Data-driven Prediction of Stem Cell Expansion Cultures

Zhaozheng Yin, Dai Fei Ker, Silvina Junkers, Takeo Kanade, Mei Chen, Lee Weiss, and Phil Campbell

Abstract—Stem cell expansion culture aims to generate sufficient number of clinical-grade cells for cell-based therapies. One challenge for ex vivo expansion is to decide the appropriate time to perform subculture. Traditionally, this decision has been reliant on human estimation of cell confluency and predicting when confluency will approach a desired threshold. However, the use of human operators results in highly subjective decision-making and is prone to inter- and intra-operator variability. Using a real-time cell image analysis system, we propose a data-driven approach to model the cell growth process and predict the cell confluency levels, signaling times to subculture. This approach has great potential as a tool for adaptive real-time control of subculturing, and it can be integrated with robotic cell culture systems to achieve complete automation.

I. INTRODUCTION

Stem cell engineering promises to revolutionize regenerative medicine by helping to repair diseased or damaged tissues and organs. Starting with the relatively small number of primary stem cells available in isolates from the body, one of the critical bioprocessing steps required by successful cell-based therapies is to generate a sufficient number of clinical-grade stem cells through ex vivo cell culture expansions [4]. However, tight control of the expansion process remains a challenge. In particular, determining the appropriate time to perform cell subculturing is important. Delayed subculturing of cells can result in cell overgrowth, which leads to loss of stem cell differentiative potential (stemness); whereas premature subculturing can lead to longer production time to achieve targeted cell yields, with associated added costs. Traditionally, the decision to subculture is based on cell confluency which is related to the cell packing densities in the culture vessel. However, estimation of cell confluency by human operators is a highly subjective task and prone to inter- and intra-operator variability [5]. Furthermore, it is not practical or cost-effective for human operators to manually observe and monitor cell cultures 24/7. Automating the decision on when to subculture cells will result in more consistent outcomes and reduce variability, leading to more efficient and reliable stem cell culture systems.

Time-lapse microscopy imaging has been used to monitor the cell growth process [3] where the degree of cell confluency level in images is used as a metric to assess the cell culture process. To augment human monitoring, we propose a data-driven approach to model the cell growth process and predict the optimal confluency for a real-time adaptive subculture system. First, time-lapse images of cells under the same culture condition are acquired to monitor the cell growth process, and to compute the cell confluency over time. These experiments are terminated without further subculture when the computed confluency exceeds a predetermined cell confluency level. These pre-recorded images with computed time series of confluency metrics serve as training data for subsequent real-time adaptive control experiments. We then build a linear subspace using principle component analysis (PCA) on the training data. When performing a new cell culture experiment with the same culture conditions as our training experiments, we project the observed confluency data onto the linear subspace to model the cell growth process and predict the future confluency. One application of our prediction approach is to notify a human operator in advance when to perform a subculture. For example, 4 hours prior to exceeding a pre-determined confluency level (e.g., 50%), the image analysis and prediction system alerts a human operator via text messaging and/or email to prepare for subculture. The goal is to help human operators expand a population of stem cells to reach a target number in an efficient manner without exceeding or being far away from the pre-determined optimal confluency level (i.e., avoiding delayed or premature subculture).

In this paper, we first introduce in Section II how we compute confluency metrics to monitor cell growth processes. Then, in Section III we present our data-driven model. The dynamic prediction on cell confluency levels is described in Section IV. In Section V we quantitatively compare our data-driven approach with other parametric models and introduce the application of our prediction system.

II. MONITORING CELL GROWTH PROCESS

During the cell culture experiment, we capture real-time phase contrast microscopy images to monitor the degree of confluency inside the field of view. The confluency metric is defined as the number of pixels occupied by cells divided by the total number of pixels in the image. For a given phase contrast image (Fig. 1a), we restore its corresponding artifact-free image without the halo or shade-off effects [6], as shown in Fig. 1b. In the restored image, cell pixels
have positive values while background pixels have near-zero values, which is amenable to image segmentation by thresholding. The thresholded binary mask is shown in Fig. 1c. The resultant cell mask overlaid on top of the original image is shown in Fig. 1d, which proves to be a good estimation of the confluency metric.

Given a time-lapse microscopy image sequence, we compute the confluency metric for each individual image. This produces time series data on confluency. As shown in Fig. 2, while stem cells keep dividing (mitosis), the confluency of the culture process increases accordingly. The small “dips” observed in the confluency curve correspond to minor changes in cell shapes over a period time.

III. Modeling Cell Growth Process

Monitoring cell growth with time-lapse microscopy imaging generates time series confluency data (e.g. Fig. 2). Parametric models on the cell growth process can be obtained by data-fitting. For example, we can fit the second-order polynomial model on the observed confluency data by

$$x(t) = p_2 t^2 + p_1 t + p_0$$

where $x = [x(0), \cdots, x(t), \cdots, x(T)]^T$ is a vector storing the observed confluency metrics from time $t = 0$ to time $t = T$, and $p = [p_1, p_2, p_3]^T$ is the parameter vector. Or, we can fit exponential model onto the data by

$$x(t) = e^{kt} + c$$

where $k$ and $c$ are the scalar parameters. All the parameters ($p, k, c$) are computed using the least square technique [1].

However, these parametric models that depend on specific cell types and culture experiments might lack practical or biological meanings. Instead, we propose a data-driven approach that models the growth process based on observed training data without assuming any specific model. We ran $N$ cell culture experiments on the same type of cells using the same culture condition to obtain the training data. Images of the cell culture experiments were captured every 5 minutes using a phase contrast microscopy imaging system, which generated $N$ time-lapse image sequences for training purposes. We computed confluency metrics for all the $N$ sequences (Fig. 3a). Since the first image of each sequence may have different degrees of confluency (i.e., the number of seed cells may be different for the $N$ sequences), we search the largest initial confluency of the $N$ curves in Fig. 3a, and then align all the $N$ curves such that they start from the same initial condition (Fig. 3b).

Then, we apply PCA [2] onto the training data using Singular Value Decomposition (SVD)

$$X = UV^T$$

where data matrix $X = [x_1, \cdots, x_N]$ stores the vectors of the confluency metrics of the $N$ image sequences, $U$ and $V$ are two orthogonal matrices, and $S$ is a diagonal matrix with rank-ordered singular values (Fig. 4). We choose the column vectors of $V$ that correspond to the first $K$ (e.g. $K = 2$) largest singular values to span a linear subspace for our data-driven modeling.

For a new cell culture experiment having the same type of cells and the same culture condition as our training experiments, we monitor its culture process and compute the observed confluency, $z$. The culture process can be modeled in our trained linear subspace by

$$y = \sum_{k=1}^{K} a_k v_k$$

where $a_k$ are the coefficients computed from the training data.
when predicting the confluency level for a new cell culture experiment, we first compute the coefficients based on the observed data with the least error on the culture process.

where \( v_k \) denotes the \( k \)th principle vector in \( V \), the coefficient \( a_k \) is computed by

\[
a_k = v_k^T \cdot z.
\]  

(5)

The performance of the modeling is evaluated by the sum of absolute difference between the modeled culture process, \( y \), and the observed culture process, \( z \),

\[
Err = \sum_{t = 0}^{T} |y(t) - z(t)|.
\]  

(6)

Compared to the two parametric models (polynomial and exponential), the data-driven model fits the observed data with the least error on the culture process in Fig. 5.

IV. PREDICTING CELL CONFLUENCY

Our goal is to accurately predict the cell confluency at a future time \( t + L \) based on the observed confluency data from time 0 till time \( t \), where \( L \) is the prediction time lag. When \( L = 1 \), we predict the confluency at the next frame. When \( L = 48 \), we predict the confluency 4 hours later (images are captured every 5 minutes, and the time unit is represented by the image index.) In this section, the data-driven model (Eq. 4) is further extended to dynamic prediction. Denote time-dependent data matrix \( X^{(t)} = [x_1^{(t)}, \cdots, x_N^{(t)}] \) where \( x_i^{(t)} = [x_i(0), \cdots, x_i(t)]^T \) (i.e., \( x_i^{(t)} \) is the observed time series confluency of sequence \( i \) from time 0 till time \( t \)), we perform SVD

\[
X^{(t)} = U^{(t)}S^{(t)}V^{(t)T}
\]  

(7)

on all the \( t \)'s (\( t = 0, \cdots, T \)). Thus, for any time index \( t \), we get a set of \( K \) principle components, \( \{v_1^{(t)}, \cdots, v_K^{(t)}\} \).

When predicting the confluency level for a new cell culture experiment, we first compute the coefficients based on the current observed time series data, \( z^{(t)} = [z(0), \cdots, z(t)]^T \),

\[
a_k^{(t)} = v_k^{(t)T} \cdot z^{(t)}
\]  

(8)

then the confluency at time \( t + L \) is predicted by

\[
z^{(t+L)}(t + L) = \sum_{k=1}^{K} a_k^{(t)} v_k^{(t+L)}(t + L)
\]  

(9)

Using the evaluation criterion in Eq. 6, we compare the data-driven prediction method to the other two predictions using parametric models. As shown in Fig. 6, when predicting the confluency in the temporal domain with a small time lag, all three prediction methods work reasonably well and the data-driven prediction achieves the least prediction error. When the prediction time lag \( (L) \) increases, the error of all the prediction methods increase (Fig. 7). In particular, the prediction by a polynomial model is quite unstable at the beginning when there is not enough data for model fitting (Fig. 7b). The data-driven prediction still achieves the least prediction error for the larger prediction lag.

We further quantitatively evaluate how well the three prediction methods can predict future confluency by changing the time lag from \( L = 1 \) (5 minutes) to \( L = 96 \) (8 hours). As shown in Fig. 8, the data-driven prediction outperforms the other two methods consistently with the least prediction error, and the prediction by data-driven or exponential model is much more stable than the prediction by polynomial model as the time lag increases.

V. EXPERIMENTS

We recorded a total of 48 image sequences under four different cell culture conditions with sample images shown in Fig. 9. The images were captured every 5 minutes and each sequence consists of 1000 images at the resolution of 1392*1040 pixels. Under each culture condition, we have
12 image sequences. We use the “leave-one-out” strategy to evaluate the prediction performance. After selecting one out of the 12 sequences, the remaining 11 sequences undergo PCA analysis to obtain the principle components (Eq. 7). Then, we run the prediction (Eq. 9) on the selected sequence and compare the prediction with the observation using Eq. 6. We repeat the “leave-one-out” evaluation for each of the 12 sequences and use the summation of all the prediction errors as the final evaluation criterion on the 12 sequences.

We propose a data-driven approach to model the cell growth process and predict when a pre-determined cell confluency threshold will be exceeded, requiring cells to be subcultured. Compared to the typical parametric models for predicting cell growth, our data-driven approach learns the cell growth model from a training set of cell culture experiments and achieves higher prediction accuracy on cell culture experiments that have the same culture condition as training experiments. This data-driven approach has great potential as a tool for adaptive real-time control of subculturing, and it can be integrated with robotic cell culture systems to achieve complete automation.

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>THE PREDICTION ERROR OF THREE METHODS WITH $L = 24$.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control</td>
</tr>
<tr>
<td>Data driven</td>
<td>122.2</td>
</tr>
<tr>
<td>Polynomial</td>
<td>146.2</td>
</tr>
<tr>
<td>Exponential</td>
<td>148.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE II</th>
<th>THE PREDICTION ERROR OF THREE METHODS WITH $L = 48$.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control</td>
</tr>
<tr>
<td>Data driven</td>
<td>136.0</td>
</tr>
<tr>
<td>Polynomial</td>
<td>259.5</td>
</tr>
<tr>
<td>Exponential</td>
<td>164.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE III</th>
<th>THE PREDICTION ERROR OF THREE METHODS WITH $L = 96$.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control</td>
</tr>
<tr>
<td>Data driven</td>
<td>133.4</td>
</tr>
<tr>
<td>Polynomial</td>
<td>631.7</td>
</tr>
<tr>
<td>Exponential</td>
<td>188.4</td>
</tr>
</tbody>
</table>

VI. CONCLUSION

Determining the appropriate time to perform subculture is important to optimize the process of stem cell expansion. We monitor the process of cell growth by computing the degree of cell confluency in phase-contrast microscopy images. Based on the cell confluency measurements, we propose a data-driven approach to model the cell growth process and predict when a pre-determined cell confluency threshold will be exceeded, requiring cells to be subcultured. Compared to the typical parametric models for predicting cell growth, our data-driven approach learns the cell growth model from a training set of cell culture experiments and achieves higher prediction accuracy on cell culture experiments that have the same culture condition as training experiments. This data-driven prediction has great potential as a tool for adaptive real-time control of subculturing, and it can be integrated with robotic cell culture systems to achieve complete automation.

REFERENCES