BIOINFORMATICS

Vol. 27 ISMB 2011, pages i295–i303
doi:10.1093/bioinformatics/btr244

A generalized model for multi-marker analysis of cell cycle progression in synchrony experiments

Michael B. Mayhew[1,*], Joshua W. Robinson[2], Boyoun Jung[3], Steven B. Haase[3,4] and Alexander J. Hartemink[2,4,*]

[1]Program in Computational Biology and Bioinformatics, [2]Department of Computer Science, [3]Department of Biology and [4]Center for Systems Biology, Institute for Genome Sciences and Policy, Duke University, Durham, NC 27708, USA

ABSTRACT

Motivation: To advance understanding of eukaryotic cell division, it is important to observe the process precisely. To this end, researchers monitor changes in dividing cells as they traverse the cell cycle, with the presence or absence of morphological or genetic markers indicating a cell's position in a particular interval of the cell cycle. A wide variety of marker data is available, including information-rich cellular imaging data. However, few formal statistical methods have been developed to use these valuable data sources in estimating how a population of cells progresses through the cell cycle. Furthermore, existing methods are designed to handle only a single binary marker of cell cycle progression at a time. Consequently, they cannot facilitate comparison of experiments involving different sets of markers.

Results: Here, we develop a new sampling model to accommodate an arbitrary number of different binary markers that characterize the progression of a population of dividing cells along a branching process. We engineer a strain of *Saccharomyces cerevisiae* with fluorescently labeled markers of cell cycle progression, and apply our new model to two image datasets we collected from the strain, as well as an independent dataset of different markers. We use our model to estimate the duration of post-cytokinetic attachment between a *S.cerevisiae* mother and daughter cell. The Java implementation is fast and extensible, and includes a graphical user interface. Our model provides a powerful and flexible cell cycle analysis tool, suitable to any type or combination of binary markers.

Availability: The software is available from: http://www.cs.duke.edu/~amink/software/cloccs/.

Contact: michael.mayhew@duke.edu; amink@cs.duke.edu

1 INTRODUCTION

Cell division is a process fundamental to the growth, development and reproduction of every living organism. In the case of the budding yeast, *Saccharomyces cerevisiae*, cell division entails a complex and highly regulated series of morphological and genetic changes (Fig. 1). To better understand these changes—and thereby the nature of budding yeast cell cycle progression—researchers track the status of certain cellular features that mark progress through the cell division cycle.

One way to monitor cell cycle progression is with populations of dividing cells. Such approaches entail synchronizing a population of cells in culture at some discrete cell cycle stage, releasing the population into the cell division cycle and collecting independent samples from the culture after release. Under such an approach, marker status for a sample of cells can then be quantified at each time point.

Fluorescent microscope images provide a rich source of marker data and have become increasingly common (Aikawa *et al.*, 2007; Dickinson, 2006; Harder *et al.*, 2006; Stacey and Hitomi, 2008). Microscope images provide the means to track large numbers of cellular features. Fluorescence images, for example, can reveal the localization and status of organelles, protein complexes and single proteins (Huh *et al.*, 2003). In a cell cycle context, microscope images permit the visualization and quantitation of multiple cell-cycle-regulated markers, offering a more detailed view of the relationships and dependencies between subprocesses of cell division. However, few methods currently exist to extract from these image data reliable estimates of cell cycle event timing or of cell cycle progression dynamics in the population.

Furthermore, the types and combinations of markers—image-derived or otherwise—can vary across studies. In the case of budding yeast, many groups look for the presence of a bud (the precursor of the soon-to-be-formed daughter cell). However, additional marker data can run the gamut from flow cytometric measurements of DNA content (Haase and Lew, 1997) to the aforementioned features derived from microscopy images. This lack of consensus in the markers under observation complicates the comparison of cell cycle analyses across studies. Thus, new general methods are required to jointly model marker data, especially when the type or combination of markers changes from one study to another.

We present a powerful, general model of cell cycle progression in a population of cells estimated from any number or combination of binary-valued markers. We build on a branching process framework called CLOCCS (Characterizing Loss of Cell Cycle Synchrony), previously developed in our lab for the purpose of describing populations undergoing cell division (Orlando *et al.*, 2007, 2009). However, in this article, we develop a completely new sampling model that is suited to estimating parameters from any experiment involving some combination of binary markers of progression. In addition, our sampling model is designed to address subtle aspects of observed data that crop up in real experiments. For instance, some small but non-negligible number of cells may be dead or halted during the experiment (and thus not progressing through the cell cycle like the other cells); our sampling model is capable of estimating the number of such cells.

We demonstrate the utility of our new sampling model on two datasets of differential interference contrast (DIC) and fluorescence

*To whom correspondence should be addressed.

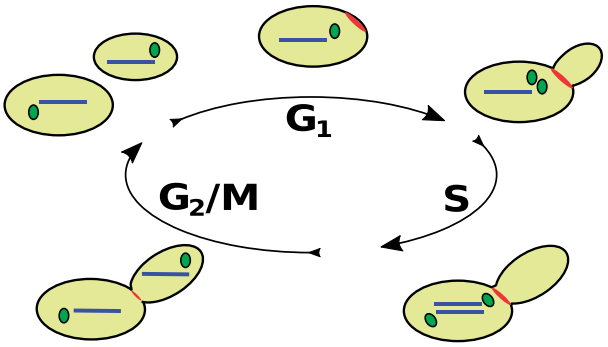

Fig. 1. Early in the cell cycle, a budding yeast mother cell undergoes a period of growth (G_1). Just prior to the time of transition from G_1 to the period of DNA replication (S phase), the myosin ring appears at the site of bud formation (red structure); the bud becomes visible shortly thereafter. Around the same time, the spindle pole body (SPB; green structure) duplicates. Now in S phase, the mother cell replicates its genome (blue bars) within the nucleus (nucleus not shown). Concurrently, the two SPBs start to separate from one another, forming a short mitotic spindle. In G_2/M phase, the two SPBs separate further, forming a long mitotic spindle and pulling the nuclei containing replicated chromosomes into the mother and daughter cells, respectively. Following mitosis, the cell undergoes cytokinesis in which the myosin ring constricts to separate the two cytoplasms and then breaks down. Enzymatic processes must synthesize mother and daughter cell walls before the two cells can separate.

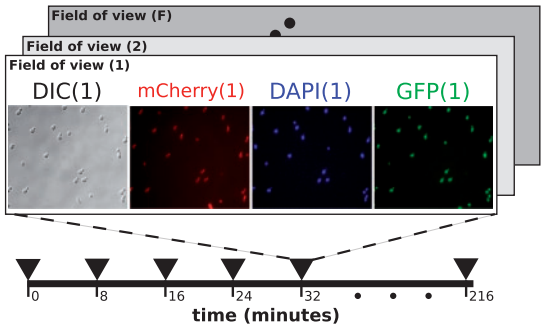

Fig. 2. Samples from a dividing budding yeast culture were collected at fixed intervals (here, 8 min) for a predetermined duration. We acquired DIC and fluorescence microscopy images for one field of view of the sample. We then repositioned the microscope to acquire images for another field of view of the sample. We continued this procedure a variable number of times (F) for each sample, stopping when we acquired images of ∼100 cells per time point.

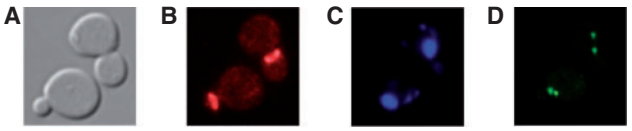

Fig. 3. Shown are DIC (**A**), red fluorescence (**B**), blue fluorescence (**C**) and green fluorescence (**D**) images of two dividing budding yeast cells selected from a single field of view. Intense spots in the red fluorescence image represent the myosin rings while intense blobs in the blue fluorescence image represent nuclei. The small punctate dots in the green fluorescence image are the SPBs. In this study, the marker data used as input to CLOCCS were derived by visual inspection from images like these. In the images shown, for example, there are two budded cells, two cells with myosin rings, two cells with short mitotic spindles and no cells with long mitotic spindles. The brightness and contrast of the images were slightly modified with iPhoto (Apple Computer, Inc.) to increase visibility.

microscopy images we collected in our lab from populations of dividing *Saccharomyces cerevisiae* cells over time. We use these images to track the presence or absence of a set of morphological markers of cell cycle progression (Figs 1–3). We then use the model to precisely infer from population-level data the duration of the attachment period (the length of time following cytokinesis required for the separation of the daughter cell wall from the mother cell wall) in yeast cells synchronized by different protocols. We also show the flexibility of the model by additionally applying it to a recent dataset of completely different binary markers compiled by a different lab (Granovskaia *et al.*, 2010). The extended model and accompanying software represent a powerful and general tool for the analysis of cell cycle progression dynamics.

2 MODEL DEVELOPMENT

As mentioned above, we build on CLOCCS, a branching process framework for describing a population of synchronized cells undergoing cell division. We therefore devote the next subsection to providing relevant background on CLOCCS. The remainder of the subsections that follow describe the new sampling model we developed in this work to estimate cell cycle progression parameters from any number or combination of binary-valued markers.

2.1 CLOCCS model construction

Working with population-level data has its drawbacks, as the loss of synchrony during a time course will confound any measurements: due to synchrony loss, each time point represents a mixture of cells at different positions in the cell cycle. The three main sources of asynchrony over time are (i) imperfect initial synchrony in the population, (ii) cell-to-cell variability in rates of cell cycle progression and (iii) differences in cell division times that are specific to certain cell types (e.g. in the case of budding yeast, daughter cells typically take longer to complete G_1). The CLOCCS framework has parameters for characterizing the contributions of all three sources of synchrony loss.

CLOCCS parameters can be grouped into three sets. One set of parameters specifies the geometry of the branching process underlying cell division. These parameters describe the duration of the cell cycle (λ) and the length of a daughter-cell-specific delay in G_1 phase (δ). A second set of parameters specifies how a dividing cell population moves across this branching process. These parameters describe the initial asynchrony of the population (μ_0, σ_0^2) as well as the 'velocity' of cell cycle progression in the population (μ_v, σ_v^2).[1]

The third set of parameters, in contrast to the first two sets of parameters, depends on the observations. These parameters are specific to the morphological marker(s) under study, and demarcate subintervals of the cell cycle during which a morphological marker is present. In previous versions of CLOCCS, these sampling model parameters were limited to budding (Orlando *et al.*, 2007) and flow cytometric data (Orlando *et al.*, 2009). For example, in the case of budding, the parameter bud^+ is used to specify the fraction of the cell cycle that has passed when the bud first appears (e.g. 0.15); only a single parameter is required since the bud disappearing coincides with the end of the cell cycle and is thus always 1.

[1]We assume that the mean cell cycle velocity μ_v is constant over time, and equal to one cell cycle unit/minute. We also assume that the velocity distributions of the mother and daughter cells are the same.

Because a population of cells loses synchrony over time, a single sample at any time point represents a mixture of cells of different genealogical ages in different phases of the cell cycle. To model the mixture of effects contributed by these subpopulations of cells, we treat cells as members of different cohorts. A cohort is indexed by its generation (g) and its reproductive instance (r). Put simply, g is the number of daughter-cell-specific delays we must take into account up to a given time point, and r is the index of the cell division that produced the current cohort. In this formulation, each cohort contributes different amounts of probability mass to different parts of the cell cycle time line. Marginalizing over the cohorts—essentially taking the weighted sum of these different probability mass contributions—allows us to compute the probability of a randomly sampled cell from the population being at position P_t along the cell cycle time line as:

$$\Pr(P_t \mid \theta, t) = \sum_{\mathcal{C}} \Pr(P_t \mid \theta, g, r, t) \Pr(g, r \mid \theta, t) \quad \text{where} \tag{1}$$

$$\Pr(g, r \mid \theta, t) = M_\theta(g, r, t) / Q_\theta(t) \tag{2}$$

$$M_\theta(g, r, t) = \begin{cases} 1 & g = r = 0 \\ \left(1 - \Phi\left(\frac{-\delta - \mu_{grt}}{\sigma_t}\right)\right) \cdot \binom{r-1}{g-1} & 1 \le g \le r \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

and

$$Q_\theta(t) = \sum_{\mathcal{C}} M_\theta(g, r, t) \tag{4}$$

Here, $\mathcal{C}$ represents all possible cohorts, $\theta = \{\mu_0, \sigma_0^2, \sigma_v^2, \lambda, \delta\}$ is the set of all CLOCCS branching process parameters, $\mu_{grt} = -\mu_0 + t - r \cdot \lambda - g \cdot \delta$, $\sigma_t = \sqrt{\sigma_0^2 + t^2 \cdot \sigma_v^2}$ and Φ is the standard normal cumulative density function. For a more rigorous treatment of the statistical details of the CLOCCS model, we refer the reader to Orlando *et al.* (2009).

2.2 General sampling model for binary marker data

Because of the challenge posed by modeling an arbitrary number of different binary markers of cell cycle progression, we need to devise a new, general sampling model. This sampling model will tie the presence or absence of each marker to a cell's progression through the cell cycle. We introduce two new parameters for each marker. These two new parameters (e.g. m^+ and m^-) represent the start and end time, respectively, of the subinterval in which marker m is observed. Assuming a marker is initially undetectable, it eventually becomes detectable some fraction of the way through the cell cycle, m^+, and then undetectable again at m^-, some time before the end of the cell cycle.

Further assume that we observe M markers, m_j, $j = 1, \ldots, M$, where every marker now has its own start time, ${m_j}^+$, and end time, ${m_j}^-$. Since markers are initially undetectable, $0 < {m_j}^+ < {m_j}^- \le 1$. Let $\mathcal{M}$ be the set of marker-specific parameters $\mathcal{M} = \{m_1^+, m_1^-, m_2^+, m_2^-, \ldots, m_M^+, m_M^-\}$. At least one element of $\mathcal{M}$ must be a non-zero constant, or the problem becomes ill-posed and a unique solution no longer exists. We can thus use budding index data as one of our markers, where bud^- is equal to 1, but other choices are possible.

Assume that samples are drawn at T time points, t_i, $i = 1, \ldots, T$, and that n_i cells are counted at time t_i. Let $m_{ji} = 1$ if the j-th marker is visible in a living cell at time t_i and $m_{ji} = 0$ otherwise. The event that $m_{ji} = 1$ indicates that the position of a cell at time t_i, P_{t_i}, falls on the lifeline within the interval $[(c + {m_j}^+)\lambda, (c + {m_j}^-)\lambda]$ for some cell cycle indexed by integer $c \ge 0$; the probability of this event is defined by the CLOCCS model.

Thus, we compute the probability of the j-th marker being visible at time t_i, $\Pr(m_{ji} = 1 \mid {m_j}^+, {m_j}^-, \theta, t_i)$ by again marginalizing over the cohorts. Taking $\xi = \{{m_j}^+, {m_j}^-, \theta, t_i\}$, we let

$$\Pr(m_{ji} = 1 \mid \xi) = \sum_{\mathcal{C}} \Pr(m_{ji} = 1 \mid \xi, g, r) \Pr(g, r \mid \theta, t_i) \tag{5}$$

where $\Pr(m_{ji} = 1 \mid \xi, g, r)$ represents the probability that a cell from cohort $\{g, r\}$ has a visible j-th marker at time t_i. For the progenitor cohort ($g = r = 0$), letting $\mu_{0t} = -\mu_0 + t_i$ and $\gamma_t(x) = \Phi\left(\frac{x - \mu_{0t}}{\sigma_t}\right)$, we have

$$\Pr(m_{ji} = 1 \mid \xi, g, r) = \sum_{c=0}^{C} \left[\gamma_t(\lambda(c + {m_j}^-)) - \gamma_t(\lambda(c + {m_j}^+))\right] \tag{6}$$

where $\sigma_t = \sqrt{\sigma_0^2 + t_i^2 \cdot \sigma_v^2}$. For subsequent cohorts ($1 \le g \le r$), letting $\mu_{grt} = -\mu_0 + t_i - r \cdot \lambda - g \cdot \delta$ and $\eta_{grt}(x) = \Phi\left(\frac{x - \mu_{grt}}{\sigma_t}\right)$, we have

$$\Pr(m_{ji} = 1 \mid \xi, g, r) = \sum_{c=0}^{C} \left[\frac{\eta_{grt}(\lambda(c + {m_j}^-)) - \eta_{grt}(\lambda(c + {m_j}^+))}{\left(1 - \eta_{grt}(-\delta)\right)}\right]$$

2.3 Accounting for cells with markers present during synchronization recovery

Synchronization methods are not perfect. At early time points, the cell population may include a small but non-negligible fraction of cells that have progressed in the cell cycle beyond the point of sensitivity to the particular synchronization method. Thus, this fraction of cells may have certain markers present before the model-specified beginning of the first cell cycle. We call this proportion of cells with initially present markers 'early cells'. In the model, cells with a visible j-th marker represent a proportion ($E_j{}^+$) of all cells positioned prior to ${m_j}^+$ of the first cell cycle. By definition, early cells belong to the progenitor cohort ($g = r = 0$). Thus, the probability of observing the j-th marker from these cells is calculated as before, but with the addition of a probability mass contribution from the early cells:

$$\Pr(m_{ji} = 1 \mid \xi, E_j{}^+) = \sum_{c=0}^{C} \left[\gamma_t(\lambda(c + {m_j}^-)) - \gamma_t(\lambda(c + {m_j}^+))\right] + E_j{}^+ \gamma_t(\lambda({m_j}^+))$$

Here, as previously, $\xi = \{{m_j}^+, {m_j}^-, \theta, t_i\}$ and $\gamma_t(x) = \Phi\left(\frac{x - \mu_{0t}}{\sigma_t}\right)$ with $\mu_{0t} = -\mu_0 + t_i$ and $\sigma_t = \sqrt{\sigma_0^2 + t_i^2 \cdot \sigma_v^2}$.

2.4 Accounting for dead or halted cells

Some synchronization protocols place significant stress on a population of cells. As a result, a small percentage of cells may die or halt cell cycle progression before release from synchrony (henceforth, we refer to either kind of cell simply as dead). Each marker in a dead cell may be either present or absent, but cannot change over the course of the experiment. The probability of observing a marker includes contributions from living cells with periodically present markers and dead cells with constantly present markers. Let $v_{ji} = 1$ if the j-th marker is present in a cell at time t_i and $v_{ji} = 0$ otherwise. Let

$$\Pr(v_{ji} = 1 \mid \xi, E_j{}^+, D, D_j{}^+) = D_j{}^+ \cdot \frac{D}{Q_\theta(t)} + \Pr(m_{ji} = 1 \mid \xi, E_j{}^+) \cdot \left(1 - \frac{D}{Q_\theta(t)}\right),$$

where D is the fraction of cells that are dead after synchronization and $D_j{}^+$ is the fraction of dead cells with a visible j-th marker. We assume that the number of dead cells remains constant over the course of the experiment and that as the cell population size increases, the relative proportion of dead cells decreases. Thus, we divide the initial proportion of dead cells D by $Q_\theta(t)$.

2.5 Joint likelihood of multiple markers

We model marker presence as a binomial random variable with the success probability p_{ji} defined as $\Pr(v_{ji} = 1 \mid \xi, E_j{}^+, D, D_j{}^+)$ and the number of trials n_i as the number of cells counted at time t_i. As we only have access to population-level statistics rather than statistics from individual cells, we assume that the presence of every marker is conditionally independent of

the others given the CLOCCS branching process parameters θ, marker parameters $\mathcal{M}$ and time points. We also assume that the sample acquisition times for a single marker are independent. Therefore, the final likelihood is

$$L(\mathcal{M},\theta,\mathcal{E},\mathcal{D})=\prod_{i=1}^{T}\prod_{j=1}^{M}\binom{n_i}{n_{ji}}p_{ji}^{n_{ji}}(1-p_{ji})^{n_i-n_{ji}},$$

where $p_{ji}=\mathsf{Pr}(v_{ji}=1\,|\,\xi,E_j{}^+,D,D_j{}^+)$ is the probability of marker j being present at time t_i, and n_{ji} is the number of cells with a visible j-th marker at time t_i. The set $\mathcal{E}$ contains all the $E_j{}^+$s, and the set $\mathcal{D}$ contains D along with all the $D_j{}^+$s.

2.6 Prior specification and Markov chain Monte Carlo parameter inference

All priors are selected to be slightly informative, since synchronization protocols and experimental conditions greatly affect posterior parameter estimates. The priors for the CLOCCS parameters θ were taken from Orlando *et al.* (2009). Thus, $\mu_0\sim\mathsf{Exp}(1/78.2)$, $\lambda\sim\mathsf{N}(78.2,18.2^2)$, $\sigma_0\sim\mathsf{Inv\text{-}Gamma}(2,78.2/3)$ and $\sigma_\nu\sim\mathsf{Inv\text{-}Gamma}(12,1)$.

One change is that we use a gamma prior on δ instead of the exponential prior mentioned in Orlando *et al.* (2009). The daughter-specific G_1 delay can last anywhere from 10 to 60 min depending on the synchronization method (Hartwell and Unger, 1977; Lord and Wheals, 1983). An exponential prior places significant weight at values near zero, which does not match our prior knowledge about the biological value of δ; therefore, we use a more appropriate diffuse gamma prior with a mean of 30 [$\delta\sim\mathsf{Gamma}(10,1/3)$][2].

For each of the marker-specific parameters, we used the slightly informative priors $m_j{}^+\sim\mathsf{Beta}(2,4)$ and $m_j{}^-\sim\mathsf{Beta}(4,2)$ since in each case we know $m_j{}^+$ takes place before $m_j{}^-$. We treat budding specially, where $bud^+\sim\mathsf{Beta}(2.4,17.6)$ and $bud^-=1$. For each marker, we expect only a small number of cells to have that marker present during synchronization recovery. Thus, we place a slightly informative prior on each $E_j{}^+$ to be small [$E_j{}^+\sim\mathsf{Beta}(1,20)$]. Similarly, we assume most cells survive synchronization, so we place an informative prior on the percentage of dead cells to be very small [$D\sim\mathsf{Beta}(1,50)$]. Since we have no knowledge about which markers might be visible in a dead cell, we use uninformative priors for each [$D_j{}^+\sim\mathsf{Beta}(1,1)$].

The parameters for each binary marker may be constrained relative to one another. Due to the nature of how long spindles are defined, m^- for short spindles and m^+ for long spindles are defined to be equal. Additionally, one could imagine constraining other marker-specific parameters. For example, since myosin rings are visible before bud emergence, we could require m^+ for the myosin rings to occur before m^+ for budding.

To infer parameters, we use a simulated annealing algorithm (Kirkpatrick *et al.*, 1983) to initialize a random walk Metropolis algorithm (Gilks *et al.*, 1996; Metropolis *et al.*, 1953). The simulated annealing initialization step was run for 20 000 iterations while the burn-in period of the Metropolis algorithm lasted another 50 000 iterations. The Metropolis algorithm was tuned to mix well and was run for a further 200 000 iterations to obtain posterior parameter estimates. Plots of posterior samples appeared stationary, and the Raftery and Lewis diagnostic (Raftery and Lewis, 1992) indicated that the sample is sufficient to estimate the 0.025th quantile of any marginal posterior to within 0.01 with a probability of 0.95.

2.7 Software development

We developed a multi-platform software tool with a user-friendly GUI for inferring CLOCCS parameters from user-supplied data. The software supports parameter inference using any combination of binary markers of cell cycle progression (as well as flow cytometry data if that is available). The GUI can visualize and record modeling results (for example, all later figures were produced with the GUI). Other features include options for controlling the prior parameter distributions, generating posterior histograms for every inferred parameter, and exporting images in `jpg`, `png` or `pdf` format. The software implementation was designed to be fast and easily extensible, making it quite simple to modify the CLOCCS model to handle different combinations of cell cycle markers. On an Intel Core 2 Duo @ 3.0 GHz, running 300 000 iterations for a model with four markers takes roughly 5 min to complete; in practice, we typically see convergence in approximately 10 000 iterations (10 s).

[2]Here $1/3$ is a rate parameter rather than a scale parameter.

3 EXPERIMENTAL METHODS

3.1 Strain construction

We used a previously constructed haploid yeast strain (SBY408; *bar1*; *SPC42-GFP-TRP1-Zeo*^{R}) derived from strain BF264-15DU (*MATa*; *ade1*; *his2*; *leu2-3,112*; *trp1-1*; *ura3*Δ*ns*) (Richardson *et al.*, 1992) in which a constituent protein of the spindle pole body, Spc42, had been fluorescently labeled with green fluorescent protein (GFP). To fluorescently label Myo1, a myosin ring protein, we amplified the plasmid pFA6a-mCherry-kanMX6 (Shaner *et al.*, 2004; Sheff and Thorn, 2004) using forward primer (5′AAATATTGATAGTAACAATGCACAGAGTAAAATTTTCAGTCGGA TCCCCGGGTTAATTAA3′) and reverse primer (5′CTTTATTTTATGTAC CACCTTTAAGACTACTATCGAAGGAGAATTCGAGCTCGTTTAAAC3'). The PCR products were then used to transform SBY408, with integration of the mCherry fluorescent protein at the C-terminus of Myo1 occurring by homologous recombination, as previously described (Longtine *et al.*, 1998). The resulting haploid yeast strain was labeled SBY1404.

3.2 Time-course microscopy and marker quantification

Yeast cultures were grown in YEPD medium (1% yeast extract, 2% bacto-peptone, 0.012% adenine, 0.006% uracil, 2% dextrose) at 30°C. For α-factor treatment, cultures were grown up overnight, resuspended in fresh media in the morning, and left to proliferate for a couple of hours to reach log-phase. Then, the mating pheromone α-factor was added to the culture for a final concentration of 50 ng/ml. After 2 h, a sample of the culture was checked for the absence of buds with a light microscope. Pheromone was removed by centrifugation and the synchronized population of G_1-arrested cells were resuspended in fresh YEPD medium and grown at 30°C. Every 10 min, 200 µl samples were taken, resulting in a total of 20 time points. For centrifugal elutriation, small daughter cells were isolated from a log-phase culture, released into YEPD + 1 M sorbitol and grown at 30°C as previously described (Orlando *et al.*, 2008). Every 8 min, 300 µl samples were taken, resulting in a total of 28 time points. All samples were kept on ice for 5 min prior to fixation with 2% paraformaldehyde. Samples were then washed with PBS and stored in the dark at 4°C in 30% glycerol. DNA was stained with 1 µg/ml 4,6′-diamino-2-phenylindole dihydrochloride (DAPI; Roche Diagnostics, Indianapolis, IN, USA). Cells from each time point were imaged with a Zeiss Axio Imager widefield fluorescence microscope with either a 100× or 63× objective and standard filter sets (Thornwood, NY, USA). Image acquisition was performed with a Hamamatsu Orca ER monochrome cooled CCD camera with IEEE (Bridgewater, NJ, USA) and MetaMorph (Universal Imaging, Downington, PA, USA).

Cell characteristics were manually quantified on a population level. For each time point, the number of cells with a bud, with a myosin ring, with a short mitotic spindle and with a long mitotic spindle were recorded. Budding was quantified using a light microscope, while we quantified myosin ring and SPB status with a fluorescent microscope. For each feature, we counted all cells across the multiple fields of view of a given time point to get the number of cells in the sample with that feature (Figs 2 and 3). A cell was counted as having a short mitotic spindle if two SPBs were visible inside the mother cell; it was counted as having a long mitotic spindle if a myosin ring was between the two SPBs.

3.3 Marker data regarding division status and location of nucleus

To demonstrate the applicability of the model to any kind of binary marker data, we analyzed a different set of marker data collected by Granovskaia *et al.* (2010). Rather than elutriating cells or treating them with mating pheromone, the authors subjected populations of cells to a genetic arrest-release protocol in which all cells were temperature-sensitive mutants (*cdc28-13*) of the budding yeast cyclin-dependent kinase, Cdc28. The kinase is a key regulator necessary for cell cycle initiation and completion. At 38°C (restrictive temperature), kinase activity is inhibited and the population of cells arrests at the end of G_1 phase. At 25°C (permissive temperature), kinase activity is restored and the population of cells proceeds into the cell cycle. For more experimental details, please refer to Granovskaia *et al.* (2010). In addition to measurements of budding, the authors collected measurements of the fractions of dividing nuclei and of nuclei at the bud neck (the interface between the mother cell and the bud).

4 RESULTS

Using a multimodal imaging platform, we collected DIC and fluorescence microscopy images of four markers of budding yeast cell cycle progression; namely, myosin rings, buds, short mitotic spindles and long mitotic spindles. We used the DIC images to measure budding and the fluorescence images to detect the other markers (see Section 3.2). We processed the images visually, counting cells with each marker across all fields of view at each time point. We compiled two such image datasets on populations of budding yeast cells synchronized either by treatment with the mating pheromone α-factor or by centrifugal elutriation. In the former method of synchronization, cells arrest at the point of cell cycle entry in late G_1 known as START (Hartwell *et al.*, 1974). Centrifugal elutriation on the other hand selects cells based on their size, isolating the smallest cells.

4.1 Model-based analysis of cell cycle progression data derived from microscopy images

The CLOCCS model fits and inferred cell cycle subintervals of each marker for the two image datasets are shown in Figures 4 and 5. In both cases, the posterior fits track the data well. Furthermore, the order of cell cycle events reflected in the order of peaks of marker visibility (Fig. 4) and the positioning of the marker-specific subintervals (Fig. 5) matches current biological knowledge of the budding yeast cell cycle. With minimal prior constraints ($ss^- = ls^+$; Table 1) on the order of each pair of marker-specific parameters, CLOCCS infers the formation of the myosin ring followed by emergence of the bud and appearance of the short and then long mitotic spindle (Bi *et al.*, 1998; Hartwell, 1974). Another CLOCCS-inferred parameter, the average time at which a cell recovers from synchronization and enters the cell cycle (μ_0), recapitulates analytical and experimental observations that elutriated cells take longer to recover on average than cells treated with α-factor [Table 1; (Bellí *et al.*, 2001; Orlando *et al.*, 2007)]. Parameter estimates for both datasets were comparable, demonstrating the robustness of the model to different synchronization methods. In particular, estimates of cell cycle duration (λ) are consistent with empirical evidence [Table 1; 96.9 min from Lord and Wheals (1981)]. The proportion of dead cells estimated for both the α-factor and elutriation datasets was close to zero, while estimates for the proportion of early cells differed between the two datasets. More specifically, the α-factor-treated cells showed a significant proportion of early cells with myosin rings, but proportions with later markers were close to zero. In contrast, estimates from the elutriated cells revealed a significant proportion of early cells with myosin rings present, a slightly smaller proportion of budded early cells, and smaller proportions still with short and long mitotic spindles. These estimates reflect the timing of appearance of these markers as well as our expectation that early cells are not too far advanced in their cell cycle progression. The model was able to identify this relationship between the proportions of early elutriated cells with different markers without the specification of any constraints on inference of the parameters.

4.2 Estimating the attachment period between mother and daughter cells

Monitoring different combinations of cell cycle markers facilitates computation of different cell cycle statistics, giving the marker-specific parameters biological interpretations. To illustrate, we used

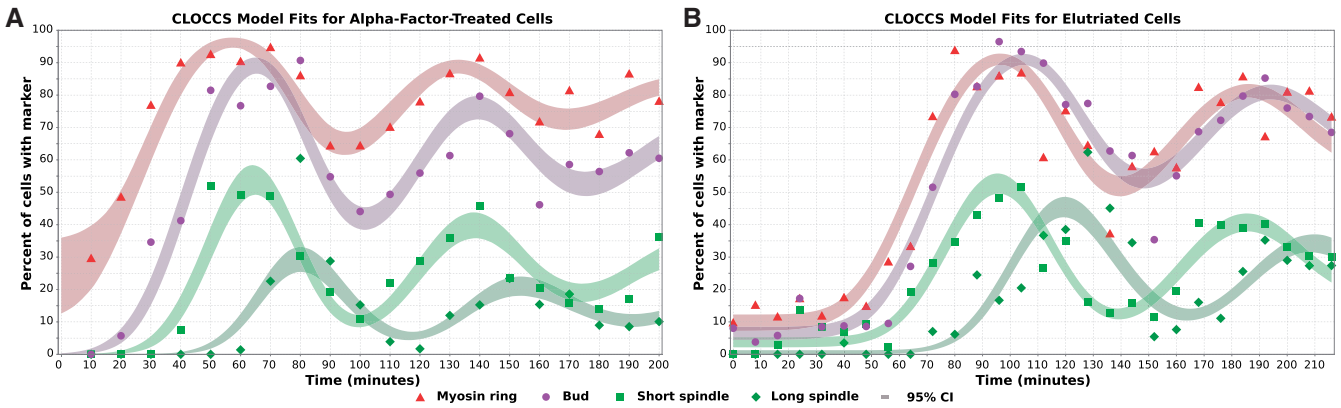

Fig. 4. Model fits for α-factor-treated cells **(A)** and elutriated cells **(B)** using four different binary markers collected at regular intervals over two to three cell cycles. Shaded bands represent 95% credible intervals for posterior inferences. Consistent with previous analytical and experimental observations, the time required for a yeast population to recover from synchronization and enter the cell cycle is longer following elutriation than treatment with α-factor (Bellí *et al.*, 2001; Orlando *et al.*, 2007). This is reflected in the rightward shift of the posterior curves for the elutriated cells relative to the posterior curves for the α-factor-treated cells.

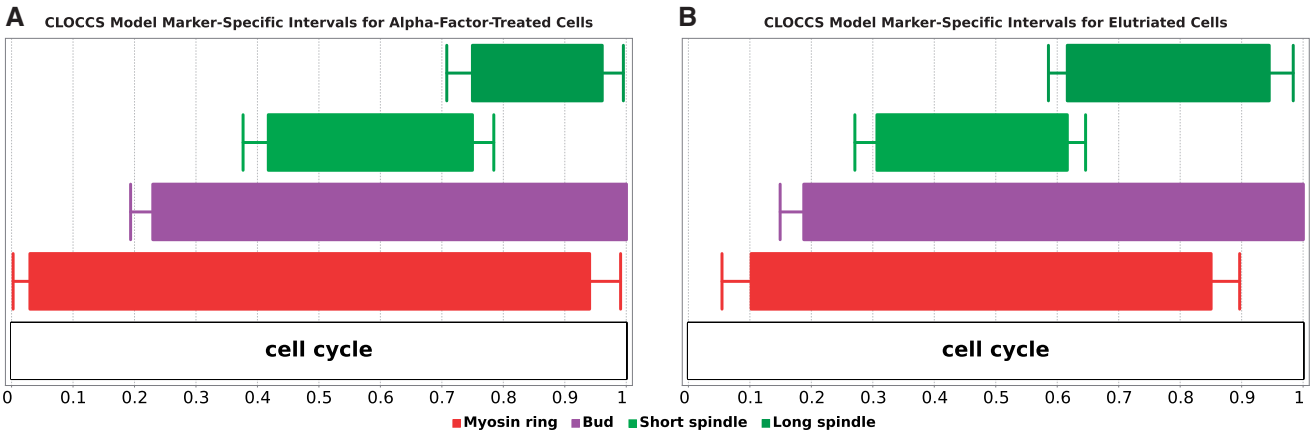

Fig. 5. Inferred marker-specific cell cycle intervals for α-factor-treated cells (**A**) and elutriated cells (**B**) are projected onto a normalized cell cycle time line, during which each of four binary markers are visible. These intervals are derived from the marker-specific parameters, m_j^+ and m_j^-, for myosin rings (red), buds (purple), short spindles (light green) and long spindles (dark green). The whiskers on each bar represent 95% credible intervals for the marker-specific parameter estimates; m_j^- for budding is set to 1 and therefore has no associated credible interval. These figures reflect known budding yeast cell cycle morphology; namely, myosin ring formation precedes bud emergence which is followed by short and then long spindle formation.

Table 1. Marginal posterior summaries for α-factor-treated and elutriated datasets

	α-factor treatment		Elutriation	
	Mean	95% CI	Mean	95% CI
μ_0	−25.643	(−29.517, −21.060)	−57.224	(−61.616, −53.334)
δ	14.681	(7.085, 20.939)	10.871	(4.935, 18.683)
σ_0	13.581	(11.952, 14.911)	16.577	(14.896, 18.220)
σ_v	0.091	(0.059, 0.122)	0.095	(0.074, 0.112)
λ	68.743	(64.225, 74.125)	86.790	(81.496, 91.115)
mr^+	0.039	(0.002, 0.099)	0.104	(0.054, 0.148)
mr^-	0.941	(0.886, 0.990)	0.852	(0.804, 0.901)
bud^+	0.237	(0.196, 0.281)	0.189	(0.150, 0.222)
bud^-	1.000		1.000	
ss^+	0.423	(0.383, 0.463)	0.308	(0.269, 0.343)
ss^-, ls^+	0.752	(0.713, 0.786)	0.617	(0.584, 0.648)
ls^-	0.961	(0.913, 0.995)	0.944	(0.900, 0.985)
E^+_{mr}	0.223	(0.065, 0.341)	0.096	(0.072, 0.122)
E^+_{bud}	0.000	(0.000, 0.001)	0.064	(0.044, 0.086)
E^+_{ss}	0.000	(0.000, 0.000)	0.036	(0.021, 0.053)
E^+_{ls}	0.000	(0.000, 0.000)	0.005	(0.000, 0.013)

mr = myosin ring, ss = short spindle, ls = long spindle. Estimates for D, D^+_{mr}, D^+_{bud}, D^+_{ss} and D^+_{ls} were essentially 0 in both experiments and thus omitted.

the parameters denoting the breakdown of the myosin ring mr^- and the end of budding bud^- to estimate the duration of time between the end of cytokinesis and the separation of the daughter and mother cell walls. By way of background, when we quantify budded cells, we are actually looking at two subpopulations of cells; one in which the mother cell has not yet undergone cytokinesis and another in which the mother and daughter cell have distinct cytoplasms but share a cell wall (an organelle of the budding yeast that surrounds the standard eukaryotic cell membrane). The period between the end of cytokinesis and cell wall separation is sometimes called the attachment period (ap) and, based on our CLOCCS parameters, can be calculated as $ap = \lambda \cdot (bud^- - mr^-)$. We found an attachment period for elutriated cells of ∼13 min and for the α-factor-treated cells of ∼4 min. Both estimates are in line with the previously experimentally derived attachment period estimates of ∼10 min (Lord and Wheals, 1981).

4.3 Application of CLOCCS to cell cycle progression data of Granovskaia *et al.*

With the new general sampling model, CLOCCS is suited to any number or type of binary marker data. To demonstrate this feature of CLOCCS, we analyzed a different dataset of binary marker data recently compiled by Granovskaia *et al.* (2010). Besides recording the presence or absence of buds, Granovskaia and colleagues monitored two binary markers not present in our image datasets: the presence or absence of dividing nuclei and nuclei at the bud neck. In addition to using different binary markers, the authors used a genetics-based protocol to synchronize cell populations (Section 3.3) rather than α-factor treatment or centrifugal elutriation. Our inferences of m_j^+ and m_j^- for the two nucleic markers are constrained such that budding precedes the appearance of nuclei at the bud neck ($bud^+ < nucneck^+$) and nuclei only divide after appearing at the bud neck ($nucneck^+ < nucdiv^+$). The CLOCCS model fits and inferred marker-specific intervals are shown in Figure 6.

Most parameter estimates from the Granovskaia *et al.* (2010) data are comparable with those derived from our own experiments (Tables 1 and 2). The inferred intervals also recapitulate the empirical observation that the nucleus does not spend much time at the bud neck prior to nuclear division (Lord and Wheals, 1981). We note that, as is not the case for our own data, a non-negligible proportion of dead cells are inferred (Table 2). However, due to the uncertainty associated with the D_j^+ parameters in this fraction of cells (especially $D^+_{nucneck}$ with a 95% confidence interval (CI) between 0.281 and 0.998), interpretation of these estimates is

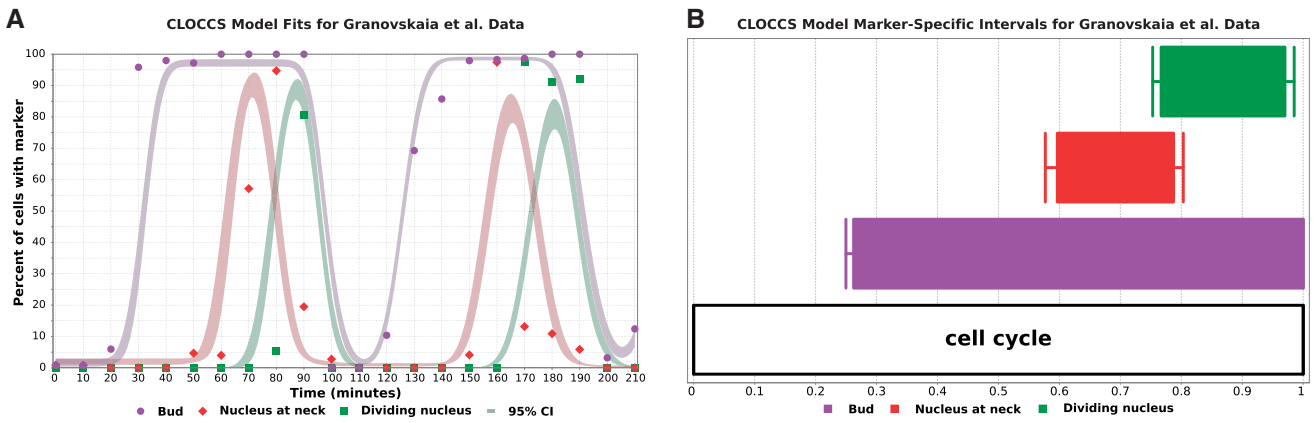

Fig. 6. CLOCCS model fits **(A)** and marker-specific cell cycle intervals **(B)** for the Granovskaia *et al.* (2010) data. Similar to Figures 4 and 5, the graphs depict posterior model fits and marker-specific subintervals for the three binary markers observed in Granovskaia *et al.* (2010).

Table 2. Marginal posterior summaries for Granovskaia *et al*. dataset

	Mean	95% CI
μ_0	−7.760	(−9.319, −6.193)
δ	3.466	(1.872, 5.293)
σ_0	4.749	(4.009, 5.417)
σ_v	0.022	(0.019, 0.027)
λ	92.202	(90.934, 93.435)
bud^+	0.262	(0.250, 0.275)
bud^-	1.000	
$nucneck^+$	0.597	(0.577, 0.617)
$nucneck^-$	0.787	(0.771, 0.802)
$nucdiv^+$	0.767	(0.753, 0.781)
$nucdiv^-$	0.969	(0.954, 0.984)
D	0.030	(0.017, 0.052)
D^+_{bud}	0.108	(0.000, 0.468)
$D^+_{nucneck}$	0.679	(0.281, 0.998)
E^+_{bud}	0.010	(0.000, 0.030)

nucneck = nucleus at bud neck, *nucdiv* = dividing nucleus. Estimates for D^+_{nucdiv}, $E^+_{nucneck}$ and E^+_{nucdiv} were essentially 0 and thus omitted.

difficult. Estimates of the fraction of early cells with either nuclei positioned at the bud neck or dividing nuclei were essentially zero, while estimates of early budded cells were around 1% (Table 2). These results suggest a relatively synchronous initial cell population with the initial synchrony potentially due to the genetics-based method of synchronization used by the authors (Granovskaia *et al.*, 2010). This analysis illustrates how the model can easily accommodate any binary markers of cell cycle progression.

4.4 Effect of early and dead cells on CLOCCS model

Accounting for early cells as well as dead cells required the introduction of two additional components to the model (see Sections 2.3 and 2.4). To assess the relative extent to which these two components improve our modeling performance, we fit the data derived from α-factor-treated cells with four versions of the model. In the *simple* model, neither early nor dead cells are accounted for. In the other three models, one or both of the two types of cells were incorporated (Fig. 7).

Inclusion of dead cells in the model did not significantly improve fits to the data (panels B and D of Figure 7 are indistinguishable from the corresponding panels A and C). In contrast, incorporating the early cells into the model improved fits to the α-factor data, particularly the number of myosin rings counted at the early time points, though with some increased uncertainty owing to the fact that fitting the E^+_{mr} parameter relies on only the earliest time points. The dead-cell-related parameters effectively allow vertical shifts in the posterior fits to the marker data. The $D_j{}^+$ parameters are slightly more constrained since they are weighted by a shared D. Thus, the proportion of dead cells influences the model fits for all markers and at all time points. Conversely, the early cell-related parameters are not so constrained in that a single parameter is inferred for each marker and is governed mostly by the earliest time points. The lack of significant change in the budding, short spindle and long spindle curves upon incorporation of the early cells (Fig. 7C and D) is as expected, since α-factor-treated cells arrest at START and thus may possess a visible myosin ring, but are not expected to have budded or migrated their SPBs.

Although figures are not shown, a similar observation holds for elutriated cells, as can be seen from the parameters in Table 1. The inclusion of dead cells adds little in this experiment, but a small number of elutriated cells have already assembled myosin rings, and smaller numbers have even budded and established a short spindle. Again, this is consistent with the biology of the synchronization procedure as cells are not arrested at START, but are rather simply selected to be small. It should be mentioned that the dead cell parameters did play a role in the data of Granovskaia *et al.*, where the temperature-sensitive mutant arrest may leave a few cells halted even after return to a permissive temperature.

5 DISCUSSION

One obvious extension of the model framework would be to obtain a joint measurement of the marker status of each cell, rather than independently assaying the presence of each marker as we are now. Knowing the joint marker status of each cell would permit the modeling of markers as a multivariate binomial random variable, and thus allow us to assess correlations between the presence of different markers, which might be particularly important when

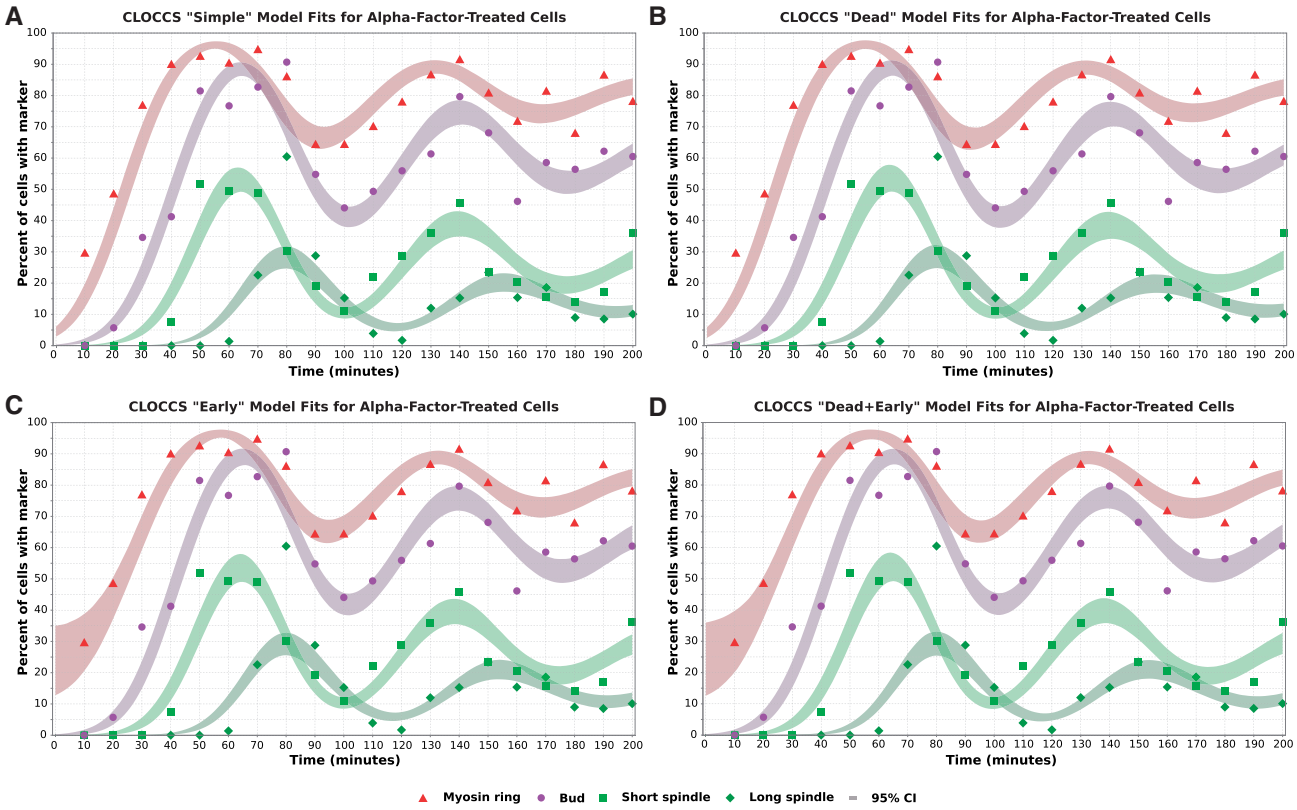

Fig. 7. (**A**) Shown are fits with the α-factor-treated cells using a model not accounting either for cells with visible markers during synchronization recovery or for dead cells. Shown in (**B**) and (**C**) are fits with a model accounting only for dead cells or for early cells, respectively. The model fits in (**D**) accounted for both early and dead cells.

studying various kinds of mutant strains in which the timings of cell cycle events have been disrupted or shuffled.

Doing this would be significantly easier if marker status were to be quantified automatically rather than compiled manually. We are currently exploring publicly available image processing tools like CellProfiler (Carpenter *et al.*, 2006) as well as development of custom software routines to obtain the joint marker status for each cell.

Besides leveraging image processing algorithms to speed up data acquisition, we are also exploring techniques to enrich the precision of our data, moving from binary markers to markers with more than two states (n-ary). For example, rather than simply assessing the presence or absence of a bud, we could obtain n-ary measurements of the approximate volume of the bud. The general sampling model we have presented here can be easily extended to incorporate these types of markers by modeling the markers as multinomial random variables instead of binomial random variables. Estimates from n-ary markers would allow us to analyze high-resolution details of a marker's behavior over time (e.g. bud volume increase in S phase versus G_2/M phase).

We found in our previous work that incorporation of more markers not only leads to more accurate parameter estimates, but also allows a better characterization of the progression dynamics of smaller and smaller subintervals of the cell cycle (Orlando *et al.*, 2009). This result is of great biological importance as the incorporation of more markers can vastly improve the temporal resolution of experimental determination of cell cycle events. This improved temporal resolution carries the potential for a better understanding of the genetic and molecular underpinnings of cell cycle progression.

In essence, our method aims to uncover the dependencies between subcellular processes as well as the temporal ordering of the events associated with these processes. Thus, while the model we have presented was applied in the context of a cell cycle analysis in *S.cerevisiae* cells, the principles of the model are applicable to other biological processes such as development (Dickinson, 2006) or disease progression (Weissleder and Pittet, 2008; Wolf *et al.*, 2007) in a range of different organisms, tissues and cells. In this way, the model represents a general framework for future methods development and an important step toward generating more comprehensive, systems-level views of complex biological processes.

6 CONCLUSION

We have presented a new, general sampling model for observations of any type or combination of binary-valued markers in the context of analyzing the cell cycle progression of dividing cell populations. This model is the first approach capable of estimating from multiple binary-valued markers the cell cycle progression dynamics and cell cycle event timing of a dividing cell population. Furthermore,

in facilitating analysis of cell cycle progression dynamics across different strains and experiments, this model represents a flexible framework for the analysis, comparison and visualization of population cell cycle progression.

ACKNOWLEDGEMENTS

The authors would like to thank members of the Hartemink and Haase labs for useful discussions and comments, especially Xin Guo, Andreas Pfenning and Diana Fusco.

Funding: This work was funded in part by a DARPA grant [HR0011-09-1-0040 to A.J.H.] and an NIH grant [P50-GM081883-01 to S.B.H. and A.J.H. *inter alia*].

Conflict of Interest: none declared.

REFERENCES

Aikawa,E. *et al*. (2007) Multimodality molecular imaging identifies proteolytic and osteogenic activities in early aortic valve disease. *Circulation*, **115**, 377–386.

Bellí,G. *et al*. (2001) Osmotic stress causes a G_1 cell cycle delay and downregulation of Cln3/Cdc28 activity in *Saccharomyces cerevisiae*. *Mol. Microbiol.*, **39**, 1022–1035.

Bi,E. *et al*. (1998) Involvement of an actomyosin contractile ring in *Saccharomyces cerevisiae* cytokinesis. *J. Cell Biol.*, **142**, 1301–1312.

Carpenter,A.E. *et al*. (2006) CellProfiler: Image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.*, **7**, R100.

Dickinson,M.E. (2006) Multimodal imaging of mouse development: Tools for the postgenomic era. *Dev. Dynamics*, **235**(9), 2386–2400.

Gilks,W.R. *et al*. (1996) Introducing Markov chain Monte Carlo. In Gilks,W.R. *et al*. (eds) *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London.

Granovskaia,M.V. *et al*. (2010) High-resolution transcription atlas of the mitotic cell cycle in budding yeast. *Genome Biol.*, **11**, R24.

Haase,S.B. and Lew,D.J. (1997) Flow cytometric analysis of DNA content in budding yeast. *Methods Enzymol.*, **283**, 322–332.

Harder,N. *et al*. (2006) Automated analysis of the mitotic phases of human cells in 3D fluorescence microscopy image sequences. In Larsen,R. *et al*. (eds), *Medical Image Computing and Computer-Assisted Intervention, MICCAI 2006*, Vol. 1 of *Lecture Notes in Computer Science*, Springer, pp. 840–848.

Hartwell,L.H. *et al*. (1974) Genetic control of the cell division cycle in yeast. *Science*, **183**, 46–51.

Hartwell,L.H. (1974) *Saccharomyces cerevisiae* cell cycle. *Bacteriol. Rev.*, **38**, 164–198.

Hartwell,L.H. and Unger,M.W. (1977) Unequal division in *Saccharomyces cerevisiae* and its implications for the control of cell division. *J. of Cell Biol.*, **75**, 422–435.

Huh,W.-K. *et al*. (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691.

Kirkpatrick,S. *et al*. (1983) Optimization by simulated annealing. *Science*, **220**, 671–680.

Longtine,M.S. *et al*. (1998) Additional modules for versatile and economical PCR-based gene deletion and modification in *Saccharomyces cerevisiae*. *Yeast*, **14**, 953–961.

Lord,P.G. and Wheals,A.E. (1981) Variability in individual cell cycles of *Saccharomyces cerevisiae*. *J. Cell Sci.*, **50**, 361–376.

Lord,P.G. and Wheals,A.E. (1983) Rate of cell cycle initiation of yeast cells when cell size is not a rate-determining factor. *J. Cell Sci.*, **59**, 183–201.

Metropolis,N. *et al*. (1953) Equations of state calculated by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.

Orlando,D.A. *et al*. (2007) A probabilistic model for cell cycle distributions in synchrony experiments. *Cell Cycle*, **6**, 478–488.

Orlando,D.A. *et al*. (2008) Global control of cell-cycle transcription by coupled CDK and network oscillators. *Nature*, **453**, 944–947.

Orlando,D.A. *et al*. (2009) A branching process model for flow cytometry and budding index measurements in cell synchrony experiments. *Ann. Appl. Stat.*, **3**, 1521–1541.

Raftery,A.E. and Lewis,S.M. (1992) One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Stat. Sci.*, **7**, 493–497.

Richardson,H. *et al*. (1992) Cyclin-B homologs in *Saccharomyces cerevisiae* function in S phase and in G2. *Genes Dev.*, **6**, 2021–2034.

Shaner,N.C. *et al*. (2004) Improved monomeric red, orange and yellow fluorescent proteins derived from *Discosoma* sp. red fluorescent protein. *Nature Biotechnol.*, **22**, 1567–1572.

Sheff,M.A. and Thorn,K.S. (2004) Optimized cassettes for fluorescent protein tagging in *Saccharomyces cerevisiae*. *Yeast*, **21**, 661–670.

Stacey,D.A. and Hitomi,M. (2008) Cell cycle studies based upon quantitative image analysis. *Cytometry Part A*, **73**, 270–278.

Weissleder,R. and Pittet,M.J. (2008) Imaging in the era of molecular oncology. *Nature*, **452**, 580–589.

Wolf,K. *et al*. (2007) Multi-step pericellular proteolysis controls the transition from individual to collective cancer cell invasion. *Nat. Cell Biol.*, **9**, 893–904.