

Investigating tunable experiment variable effects on hiPSC-CMs maturation via unsupervised learning

Shenbageshwaran Rajendiran, Mohammadjafar Hashemi, Ferdous Frinklea,
Nathan Young, Elizabeth Lipke, Selen Cremaschi*

*Department of Chemical Engineering, Auburn University, Auburn, AL 36849, USA
selen-cremaschi@auburn.edu*

Abstract

Cardiomyocytes (CMs) are heart cells responsible for heart contraction and relaxation. CMs can be derived from human induced pluripotent stem cells (hiPSCs) with high yield and purity. Mature CMs can potentially replace dead and dysfunctional cardiac tissue and be used for screening cardiac drugs and toxins. However, hiPSCs-derived CMs (hiPSC-CMs) are immature, which limits their utilization. Therefore, it is crucial to understand how experimental variables, especially tunable ones, of hiPSC expansion and differentiation phases affect the hiPSC-CM maturity stage. This study applied clustering algorithms to day 30 cardiac differentiation data to investigate if any maturity-related cell features could be related to the experimental variables. The best models were obtained using k-means and Gaussian mixture model clustering algorithms based on the evaluation metrics. They grouped the cells based on eccentricity and elongation. The cosine similarity between the clustering results and the experimental parameters revealed that the Gaussian mixture model results have strong similarities of 0.88, 0.94, and 0.93 with axial ratio, diameter, and cell concentration.

Keywords: hiPSC encapsulation, CM differentiation, hiPSC-CM maturation, clustering techniques, K Means clustering

1. Introduction

Cardiovascular disease (CVD) is the leading cause of death worldwide (Ahmed et al., 2020). A heart attack can cause the loss of more than one billion heart cells, initiating blood flow overload and overstretching on viable cardiac cells, potentially leading to death. Because the human heart has limited regenerative capacity, it cannot replace damaged cells. Engineered heart tissue is a potential alternative for heart failure due to the difficulties associated with Cardiac Implantable Electronic devices (CIEDs) and heart transplants, two available therapies (Kempf et al., 2016). Human-induced pluripotent stem cells (hiPSCs) can be differentiated into CMs with the potential to produce therapeutic CMs (Tani et al., 2022). However, CMs produced by current differentiation protocols are immature. Immature CMs differ substantially from mature ones. For example, immature CMs have underdeveloped mitochondria, limited fatty acid oxidation capacity, less elongated cells, and disorganized sarcomeres and myofibrils. Research is ongoing to develop protocols that lead to mature hiPSC-CMs (Hamledari et al., 2022). This paper applied clustering algorithms to group hiPSC-CMs based on maturity-relevant features and investigated the relationship between clustering results and tunable experimental variables of the differentiation protocol with the aim of identifying variables that produce more matured CMs. Clustering, instead of classification, was employed

because the hiPSC-CMs cannot be labeled to train a classifier. The CMs were produced through hiPSC hydrogel encapsulation and direct differentiation within a 3D-engineered tissue microenvironment (Chang et al., 2020). The unbiased data were collected from 17 different batches, with each batch having different experimental variables, such as axial ratio (AR), cell concentration, PEG-fibrinogen (PF) concentration, and microspheroid size (i.e., diameter). Eight clustering algorithms from Scikit-learn (Fabian Pedregosa) were applied to day 30 cardiac differentiation data. The maturity-relevant features considered for clustering were cell area, cell circularity, eccentricity, elongation, sarcomere length, sarcomere organization score, and orientation index. The k-means and Gaussian mixture clustering algorithms yielded the best clusters based on evaluation metrics.

2. Experimental Procedure for hiPSC-CM Production and Data Collection

2.1. HiPSC Culture, Encapsulation, and Cardiomyocyte Differentiation

Un-Arc 16 Facs II (Shinnawi et al., 2015) was cultured on Geltrex (Gibco) with E8 media. The hiPSCs were resuspended in PF precursor solution at 30-60 million cells/mL and encapsulated within PF by using a novel microfluidic system as described previously (Finklea et al., 2021). Microspheroids with different sizes and ARs were produced (Seeto et al., 2019). The D-optimal experimental design was used, assuming Gaussian process model. Microspheroids were cultured for an additional 2 days in E8 media with daily media changes (days -2 and -1). To analyze the initial size and AR of the microspheroids, the autofluorescence of the photoinitiator Eosin Y in PF was captured using the FITC filter on the Nikon Eclipse Ti fluorescence microscope at low magnification. Standard plugins in ImageJ were used for quantification. To initiate cardiac differentiation on day 0, microspheroids were transferred to chemically defined cardiac differentiation media (CDM3) supplemented with CHIR99021 (5 – 7.5 μ M, STEMCELL Technologies). Exactly 24 h later, the media was exchanged for CDM3 supplemented with 5 μ M IWP2 (STEMCELL Technologies). Fresh CDM3 was added on days 3, 5, 7, and 10; following day 10, microspheroids were cultured in RPMI/B27 (Gibco). The differentiation outcomes were assessed using flow cytometry on day 10 using the primary antibodies, such as cTnT (Invitrogen) and MF20 (DSHB), and secondary antibody (1:300, AlexaFluor 647 goat anti-mouse IgG (ThermoFisher)).

2.2. Microspheroid dissociation and replating

Microspheroids on day 30 of differentiation were dissociated by incubating in a Collagenase-B (1 mg/mL, Roche) supplemented by DNase (0.05 mg/mL, Worthington) in PBS dissociation solution at 37 °C for about 8 minutes. The cells were plated on a Matrigel-coated coverslip in RPMI 20 medium (RPMI 1640 medium with 20% FBS, Atlanta Biologicals) supplemented with 5 μ M RI for 2 days before starting staining. After fixation, the cells were washed with PBS, permeabilized with PBS-T containing 0.2% Triton X-100 in PBS for 30 minutes, and blocked with 10% FBS in PBS blocking buffer for 45 minutes at room temperature. Cells were incubated in primary antibody (aSA) overnight at 4 °C. After washing, cells were incubated in the secondary antibody (Alexa Fluor 568, 1:200), and nuclei stained with Bisbenzimidazole Hoechst 33342 (MilliporeSigma) for 1 hour at room temperature. All samples were visualized with Nikon Eclipse TE2000 Inverted Microscopes equipped with a Nikon A1 plus Confocal Microscope System. To assess hiPSC-CMs cell morphology and sarcomere structure, SarcOmere Texture Analysis (Sutcliffe et al., 2018) was used for immunofluorescent confocal images.

3. Clustering Algorithms

3.1. K-means Clustering

Given the number of clusters ($|C|$), the algorithm (Likas et al., 2003) randomly initializes $|C|$ cluster centroids ($\mu_j, j = 1, 2, \dots, |C|$) and assigns a class membership to each point ($x_i, i = 1, 2, \dots, n$) based on its closest centroid. Then, the centroids are reestimated using Eq. (1). Each point is reassigned a new class membership based on the new centroids. The steps are repeated until class memberships stop changing.

$$\sum_{i=0}^n \min_{\mu_j \in C} (||x_i - \mu_j||^2) \quad (1)$$

3.2. Gaussian Mixture Model Clustering

The Gaussian mixture model (Yang et al., 2012) assumes that all data points come from a finite number of Gaussian distributions with unknown parameters. The distribution parameters are estimated by the expectation maximization (EM) algorithm. The EM algorithm first assigns the mean and variance of the distributions either randomly or based on the centroids of k-means clustering results. The probability of a point belonging to a cluster is calculated using the distribution parameters. Cluster means and variances are improved by maximizing the likelihood of the data given those parameters, and the procedure is repeated until the mean and variance of each distribution stop changing.

3.3. Agglomerative Clustering

Agglomerative clustering (Murtagh & Legendre, 2014) is a hierarchical algorithm that uses a bottom-up approach. Each data point is initially considered a “cluster.” The algorithm proceeds by successively merging clusters using a selected linkage criterion. Criterion using ward, complete, average, and single linkage minimizes the sum of squared distances within all clusters, the maximum distance between observations of pairs of clusters, the average of the distances between all observations of pairs of clusters, and the distance between the closest observations of pairs of clusters, respectively.

4. Metrics Used for Evaluating Clustering Results

The three evaluation metrics used to assess the cluster results were Silhouette score, Calinski-Harabasz index (CH), and the Davies-Bouldin index (DB). Based on the results of these metrics, the best three cluster models were selected. Silhouette score (Shahapure & Nicholas) for a data point, s , is calculated using the distance between that point and all other points in the same and nearest clusters. The average silhouette score of all the data points gives the Silhouette score for the clusters. Silhouette score is between -1 and 1, with 1 indicating that the clusters are well separated. The CH (Maulik & Bandyopadhyay, 2002) is the ratio of the sum of dispersion between and within clusters for all clusters. The CH is higher for dense and separated clusters. The DB (Maulik & Bandyopadhyay, 2002) is defined as the average similarity between each cluster C_a and its most similar one C_b . The similarity score, R_{ab} , given in Eq. (2), is calculated between cluster a and b , d_a is the average distance between each point in cluster a and its centroid, and D_{ab} is the distance between cluster centroids a and b . The DB is calculated by Eq. (3). Values closer to zero indicate better clustering.

$$R_{ab} = \frac{d_a + d_b}{D_{ab}} \quad (2)$$

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{i \neq j} R_{ab} \quad (3)$$

5. Results and Discussion

5.1. Clustering Results

Five hundred thirteen (513) hiPSC-CMs from 17 batches were analyzed to obtain cell morphological features, such as cell area, eccentricity, circularity, elongation, and sarcomere properties, such as sarcomere length, orientation index, and organization score. Only k-means, Gaussian mixture model, and agglomerative clustering algorithms were found to be applicable to the dataset. The agglomerative clustering algorithm grouped almost all hiPSC-CMs in one cluster. These clustering results were not further analyzed. Because this study aims to investigate the relationship between mature and immature hiPSC-CMs and experimental variables, the number of clusters was set to two.

Based on the evaluation metrics, k-means clustering yielded the best model with a Silhouette score of 0.17, CH of 107, and DB of 1.99. Gaussian mixture model clustering with spherical, full, and diagonal covariance yielded the following models in that order. The evaluation metrics for these models were similar to those for the k-means algorithm. Analyzing the two clusters yielded by the k-means and Gaussian mixture model algorithms revealed that the hiPSC-CMs were clustered based on eccentricity and elongation. Mature CMs typically have an elliptical shape with larger eccentricity and elongation values (Karbassi et al., 2020).

The k-means algorithm grouped hiPSC-CMs with eccentricity values between 0.7 and 1 and elongation between 2 and 10 in one cluster, which we named the “Mature” cluster, and the remaining cells in another, we named the “Immature” cluster. The Gaussian mixture model algorithms yielded clusters with the same cell property ranges. The overall clustering results suggest that eccentricity and elongation were significant features for separating mature hiPSC-CMs from immature cells by unsupervised learning.

5.2. Relationship between hiPSC-CM Maturity and Tunable Experimental Variables

To investigate the potential relationship between hiPSC-CM maturity and tunable experimental variables, we calculated the percentage of hiPSC-CMs clustered as “Mature” in each of the 17 batches. A set of tunable experimental variables, namely AR, diameter, cell concentration PEG concentration, and PF concentration, were used in each batch. Then, the Cosine similarity (Xia et al., 2015) between this percentage and each tunable experimental variable for each clustering result was calculated. Cosine similarity is a measure of similarity between two vectors. Using cosine similarity, we quantify how similar the percentage of mature CMs to the experimental variables. The resulting similarity values are plotted in Figure 1. The PEG and PF concentration results were not considered because there were only three values for these variables. Figure 1 illustrates that the strongest similarity metric values were observed for the results obtained using the Gaussian mixture model clustering, with a cosine similarity of 0.88 with AR, 0.94 with diameter, and 0.93 with the cell concentration. The cosine similarity values for the results obtained with the k-means algorithm are similar, with a cosine similarity of 0.86 with AR, 0.92 with diameter, and 0.91 with cell concentration.

Figure 2 plots the percentage of “Mature” hiPSC-CMs in each batch calculated using Gaussian mixture model clustering results versus AR (Figure 2(a)), diameter (Figure 2(b)), and cell concentration (Figure 2(c)). Figures 2(a), 2(b), and 2(c) suggest that there is a relationship between the percentage of “Mature” hiPSC-CMs and AR, diameter, and cell concentration. Three of the five batches with ARs less than two have less than 45 % of hiPSC-CMs clustered as “Mature”. In contrast, batches with an AR greater than two have at least 45% hiPSC-CMs clustered as “Mature”, except for one. Figure 2(c) shows that when the concentration is equal to 60 million cells/mL, the percentage of mature cells is greater than 50%. These observations suggest the existence of non-linear and

potentially complex relationships between the mature hiPSC-CMs and the ARs, diameters, and cell concentration, which should be further investigated.

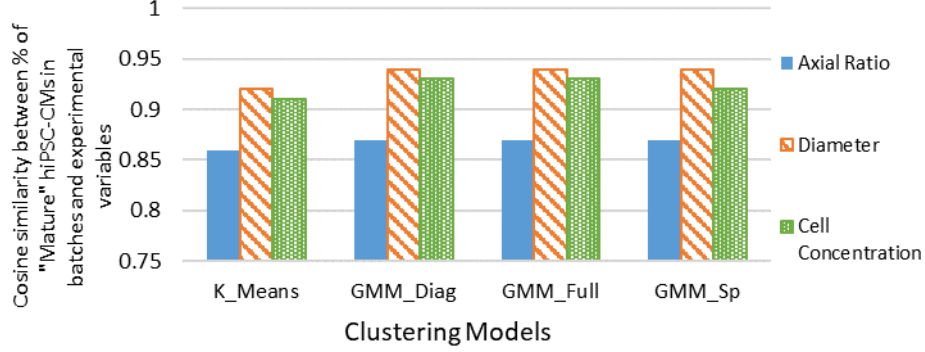


Figure 1 – Cosine similarity between the percentage of “Mature” hiPSC-CMs in batches based on the clustering results and the experimental parameters. K_Means – k-means clustering, GMM_Diag, GMM_Full, and GMM_Sp – Gaussian mixture model with diagonal, full, and spherical covariance function.

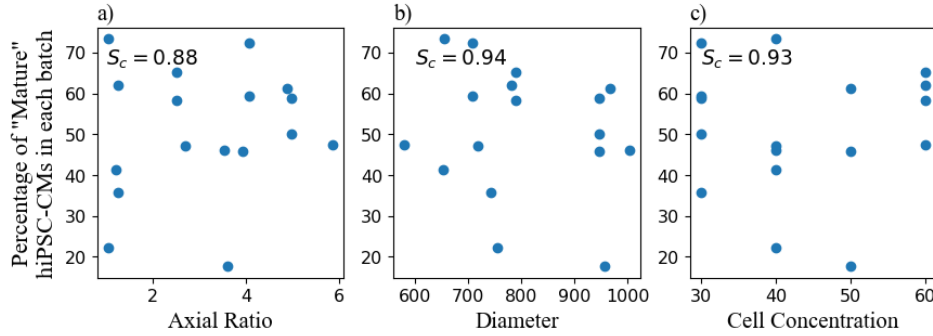


Figure 2 - Percentage of mature hiPSC-CMs clustered by k means clustering vs. (a) AR (b) Diameter (c) Cell concentration. S_c – Cosine similarity between the percentage of mature CMs and each experimental parameter.

6. Conclusions and Future Directions

Using day 30 cardiac differentiation data, we identified eccentricity and elongation as the significant features for clustering hiPSC-CMs into two groups via k-means and Gaussian mixture model clustering algorithms. The results of the Gaussian mixture and k-means models suggested that there is a concurrent relationship between ARs, diameter, and cell concentration, all experimentally tunable variables, and hiPSC-CMs with higher eccentricity and elongation, which are associated with maturity. Future work will investigate using different clustering approaches that relax the compact set assumption for determining better “Mature” cell cluster boundaries. Then, the relationships between the experimental variables and mature hiPSC-CMs will be modeled to aid in directing hiPSC differentiation towards mature CMs.

7. Acknowledgments

This work is funded by NSF grant number 2135059.

References

- Ahmed, R. E., Anzai, T., Chanthra, N., & Uosaki, H. (2020). A Brief Review of Current Maturation Methods for Human Induced Pluripotent Stem Cells-Derived Cardiomyocytes. *Front Cell Dev Biol*, 8, 178.
- Chang, S., Finklea, F., Williams, B., Hammons, H., Hodge, A., Scott, S., & Lipke, E. (2020, Jul). Emulsion-based encapsulation of pluripotent stem cells in hydrogel microspheres for cardiac differentiation. *Biotechnol Prog*, 36(4), e2986.
- Fabian Pedregosa, G. V., Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay;. *Scikit-learn: Machine Learning in Python*. <https://scikit-learn.org/stable/about.html#citing-scikit-learn>
- Finklea, F. B., Tian, Y., Kerscher, P., Seeto, W. J., Ellis, M. E., & Lipke, E. A. (2021, Jul). Engineered cardiac tissue microsphere production through direct differentiation of hydrogel-encapsulated human pluripotent stem cells. *Biomaterials*, 274, 120818.
- Hamledari, H., Asghari, P., Jayousi, F., Aguirre, A., Maaref, Y., Barszczewski, T., Ser, T., Moore, E., Wasserman, W., Klein Geltink, R., Teves, S., & Tibbits, G. F. (2022). Using human induced pluripotent stem cell-derived cardiomyocytes to understand the mechanisms driving cardiomyocyte maturation. *Front Cardiovasc Med*, 9, 967659
- Karbassi, E., Fenix, A., Marchiano, S., Muraoka, N., Nakamura, K., Yang, X., & Murry, C. E. (2020, Jun). Cardiomyocyte maturation: advances in knowledge and implications for regenerative medicine. *Nat Rev Cardiol*, 17(6), 341-359.
- Kempf, H., Andree, B., & Zweigerdt, R. (2016, Jan 15). Large-scale production of human pluripotent stem cell derived cardiomyocytes. *Adv Drug Deliv Rev*, 96, 18-30.
- Kerscher, P., Kaczmarek, J. A., Head, S. E., Ellis, M. E., Seeto, W. J., Kim, J., Bhattacharya, S., Suppiramaniam, V., & Lipke, E. A. (2017, Aug 14). Direct Production of Human Cardiac Tissues by Pluripotent Stem Cell Encapsulation in Gelatin Methacryloyl. *ACS Biomater Sci Eng*, 3(8), 1499-1509.
- Likas, A., Vlassis, N., & J. Verbeek, J. (2003). The global k-means clustering algorithm. *Pattern Recognition*, 36(2), 451-461.
- Maulik, U., & Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12), 1650-1654.
- Murtagh, F., & Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification*, 31(3), 274-295.
- Seeto, W. J., Tian, Y., Pradhan, S., Kerscher, P., & Lipke, E. A. (2019). Rapid Production of Cell-Laden Microspheres Using a Flexible Microfluidic Encapsulation Platform. *Small*, 15(47), 1902058.
- Shahapure, K. R., & Nicholas, C. (2020). Cluster Quality Analysis Using Silhouette Score.
- Shinnawi, R., Huber, I., Maizels, L., Shaheen, N., Gepstein, A., Arbel, G., Tijssen, A. J., & Gepstein, L. (2015, Oct 13). Monitoring Human-Induced Pluripotent Stem Cell-Derived Cardiomyocytes with Genetically Encoded Calcium and Voltage Fluorescent Reporters. *Stem Cell Reports*, 5(4), 582-596.
- Sutcliffe, M. D., Tan, P. M., Fernandez-Perez, A., Nam, Y. J., Munshi, N. V., & Saucerman, J. J. (2018, Jan 19). High content analysis identifies unique morphological features of reprogrammed cardiomyocytes. *Sci Rep*, 8(1), 158.
- Tani, H., Tohyama, S., Kishino, Y., Kanazawa, H., & Fukuda, K. (2022, Mar). Production of functional cardiomyocytes and cardiac tissue from human induced pluripotent stem cells for regenerative therapy. *J Mol Cell Cardiol*, 164, 83-91.
- Xia, P., Zhang, L., & Li, F. (2015). Learning similarity with cosine similarity ensemble. *Information Sciences*, 307, 39-52.
- Yang, M.-S., Lai, C.-Y., & Lin, C.-Y. (2012). A robust EM clustering algorithm for Gaussian mixture models. *Pattern Recognition*, 45(11), 3950-3961.