33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

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 pre-determined 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

This work was supported by NIH grants RO1EB004343 & RO1EB007369, CMU Cell Image Analysis Consortium and Intel Labs budget.

Zhaozheng Yin, Silvina Junkers, Takeo Kanade and Lee Weiss are with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA15213 USA (Phone: 412-268-3016; e-mail: {yinz,sjunkers,tk,lew}@cs.cmu.edu).

Dai Fei Ker is with the Biological Sciences Department, Carnegie Mellon University, Pittsburgh, PA15213 USA (e-mail: eker@andrew.cmu.edu).

Mei Chen is with Intel Science and Technology Center, Pittsburgh, PA15213, USA (e-mail: mei.chen@intel.com).

Phil Campbell is with the Institute of Complex Engineering Systems, Carnegie Mellon University, Pittsburgh, PA15213 USA (e-mail: pcampbel@cs.cmu.edu).

978-1-4244-4122-8/11/$26.00 ©2011 IEEE

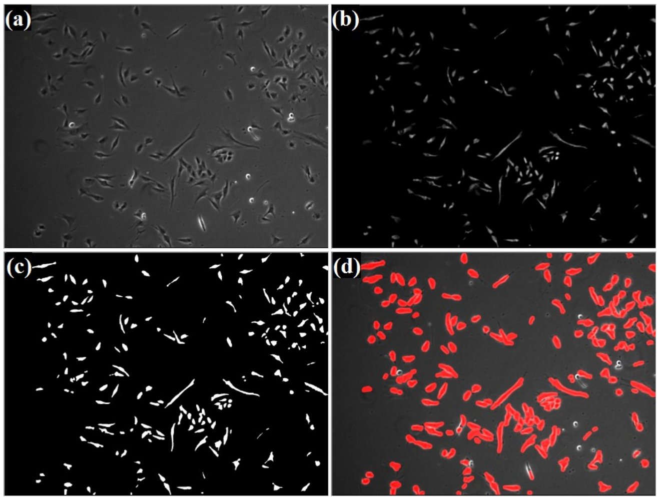


Fig. 1. Compute confluency. (a) A phase contrast microscopy image; (b) Restored image without halo or shade-off artifacts; (c) The segmented cell masks by globally thresholding the restored image; (d) Segmentation results (red) overlaid on the original image.

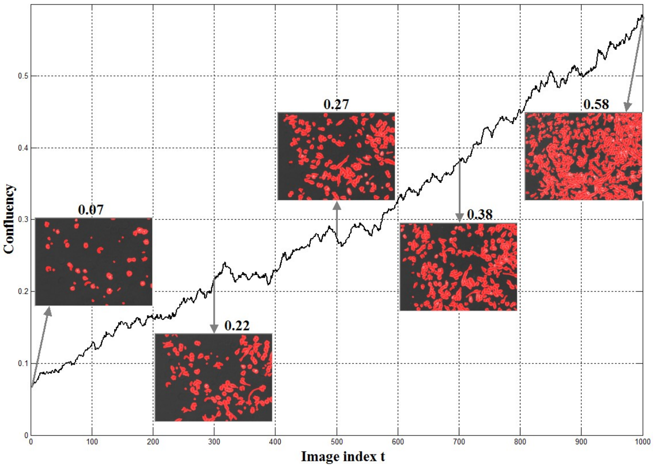


Fig. 2. The confluency increases during the culture process. Five sample images overlaid with segmented cell masks (red) show the confluency level at five time instants.

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

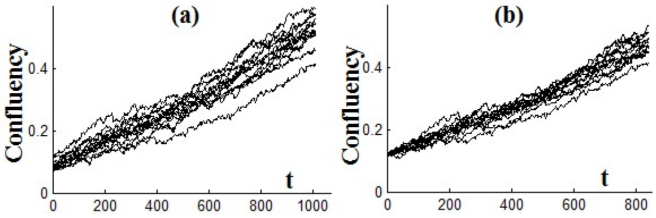


Fig. 3. Align time series confluency curves. (a) The original confluency metrics of N time-lapse image sequences; (b) Aligned curves such that their cell culture processes start from the same initial confluency.

$$\mathbf{x}(t) = \mathbf{p}_2 t^2 + \mathbf{p}_1 t + \mathbf{p}_0 \tag{1}$$

where $\mathbf{x} = [\mathbf{x}(0), \cdots, \mathbf{x}(t), \cdots, \mathbf{x}(T)]^T$ is a vector storing the observed confluency metrics from time $t = 0$ to time $t = T$, and $\mathbf{p} = [\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3]^T$ is the parameter vector. Or, we can fit exponential model onto the data by

$$\mathbf{x}(t) = e^{kt} + c \tag{2}$$

where k and c are the scalar parameters. All the parameters $(\mathbf{p}, k, c)$ 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)

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T \tag{3}$$

where data matrix $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]$ stores the vectors of the confluency metrics of the N image sequences, $\mathbf{U}$ and $\mathbf{V}$ are two orthogonal matrices, and $\mathbf{S}$ is a diagonal matrix with rank-ordered singular values (Fig. 4). We choose the column vectors of $\mathbf{V}$ 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, $\mathbf{z}$. The culture process can be modeled in our trained linear subspace by

$$\mathbf{y} = \sum_{k=1}^{K} a_k \mathbf{v}_k \tag{4}$$

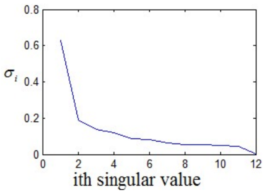


Fig. 4. The variance (singular value) of each principle component.

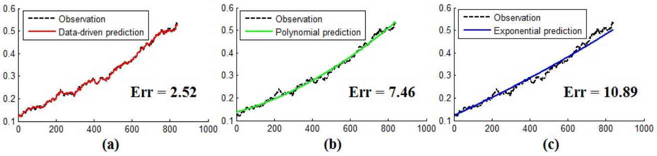


Fig. 5. Modeling cell growth by three methods: (a) Data-driven; (b) Polynomial, and (c) Exponential. The data-driven model fits the observed data with the least error on the culture process.

where $\mathbf{v}_k$ denotes the kth principle vector in $\mathbf{V}$, the coefficient a_k is computed by

$$a_k = \mathbf{v}_k^T \mathbf{z}. \tag{5}$$

The performance of the modeling is evaluated by the sum of absolute difference between the modeled culture process, $\mathbf{y}$, and the observed culture process, $\mathbf{z}$,

$$\text{Err} = \sum_{t=0}^{T} |\mathbf{y}(t) - \mathbf{z}(t)|. \tag{6}$$

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 $\mathbf{X}^{(t)} = [\mathbf{x}_1^{(t)}, \cdots, \mathbf{x}_N^{(t)}]$ where $\mathbf{x}_i^{(t)} = [\mathbf{x}_i(0), \cdots, \mathbf{x}_i(t)]^T$ (i.e., $\mathbf{x}_i^{(t)}$ is the observed time series confluency of sequence i from time 0 till time t), we perform SVD

$$\mathbf{X}^{(t)} = \mathbf{U}^{(t)} \mathbf{S}^{(t)} {\mathbf{V}^{(t)}}^T \tag{7}$$

on all the t's ($t = 0, \cdots, T$). Thus, for any time index t, we get a set of K principle components , $\{\mathbf{v}_1^{(t)}, \cdots, \mathbf{v}_K^{(t)}\}$.

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

$$a_k^{(t)} = {\mathbf{v}_k^{(t)}}^T \mathbf{z}^{(t)} \tag{8}$$

then the confluency at time $t + L$ is predicted by

$$\mathbf{z}^{(t+L)}(t + L) = \sum_{k=1}^{K} a_k^{(t)} \mathbf{v}_k^{(t+L)}(t + L) \tag{9}$$

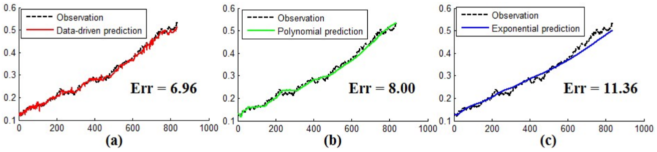


Fig. 6. Predicting the confluency in the next image using three prediction methods: (a) Data-driven; (b) Polynomial, and (c) Exponential. The data-driven model has the least prediction error on the culture process.

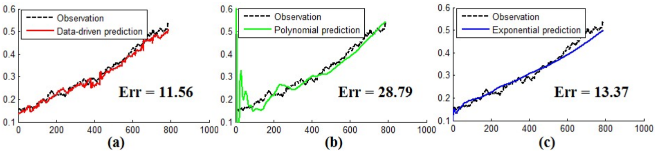


Fig. 7. Predicting the confluency 4 hours later using three prediction methods: (a) Data-driven; (b) Polynomial, and (c) Exponential. The data-driven model is more stable compared to the parametric models and it has the least prediction error on the culture process.

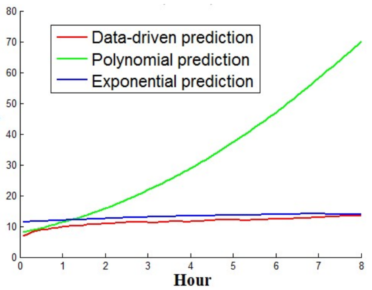


Fig. 8. The prediction error of three methods regarding to different prediction time lags. The data-driven prediction is stable and it outperforms the other two parametric methods consistently with the least prediction error.

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

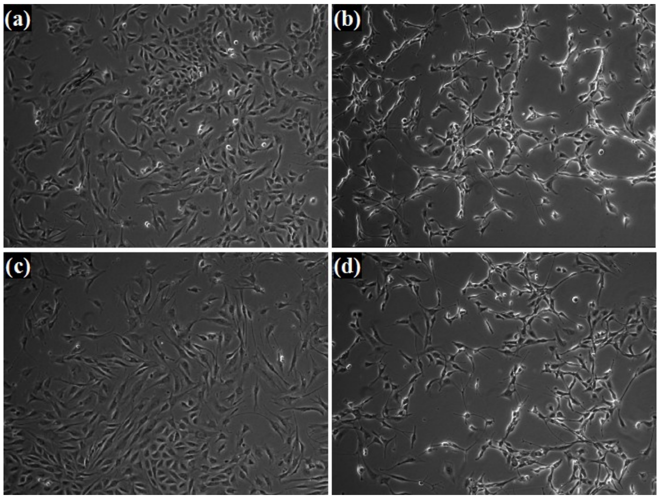


Fig. 9. Sample images from four cell culture conditions. (a) Control; (b) With FGF2; (c) With BMP2; (d) With FGF2+BMP2.

TABLE I

THE PREDICTION ERROR OF THREE METHODS WITH $L = 24$.

	Control	FGF2	BMP2	FGF2+BMP2
Data driven	122.2	68.7	154.0	137.6
Polynomial	146.2	105.3	179.6	168.2
Exponential	148.9	97.4	309.0	211.2

TABLE II

THE PREDICTION ERROR OF THREE METHODS WITH $L = 48$.

	Control	FGF2	BMP2	FGF2+BMP2
Data driven	136.0	74.4	169.6	157.8
Polynomial	259.5	222.7	304.8	315.6
Exponential	164.7	112.6	340.9	235.3

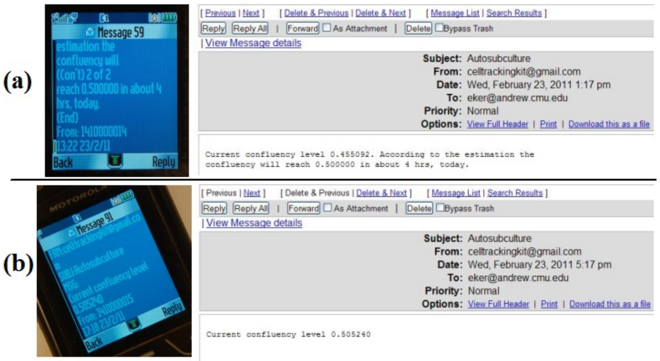
TABLE III

THE PREDICTION ERROR OF THREE METHODS WITH $L = 96$.

	Control	FGF2	BMP2	FGF2+BMP2
Data driven	153.4	78.4	192.6	178.9
Polynomial	631.7	628.9	714.8	809.7
Exponential	188.4	136.3	400.5	278.7

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. As shown in Tables 1, 2 and 3, the data-driven prediction achieves the least error at confluency prediction over all the four culture conditions for different prediction time lags.

The data-driven prediction on cell culture process is useful for automating the decision process for determining when to perform subculture. A human operator first runs several experiments to culture the cells until they reach a pre-determined cell confluency level for subculture. The recorded image sequences corresponding to these experiments will be used to build the data-driven model in Eq. 4 and compute the time series principle components in Eq. 7. Using the same type of cells and under the same culture condition, the human operator starts the recursive cell culture/subculture process whose goal is to culture a sufficient number of cells. In the meantime, the human operator sets up the image analysis and prediction system such that it can notify him/her h (e.g. $h = 4$) hours prior to exceeding a pre-determined cell confluency level, to prepare for subculture. Fig. 10 shows a successful cell culture experiment by the advance notification.


(a)
Message 59
Back
Reply
[Previous | Next] [Delete & Previous | Delete & Next] [Message List | Search Results]
Reply
Reply All
Forward
□ As Attachment
Delete
□ Bypass Trash
| View Message details
View Full Header | Print | Download this as a file
(b)
MOTOROLA
Message 91
Back
Reply
[Previous | Next] [Delete & Previous | Delete & Next] [Message List | Search Results]
Reply
Reply All
Forward
□ As Attachment
Delete
□ Bypass Trash
| View Message details
View Full Header | Print | Download this as a file

(a)

estimation the
confluency will
((Con't) 2 of 2
reach 0.500000 in about 4
hrs, today.
(End)
From: 1410000014
13:22 23/2/11

Subject: Autosubculture
From: celltrackingkit@gmail.com
Date: Wed, February 23, 2011 1:17 pm
To: eker@andrew.cmu.edu
Priority: Normal

```
Current confluency level 0.455092. According to the estimation the
confluency will reach 0.500000 in about 4 hrs, today.
```

(b)

FRM:celltrackingkit@gmail.co
m
SUBJ:Autosubculture
MSG:
Current confluency level
0.505240
From: 1410000015
17:18 23/2/11

Subject: Autosubculture
From: celltrackingkit@gmail.com
Date: Wed, February 23, 2011 5:17 pm
To: eker@andrew.cmu.edu
Priority: Normal

```
Current confluency level 0.505240
```

Fig. 10. Advance notification for cell culture. (a) A human operator was notified by text message and email 4 hours prior to exceeding a pre-determined cell confluency level; (b) Confirmation text message and email were sent when the cell confluency level approached the pre-determined threshold.

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 realtime control of subculturing, and it can be integrated with robotic cell culture systems to achieve complete automation.

REFERENCES

[1] C. Bishop,"Pattern Recognition and Machine Learning," Springer, 2006.

[2] R. Duda, P. Hart, and D. Stork, "Pattern Classification," Wiley, 2001.

[3] M. Kino-oka, and J.E. Prenosil, "Development of an On-Line Monitoring System of Human Keratinocyte Growth by Image Analysis and Its Application to Bioreactor Culture," *Biotechnol. Bioeng.*, 67(2): 234-9, 2000.

[4] Y. Liu, P. Hourd, A. Chandra, and D. J. Williams, "Human Cell Culture Process Capability: a Comparison of Manual and Automated Production," *J. Tissue Eng. Regen. Med.*, 4:45-54, 2010

[5] F. S. Veraitch, R. Scott, J. Wong, Gary. J. Lye, and C. Mason, " The Impact of Manual Processing on the Expansion and Directed Differentiation of Embryonic Stem Cells" *Biotechnology and Bioengineering*, 99(5): 1216-1229, 2008.

[6] Z. Yin, K. Li, T. Kanade and M. Chen, "Understanding the Optics to Aid Microscopy Image Segmentation," *Proceedings of the 13th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2010.