Parametric modeling of cellular state transitions as measured with flow cytometry

Saumyadipta Pyne  
Department of Medical Oncology,  
Dana-Farber Cancer Institute,  
Harvard Medical School,  
Boston, MA 02115, USA  
Saumyadipta_Pyne@dfci.harvard.edu

Steven B. Haase  
Department of Biology,  
Duke University,  
Durham, North Carolina, USA  
shaase@duke.edu

Hsiu J. Ho  
Department of Statistics,  
Tunghai University,  
Taichung 407, Taiwan  
hsiujho@gmail.com

Tsung I. Lin  
Department of Applied Mathematics  
and Institute of Statistics,  
National Chung Hsing University,  
Taichung 402, Taiwan  
tilin@amath.nchu.edu.tw

Abstract—Gradual or sudden transitions among different states as exhibited by cell populations in a sample under particular conditions or stimuli can be detected and profiled by flow cytometric time-course data. Such temporal profiles however often contain non-Gaussian features due to transient states, and thus present unique modeling challenges. We propose a precise and parametric modeling method based on finite mixtures of skew $t$-Normal distributions that are robust against non-Gaussian features caused by asymmetry and outliers in data. Further, we present a new greedy EM algorithm for fast and optimal model selection. The parsimonious approach of our greedy algorithm allows us to detect the genuine dynamic variation in the key features as and when they appear in time course data. Thus our model parameters can provide precise characteristics of cellular state transition. We applied our method to learn the temporal features of yeast cell cycle progression based on knockout of S-phase triggering cyclins Clb5 and Clb6, and to statistically compare the delay phenotypes due to differential regulation of the two cyclins.

Keywords—finite mixture model; skew $t$-Normal distribution; greedy EM; state transition; flow cytometry; cell cycle

I. INTRODUCTION

Flow Cytometry is among the most widely used platforms in biomedical research and clinical labs. It is used for investigation of a variety of biological problems at the single cell level. Classical applications of flow cytometry include quantitative measurements of DNA content and cell cycle progression. It has also proved to be an useful tool for studies of dynamic cellular properties such as differentiation, proliferation and apoptosis especially in the contexts of stem cells and cancer. Such applications also make flow cytometry the ideal platform for the purpose of identifying the myriad states and functions in different specimens that vary over time under particular conditions and stimuli.

Typically, a flow sample is stained with fluorescent dyes, possibly attached to antibodies, and per cell events such as the expression of a cell-surface marker or the DNA content are measured in terms of fluorescence intensity. The distribution of these events are then plotted or modeled statistically for identification of important features in the sample (e.g. [1]). While developments in computational cytomics have produced many modeling techniques, several important problems have not yet been addressed adequately. One such issue involves precise parametric modeling of dynamic profiles such that the model parameters can characterize the transition of the populations in a sample through different cellular states. The scenario could be complicated further by the adoption of different trajectories by different subpopulations. Indeed a rigorous algorithm for modeling cellular state transitions can not only automate the traditionally manual approach, but also extend it to increasingly complex and high-throughput experiments.

Many major cytometric studies have highlighted the importance of characterizing temporal profiles at single cell resolution, e.g. cell cycle expression kinetics (e.g. [2, 3]), differential phospho-proteomic expression in particular sub-populations and cohorts (e.g. [4]), dynamics of cell differentiation into distinct lineages (e.g. [5]). Clearly, formulation of a cellular state-space and the transitions therein can help us model such temporal cytometric profiles with rigor, and study the changes in parametric detail. Precise probabilistic modeling of sample distributions at each stage can automatically yield features that are both statistically well-defined and biologically insightful [1].

Temporal profiling of state transitions can, however, present unique modeling challenges. Often the transient states produce non-Gaussian features such as skewed sub-populations caused by rush or delay in progression from one state state to another. Intermediate states also produce outliers that cannot be definitively identified with the more distinctive states. Moreover certain metastable states may appear only inconsistently in a given time course. Thus a
probabilistic model that enables precise detection of characteristic features in time course profiles may be the best way to represent the underlying state-space, and also reveal the sudden or gradual nature of the transition. In terms of the distribution of events in a flow sample, features characteristic of different states may be defined in terms of varying size (say, percentage of cells in a peak or cluster), location (such as mean or mode) or significance (peak density). While traditionally such changes were detected with manual or non-parametric techniques, several model-based frameworks have recently been applied with success, e.g. [1, 6, 7].

In the present study, we address the problem of learning all the features in a flow sample with a finite mixture model (STNMIX) of skew $t$-Normal distributions. Importantly, we also present a new greedy Expectation-Maximization (EM) algorithm for fitting the STNMIX model. It starts with a minimum number of components and sequentially inserts a new component to the mixture until convergence is achieved. This parsimonious approach allows us to detect the dynamic appearance and disappearance of transient features that are characteristic of most state transitions. In addition, our approach has several distinct advantages. First, the fitted probability density function rigorously specifies the significance of every feature, which allows us to identify the significant ones in the data. Asymmetric and heavy-tailed skew $t$-Normal components model the data precisely even in the presence of outliers or skewed shapes. Indeed, the robust components, along with the parsimonious fitting strategy, help us determine the optimal number of cell populations with accurately estimated sizes and locations. Moreover, parameters unique to our model such as skewness can help in characterizing phenotypes (such as delay phenotypes in gene knockout experiments) that directly correspond to interesting cellular states and functions. Finally, by design, our STNMIX model is computationally faster to fit than the skew $t$ mixture (STMX) model [1, 6] without sacrificing precision or rigor. Ho et al. [7] summarized the differences between the STMX and STNMIX models and showed the implementation of the STNMIX model is generally much simpler and faster than that of STMX model.

We applied our method to profile and compare the DNA distributions over the course of cell cycle progression in two mutant strains of budding yeast *Saccharomyces cerevisiae*. In late G1-phase, while both Clb5 and Clb6 activate Cdc28p to promote initiation of DNA synthesis, the exact mechanisms and extents of regulating this transition from G1 to S phase are distinct for the two cyclins [2]. In particular, Clb5 knockout causes a more prominent S phase defect during cell cycle progression in yeast cells than Clb6 knockout. Since DNA replication happens in S phase, we studied the dynamics of transition from the start and end states corresponding of one and two copies of the chromosomes (respectively, G1 and G2-M phases) while passing through intermediate states corresponding to S phase delay in the mutants. Notably, while genetic mutations are long known to produce delay phenotypes in cell cycle progression, our skew $t$-Normal components can uniquely and formally detect such lagging or hastening trends in subpopulations, which is clearly not possible with traditional Gaussian kernels [3]. Moreover, our modeling results could be used for hypothesis testing about separability among components, thereby offering insights into the discrete (switch-like) or continuous (spectrum-like) nature of the state transition, leading to observation of multistable and multistep dynamics [5]. Finally, our fast greedy algorithm can provide automated analysis for high-throughput screens as well as enable objective comparison of multiple time-course experiments.

II. Results

We fitted our new skew $t$-Normal mixture model (see Methods) to flow samples from two different courses of 10 time-points each. The two time-courses studied cell cycle progression in yeast cells with knockout of S-phase triggering cyclins Clb5 (Clb5Δ) and Clb6 (Clb6Δ3P), and span more than one cell cycle period (with respect to wild-type yeast cells dividing under the same protocol). The fitted mixture models identified two or three components in every sample, which typically corresponded to the 1C and 2C peaks before and after DNA synthesis, along with subpopulations in the intermediate S-phase, thus characterizing an overall spectrum of profiles of different state transitions (Fig. 1).

We note that our fitted models gave a highly accurate, reliable and smooth representation of the noisy cytometric time-course data (Fig. 2). The orange curve depicts the fitted profile of the DNA histogram at time $t = 25$ min. The individual components in the model are shown as black dotted curves. To determine the precision of our model (STNMIX), we computed log-likelihood maxima $\ell_{\text{max}}$, BIC values, CPU times (CT; in seconds), the distances $D_n$ and $p$-values of Kolmogorov-Smirnov (K-S) test, and compared in Table I with the same for four competing 2-component mixture models (of normal, $t$, skew normal, and skew $t$). According to BIC, clearly the 2-component STNMIX model with equal dfs has the best fit. In fact, among other skew models, the CPU time is substantially reduced for STNMIX suggesting its ease of fit. Finally, given its smallest K-S distance, we conclude that STNMIX achieves the most precise modeling in terms of both the count and asymmetry of components in the given data. Besides precise identification of the features in a particular sample, the models are also used for objective comparison of profiles both within and across time-courses. For instance, the distinct skew in some of the components directly corresponded to delay phenotypes (like lag) for certain cell subpopulations in the sample (note the lag in the left component in Fig. 2). In particular, we computed the Gap statistic [8] as a measure of dispersion of cellular events between the two extreme states corresponding to the
1C and 2C peaks. Tested against a reference distribution of data with no clustering, the GDF statistics support the biological observation of Jackson et al. [2] that the Ctb5 mutant shows more pronounced S-phase delay phenotypes than the Ctb6 mutant and hence has less well-separated components in mid-cell cycle (e.g. \( t = 25 \)). The contrast between the samples in terms of cells showing a slower state transition from 1C to 2C may be observed in Table II for different time-points. Finally, we can monitor the gradual variation in the key features at each successive time-point to gain insights into the differential regulation of the S-phase by the cyclins Ctb5 and Ctb6 (Fig. 3).

III. METHODS AND MATERIALS

For precise, parametric and fast modeling of temporal profiles, we describe the skew t-Normal mixture model. To simplify notation throughout this paper, we let \( \phi(\cdot) \) and \( \Phi(\cdot) \) denote the probability density function (pdf) and the cumulative distribution function (cdf) of the standard normal distribution, respectively.

For \( t(x|\xi,\sigma^2) = \frac{1}{\sqrt{\pi \sigma^2}} \exp\left(\frac{-(x-\xi)^2}{\nu \sigma^2}\right) \),

\[ t(x|\xi,\sigma^2,\nu) = \frac{1}{\sqrt{\pi \nu \sigma^2}} \exp\left(\frac{-(x-\xi)^2}{\nu \sigma^2}\right) \]

denote the pdf of the t distribution with location \( \xi \), scale \( \sigma^2 \) and degrees of freedom (df) \( \nu \), and \( t(x|\nu) \) simply for the case when \( \xi = 0 \) and \( \sigma = 1 \); and let \( \Gamma(\alpha,\beta) \) be the gamma distribution with density \( g(x|\alpha,\beta) \propto x^{\alpha-1} \exp(-\beta x) \). We start by defining the STN distribution and then studying some further properties.

As introduced by Gómez et al. [9], a random variable \( Y \) is said to follow the STN with location parameter \( \xi \in \mathbb{R} \), scale parameter \( \sigma^2 \in (0,\infty) \), skewness parameter \( \lambda \in \mathbb{R} \) and degrees of freedom \( \nu \in (0,\infty) \) if it has the density

\[ \psi(y) = 2t(y|\xi,\sigma^2,\nu)\Phi\left(\frac{y - \xi}{\sigma}\right). \tag{1} \]

We shall write \( Y \sim STN(\xi,\sigma^2,\lambda,\nu) \) if \( Y \) has the density of (1).

Ho et al. [7] give following hierarchical representation of STN to establish an EM-type algorithm [10].

\[ Y|\gamma,\tau \sim N\left(\xi + \frac{\sigma \lambda}{\tau + \lambda^2} \gamma, \frac{\sigma^2}{\tau + \lambda^2}\right), \]
\[ \gamma|\tau \sim TN\left(0, \frac{1}{\tau}; (0,\infty)\right), \]
\[ \tau \sim \Gamma(\nu/2,\nu/2), \tag{2} \]

where \( TN(\mu,\sigma^2;(a,b)) \) represents the truncated normal distribution for \( N(\mu,\sigma^2) \) lying within the truncated interval \( (a,b) \).

Consider \( n \) independent random variables \( Y_1, \ldots, Y_n \), which are taken from a mixture of STN distributions. The pdf of a \( g \)-component STNMIX model is

\[ f(y_j|\Theta_g) = \sum_{i=1}^{g} w_i \psi(y_j|\theta_i), \tag{3} \]

where \( w_i \)'s are mixing proportions which are constrained to be positive and \( \sum_{i=1}^{g} w_i = 1 \).

Based on (2), a practical ECM/ECME algorithm [11, 12] proceeds by Ho et al. [7] as follows:

**E-step:** Given \( \Theta_g = \Theta_g^{(h)} \), compute following \( \hat{z}(h), \hat{\kappa}(h) \), \( \hat{\theta}_i(h) \), and \( \hat{\gamma}_{1ij}(h) \), for \( i = 1, \ldots, g \) and \( j = 1, \ldots, n \).

\[ \hat{z}(h)_{ij} = \frac{\hat{w}_i^0(\psi(y_j|\hat{\theta}_i^{(h)}))}{f(y_j|\Theta^{(h)})}, \quad \hat{\gamma}_{ij}(h) = \frac{\hat{\nu}_i^0(1 + \hat{\nu}_i^{(h)})}{\hat{\nu}_i^{(h)} + \hat{\nu}_i^{(h)}}, \]
\[ \hat{\kappa}(h)_{ij} = \text{DG}\left(\frac{\hat{\nu}_i^{(h)} + 1}{2} - \log\left(\frac{\hat{\nu}_i^{(h)} + \hat{w}_i^{(h)}}{2}\right)\right), \]
\[ \hat{\gamma}_{1ij}(h) = \hat{\lambda}_i^0 \hat{u}_{ij} + \Phi(\hat{\lambda}_i^0 \hat{u}_{ij}) \]

where \( \hat{u}_{ij}^{(h)} = (y_j - \hat{\xi}_i^{(h)})/\hat{\sigma}_i^{(h)} \).

**CM-step:** Update the estimation by

\[ \hat{w}_i^{(h+1)} = \hat{\pi}_i^{(h)} / n, \]
\[ \hat{\xi}_i^{(h+1)} = \frac{\hat{b}_i^{(h)}}{\hat{n}_i^{(h)}}, \]
\[ \hat{\nu}_i^{(h+1)} = \sum_{j=1}^{n} z_{ij}^{(h)} \hat{\gamma}_{ij}(h) \frac{(y_j - \hat{\xi}_i^{(h)})^2}{\hat{\nu}_i^{(h)}}, \]
\[ \hat{\lambda}_i^{(h+1)} = \sum_{j=1}^{n} z_{ij}^{(h)} \hat{u}_{ij}^{(h)}, \]
\[ \hat{\gamma}_{1ij}^{(h+1)} = \arg \max \left\{ \frac{\nu_i^{(h)}}{2} \log \left(\frac{\nu_i^{(h)}}{2}\right), \right. \]
\[ \left. - \log \Gamma\left(\frac{\nu_i^{(h)}}{2} + \frac{\nu_i^{(h)}}{2}\right) \right\}, \]

where \( \hat{\gamma}_{1ij}^{(h)} = \sum_{j=1}^{n} z_{ij}^{(h)} \hat{\gamma}_{ij}(h) \frac{(y_j - \hat{\xi}_i^{(h)})}{\hat{\nu}_i^{(h)} + \hat{\nu}_i^{(h)} + \hat{\nu}_i^{(h)}} \) and \( \hat{\nu}_i^{(h)} \).

If the dfs are assumed to be identical, say \( \nu_1 = \cdots = \nu_g = \nu \), we could update \( \hat{\nu}_i^{(h)} \) by

\[ \hat{\nu}_i^{(h+1)} = \arg \max \left\{ \sum_{j=1}^{n} z_{ij}^{(h)} \psi(y_j|\hat{\xi}_i^{(h)},\hat{\gamma}_{1ij}^{(h)},\hat{\nu}_i^{(h)},\nu) \right\}. \]

The E-step and CM/CML-steps are alternated repeatedly until a suitable convergence rule is satisfied, e.g., the Akaike acceleration based stopping criterion \( |\ell^{(h+1)} - \ell^{(h)}| < \epsilon \), where \( \ell^{(h)} \) is the observed log-likelihood evaluated at
is the asymptotic estimate of the log-likelihood at iteration \( h + 1 \) ([13]; Chap. 4.9) and \( \epsilon \) is the desired tolerance. For numerical analyses in this paper, a default value of \( \epsilon = 10^{-6} \) was used to terminate the iterations.

### IV. Greedy Learning for STN Mixtures

In this section, we present a greedy version of the EM algorithm introduced by Vlassis and Likas [14] for learning the optimum number of components in the fitting of STN-MIX models. The fundamental concept of the greedy EM algorithm is to start from a minimum number of components and sequentially insert a new component to the mixture until convergence is achieved. The stopping criterion can be the specification of obtaining a pre-specified maximum number of components or attaining the pre-specified convergence tolerance.

Suppose that a new component \( \psi(y_j | \theta_{g+1}) \) is added to a \( g \)-component \( f(y_j | \Theta_g) \). The resulting mixture takes the form of

\[
f(y_j | \Theta_{g+1}) = (1 - a) \ f(y_j | \Theta_g) + a \ \psi(y_j | \theta_{g+1}),
\]

where \( 0 < a < 1 \) and \( \Theta_{g+1} = (\Theta_g, a, \theta_{g+1}) \) with \( \theta_{g+1} \) being the added parameters \( (\xi_{g+1}, \sigma^2_{g+1}, \lambda_g, \nu_{g+1}) \).

Given an old mixture \( f(y_j | \Theta_g) \), the weight \( a \) and \( \theta_{g+1} \) are optimally chosen to maximize the new log-likelihood

\[
\mathcal{L}_{g+1} = \sum_{j=1}^{n} \log f(y_j | \Theta_{g+1})
\]

or

\[
\mathcal{L}_{g+1} = \sum_{j=1}^{n} \log [(1 - a) \ f(y_j | \Theta_g) + a \ \psi(y_j | \theta_{g+1})]. \tag{4}
\]

To find the optimal solution in (4), we start by performing a local search with for the newly inserted component. This gives rise to the following partial EM steps where \( \hat{\theta} \) denotes and the partial ML estimates of \( \theta \). For notational simplicity, the subscript \((g + 1)\) is suppressed in the below Partial E-step.

**Partial E-step:** Calculating the conditional expectation of latent variables at the \( k \)th iteration, this yields

\[
\tilde{z}_j^{(k)} = \frac{\hat{a}^{(k)} \psi(y_j | \theta^{(k)})}{(1 - \hat{a}^{(k)}) f(y_j | \Theta_g^{(k)}) + \hat{a}^{(k)} \psi(y_j | \theta^{(k)})},
\]

\[
\tilde{x}_j^{(k)} = \frac{\tilde{z}_j^{(k)}}{\hat{\gamma}_j}, \quad \tilde{\lambda}_j^{(k)} = \frac{\tilde{z}_j^{(k)} \tilde{\kappa}_j^{(k)}}{\Phi(\tilde{\lambda}_j^{(k)} \tilde{\kappa}_j^{(k)})},
\]

\[
\hat{\sigma}_j = \mathrm{DG} \left( \frac{\hat{\rho}(k) + 1}{2} \right) - \log \left( \frac{\hat{\rho}(k) + \tilde{\mu}_j^{(k)}}{2} \right),
\]

where \( \tilde{\mu}_j^{(k)} = (y_j - \tilde{\xi}^{(k)}) / \hat{\sigma}(k) \).

**Partial M-step:** Updating the new parameters in \((a, \theta_{g+1})\), we get

\[
\hat{a}^{(k+1)} = \frac{\sum_{j=1}^{n} \tilde{z}_j^{(k)}}{n},
\]

\[
\hat{\lambda}_j^{(k+1)} = \frac{\tilde{b}_j^{(k)}}{\tilde{\lambda}_j^{(k)}}, \quad \hat{\sigma}_j^{(k+1)} = \frac{\sum_{j=1}^{n} \tilde{z}_j^{(k)} \tilde{\lambda}_j^{(k)} + \tilde{\lambda}_j^{(k)} \sum_{j=1}^{n} \tilde{z}_j^{(k)}}{\sum_{j=1}^{n} \tilde{z}_j^{(k)}},
\]

\[
\hat{\kappa}_j^{(k+1)} = \frac{\sum_{j=1}^{n} \tilde{z}_j^{(k)} \tilde{\lambda}_j^{(k)} \tilde{\mu}_j^{(k)}}{\sum_{j=1}^{n} \tilde{z}_j^{(k)}},
\]

where \( \tilde{\mu}_j^{(k)} = (y_j - \tilde{\xi}^{(k)}) / \hat{\sigma}(k) \), with ideas similar to Vlassis and Likas [14], we provided a global search strategy for extracting proper parameter initialization for \( \Theta_{g+1} \). With ideas similar to Vlassis and Likas [14], we provided a global search strategy for extracting proper parameter initialization for \( \Theta_{g+1} \). We also provide a local maximum of \( \ell_{g+1}(a_0) \) evaluated at \( a = a_0 \). It can be deduced from (5) that a local maximum of \( \ell_{g+1} \) around \( a_0 = 0.5 \) is given by

\[
\hat{\ell}_{g+1} = \sum_{j=1}^{n} \log \left( f(y_j | \Theta_g) + \psi(y_j | \theta_{g+1}) \right)
\]

or

\[
\hat{\ell}_{g+1} = \left[ \sum_{j=1}^{n} \delta_j \right] + \frac{1}{\hat{\sigma}_j^{(k)}} \left( \sum_{j=1}^{n} \delta_j \right) - \frac{1}{\hat{\sigma}_j^{(k)}} \left( \sum_{j=1}^{n} \delta_j \right), \tag{6}
\]

where \( \delta_j = \frac{f(y_j | \Theta_g) - \psi(y_j | \theta_{g+1})}{f(y_j | \Theta_g) + \psi(y_j | \theta_{g+1})} \).

So the optimal value of \( a \) can be calculated as

\[
\hat{a} = \frac{1}{2} \left[ 1 - \frac{\sum_{j=1}^{n} \delta_j}{\sum_{j=1}^{n} \delta_j} \right]. \tag{7}
\]

Following the suggestion of Li and Barron [15], one may set \( \hat{a} = 0.5 \) for \( g = 1 \) and \( \hat{a} = 2/(g + 1) \) for \( g \geq 2 \) as a default recommendation when the estimated value (7) fall outside the range of \((0, 1)\).
In our global search, a convenience choice of $\sigma_{g+1}^{2(0)}$ is $n^{-1/5}$ times half of the sample variance $s^2_g$, whereas $\lambda_{g+1}$ and $\nu_{g+1}^{2(0)}$ are always fixed at 0 and 10, respectively. For the initial choice of $\xi_{g+1}$, we search over the 5th, 10th, 15th, ... 95th quantiles of $y$ and set $\xi_{g+1}^{(0)}$ to the one that maximizes (6).

The implementation of the greedy EM algorithm is summarized below.

1) Start with $g = 1$ and compute the ML estimates of the single-component STNMIX model via the ECME algorithm.
2) If $g > 1$, estimate $\Theta_g$ via the EM-type algorithms.
3) Perform a global search to find a proper initialization of $a$ and $\xi_{g+1}$.
4) Apply the partial EM-steps until convergence. For instance, $|\hat{L}_{g+1}^{(k)} - \hat{L}_{g+1}^{(k-1)} - 1| < 10^{-6}$.
5) If $\hat{L}_{g+1} \leq L_g + m$ then terminate, where $m > 0$ is a penalty term. Otherwise allocate the new component to the model and go to 2. Set $g = g + 1$.

Given $r$ candidates (here are 19 quantiles of sample), the time complexity of our greedy EM algorithm is $O(ngr)$. If overall sample was considered as candidates in the global search, then the run time is the same as Vlassis and Likas [14].

V. CELL CYCLE DATA

For details of the yeast cell cycle experiments and time-course data analyzed by the present study, see Jackson et al. [2].

VI. CONCLUSION

In this study, we described a precise and parametric method for learning the characteristic features of cellular state transition in temporal flow cytometric profiles using skew $t$-Normal mixture model. We also presented a new greedy EM algorithm for fast and optimal model selection. The parsimonious approach of our greedy algorithm allows us to detect the variation in the features that appear and disappear at different points of time thereby offering statistical characterization of the overall nature of state transition. The proposed algorithmic framework is general and may be applied to other similar domains.

REFERENCES

[1] S. Pyne, X. Hu, K. Wang, E. Rossin, T.I. Lin, C. Baecher-Allen, G.J. McLachlan, P. Tamayo, D.A. Hafler, P.L. De Jager and J.P. Mesirov, “Automated high-dimensional flow cytometric data analysis,” Proc. Natl. Acad. Sci. USA, vol. 106, pp. 8519–8524, 2009.

[2] L.P. Jackson, S.I. Reed, and S.B. Haase, “Distinct Mechanisms Control the Stability of the Related S-Phase Cyclins Clb5 and Clb6”, Molecular and Cellular Biology, vol. 26, pp. 2456–2466, 2006.

[3] W. Niu, Z. Li, W. Zhan, V.R. Iyer and E.M. Marcotte, “Mechanisms of cell cycle control revealed by a systematic and quantitative overexpression screen in S. cerevisiae”, PLoS Genetics, vol. 4, e1000120, 2008, doi:10.1371/journal.pgen.1000120.

[4] J.M. Irish, J.H. Myklebust, A.A. Alizadeh, R. Houot, J.P. Sharman, D.K. Czerwinski, G.P. Nolan and R. Levy, “AB-cell signaling networks reveal a negative prognostic human lymphoma cell subset that emerges during tumor progression,” Proc. Natl. Acad. Sci. USA, vol. 107, pp. 12747–12754, 2010.

[5] H.H. Chang, P.Y. Oh, D.E. Ingber and S. Huang, “Multistable and multistep dynamics in neutrophil differentiation,” BMC Cell Biol., vol. 11, p. 11, 2006.

[6] S. Frühwirth-Schnatter and S. Pyne, “Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-$t$ distributions”, Biostatistics, vol. 11, pp. 317–336, 2010.

[7] H.J. Ho, S. Pyne and T.I. Lin, “Maximum likelihood inference for mixtures of skew Student-t-normal distributions through practical EM-type algorithms,” Statistics and Computing, DOI: 10.1007/s11222-010-9225-9.

[8] R. Tibshirani, G. Walther and T. Hastie, “Estimating the number of clusters in a data set via the gap statistic,” Journal of the Royal Statistical Society, Series B, vol. 63, pp. 411–423, 2001.

[9] H.W. Gómez, O. Venegas and H. Bolfarine, “Skew-symmetric distributions generated by the distribution function of the normal distribution,” Environmetrics, vol. 18, pp. 395–407, 2007.

[10] A.P. Dempster, N.M. Laird and D.B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm (with discussion),” Journal of the Royal Statistical Society, Series B, vol. 39, pp. 1–38, 1977.

[11] X.L. Meng and D.B. Rubin, “Maximum likelihood estimation via the ECM algorithm: a general framework,” Biometrika, vol. 80, pp. 267–78, 1993.

[12] C.H. Liu and D.B. Rubin, “The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence,” Biometrika, vol. 81, pp. 633–648, 1994.

[13] G.J. McLachlan and T. Krishnan, The EM algorithm and extensions, 2nd ed., John Wiley & Sons, New York, 2008.

[14] N. Vlassis and A. Likas, “A greedy EM Algorithm for Gaussian mixture learning,” Neural Proc. Lett., vol. 15, pp. 77–87, 2002.

[15] Li, J.Q. and Barron, A.R. Mixture density estimation. In: Advances in Neural Information Processing Systems 12, The MIT Press, 2000.

[16] Z. Darzynkiewicz, H. Crissman and J.W. Jacobberger, “Cytometry of the cell cycle: cycling through history”, Cytometry A, vol. 58, pp. 21–32, 2004.
Table I
DETAILS OF COMPETING MODELS FOR CLB5Δ DATA AT \( t = 25 \) MIN.

| Model | NMIX | TMIX | SNMIX | STMIX | STNMIX |
|-------|------|------|-------|-------|--------|
| \( \hat{\ell}_{\text{max}} \) | -7040.90 | -6916.52 | -6992.27 | -6918.17 | -6916.52 |
| BIC   | 14127.73 | 13906.53 | 13909.82 | 13906.53 | 13906.53 |
| CT    | 0.91  | 21.20  | 35.18  | 120.76 | 21.20  |
| \( D_{n} \) | 0.01466 | 0.00746 | 0.01355 | 0.00766 | 0.00746 |
| \( p \)-value | 0.03050 | 0.64900 | 0.01550 | 0.61300 | 0.64900 |

Table II
MEASURING DISPERSION OF EVENTS AT EACH TIME POINT. GAP STATISTICS FOR CLB5Δ (GAP1) AND CLB6Δ3P (GAP2) AND ASSOCIATED STANDARD ERRORS.

| Time | Gap1 | Gap2 | SE1  | SE2  |
|------|------|------|------|------|
| 0    | 0.689| -0.170| 0.016| 0.016|
| 10   | 0.436| -0.335| 0.016| 0.019|
| 20   | 0.022| -1.245| 0.012| 0.016|
| 25   | 0.203| -0.789| 0.013| 0.018|
| 30   | -0.338| -0.164| 0.016| 0.015|
| 35   | -0.439| -0.223| 0.013| 0.014|
| 40   | -0.371| -0.403| 0.015| 0.015|
| 50   | 0.281| 0.233 | 0.015| 0.015|
| 60   | 0.510| 0.096 | 0.016| 0.014|
| 75   | -1.550| 0.100 | 0.013| 0.014|

Figure 1. Cell cycle time-course profiles. Overall spectrum of temporal profiles based on STNMIX modeling of flow cytometric DNA content data.

Figure 2. Modeling a temporal flow cytometric profile. DNA distribution for Clb5Δ at \( t = 25 \) min, as depicted by a histogram, is modeled with a 2-component skew t-Normal mixture. The orange curve shows the fitted profile while the underlying components are shown in dotted curves.

Figure 3. Comparison of time-course profiles. The orange-red and green-blue curves represent DNA distributions of Clb5Δ and Clb6Δ3P cells respectively. The time-points in minutes and budding information are indicated.