Computers in Biology and Medicine 94 (2018) 55–64

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Critical texture pattern feature assessment for characterizing colonies of induced pluripotent stem cells through machine learning techniques

Muthu Subash Kavitha [a], Takio Kurita [b], Byeong-Cheol Ahn [a,*]

[a] Department of Nuclear Medicine, Kyungpook National University School of Medicine and Hospital, Daegu, South Korea
[b] Graduate School of Engineering, Hiroshima University, Hiroshima, Japan

ARTICLE INFO

Keywords:
Texture features
Induced pluripotent stem cells
Machine learning
Colony region
Regression

ABSTRACT

The objectives of this study are to assess various automated texture features obtained from the segmented colony regions of induced pluripotent stem cells (iPSCs) and confirm their potential for characterizing the colonies using different machine learning techniques. One hundred and fifty-one features quantified using shape-based, moment-based, statistical and spectral texture feature groups are extracted from phase-contrast microscopic colony images of iPSCs. The forward stepwise regression model is implemented to select the most appropriate features required for categorizing the colonies. Support vector machine (SVM), random forest (RF), multilayer perceptron (MLP), decision tree (DT), and adaptive boosting (Adaboost) classifiers are used with ten-fold cross-validation to evaluate the texture features within each texture feature group and fused-features group to characterize healthy and unhealthy colonies of iPSCs. Overall, based on the classification performances of the four texture feature groups using the five classifier models, statistical features always exhibit a high predictive capacity (>87.5%). However, the classification performance using fused texture patterns with statistical, shape-based, and moment-based features was found to be robust and reliable with fewer false positive and false negative values compared to the features when either one is used for the classification of colonies of iPSCs. Furthermore, the results showcase that the SVM, RF and Adaboost classifiers deliver better classification performances than DT and MLP. Our findings suggest that the proposed automated fused statistical, shape-based, and moment-based texture pattern features trained with machine learning techniques are potentially more appropriate and helpful to biologists for characterizing colonies of stem cells.

1. Introduction

Induced pluripotent stem cells (iPSCs) are pluripotent stem cells generated from adult cells, and their generation by reprogramming offers tremendous potential for the development of therapies for numerous diseases and disease modeling [1]. However, constant control and investigation of the characteristics of iPSCs are important for subsequent drug therapy and clinical transplantations. Several methods have been reported for the quality determination of stem cells. Subjective approaches include auditory feedback with four types of sonification methods [2] and protocol-based subjective differentiation [1]. In addition to these purely subjective methods, several semi-automatic computer software tools have been developed [3,4]. Specifically, these software tools analyze the iPSC categories in relation to the morphological characteristics of the iPSC colonies. To reduce the probability of errors associated with manual estimation, several researchers have developed automated systems to evaluate human iPSCs [5,6]. Most of these automated methods aim to identify healthy colonies of iPSCs using intensity-based biomarkers from a manually optimized learning set constructed from patches of cells in the colonies. A completely automated system that considers the texture pattern features of iPSC colonies may be able to accurately characterize microscopic stem cells and confirm additional texture features.

Several image-based textural feature analysis methods have been developed and successfully applied in various biomedical applications [7,8]. Investigators have tested texture feature patterns of a variety of stem cells for identifying their colonies [9,10]. A recent study examined the potential of the biomarkers of embryonic stem cells using video images for distinguishing between the qualities of differentiated and undifferentiated stem cells [9]. Although such extracted microscopic stem cell biomarkers exhibit variations in gray-level intensities, they ignore the spatial relationships between the reference and neighbor pixels of a

* Corresponding author.
E-mail address: abc2000@knu.ac.kr (B.-C. Ahn).

https://doi.org/10.1016/j.compbiomed.2018.01.005
Received 3 August 2017; Received in revised form 17 January 2018; Accepted 17 January 2018

colony region [6]. High-order textural features have also been examined to differentiate between the domed and flat structures of the embryonic stem cells of mice; however, they failed to determine the degree of complexity and local intensity variations in the object structure [10]. The above-mentioned findings, though promising, decisions based on multiple feature sets with various classification algorithms are often more robust and can lead to global decisions [11].

The main purpose of this study was to analyze the effectiveness of numerous textures extracted quantitatively from segmented colony regions and evaluate these features using several machine learning classifiers for their ability to characterize iPSC colonies. The predictive power of each microscopic texture feature of an iPSC colony for classifying healthy and unhealthy colony groups was examined via the receiver operating characteristic analysis method. Five machine learning classifiers with ten-fold cross-validation were trained with the selected texture features within each texture feature group and fused-features group. In addition, we compared selected texture features within each texture feature group and fused-features group according to performance metrics to identify more optimal feature sets for the classification of colonies of stem cells.

2. Materials and methods

2.1. Data acquisition

Out of 169 iPSC data sets, 90 datasets were maintained as described elsewhere [12] and the remaining 79 data sets were purchased from American Type Culture Collection. In brief, inactive murine embryonic fibroblast (MEF) feeder cells were seeded on gelatin-coated tissue culture dishes. The iPSCs were co-cultured with the MEF feeders in an embryonic stem cell medium at 37 °C under 3% CO_2. For passaging of the iPSCs, collagenase was added to detach the colonies from the plate. The harvested colonies were triturated to generate medium-sized small fragments that were subsequently seeded on new plates together with the MEF feeders. All images were prepared under the 100× objective of the phase contrast microscope in the BioStation CT system using automatic Z-focus with a resolution of 1360 × 1024 pixels. The stem cells used in this experiment were not stained or genetically modified, thereby enabling their noninvasive analysis.

2.2. iPSC colony segmentation

In the segmentation procedure median filtering of 15 × 15-pixel size was found to be suitable as a preprocessing step to improve the quality of the image [13]. This sliding window spatial filter replaces the center value with the median value of all pixels in the window. It removes the background noise while preserving the colony region contour. Because the application of filters with different filter sizes such as average, Gaussian and Sobel filters on the original image suppress or eliminate some small colony contour features, a median filter is used to produce a smooth edge colony image without blurring for subsequent procedures. The segmentation of the colony region was achieved by adapting the k-means clustering algorithm [14], whose objective criterion is to minimize the sum of the within-cluster variations. The first step of this algorithm is the selection of the number of clusters as a priori information. The number of clusters (k) varying from 2 to 8 was evaluated in this study. Some uncertainties of cells were observed inside the colonies as well as in the surrounding feeder cells. Furthermore, if the number of clusters increased above two, we observed that the small variations between the colony contour and the feeder cells resulted in incorrect segmentation of overlapping of colony contour with their connected regions of other cells. Hence the number of clusters k = 2 was most suitable for segmenting the colony regions from the background. The next step was the allocation of every pixel according to the group with nearest centroid based on the measured minimum squared Euclidean distance. Subsequent to the grouping of similar object pixels, the algorithm recalculated the new centroids of each cluster. This process was repeated until there were no further changes in the locations of the centroids. The objective criterion function is represented as

$$P = \sum_{j=1}^{q} \sum_{i=1}^{k} \left\| u_i^j - e_j \right\|^2 \quad (1)$$

where $\left\| u_i^j - e_j \right\|^2$ is the desired distance measured between pixel u_i^j and cluster center e_j. Finally, the cluster pixels were reshaped into an image that illustrated the segmented colony region of the iPSCs. Morphological dilation using a disk-shaped structuring element with radius 2, erosion using structuring element with ones of 3, and hole filling operations to fill holes in the background pixels were then used to strengthen and smooth the colony pixel objects. A size filtering method based on a user-specified threshold was applied to remove the unwanted objects that surround a colony region. We experimentally determined that 9000 pixels were optimum to discard other objects, resulting in an image that depicted only the colony region. In the final step, the labeling of the largest connected component was estimated as a separate region that is considered as the contour of the iPSC colony used for computing the colony features to evaluate the characteristics of healthy and unhealthy colonies as shown in Fig. 1 and Fig. 2, respectively.

2.3. Texture feature measurements of iPSC colony

Overall, 151 texture features were estimated to determine the characteristics of the iPSC colonies as shown in Table 1. Based on the shape-based, moment-based, statistical, and spectral features, the texture features were quantitatively measured in the colony regions of the iPSCs. Briefly, the shape-based features were computed from the geometrical area of the colony within a closed boundary. These describe the appearance of the colony region, but do not consider the gray-level distribution of the enclosed boundary region. The shape-based features included area, perimeter, equivalent diameter, eccentricity, solidity, and the major and minor axes of the ellipse. The moment-based features were estimated from the gray-level values of the colony region and included local statistics (LS) and local neighborhood statistical (LNS) features. Local statistics (mean, variance, skewness, and kurtosis) were used to calculate the features based on each gray level intensity of a Gabor filter bank with four orientations (0°, 45°, 90°, and 135°) and four scales (7 × 7, 9 × 9, 11 × 11, and 13 × 13) and accordingly, 64 features were estimated. The LNS features (mean, variance, skewness, kurtosis, energy, and entropy) estimated the neighborhood intensity values that were normalized relative to the position of each pixel of the enclosed colony region, resulting in 36 normalized features. The statistical features extracted from the gray-level values of the colony object included 13 gray-level co-occurrence matrices (GLCMs) and four statistical feature matrices (SFMs) [15,16]. The GLCM features were computed in four directions (0°, 45°, 90°, and 135°) at distance d = 1. The SFM features of coarseness, contrast, period, and roughness were calculated as 4 × 4 matrices. The average and maximum values of the GLCM and SFM were calculated, and thus, 26 and 8 features, respectively were considered. The spectral features showed the periodic structures of the enclosed colony regions. They described the local maximum at the respective frequencies in the Fourier spectrum [17]. We applied the Fourier power spectrum (Fs) to divide the image intensity into two parts based on the radial distance and orientation. In addition, self-similarity of the object was indicated by computing four fractal dimension (FD) features by including all possible grid sizes (2, 4, 8, 16, 32, 64, and 128) as discussed in Ref. [18]. The average and maximum values of FDs were considered, and consequently, eight FD features were used.

2.4. Feature selection by stepwise regression algorithm

The goal of feature selection is to select effective features from the