Motivations

- Classification of human embryonic stem-cell-derived cardiomyocytes (hESC-CMs) is relevant to reduce risks in their application to cardiac regenerative medicine and to enrich drug screen analyses.
- There is a lack of labels in hESC-CM domain because it is not clear how adult CM phenotypes are expressed in hESC-CM APs populations.
- State-of-the-art methods are computationally expensive [1].

- RNNs significantly reduce computational cost, but at the expense of accuracy [2].

Contributions

- Our goal is to classify hESC-CM APs considering the domain shift between embryonic and adult CMs when training the RNNs.
- We apply, for the first time, the concept of domain adaptation to address the domain shift between hESC-derived cells and adult cells throughout the differentiation process.
- Our approach preserves computational advantages of RNN-based approaches and it outperforms the state of the art in terms clustering quality and inter-dataset generalization.

Problem Formulation

- Set of unlabeled hESC-CM APs \( \Omega = \{x_i^e\}_i \): each \( x_i^e \) is a time-series of length \( K \).
- Set of labeled adult CM APs \( \Omega_y = \{x_j^y\}_j \); each \( x_j^y \) is a time-series of length \( K \) labeled as atrial (\( y = 0 \)) or ventricular (\( y = 1 \)).
- Problem: Assign a label \( y' \) to a new \( x^y \), where \( y' = 0 \) denotes atrial-like and \( y' = 1 \) denotes ventricular-like.

Assumptions:

- \( \mathbb{P}(x | \text{embryonic}) \neq \mathbb{P}(x | \text{adult}) \).
- \( \mathbb{P}(y | x, \text{embryonic}) = \mathbb{P}(y | x, \text{adult}) \) (covariate shift assumption).

Network Architecture

- RNN with LSTM units
- Feature extractor (LSTM)

\[ x \rightarrow \varphi_{\varphi}(x) = h(x, K) \]

- Output layer (Sigmoid)

\[ y = \varphi(y, \varphi)(x^y W + b) \]

where \( \varphi(z) = \frac{1}{1 + e^{-z}} \).

Loss Function

To address lack of labeled hESC-CM data and domain shift we proposed a semi-supervised learning approach whose loss consists of three terms.

\[ \frac{1}{N} \sum_{i=1}^{N} \ell_{\text{c}}(\hat{y}_i, y_i^y) + \frac{\lambda}{N_L} \sum_{i=1}^{N_L} \ell_{\text{c}}(\hat{y}_i, y_i^y) + \alpha \text{MMD}^2(\{x_i^e, y_i\}_i, \{x_i^y, y_i\}_i) \]

Supervised term: binary cross-entropy for classifying adult CMs.

\[ \ell_{\text{c}}(\hat{y}_i, y_i^y) = - \log(\hat{y}_i) + \log(1 - \hat{y}_i) \]

Unsupervised term: contrastive loss on hESC-CM APs that encourages similar predictions for similar embryonic CMs.

\[ \ell_{\text{c}}(\hat{y}_i, y_i^e) = s_{x_i^e, x_j^e} \cdot \ell_{\text{c}}(\hat{y}_i, y_i^e) + (1 - s_{x_i^e, x_j^e}) \cdot \ell_{\text{c}}(1(\hat{y}_i), 1(y_i^e)) \]

where \( s_{x_i^e, x_j^e} = \exp(-\|x_i^e - x_j^e\|^2) \in [0, 1] \) represents the similarity between \( x_i^e \) and \( x_j^e \).

Domain adaptation term: Maximum Mean Discrepancy [5] between both domains

\[ \alpha \text{MMD}^2(\Omega, \Omega_y) = \sum_{i=1}^{N} \sum_{j=1}^{N} K(x_i^e, x_j^y) + \sum_{i=1}^{N} \sum_{j=1}^{N} K(x_i^y, x_j^y) - 2 \sum_{i=1}^{N} \sum_{j=1}^{N} K(x_i^e, x_j^y) \]

with \( K(x_i^e, x_j^y) = \exp(-\frac{\|x_i^e - x_j^y\|^2}{\sigma^2}) \) Gaussian kernel.

Clustering Quality Index

- The Davies-Bouldin Index (DBI) between two clusters \( \Omega = \{x_1^e\} \mid y_1^e < 0.5 \) and \( \Omega = \{x_2^e\} \mid y_2^e > 0.5 \) as the ratio between intra-cluster dispersion and distance between clusters.

\[ \text{DBI} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} \|x_i^e - x_j^e\|^2}{\sum_{i=1}^{N} \|x_i^e - x_{\text{mean}}^e\|^2} \]

where \( S_P \) is the mean distance from elements of class \( y \) to the average signal of the same class, and \( M_P \) is the distance between the average signals of both classes.

Experiments

- Dataset:
  - Adult CM APs: 1600 synthetic adult CMs generated using the O’Hara-Rudy ventricular model (ORd) [6] and the Nygren atrial model [7].
  - Unlabeled hESC-CM APs: 9490 hESC-CM APs obtained from 9 cell aggregates [3].
  - Labeled hESC-CM APs: 52 hESC-CM APs obtained from single cell recordings [4].

- Implementation Details:
  - Keras with TensorFlow backend. RMSProp optimizer, batches of 3 adult APs and 16 hESC-CM APs (100 batches, reported 10-fold crossvalidation performance).

- Baselines:
  - I-NN-Metamorphosis. 1-Nearest-Neighbor method with metamorphosis distances [1].
  - Sup-LSTM. Supervised learning \( \lambda = 0 \) and \( \gamma = 0 \).
  - Semi-M-LSTM. Semi-supervised learning \( \lambda = 0.1 \) (metamorphosis) and domain adaptation \( \gamma = 0.2 \).

- Two cases studied:
  - DA-Sup-LSTM. Supervised learning \( \lambda = 0 \) with domain adaptation \( \gamma = 1 \).
  - DA-Semi-M-LSTM. Semi-supervised learning \( \lambda = 0.1 \) (metamorphosis) and domain adaptation \( \gamma = 5 \).

Metrics:

- Clustering quality (DBI) measured in the unlabeled hESC-CM APs dataset, and accuracy measured in the labeled hESC-CM APs dataset (no retraining).

Conclusion

- Domain adaptation concepts have been shown to be useful in the context of hESC-CMs.
- The proposed method outperforms the state of the art not only in terms of clustering quality (median DBI 0.2197 vs 0.2297), but also in terms of computational efficiency (inference time of 0.4 secs vs 12 secs with comparable resources) and inter-dataset generalization (median Acc 99.04% vs 96.15%).

Acknowledgments:

The authors thank Dr. Giann Gorose for insightful discussions, and Dr. Renjun Zhu and Prof. Leslie Tung for providing the hESC-CMs dataset.

References


