Ho et al. BMC Bioinformatics 2012, 13(Suppl 5):S5
http://www.biomedcentral.com/1471-2105/13/S5/S5

BMC Bioinformatics

RESEARCH Open Access

Parametric modeling of cellular state transitions as measured with flow cytometry

Hsiu J Ho[1], Tsung I Lin[1,2], Hannah H Chang[3,4,5], Steven B Haase[6], Sui Huang[7], Saumyadipta Pyne[8,9*]

From First IEEE International Conference on Computational Advances in Bio and medical Sciences (ICCABS 2011)
Orlando, FL, USA. 3-5 February 2011

Abstract

Background: Gradual or sudden transitions among different states as exhibited by cell populations in a biological sample under particular conditions or stimuli can be detected and profiled by flow cytometric time course data. Often such temporal profiles contain features due to transient states that present unique modeling challenges. These could range from asymmetric non-Gaussian distributions to outliers and tail subpopulations, which need to be modeled with precision and rigor.

Results: To ensure precision and rigor, we propose a parametric modeling framework StateProfiler 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 in StateProfiler 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. We also present a procedure to construct a well-fitted profile by merging any redundant model components in a way that minimizes change in entropy of the resulting model. This allows precise profiling of unusually shaped distributions and less well-separated features that may appear due to cellular heterogeneity even within clonal populations.

Conclusions: By modeling flow cytometric data measured over time course and marker space with StateProfiler, specific parametric characteristics of cellular states can be identified. The parameters are then tested statistically for learning global and local patterns of spatio-temporal change. We applied StateProfiler to identify the temporal features of yeast cell cycle progression based on knockout of S-phase triggering cyclins Clb5 and Clb6, and then compared the S-phase delay phenotypes due to differential regulation of the two cyclins. We also used StateProfiler to construct the temporal profile of clonal divergence underlying lineage selection in mammalian hematopoietic progenitor cells.

Background

Flow Cytometry is among the most widely used platforms in biomedical research and clinical labs. It is used for investigation of a wide variety of biological problems at single cell level. Classical applications of flow cytometry include quantitative measurements of DNA content and cell cycle progression [1]. It is also one of the key platforms for studying dynamic cellular properties such as differentiation, proliferation and apoptosis, especially in the contexts of stem cells and cancer [2]. Such applications make flow cytometry the ideal platform for the purpose of identifying and monitoring 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. While developments

* Correspondence: spyne@broad.mit.edu
[8]Broad Institute of MIT and Harvard University, Cambridge, MA 02142, USA
Full list of author information is available at the end of the article

in computational cytomics have produced many useful analytical methods (e.g. [3]), several important problems have not yet been addressed adequately. One such issue involves precise parametric modeling of dynamic features in temporal profiles such that the model parameters can characterize the transition of the populations in a sample through different cellular states. Often simple statistics such as population mean or size can be imprecise in the presence of unusually shaped distributions and outliers in temporal profiles. The modeling 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, which is subjective and labor-intensive, 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 for a variety of purposes such as cell cycle expression kinetics (e.g. [4,5]), pharmacodynamics (e.g. [6,7]), signaling alterations in specific subpopulations (e.g. [8,9]), dynamics of differentiation into distinct lineages (e.g. [10,11]), and so on. Clearly, mathematical formulation of a cellular state-space, and the transitions therein, can help us model a given collection of temporal flow cytometric profiles with the required rigor. Thereupon we can study the changes in features (say, in comparison with those in control profiles) and monitor trends in parametric detail. Precise probabilistic modeling of sample distributions at each stage can automatically reveal such dynamic features as emergence of a tail subpopulation or change in the skewness of a cluster that are statistically well-defined as well as biologically insightful [3].

Temporal profiling of cellular state transitions in flow data can, however, present unique modeling challenges. Often the transient states produce non-Gaussian features such as asymmetric or trailing subpopulations owing to rush or delay in progression from one state to another [5]. Intermediate states might also produce outliers that cannot be clearly distinguished from the more distinctive states. Moreover certain metastable states may appear only inconsistently in a given time course [11]. Often the transient features appear and disappear at the tails of the more prominent distributions, and may be hard to model via automation. Thus a framework that uses robust probabilistic density functions to model time course data may be the best way to represent the underlying state-space, and reveal any sudden or gradual transition therein. In terms of the distribution of events in a flow sample, characteristics of different states may be determined by variation in size (say, percentage of cells in a peak or cluster), location (such as mean or mode) or significance (peak density) of the model components. While traditionally such changes were detected with manual or non-parametric techniques, several model-based frameworks have recently been applied with success, e.g. [3,12-15].

Here we present StateProfiler, a new framework based on finite mixture models of skew *t*-Normal distributions (STNMIX) for statistical characterization of flow cytometric time course data. In particular, we present in StateProfiler a new greedy Expectation-Maximization (EM) algorithm for fitting our STNMIX model. The greedy EM algorithm starts with a minimum number of distributions (or *components*) and sequentially inserts a new component to the mixture until model convergence is achieved. This parsimonious approach allows us to detect the dynamic appearance (and disappearance) of transient features that are characteristic of many state transitions. In addition, intermediate states are known to produce spatial features in the form of distributions with unusual shapes or low separation, which can lead to overlapping components, and hence to an overestimated number of model components. For optimal model selection, we therefore also provide in StateProfiler a new procedure for merging skew *t*-Normal components that are significantly overlapping in the mixture such that the change in entropy of the resulting model is minimal. Besides profiling of unusually shaped distributions and less well-separated features, this allows StateProfiler to tackle cellular heterogeneity that exists even within clonal populations.

We applied StateProfiler to learn the temporal features of cell cycle progression in two mutant strains of budding yeast *Saccharomyces cerevisiae*. Based on knockout of S-phase triggering cyclins Clb5 and Clb6, we compared the S-phase delay phenotypes resulting from the differential regulation of the two cyclins. Also we used StateProfiler to construct the overall temporal profile of clonal divergence underlying lineage selection in mammalian hematopoietic progenitor EML cells. By comparing the fitted models at each time point, we observed a slow and non-montonic convergence of clonal outlier subpopulations to a final median state.

Results and discussion

Temporal profiling with StateProfiler has several distinct advantages. First, the skew *t*-Normal mixture fitted to the data is defined by a probability density function (pdf). This function is well-defined at any resolution and can be visualized as a smooth profile, which is, unlike kernel-based non-parametric representations, not dependent on bandwidth specification. Importantly, the pdf rigorously specifies the significance of every feature, which allows us to detect the significant ones in the profile, while ignoring the ones which are not.

StateProfiler bases its optimal modeling on 3 strategies: (1) to begin with, asymmetric and heavy-tailed STNMIX components model the data precisely even in the presence of outliers or skewed populations, further (2) the parsimonious fitting of the model with greedy EM yields accurately estimated components, and finally, (3) any redundant components are merged into a well-fitted output profile. By design, our STNMIX model is computationally faster to fit than the skew t mixture (STMIX) model [3,12,16,17] without sacrificing precision or rigor. Ho *et al*. [13] summarized the differences between the STMIX and STNMIX models and showed the implementation of the STNMIX model is generally much simpler and faster than that of STMIX model.

For temporal profiling, certain parameters of STNMIX model such as shape are uniquely suited to detect lagging or hastening trends in subpopulations (such as delay phenotypes in gene knockout experiments) that directly correspond to interesting cellular states and functions. Clearly this is neither possible with non-parametric representations nor using traditional parametric models based on Gaussian, t or other symmetric components [5]. Moreover, such shape or size parameters could be used to test for separability among components - i.e. to identify tendencies of subpopulations to move towards or away from each other without actually changing their mean locations. Parametric "snapshots" of such back-and-forth trends can shed light on the the discrete (switch-like) or continuous (spectrum-like) nature of the state transitions, leading to statistical observation of systems exhibiting multistable dynamics [10].

To illustrate some applications of StateProfiler, we analyzed two previously generated datasets for studying (a) cell division cycle and (b) cell differentiation in different species.

Cell cycle profiling

We applied StateProfiler to identify the temporal features of budding yeast cell cycle progression based on knockout of S-phase triggering cyclins Clb5 and Clb6. 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 [4]. 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. Interestingly, while genetic mutations are long known to produce delay phenotypes in cell cycle progression, few algorithms prior to StateProfiler could model the lag in the DNA distributions with precision.

We fitted STNMIX models to flow samples from two cell cycle time courses with 10 time-points each in yeast cells with knockout of Clb5 (Clb5Δ) and Clb6 (Clb6Δ3P). The time courses spanned 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 more 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 (Figure 1).

The smooth profiles of the noisy DNA histograms at every time-point are constructed with StateProfiler according to optimal change in the entropy values of the fitted model (Figure 2). For example, the entropy plot (Figure 2a, b) suggests a jump in entropy (or elbow) beyond $g = 2$ components for Clb5Δ data at t = 25 min (blue histogram in Figure 2c). The resulting 2-component profile is depicted by the orange curve in Figure 2c. The individual components involved in the model are identified and shown as black dotted curves. Their parameters could be used to detect features for purposes like sorting cells (FACS) or monitoring trends in specific subpopulations (e.g. note the lag in the left component in Figure 2).

To determine the precision of STNMIX, we computed log-likelihood maxima $\hat{\ell}_{\max}$, BIC values, and distances D_n based on Kolmogorov-Smirnov (K-S) test statistic, and compared in Table 1 with the same for four competing 2-component mixture models (of normal, t, skew normal, and skew t) known from the literature [3,18]. According to BIC, the optimal selection of the STNMIX model with equal dfs is evident (e.g. the 2-component model at t = 25). As seen from Dn, we also conclude that STNMIX achieves the most precise modeling in terms of both the count and asymmetry of components in the given data. Further, we used the models for objective comparison of profiles both within and across time-courses. We computed the Gap statistic [19] as a measure of dispersion of cellular events between the two extreme states corresponding to the 1C and 2C peaks or clusters. Tested against a reference distribution of data with no clustering, the Gap statistics support the biological observation of Jackson *et al*. [4] that the Clb5 mutant shows more pronounced S-phase delay phenotypes than the Clb6 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 2 for different time-points. Finally, we observe the gradual variation in the key features at each

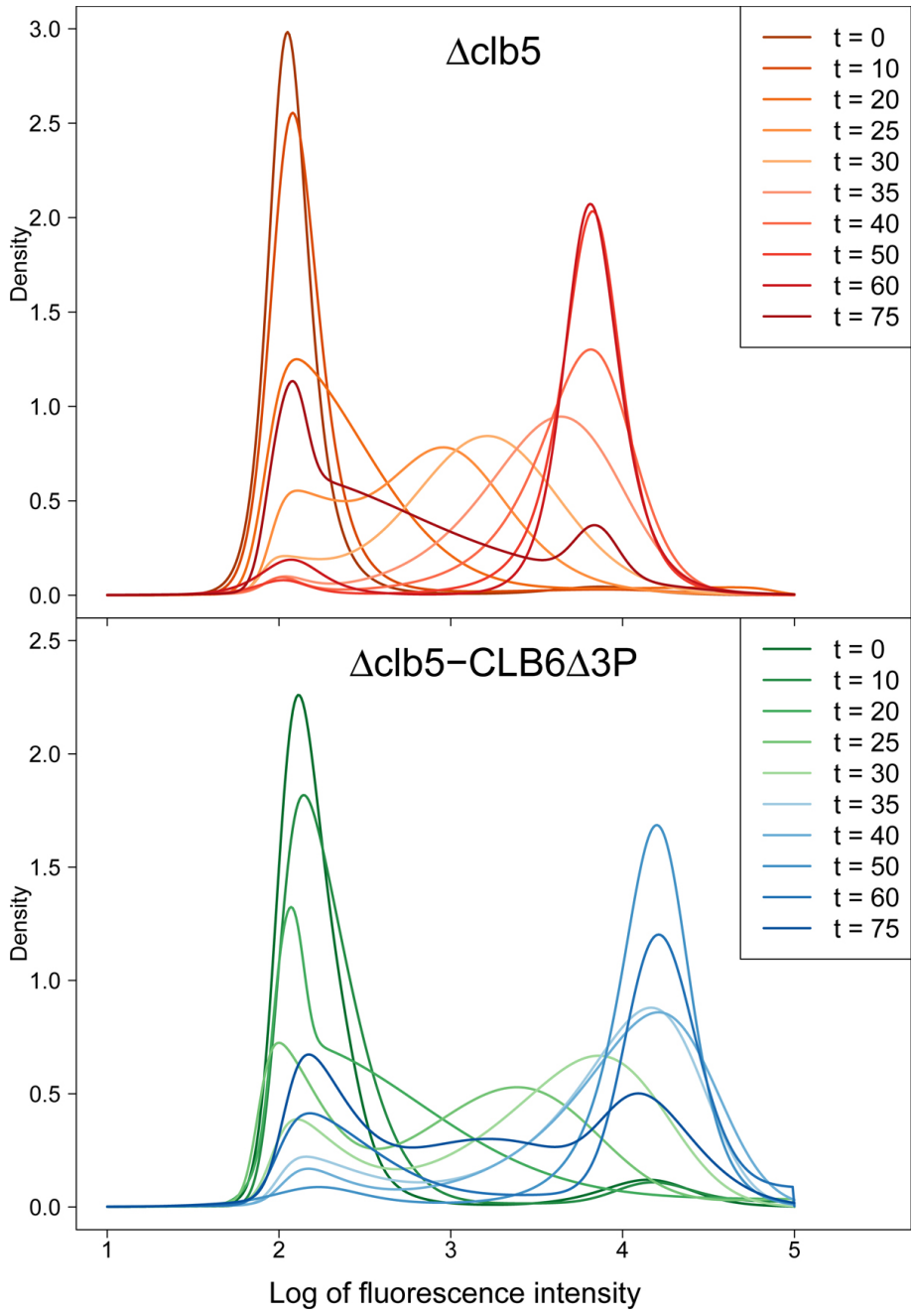

Figure 1 Cell cycle time-course profiles. Cell cycle time-course profiles. Overall spectrum of temporal profiles based on STNMIX modeling of flow cytometric DNA content data.

successive time-point to gain insights into the differential regulation of the S-phase by the cyclins Clb5 and Clb6 (Figure 3).

Cell differentiation profiling

Another key area in which flow data are extremely insightful about different state transitions is cell differentiation. In recent years, many important advances in biology have been made by studying the modes and mechanisms of differentiation especially in the context of stem cells and cancer. Stem cell differentiation has also been studied for their clinical applications such as in the field of regenerative medicine. An excellent review of the field is given in a recent text edited by Krishan *et al.* [2]. Over the course of differentiation, the profiles of expression of various markers - including

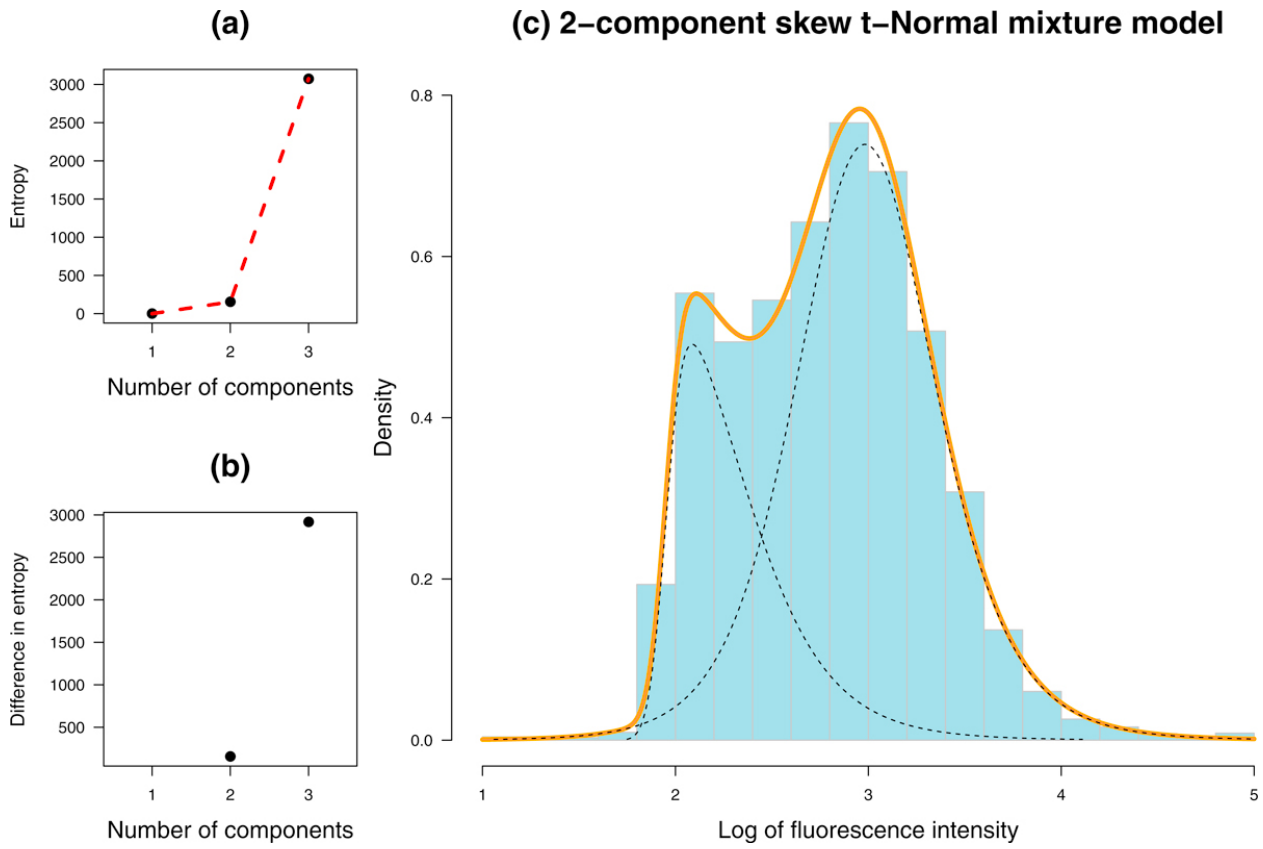

Figure 2 Modeling a temporal flow cytometric profile. Modeling a temporal flow cytometric profile. (a) Entropy values for a profile combining a given number of components (*g*) based on the results of the Greedy EM algorithm for Clb5Δ at t = 25 min. (b) Differences between successive entropy values as *g* increases. (c) 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.

those indicating stemness or commitment to a lineage - vary according to transitions of populations through unstable, metastable and eventually stable states. Often measurable phenotypic diversity appears due to cell-to-cell variability even within clonal populations, which are manifest and can be studied as outlier events or asymmetric or tail subpopulations. Sometimes these features are transient and peripheral, and could be hard to distinguish via automation. Accurate modeling of dynamic flow profiles is thus essential to identify or monitor transitional features as and when they appear (or disappear) for objective temporal characterization of the state-space components involved in differentiation.

In the present study, we analyzed clonal populations of EML cells, a multipotent mouse haematopoietic cell line that can differentiate into myeloid, erythroid, and other lineages. In a recent study, Chang *et al*. [11] measured the expression levels of the stem cell marker Sca-1 in different subpopulations of EML cells as time course data. They observed that cell-to-cell heterogeneity in this clonal progenitor population gave rise to Sca-1 outlier cells - cells that exhibit very high or low Sca-1 expression - and possessed distinct gene expression patterns. The heterogeneity could not be attributed to measurement noise or cell-cycle-dependent cell size variation. Eventually, however, each of these distinct Sca-1 subpopulations' profiles became similar to that of the median cells, thus revealing an attractor state. Yet it was noted [11] that the divergence lasted long enough to allow different propensities for either subpopulation, i.e. low and high Sca-1, to enter into a transient state that primes them for either the erythroid or the myeloid lineage, as captured by their differential expression of lineage-specific transcription factors.

For precise characterization of the dynamics by which population heterogeneity arose in this clonal population via outliers and subsided ultimately, cells with the lowest, middle and highest levels of Sca-1 expression were isolated by [11] using fluorescence-activated cell sorting (FACS). We call these subsets Sca-1^{low}, Sca-1^{mid}, and Sca-1^{high}. Following FCAS, the sorted cells were immediately stripped of the staining antibody and cultured in

Table 1 Details of competing models for Clb5 data

t	Criterion	NMIX	TMIX	SNMIX	STMIX	STNMIX
0	$\hat{\ell}_{\max}$	2539.78	2647.35	2682.44	2771.12	2759.40
	BIC	-5033.60	-5239.55	-5300.53	-5468.69*	-5445.25
	D_n	0.0413	0.0262	0.0292	0.0164†	0.0185
10	$\hat{\ell}_{\max}$	1201.11	1224.87	1357.82	1405.31	1406.80
	BIC	-2356.27	-2394.60	-2651.32	-2737.09	-2740.08*
	D_n	0.0424	0.0284	0.0312	0.0214	0.0190†
20	$\hat{\ell}_{\max}$	-5463.40	-5462.64	-4869.75	-4792.67	-4791.62
	BIC	10972.72	10980.37	9803.79	9658.80	9656.69*
	D_n	0.0758	0.0715	0.0251†	0.0264	0.0266
25	$\hat{\ell}_{\max}$	-7040.90	-6981.07	-6992.27	-6918.17	-6916.52
	BIC	14127.73	14017.26	14048.84	13909.82	13906.53*
	D_n	0.0147	0.0145	0.0155	0.0077	0.0075†
30	$\hat{\ell}_{\max}$	-7251.45	-7226.16	-7228.28	-7203.05	-7201.55
	BIC	14548.76	14507.34	14520.76	14479.46	14476.46*
	D_n	0.0218	0.0175	0.0143	0.0129	0.0110†
35	$\hat{\ell}_{\max}$	-6413.58	-6412.58	-6374.92	-6320.38	-6334.20
	BIC	12872.96	12880.12	12813.96	12714.04*	12741.69
	D_n	0.0196	0.0230	0.0136	0.0117†	0.0152
40	$\hat{\ell}_{\max}$	-4626.43	-4625.80	-4546.10	-4429.22	-4461.12
	BIC	9298.55	9306.44	9156.18	8931.56*	8995.37
	D_n	0.0338	0.0306	0.0170	0.0123†	0.0184
50	$\hat{\ell}_{\max}$	-1286.86	-1121.26	-1286.53	-1093.80	-1086.34
	BIC	2619.40	2297.35	2637.03	2260.70	2245.79*
	D_n	0.0222	0.0145	0.0218	0.0139	0.0132†
60	$\hat{\ell}_{\max}$	-2016.29	-1596.75	-1835.23	-1573.97	-1568.82
	BIC	4078.18	3248.21	3734.30	3220.89	3210.59*
	D_n	0.0540	0.0203	0.0339	0.0172	0.0131†
75	$\hat{\ell}_{\max}$	-8146.57	-7810.70	-7772.60	-7770.74	-7769.87
	BIC	16393.86	15731.29	15682.55*	15688.00	15686.25
	D_n	0.0219	0.0119	0.0079†	0.0101	0.1104

Details of competing models for Clb5Δ data. Here, the notations stand for log-likelihood maxima $\hat{\ell}_{\max}$max, BIC values, and distances D_n based on Kolmogorov-Smirnov (K-S) test statistic. The abbreviation of models are the normal mixtures (NMIX), the *t* mixtures (TMIX), the skew-normal mixtures (SNMIX), the skew-*t* mixtures (STMIX) and the skew-*t*-normal mixtures (STNMIX), respectively. According to BIC and D_n, the optimal selection of the STNMIX model with equal dfs is evident for most points. The smallest values of BIC and D_n are indicated by * and †, respectively.

Table 2 Measuring dispersion of events at each time point

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

Measuring dispersion of events at each time point. Gap statistics for Clb5Δ (Gap1) and Clb6Δ3P (Gap2) and associated standard errors.

standard growth medium. Subsequently, Sca-1 fluorescence intensity were measured individually for each of the 3 subpopulations as time course data. Similar measurements were made for an original clonal population of EML cells for comparison (we call it Sca-1^{all}).

We applied the StateProfiler framework to model the flow profiles for 14-point time course data for each of the 4 populations. Often finite mixtures of Gaussians are used for modeling the theoretical subpopulation structure in such profiles [11,20]. However, using Gaussian components, precise modeling in the presence of outliers due to cell-to-cell heterogeneity is particularly difficult for clonal populations. This is because an optimal model must be able to accommodate such heterogeneity without requiring extra components, but Gaussian components with sharp tails are hardly robust against outliers. It leads to sub-optimal models with spurious subpopulations, which makes their biological interpretation difficult.

StateProfiler addressed the modeling problem in two ways. First, its skew *t*-Normal components are robust to outliers and asymmetry in the distributions. This helps in modeling transitional features even if they lead to unusually shaped or heavy tailed distributions. Second, even if redundant subpopulations were identified, the new merging procedure in StateProfiler can re-construct any significantly overlapping components in a statistically optimal fashion, i.e. to produce a combined profile by causing minimal change in entropy of the model pre- and post-reconstruction.

The dual advantages of the StateProfiler modeling algorithm allowed us to compute highly accurate profiles of Sca-1 expression in the time course datasets for the three sorted and the unsorted EML cells. The steps of the merging procedure through which an optimal structure for the model is "stitched" together are illustrated with an example in Figure 4. Finally, we compared the divergence of the 3 sorted subsets from the corresponding unsorted population using Kullback-Leibler distances between the probability density functions specifying their profiles. A visual comparison of the profiles is shown in Figure 5. The trend of decreasing divergence, as the 3 sorted profiles become similar to the unsorted profile with progression of time, is shown in Figure 6.

StateProfiler's parametric characterization can reveal various features and trends of interest in terms of specific parameters. For instance, we observe that by 3 days, both Sca-1^{mid} and Sca-1^{high} have already started to resemble the unsorted population, and by 6 days, they actually have their own low Sca-1 tails. Another trend of possible interest is the slow but continuous fluctuation in the proportion of low Sca-1 outliers in the unsorted population.

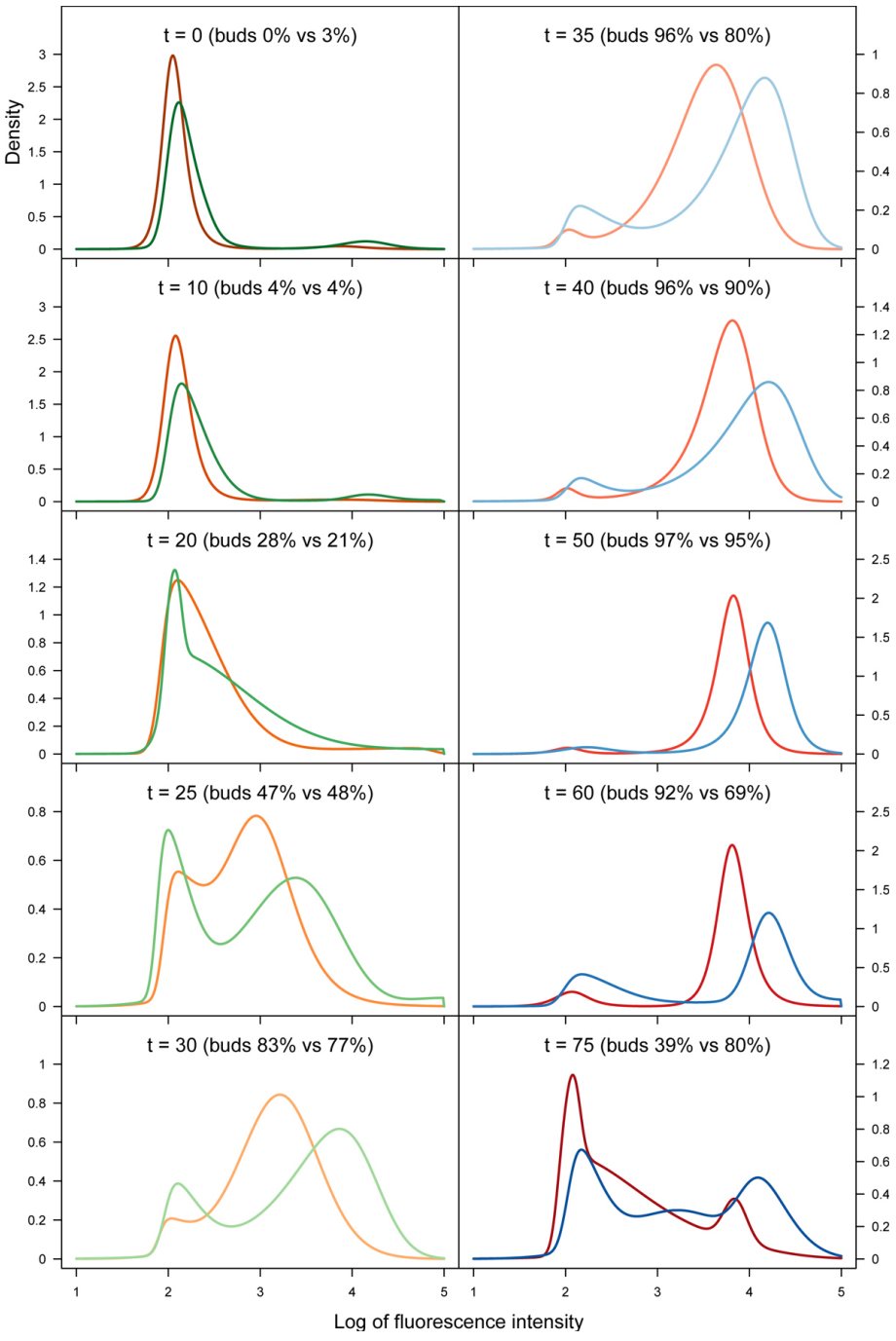

Figure 3 Comparison of time-course profiles. Comparison of time-course profiles constructed with StateProfiler. The orange-red and green-blue curves represent DNA distributions of Clb5Δ and Clb6Δ3P cells respectively. The time-points in minutes and budding information are indicated.

Finally, it appears that the eventual stable state when the 3 profiles finally coincide is reached at a point of time much later than 9 days, as suggested by [11], and takes probably double that time (432 h). In the mean time, as we see in Figure 6, the states might continue to drift closer and apart as in a dynamical system exhibiting multistable behaviour. If indeed the departure from the average state has biological functionality in the priming of cell fate commitment, then a non-monotonic, delayed restoration of the underlying molecular mechanisms may be justified by having more than a few cells with random fluctuation and call for further investigation.

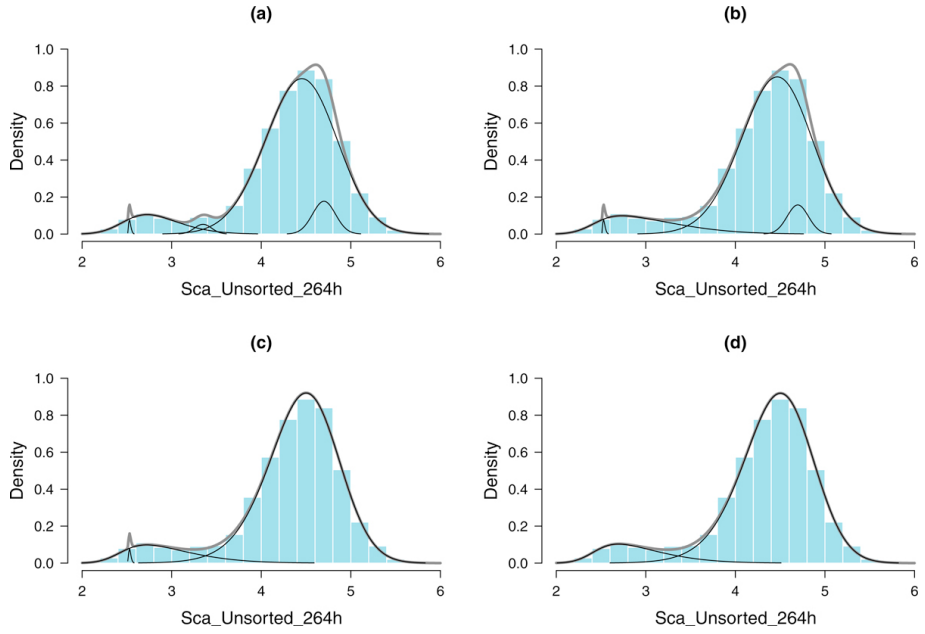

Figure 4 An example of merging mixture components. The Sca-1 expression data for the unsorted population of EML cells at 264 h is shown in the histogram. At each step of the merging algorithm, the fitted profile is shown as a thick grey curve, and the individual components in think black curves. (a) Initial profile computed by Greedy EM with $g = 5$, Entropy = 2351. (b) Merged profile with $g = 4$, Entropy = 573. Combining a group of components in the left significantly reduces entropy. (c) Merged profile with $g = 3$, Entropy = 297. (d) The final merged profile with $g = 2$ components and Entropy = 48.

Conclusions

In this study, we described StateProfiler, a framework to construct temporal profiles with flow data, which can facilitate parametric modeling of cellular state transitions Towards this, we presented 3 key features of the framework. First, we described a finite mixture of skew t-Normal distributions. Second, we 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 as and when they appear and disappear at different points of time thereby offering a parametric characterization of the overall nature of state transition. Third, we designed a mixture merging procedure for ensuring robust estimation of the fitted profile. The code implementing the framework is available from the authors upon request. Indeed the proposed framework is effective, general and may be applied to other similar domains.

Methods and materials

Mixtures of skew Student-*t*-normal distributions

We describe the skew t-Normal mixture model (STNMIX) of StateProfiler. To simplify notation, we let $\varphi(.)$ and $\Phi(.)$ denote the probability density function (pdf) and the cumulative distribution function (cdf) of the standard normal distribution, respectively. Let

$$t(x|\xi,\sigma^2,\nu) = \frac{\Gamma(\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\pi\nu}\sigma}\left(1+\frac{(x-\xi)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2}$$

denote the pdf of the t distribution with location ξ, scale σ^2 and degrees of freedom (df) ν, 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 note further properties.

As introduced by Gómez *et al.* [21], 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 $\nu \in (0,\infty)$ it is has the density

$$\psi(\gamma) = 2t(\gamma|\xi,\sigma^2,\nu)\Phi\left(\lambda\frac{\gamma-\xi}{\sigma}\right). \quad (1)$$

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

Ho *et al.* [13] give following hierarchical representation of STN to establish an EM-type algorithm [22].

$$\begin{aligned} Y|\gamma,\tau &\sim N\left(\xi+\frac{\sigma\lambda}{\tau+\lambda^2}\gamma, \frac{\sigma^2}{\tau+\lambda^2}\right), \\ \gamma|\tau &\sim TN\left(0, \frac{\tau+\lambda^2}{\tau}; (0,\infty)\right), \\ \tau &\sim \Gamma(\nu/2, \nu/2), \end{aligned} \quad (2)$$

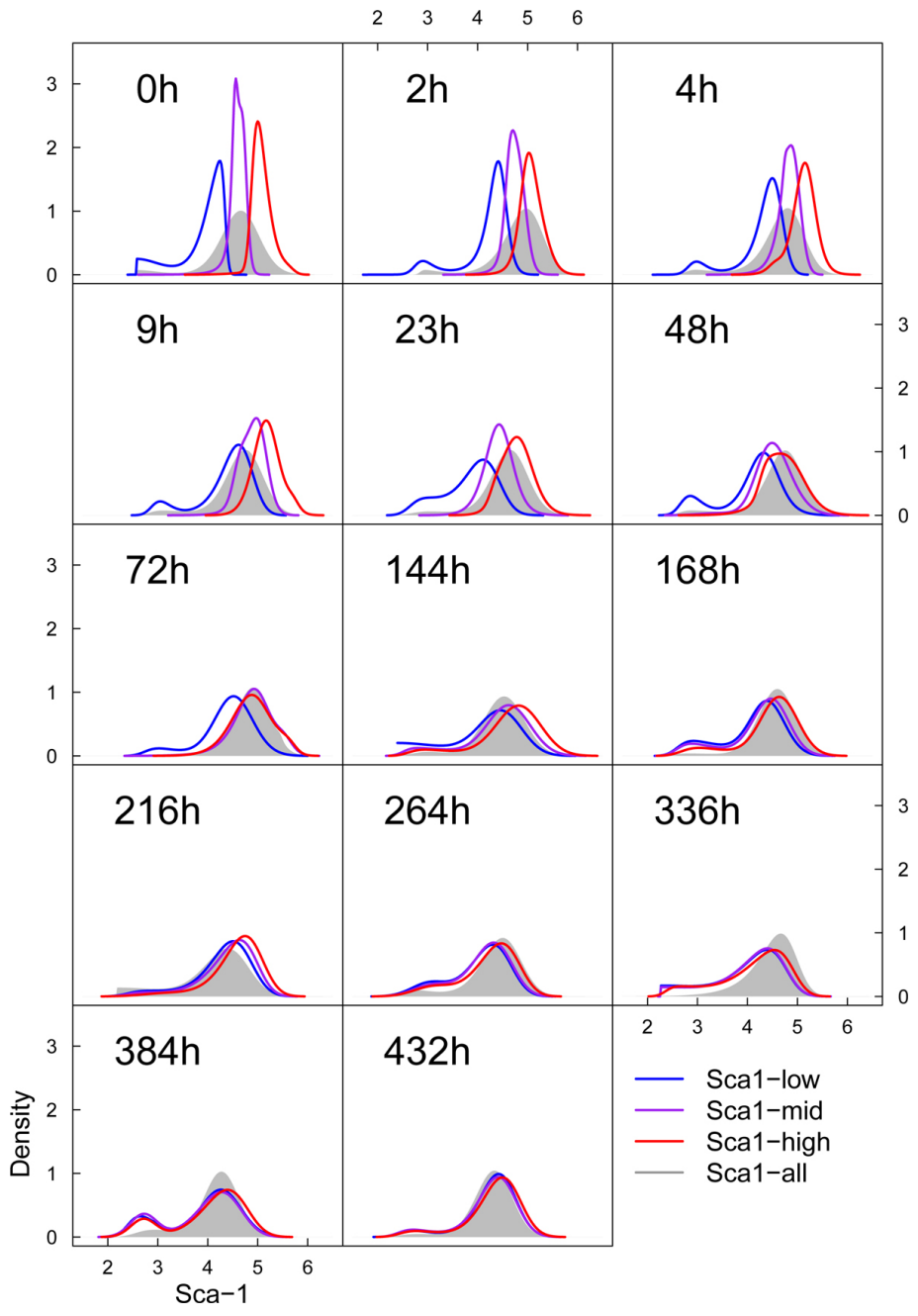

Figure 5 Comparison of time-course profiles. The temporal profiles of the 3 sorted subsets and the unsorted clonal population are constructed with StateProfiler, and plotted for visual comparison.

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

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

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

Where w_i's are mixing proportions which are constrained to be positive and $\sum_{i=1}^{g} w_i = 1$, $\psi(y_j|\theta_i)$ is the STN density defined in (1) and $\Theta_g = (w_1,..., w_{g-1}, \theta_1,..., \theta_g)$

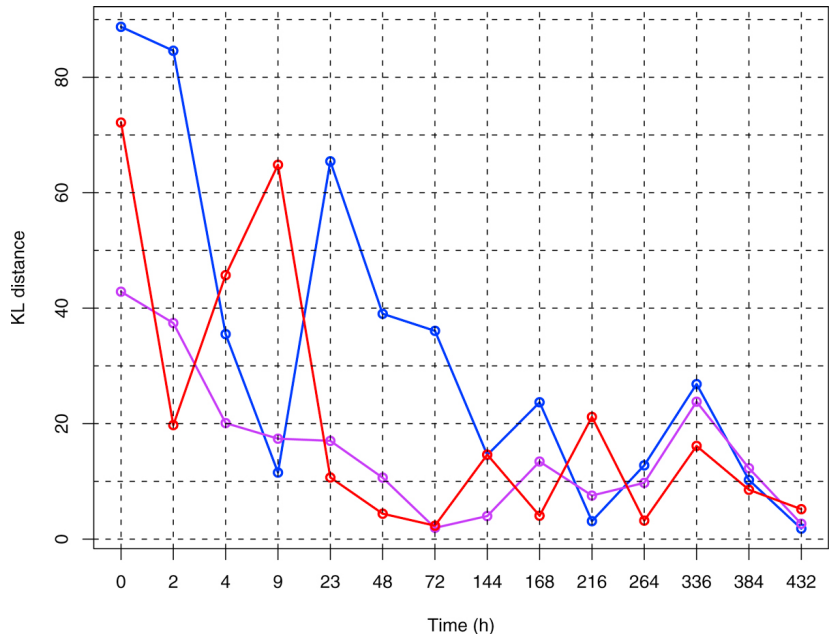

Figure 6 The trend of convergence to the unsorted profile. Kullback-Leibler (KL) distance of the profiles for each of the 3 sorted subpopulations from the unsorted profile at a given time-point. While the distances decrease with time, the trend is slow and does not appear to be monotonic.

represents all unknown parameters. Note that the component vector θ_i consists of $(\xi_i, \sigma_i^2, \lambda_i, \nu_i)$.

Based on (2), a practical ECM/ECME algorithm [23,24] proceeds are described by Ho *et al.* [13] as follows:

E-step: Given $\Theta_g = \hat{\Theta}_g^{(h)}$, compute following $z_{ij}^{(\hat{h})}$, $\hat{\tau}_{ij}^{(h)}$, $\kappa_{ij}^{(\hat{h})}$ and $\gamma_{1\ ij}^{(\hat{h})}$ for $i = 1,..., g$ and $j = 1,..., n$.

$$\hat{z}_{ij}^{(h)} = \frac{\hat{w}_i^{(h)} \psi(y_j|\hat{\theta}_i^{(h)})}{f(y_j|\hat{\Theta}^{(h)})}, \quad \hat{\tau}_{ij}^{(h)} = \frac{\hat{v}_i^{(h)} + 1}{\hat{v}_i^{(h)} + \hat{u}_{ij}^{2(h)}},$$

$$\hat{\kappa}_{ij}^{(h)} = \mathrm{DG}\left(\frac{\hat{v}_i^{(h)} + 1}{2}\right) - \log\left(\frac{\hat{v}_i^{(h)} + \hat{u}_{ij}^{2(h)}}{2}\right),$$

$$\hat{\gamma}_{1ij}^{(h)} = \hat{\lambda}_i^{(h)} \hat{u}_{ij}^{(h)} + \frac{\phi(\hat{\lambda}_i^{(h)} \hat{u}_{ij}^{(h)})}{\Phi(\hat{\lambda}_i^{(h)} \hat{u}_{ij}^{(h)})},$$

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

CM-step: Update the estimation by

$$\hat{w}_i^{(h+1)} = \hat{n}_i^{(h)}/n,$$

$$\hat{\xi}_i^{(h+1)} = \frac{\hat{b}_{1i}^{(h)} + \hat{\lambda}_i^{2(h)} \hat{b}_{2i}^{(h)} - \hat{\sigma}_i^{(h)} \hat{\lambda}_i^{(h)} \hat{b}_{3i}^{(h)}}{\sum_{j=1}^n \hat{z}_{ij}^{(h)} \hat{\tau}_{ij}^{(h)} + \hat{\lambda}_i^{2(h)} \hat{n}_i^{(h)}},$$

$$\hat{\sigma}_i^{2(h+1)} = \frac{1}{\hat{n}_i^{(h)}} \sum_{j=1}^n \hat{z}_{ij}^{(h)} \hat{\tau}_{ij}^{(h)} (y_j - \hat{\xi}_i^{(h+1)})^2,$$

$$\hat{\lambda}_i^{(h+1)} = \frac{\sum_{j=1}^n \hat{z}_{ij}^{(h)} \hat{\gamma}_{1ij}^{(h)} \hat{u}_{ij}^{(h+1)}}{\sum_{j=1}^n \hat{z}_{ij}^{(h)} \hat{u}_{ij}^{2(h+1)}},$$

$$\hat{v}_i^{(h+1)} = \arg\max_{v_i} \left\{ \frac{v_i}{2} \log\left(\frac{v_i}{2}\right) - \log\Gamma\left(\frac{v_i}{2}\right) + \left(\frac{v_i}{2}\right) \hat{b}_{4i}^{(h)} \right\},$$

where $\hat{n}_i^{(h)} = \sum_{j=1}^n \hat{z}_{ij}^{(h)}$, $\hat{b}_{1i}^{(h)} = \sum_{j=1}^n \hat{z}_{ij}^{(h)} \hat{\tau}_{ij}^{(h)} y_j$, $\hat{b}_{3i}^{(h)} = \sum_{j=1}^n \hat{z}_{ij}^{(h)} \hat{\gamma}_{1ij}^{(h)}$, $\hat{b}_{3i}^{(h)} = \sum_{j=1}^n \hat{z}_{ij}^{(h)} \hat{\gamma}_{1ij}^{(h)}$,

$\hat{u}_{ij}^{(h+1)} = (y_j - \hat{\xi}_i^{(\hat{h}+1)})/\sigma_i^{(\hat{h}+1)}$. and

$\hat{u}_{ij}^{(h+1)} = (y_j - \hat{\xi}_i^{(\hat{h}+1)})/\sigma_i^{(\hat{h}+1)}$.

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

$$\hat{v}^{(h+1)} = \arg\max_{v} \left\{ \sum_{j=1}^n \log\left\{ \sum_{i=1}^g \hat{w}_i^{(h+1)} \times \psi\left(y_j|\hat{\xi}_i^{(h+1)}, \hat{\sigma}_i^{2(h+1)}, \hat{\lambda}_i^{(h+1)}, v\right) \right\} \right\}.$$

The E-step and CM/CML-steps are alternately repeated until a suitable convergence rule is satisfied, e.g., the Aitken acceleration based stopping criterion $\ell^{(h+1)}$ where $\ell^{(h+1)}$ is the observed log-likelihood evaluated at $\hat{\Theta}_g^{(h)}$, $\ell_\infty^{(h+1)}$ is the asymptotic estimate of the log-likelihood at iteration $h + 1$ (see [18]; Chap. 4.9) and ε is the desired tolerance. For numerical analyses in this paper, a default value of $\varepsilon = 10^{-6}$ was used to terminate the iterations.

Greedy learning for STN mixtures

In this section, we present a new greedy version of the EM algorithm to determine the optimum number of components in the fitting of STNMIX models. The greedy EM approach was first introduced by Vlassis and Likas [25]. 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 a pre-specified maximum number of components or a pre-specified convergence tolerance.

Suppose 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 θ_{g+1} being the added parameters $(\xi_{g+1}, \sigma^2_{g+1}, \lambda_{g+1}, \nu_{g+1})$. Given an old mixture $f(y_j|\hat{\Theta}_g)$, the weight a and θ_{g+1} are optimally chosen to maximize the new log-likelihood

$$\begin{aligned}\mathcal{L}_{g+1} &= \sum_{j=1}^{n} \log f(y_j|\Theta_{g+1}) \\ &= \sum\nolimits_{j=1}^{n} \log[(1-a)f(y_j|\hat{\Theta}_g) + a\psi(y_j|\theta_{g+1})].\end{aligned} \tag{4}$$

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 $\tilde{\theta}$ denotes and the partial ML estimates of θ. For notational simplicity, the subscript $(g + 1)$ is suppressed below in the Partial E-step.

Partial E-step: Calculating the conditional expectation of latent variables at the kth iteration, this yields

$$\begin{aligned}\tilde{z}_j^{(k)} &= \frac{\tilde{a}^{(k)}\psi(y_j|\tilde{\theta}^{(k)})}{(1-\tilde{a}^{(k)})f(y_j|\hat{\Theta}_g^{(k)}) + \tilde{a}^{(k)}\psi(y_j|\tilde{\theta}^{(k)})}, \\ \tau_j^{\hat{(k)}} &= \frac{\tilde{v}^{(k)}+1}{\tilde{v}^{(k)}+\tilde{u}_j^{2(k)}},\ \tilde{\gamma}_{1j}^{(k)} = \tilde{\lambda}^{(k)}\tilde{u}_j^{(k)} + \frac{\phi(\tilde{\lambda}^{(k)}\tilde{u}_j^{(k)})}{\Phi(\tilde{\lambda}^{(k)}\tilde{u}_j^{(k)})}, \\ \tilde{\kappa}_j^{(k)} &= \mathrm{DG}\left(\frac{\tilde{v}^{(k)}+1}{2}\right) - \log\left(\frac{\tilde{v}^{(k)}+\tilde{u}_j^{2(k)}}{2}\right),\end{aligned}$$

Where $\tilde{u}_j^{(k)} = (y_j - \tilde{\xi}^{(k)})/\tilde{\sigma}^{(k)}$.

Partial M-step: Updating the new parameters in $(a, \boldsymbol{\theta}_{g+1})$, we get

$$\begin{aligned}\tilde{a}^{(k+1)} &= \frac{\sum_{j=1}^{n}\tilde{z}_j^{(k)}}{n}, \\ \tilde{\xi}_{g+1}^{(k+1)} &= \frac{\tilde{b}_1^{(k)} + \tilde{\lambda}^{2(k)}\tilde{b}_2^{(k)} - \tilde{\sigma}^{(k)}\tilde{\lambda}^{(k)}\tilde{b}_3^{(k)}}{\sum_{j=1}^{n}\tilde{z}_j^{(k)}\tilde{\tau}_j^{(k)} + \tilde{\lambda}^{2(k)}\sum_{j=1}^{n}\tilde{z}_j^{(k)}}, \\ \tilde{\sigma}_{g+1}^{2(k+1)} &= \frac{\sum_{j=1}^{n}\tilde{z}_j^{(k)}\tilde{\tau}_j^{(k)}(y_j - \tilde{\xi}^{(k+1)})^2}{\sum_{j=1}^{n}\tilde{z}_j^{(k)}}, \\ \tilde{\lambda}_{g+1}^{(k+1)} &= \frac{\sum_{j=1}^{n}\tilde{z}_j^{(k)}\tilde{\gamma}_{1j}^{(k)}\tilde{u}_j^{(k+1)}}{\sum_{j=1}^{n}\tilde{z}_j^{(k)}\tilde{u}_j^{2(k+1)}}, \\ \tilde{\nu}_{g+1}^{(k+1)} &= \arg\max_{\nu}\left\{\frac{\nu}{2}\log\left(\frac{\nu}{2}\right)\right. \\ &\quad \left. - \log\Gamma\left(\frac{\nu}{2}\right) + \frac{\nu}{2}\tilde{b}_4^{(k)}\right\},\end{aligned}$$

Where $\tilde{u}_j^{(k+1)} = (y_j - \tilde{\xi}^{(k+1)})/\tilde{\sigma}^{(k+1)}$, $\tilde{b}_2^{(k)} = \sum_{j=1}^{n}\tilde{z}_j^{(k)}\gamma_j$, $\tilde{b}_2^{(k)} = \sum_{j=1}^{n}\tilde{z}_j^{(k)}\gamma_j$, $\tilde{b}_3^{(k)} = \sum_{j=1}^{n}\tilde{z}_j^{(k)}\tilde{\gamma}_{1j}^{(k)}$, and $\tilde{b}_4^{(k)} = \sum_{j=1}^{n}\tilde{z}_j^{(k)}(\tilde{\kappa}_j^{(k)} - \tilde{\tau}_j^{(k)})/\sum_{j=1}^{n}\tilde{z}_j^{(k)}$.

The above partial EM steps constitute a fast and simple procedure to locally seek for the maximum of $\mathcal{L}_{g+1}$. To our experience, this local search scheme is very sensitive the initialization of a and ξ_{g+1}. Similar to Vlassis and Likas [25], we provided a global search strategy for extracting proper parameter initialization for a and $\xi_{g+1}^{(0)}$. By a second-order Taylor expansion for $\mathcal{L}_{g+1}$, we obtain the following approximation:

$$\hat{\mathcal{L}}_{g+1} = \mathcal{L}_{g+1}(a_0) - \frac{[\dot{\mathcal{L}}_{g+1}(a_0)]^2}{2\ddot{\mathcal{L}}_{g+1}(a_0)}, \tag{5}$$

where $\dot{\mathcal{L}}_{g+1}(a_0)$ and $\ddot{\mathcal{L}}_{g+1}(a_0)$ are the first and second derivatives of $\mathcal{L}_{g+1}$ evaluated at $a = a_0$. It can be deduced from (5) that a local maximum of $\mathcal{L}_{g+1}$ around $a_0 = 0.5$ is given by

$$\begin{aligned}\hat{\mathcal{L}}_{g+1} = &\sum_{j=1}^{n}\log\left(\frac{f(y_j|\hat{\Theta}_g) + \psi(y_j|\theta_{g+1})}{2}\right) \\ &+ \frac{\left[\sum_{j=1}^{n}\delta_j(\theta_{g+1})\right]^2}{2\sum_{j=1}^{n}\delta_j^2(\theta_{g+1})}\end{aligned} \tag{6}$$

with

$$\delta_j(\theta_{g+1}) = \frac{f(y_j|\hat{\Theta}_g) - \psi(y_j|\theta_{g+1})}{f(y_j|\hat{\Theta}_g) + \psi(y_j|\theta_{g+1})}.$$

So the the optimal value of a can be calculated as

$$\hat{a} = \frac{1}{2}\left(1 - \frac{\sum_{j=1}^{n}\delta_j(\theta_{g+1})}{\sum_{j=1}^{n}\delta_j^2(\theta_{g+1})}\right). \tag{7}$$

Following the suggestion of Li and Barron [26], 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 $\tilde{\sigma}_{g+1}^{2(0)}$ is $n^{-1/5}$ times half of the sample variance s_y^2 whereas $\tilde{\lambda}_{g+1}^{2(0)}$ and $\tilde{\nu_{g+1}}^{2(0)}$ are always fixed at 0 and 10, respectively. For the initial choice of ξ_{g+1}, we search over the 5th, 10th, 15th, ... 95th quantiles of $\boldsymbol{y}$ and set $\tilde{\xi}_{g+1}^{(0)}$ 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 $\mathbf{\Theta}_g$ via the EM-type algorithms.
3. Perform a global search to find a proper initialization of a and ξ_{g+1}.
4. Apply the partial EM-steps until convergence. For instance, $|\hat{\mathcal{L}}_{g+1}^{(k)}/\hat{\mathcal{L}}_{g+1}^{(k-1)} - 1| < 10^{-6}$.
5. If $\hat{\mathcal{L}}_{g+1} \leq \hat{\mathcal{L}}_g + m$ 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 (we have 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 running time is similar to Vlassis and Likas [25].

Merging mixture algorithm

The greedy EM algorithm provides a convenient method for automatically selecting a number of components for a mixture model under reasonable assumptions (such as convexity of components). Yet if data have certain spatial features due to distributions with unusual shapes or low separation [8], it can lead to overlapping components, and hence to overestimation in the number of components in spite of the parsimonious approach. To augment our greedy algorithm for obtaining a robust estimate of the number of components, we extend the merging mixture approach of Baudry *et al.* [27] to skew t-Normal components. While merging techniques have been applied in the past to symmetric distributions [27,28], designing a procedure for asymmetric distributions obviates any need for spurious components that may be required for the sole purpose of modeling asymmetry, and thus avoids redundant merging.

The basic idea behind the procedure is to use the maximum merged entropy to iteratively combine two possibly overlapping clusters, until the result of combination belong a single cluster (see implementation in [28]). The steps of the merging algorithm in StateProfiler are described below.

1. Calculate the mean entropy of maximum estimation for g components as

$$\text{Ent}(g) = -\sum_{j=1}^{n}\sum_{i=1}^{g} \hat{z}_{ij} \log \hat{z}_{ij} \geq 0,$$

where $\hat{z}_{ij}$ denotes the posterior probability given $\mathbf{\Theta}_g$ fix at $\hat{\Theta}_g$.

2. Two clusters l and l' to be combined are those maximizing the criterion:

$$\left(-\sum_{j=1}^{n} \left\{ \hat{z}_{il} \log \hat{z}_{il} + \hat{z}_{il'} \log \hat{z}_{il'} \right\} + \sum_{j=1}^{n} (\hat{z}_{il} + \hat{z}_{il'}) \log(\hat{z}_{il} + \hat{z}_{il'}) \right)$$

among all possible pairs of clusters (l, l').
3. Obtain the merged entropy

$$\text{Ent}(g-1) = -\sum_{j=1}^{n} \left\{ \sum_{i \neq l,l'} \hat{z}_{ij} \log \hat{z}_{ij} + \hat{z}_{i,l\cup l'} \log \hat{z}_{i,l\cup l'} \right\},$$

where $\hat{z}_{i,l\cup l'} = \hat{z}_{il} + \hat{z}_{il'}$ is the posterior probability of the new cluster $l \cup l'$.
4. Update $\hat{z}_j$ consists of the unmerged and merged posterior probabilities.
5. Set $g = g - 1$ and go to 2. Repeat until $g = 1$.
6. A solution of number of components can be identified (i) a sudden jump or "elbow" in a plot of the entropy of clustering versus the number of clusters, or (ii) peaks in a plot of the number of clusters versus the difference in entropy.

Data and experiments

For details of the yeast cell cycle experiments and time-course data analyzed by StateProfiler, see [4]. For details of EML cell differentiation data, see [11].

Acknowledgements

TIL was partially supported by National Science Council of Taiwan (Grant NO. NSC99-2118-M-005-001-MY2).

This article has been published as part of *BMC Bioinformatics* Volume 13 Supplement 5, 2012: Selected articles from the First IEEE International Conference on Computational Advances in Bio and medical Sciences (ICCABS 2011): Bioinformatics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/supplements/13/S5.

Author details

[1]Department of Applied Mathematics and Institute of Statistics, National Chung Hsing University, Taichung 402, Taiwan. [2]Department of Public Health, China Medical University, Taichung 404, Taiwan. [3]Department of Pathology and Surgery, Children's Hospital Boston, Harvard Medical School, Boston, MA 02115, USA. [4]Program in Biophysics, Harvard University, Cambridge, MA 02139, USA. [5]MD-PhD Program, Harvard Medical School, Boston, Massachusetts 02115, USA. [6]Department of Biology, Duke University, Durham, North Carolina, USA. [7]Institute for Biocomplexity and Informatics,

University of Calgary, Calgary, Alberta T2N 1N4, Canada. [8]Broad Institute of MIT and Harvard University, Cambridge, MA 02142, USA. [9]Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA.

Authors' contributions

HJH and TIL co-developed the statistical methods and performed data analysis. SP conceived the project, designed the approach, and analyzed the results. All authors contributed to the development of the methodology and to writing the manuscript. HJH and TIL contributed equally and are the first authors as well as listed in alphabetical order.

Competing interests

The authors declare that they have no competing interests.

Published: 12 April 2012

References

1. Darzynkiewicz Z, Crissman H, Jacobberger JW: **Cytometry of the cell cycle: cycling through history.** *Cytometry A* 2004, **58**:21-32.
2. Krishan A, Krishnamurthy H, Totey S: *Applications of Flow Cytometry in Stem Cell Research and Tissue Regeneration* John Wiley & Sons Inc; 2010.
3. Pyne S, Hu X, Wang K, Rossin E, Lin TI, Maier LM, Baecher-Allan C, McLachlan GJ, Tamayo P, Hafler DA, Jager PLD, Mesirov JP: **Automated high-dimensional flow cytometric data analysis.** *Proc Natl Acad Sci USA* 2009, **106**:8519-8524.
4. Jackson LP, Reed SI, Haase SB: **Distinct mechanisms control the stability of the related S-phase cyclins Clb5 and Clb6.** *Mol Cell Biol* 2006, **26**:2456-2466.
5. Niu W, Li Z, Zhan W, Iyer VR, Marcotte EM: **Mechanisms of cell cycle control revealed by a systematic and quantitative overexpression screen in S. cerevisiae.** *PLoS Genet* 2008, **4**:e1000120, Doi:10.1371/journal.pgen.1000120.
6. Hedley DW, Chow S, Goolsby C, Shankey TV: **Pharmacodynamic monitoring of molecular-targeted agents in the peripheral blood of leukemia patients using flow cytometry.** *Toxicol Pathol* 2008, **36**:133-139.
7. Krishan A, Hamelik RM: **Flow cytometric monitoring of fluorescent drug retention and efflux.** *Methods Mol Med* 2005, **111**:149-166.
8. Kotecha N, Flores NJ, Irish JM, Simonds EF, Sakai DS, Archambeault S, Diaz-Flores E, Coram M, Shannon KM, Nolan GP, Loh ML: **Single-cell profiling identifies aberrant STAT5 activation in myeloid malignancies with specific clinical and biologic correlates.** *Cancer Cell* 2008, **14(4)**:335-343.
9. Irish JM, Myklebust JH, Alizadeh AA, Houot R, Sharman JP, Czerwinski DK, Nolan GP, Levy R: **B-cell signaling networks reveal a negative prognostic human lymphoma cell subset that emerges during tumor progression.** *Proc Natl Acad Sci USA* 2010, **107**:12747-12754.
10. Chang HH, Oh PY, Ingber DE, Huang S: **Multistable and multistep dynamics in neutrophil differentiation.** *BMC Cell Biol* 2006, **7**:11.
11. Chang HH, Hemberg M, Barahona M, Ingber DE, Huang S: **Transcriptome-wide noise controls lineage choice in mammalian progenitor cells.** *Nature* 2008, **453(7194)**:544-547.
12. Frühwirth-Schnatter S, Pyne S: **Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-*t* distributions.** *Biostatistics* 2010, **11**:317-336.
13. Ho HJ, Pyne S, Lin TI: **Maximum likelihood inference for mixtures of skew Student-*t*-normal distributions through practical EM-type algorithms.** *Stat Comput* 2012, **22**:287-299.
14. Lin TI, Lee JC, Yen SY: **Finite mixture modelling using the skew normal distribution.** *Stat Sinica* 2007, **17**:909-927.
15. Rossin E, Lin TI, Ho HJ, Mentzer SJ, Pyne S: **A framework for analytical characterization of monoclonal antibodies based on reactivity profiles in different tissues.** *Bioinformatics* 2011, **27(19)**:2746-2753.
16. Lin TI, Lee JC, Hsieh WJ: **Robust mixture modeling using the skew *t* distribution.** *Stat Comput* 2007, **17**:81-92.
17. Lin TI: **Robust mixture modeling using multivariate skew *t* distributions.** *Stat Comput* 2010, **20**:343-356.
18. McLachlan GJ, Krishnan T: *The EM algorithm and extensions* John Wiley & Sons Inc; 2008.
19. Tibshirani R, Walther G, Hastie T: **Estimating the number of clusters in a data set via the gap statistic.** *J R Stat Soc Ser B Methodol* 2001, **63**:411-423.
20. Song C, Phenix H, Abedi V, Scott M, Ingalls BP, Kaern M, Perkins TJ: **Estimating the stochastic bifurcation structure of cellular networks.** *PLoS Comput Biol* 2010, **6(3)**:e1000699.
21. Gómez HW, Venegas O, Bolfarine H: **Skew-symmetric distributions generated by the distribution function of the normal distribution.** *Environmetrics* 2007, **18**:395-407.
22. Dempster AP, Laird NM, Rubin DB: **Maximum likelihood from incomplete data via the EM algorithm (with discussion).** *J R Stat Soc Ser B Methodol* 1977, **39**:1-38.
23. Meng XL, Rubin DB: **Maximum likelihood estimation via the ECM algorithm: A general framework.** *Biometrika* 1993, **80**:267-278.
24. Liu CH, Rubin DB: **The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence.** *Biometrika* 1994, **81**:633-648.
25. Vlassis N, Likas A: **A greedy EM algorithm for Gaussian mixture learning.** *Neural Process Lett* 2002, **15**:77-87.
26. Li JQ, Barron AR: **Mixture Density Estimation.** In *Advances in Neural Information Processing Systems 12, [NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999*. The MIT Press;Solla SA, Leen TK, Müller KR 1999:279-285.
27. Baudry JP, Raftery AE, Celeux G, Lo K, Gottardo R: **Combining Mixture Components for Clustering.** *J Comput Graph Stat* 2010, **9**:332-353.
28. Finak G, Bashashati A, Brinkman R, Gottardo R: **Merging mixture components for cell population identification in flow cytometry.** *Adv Bioinformatics* 2009, **2009**:Article ID 247646, Doi:10.1155/2009/247646.

doi:10.1186/1471-2105-13-S5-S5
Cite this article as: Ho *et al.*: Parametric modeling of cellular state transitions as measured with flow cytometry. *BMC Bioinformatics* 2012 **13** (Suppl 5):S5.