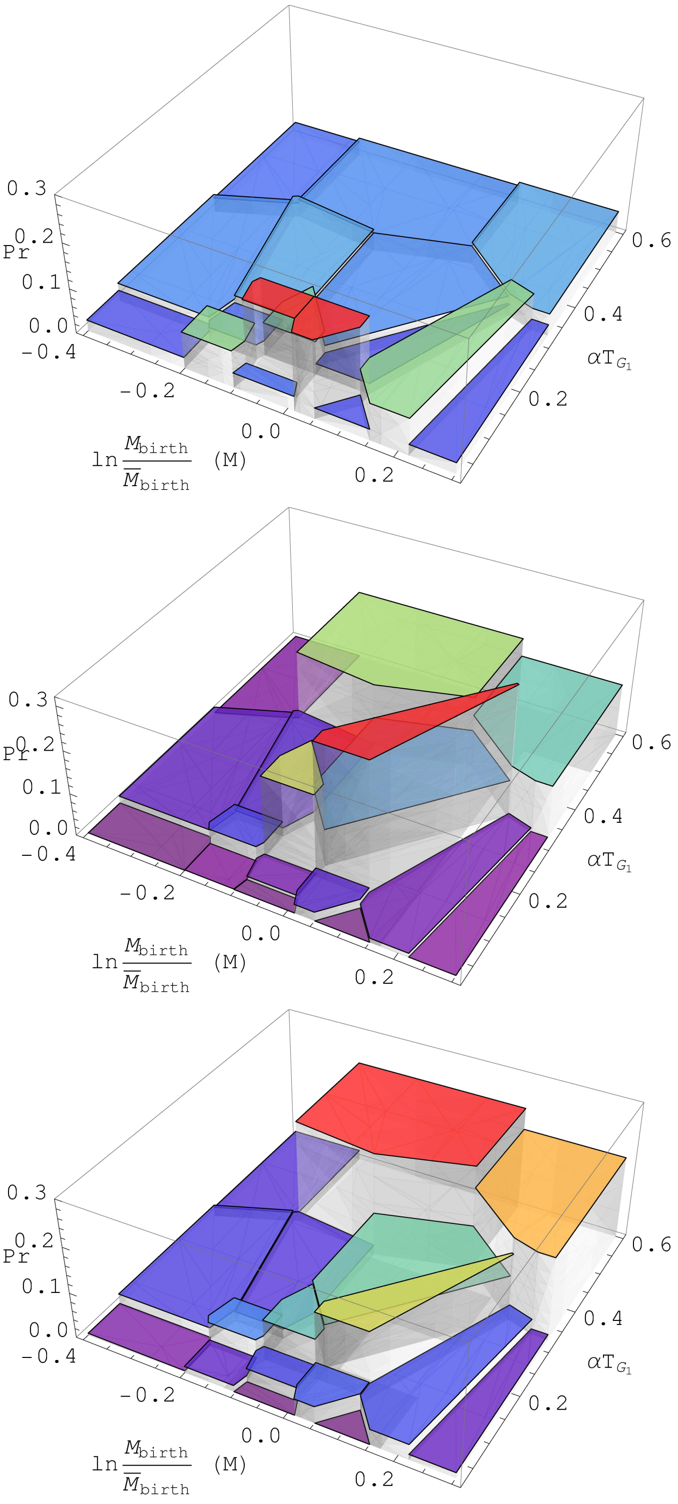

Fig. 8. Two-dimensional histogram of the joint distribution of the pair (mass at birth, duration of G_1 phase) for mother cells from the empirical data (top), from the simulation using the best point from Table 1 (middle), and from the simulation using the best point found by QNSTOP (bottom). The polygons in this display correspond to the rectangles in Figure 1, because the plotting program partitions the horizontal plane into a Voronoi diagram based on the centers of each of the rectangles in Figure 1. The height of each polygon is the relative frequency of data points lying in the corresponding rectangle.

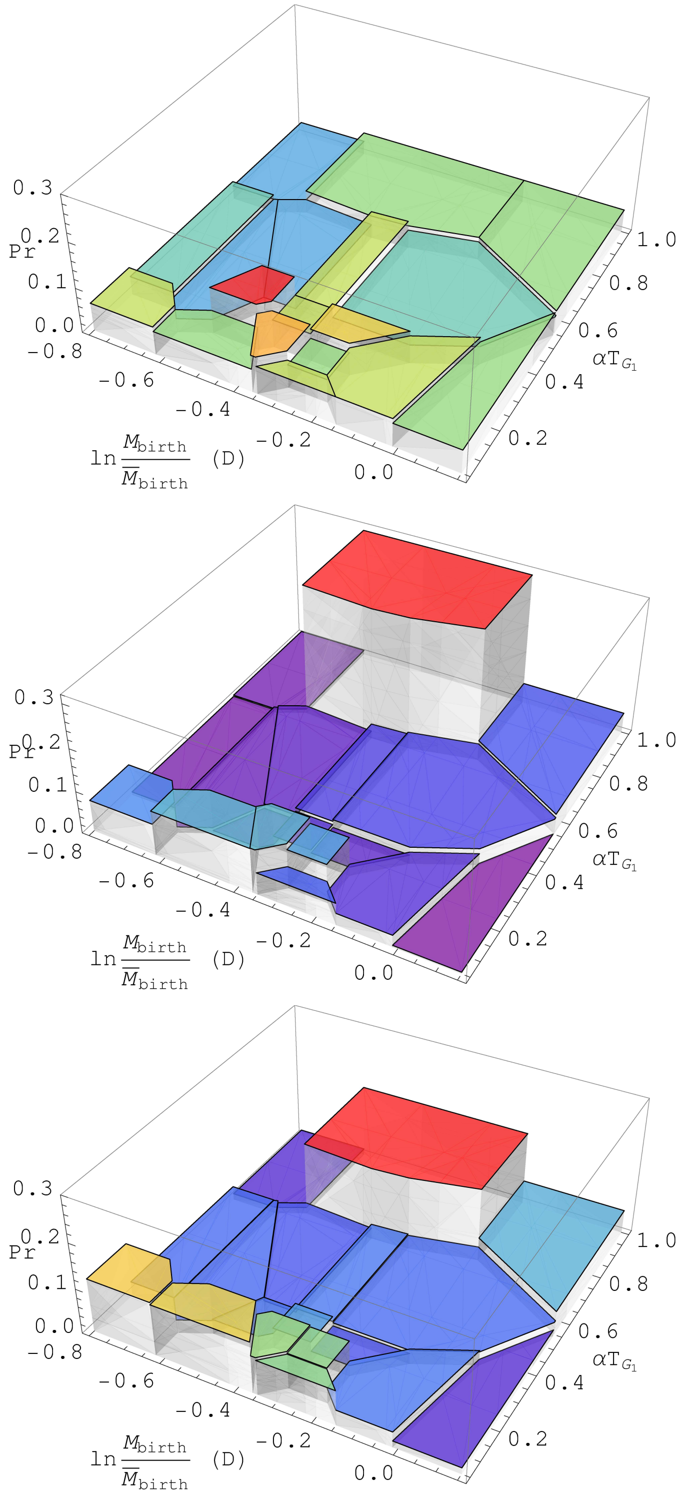

Fig. 9. Two-dimensional histogram of the joint distribution of the pair (mass at birth, duration of G_1 phase) for daughter cells from the empirical data (top), from the simulation using the best point from Table 1 (middle), and from the simulation using the best point found by QNSTOP (bottom). As in Figure 8 the polygons correspond to the rectangles in a partition of the daughter cell data (not shown here), similar to that for the mothers in Figure 1, and the height of each polygon is the relative frequency of data points lying in the corresponding rectangle.

experimental distributions of sufficient quality are rare indeed.

This work tested the efficacy of a quasi-Newton method for stochastic optimization (QNSTOP) to estimate the parameters in a system of SDEs that model the molecular interactions governing progression through the cell division cycle in budding yeast. The model has 44 independent parameters that determine the random fluctuations in molecular populations, and these fluctuations determine the variability from cell to cell of certain observable properties, such as cell cycle time, time spent in G_1 phase of the cell cycle, and cell size at birth. Di Talia et al. [3] have collected data on the distributions of these observables, and on the joint distribution of the pair (mass at birth, G_1 duration). Budding yeast cells divide asymmetrically into a large 'mother' cell and a small 'daughter' cell, so Di Talia measured separate distributions for mother-cell and daughter-cell populations. Hence, Di Talia provides sample data sets from eight different distributions.

QNSTOP can efficiently find a globally (but occasionally only locally) optimal stochastic parameter vector X by minimizing the sum of Hellinger distances $f(X) = \sum_{i=1}^{8} d_{H,i}(p,q)$ between the observed and computed probability mass functions p and q, respectively, for each of the eight different distributions. QNSTOP's fit to these distributions is considerably better than the 'best' fit found in an earlier publication [6], which used a differential evolution algorithm on an objective function that was a sum of squares of deviations between summary statistics (means and standard deviations) for the eight empirical distributions: $f(X) = 1.65$ for QNSTOP, $f(X) = 2.48$ for differential evolution. Presumably, QNSTOP is doing a better job because it is a more efficient algorithm than differential evolution and because it is using all of the information in the full distributions rather than just the summary statistics. A major conclusion of this work is that matching summary statistics and even marginal distributions does not in practice imply that the joint distributions match.

A few conclusions about the model can be drawn from the best parameter vector found by QNSTOP (Table 3). First of all, fluctuations in protein levels in the stochastic model are most sensitively dependent on the parameters $m_{min,i}$. Genes with smaller values of this parameter display larger fluctuations in protein levels. For Oguz's best parameter vector (Table 1), the noisiest gene expression is attributable to *CKI* and *PDS1*. For QNSTOP's best parameter vector, *CLN3* is, by far, the noisiest gene, which seems quite reasonable because Cln3 protein abundance is quite low in budding yeast cells and Cln3-dependent kinase activity is known to play a major role in the G_1-to-S phase transition. Secondly, in QNSTOP's best parameter vector, all mRNAs (except for *CLN3* mRNA) have degradation rate constants in the range 1.3 – 2.1 min^{-1}, which corresponds to half lives in the range 0.33 – 0.51 min. These values seem to be quite smaller than what one might expect (say, 5 min half life), but rapid turn over of mRNAs seems to be necessary to limit the magnitude of protein-level fluctuations in the stochastic model. Notice that *CLN3* mRNA has a noticeably longer half life (1.25 min) than any of the other mRNAs in the model, presumably because it is fluctuations in *CLN3* mRNA numbers that plays the most important role in determining the noisiness of the model's behavior. The fact that the model requires rapid turnover of mRNA species in order to fit the observed probability distributions of cell cycle observables suggests that the way molecular noise is incorporated into the model may be oversimplified. More elaborate models, which incorporate mRNA bursting, mRNA processing, mRNA transport, etc., will have to be explored in later publications.

Acknowledgement

The authors thank Tricity M. Andrew and Cara Simpson for programming assistance early in the project. We are deeply indebted to Dr. Stefano Di Talia of the Duke University Medical Center for supplying us with his original data on cell cycle observables of individual budding yeast cells.

References

[1] B. D. Amos, D. R. Easterling, L. T. Watson, W. I. Thacker, B. S. Castle, and M. W. Trosset, “Algorithm XXX: QNSTOP—quasi-Newton algorithm for stochastic optimization,” *ACM Trans. Math. Software*, to appear, 2017.

[2] B. S. Castle, “Quasi-Newton methods for stochastic optimization and proximity-based methods for disparate information fusion,” Ph.D. thesis, Indiana University, Bloomington, IN, 2012, http://mypage.iu.edu/˜mtrosset/Students/Brent/brent-dissertation.pdf.

[3] S. Di Talia, J. M. Skotheim, J. M. Bean, E. D. Siggia, and F. R. Cross, “The effects of molecular noise and size control on variability in the budding yeast cell cycle,” *Nature*, vol. 448, no. 7156, 947–951, 2007.

[4] T. Laomettachit, “Mathematical modeling approaches for dynamical analysis of protein regulatory networks with application to the budding yeast cell cycle and the circadian rhythm in cyanobacteria,” Ph.D. thesis, Virginia Polytechnic Institute & State University, Blacksburg, VA, 2011, http://hdl.handle.net/10919/29492.

[5] C. Oguz, T. Laomettachit, K.C. Chen, L. T. Watson, W. T. Baumann, and J. J. Tyson, “Optimization and model reduction in the high dimensional parameter space of a budding yeast cell cycle model,” *BMC Syst. Biol.*, vol. 7, 53, 2013.

[6] C. Oguz, A. Palmisano, T. Laomettachit, L. T. Watson, W. T. Baumann, and J. J. Tyson, “A stochastic model correctly predicts changes in budding yeast cell cycle dynamics upon periodic expression of *CLN2*,” *PLOS ONE*, vol. 9, no. 5, Article 96726, 1–17, 2014.