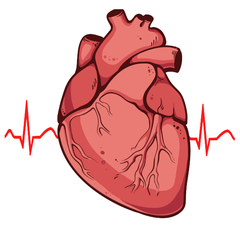
An Unsupervised Domain Adaptation Approach to Classification of Stem Cell-Derived Cardiomyocytes

Carolina Pacheco and René Vidal

Center for Imaging Science, Mathematical Institute for Data Science, Department of Biomedical Engineering, Johns Hopkins University, Baltimore, USA.

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].

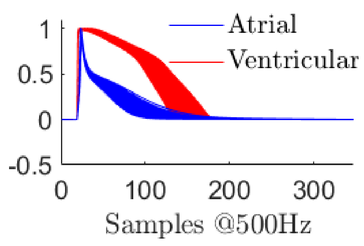


Fig. 1: APs from adult CMs (synthetic data)

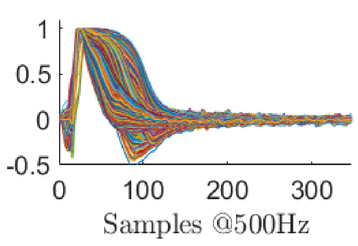


Fig. 2: APs from hESC-CMs (unlabeled dataset [3])

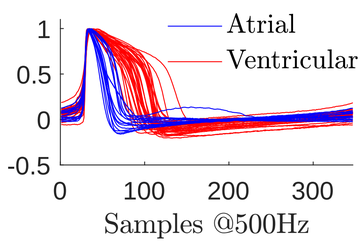


Fig. 3: APs from hESC-CMs (labeled dataset [4])

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 $\Omega_e = \{\mathbf{x}_j^e\}_{j=1}^{N_e}$: each $\mathbf{x}_j^e$ is a time-series of length K.
- Set of labeled adult CM APs $\Omega_a = \{\mathbf{x}_i^a, y_i^a\}_{i=1}^{N_a}$: each $\mathbf{x}_i^a$ is a time-series of length K labeled as atrial ($y=0$) or ventricular ($y=1$).
- Problem: Assign a label $\hat{y}^e$ to a new $\mathbf{x}^e$, where $\hat{y}^e = 0$ denotes atrial-like and $\hat{y}^e = 1$ denotes ventricular-like.
- Assumptions:
  - $\mathbb{P}\{\mathbf{x} \mid embryonic\} \neq \mathbb{P}\{\mathbf{x} \mid adult\}$.
  - $\mathbb{P}\{y \mid \mathbf{x}, embryonic\} = \mathbb{P}\{y \mid \mathbf{x}, adult\}$ (covariate shift assumption).

Network Architecture

- RNN with LSTM units
  - Feature extractor (LSTM) $\mathbf{x} \mapsto \varphi_{\mathcal{W}_\mathcal{F}}(\mathbf{x}) = h(\mathbf{x}, K)$.
  - Output layer (Sigmoid) $\varphi_{\mathcal{W}_\mathcal{F}}(\mathbf{x}) \mapsto \hat{y}$, $\hat{y} = \sigma(\varphi_{\mathcal{W}_\mathcal{F}}(\mathbf{x})^T W_c + b_c)$, where $\sigma(z) = \frac{1}{1+e^{-z}}$.

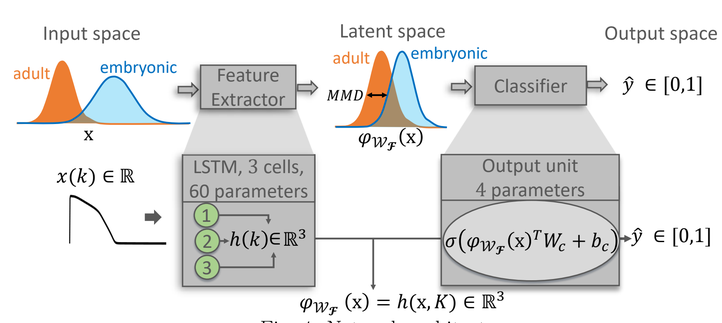


Fig. 4: Network architecture

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.

$$\frac{1-\lambda}{N_a}\left(\sum_{i=1}^{N_a} \ell_s\left(y_i^a, f_\mathcal{W}(\mathbf{x}_i^a)\right)\right) + \frac{\lambda}{N_e\left(N_e-1\right)}\left(\sum_{j=1}^{N_e}\sum_{j'\neq j} \ell_u\left(f_\mathcal{W}(\mathbf{x}_j^e), f_\mathcal{W}(\mathbf{x}_{j'}^e)\right)\right) + \gamma\widehat{\mathcal{MMD}}^2\left(\{\varphi_{\mathcal{W}_\mathcal{F}}(\mathbf{x}_i^a)\}_{i=1}^{N_a}, \{\varphi_{\mathcal{W}_\mathcal{F}}(\mathbf{x}_j^e)\}_{j=1}^{N_e}\right)$$

- Supervised term: binary crossentropy for classifying adult CMs

$$\ell_s(y_i^a, \hat{y}_i^a) = -y_i^a \log(\hat{y}_i^a) - (1-y_i^a)\log(1-\hat{y}_i^a).$$

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

$$\ell_u(\hat{y}_j^e, \hat{y}_{j'}^e) = s_{(j,j')} \cdot \ell_s(\hat{y}_j^e, \hat{y}_{j'}^e) + \left(1 - s_{(j,j')}\right) \cdot \ell_s((1-\hat{y}_j^e), \hat{y}_{j'}^e),$$

where $s_{(j,j')} = \exp\left(-\frac{d^4(\mathbf{x}_j^e, \mathbf{x}_{j'}^e)}{\sigma_s^4}\right) \in [0,1]$ represents the similarity between $\mathbf{x}_j^e$ and $\mathbf{x}_{j'}^e$

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

$$\widehat{\mathcal{MMD}}^2(\Omega_a, \Omega_e) = \sum_{i=1}^{N_a}\sum_{i'=1}^{N_a}\frac{\mathcal{K}\left(\mathbf{x}_i^a, \mathbf{x}_{i'}^a\right)}{N_a^2} + \sum_{j=1}^{N_e}\sum_{j'=1}^{N_e}\frac{\mathcal{K}\left(\mathbf{x}_j^e, \mathbf{x}_{j'}^e\right)}{N_e^2} - \sum_{i=1}^{N_a}\sum_{j=1}^{N_e}\frac{2\mathcal{K}\left(\mathbf{x}_i^a, \mathbf{x}_j^e\right)}{N_a N_e},$$

with $\mathcal{K}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$ Gaussian kernel.

Clustering Quality Index

- The Davies-Bouldin Index (DBI) between two clusters $\Omega_0 = \left\{\mathbf{x}_j^e \mid \hat{y}_j^e < 0.5\right\}$ and $\Omega_1 = \left\{\mathbf{x}_j^e \mid \hat{y}_j^e \geq 0.5\right\}$ is the ratio between intra-cluster dispersion and distance between clusters

$$DBI\left(\Omega_0, \Omega_1\right) = \frac{S_0 + S_1}{M_{01}},$$

where S_y is the mean distance from elements of class y to the average signal of the same class, and M_{01} is the distance between the average signals of both classes.

Experiments

- Dataset
  - Adult CM APs: 1600 synthetic adult APs generated using the O'hara-Rudy ventricular model (ORd) [6] and the Nygren atrial model [7].
  - Unlabeled hESC-CM APs: 6940 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
  - 1NN-Metamorphosis. 1-Nearest-Neighbor method with metamorphosis distances [1].
  - Sup-LSTM. Supervised learning $\lambda = 0$ and $\gamma = 0$ [2].
  - Semi-M-LSTM. Semi-supervised learning $\lambda = 0.1$ (metamorphosis) and $\gamma = 0$ [2].
- Results

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).

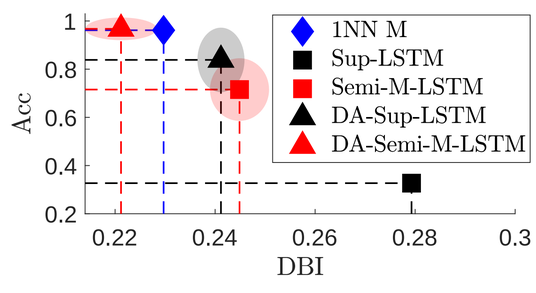


Fig. 5: Acc vs DBI. Mean performance is marked by solid-colored symbols and variability is shown by translucent ellipses.

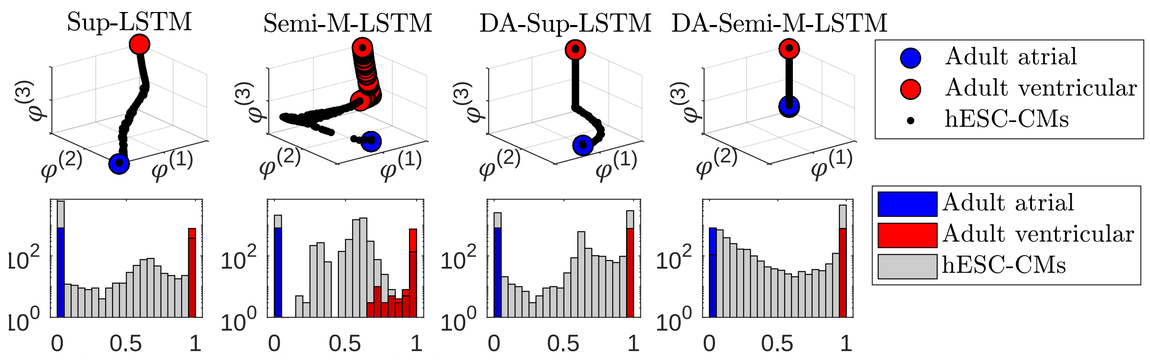


Fig. 6: Distribution of adult CMs and hESC-CMs in latent space (first row) and histogram of their projection on the one dimensional path (second row)

Conclusion

- Domain adaptation concepts have 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 Gorospe for insightful discussions, and Dr. Renjun Zhu and Prof. Leslie Tung for providing the hESC-CMs dataset.

References

[1] G. Gorospe, R. Zhu, J.-Q. He, L. Tung, L. Younes, and R. Vidal. Efficient metamorphosis computation for classifying embryonic cardiac action potentials. In *5th Workshop on Mathematical Foundations of Computational Anatomy*, 2015.

[2] C. Pacheco and R. Vidal. Recurrent Neural Networks for Classifying Human Embryonic Stem Cell-Derived Cardiomyocytes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, p.581-589, 2018.

[3] R. Zhu, M.A. Millrod, E.T. Zambidis, and L. Tung. Variability of action potentials within and among cardiac cell clusters derived from human embryonic stem cells. *Scientific Reports*, vol. 6, p.18544, 2016.

[4] J.Q He, Y. Ma, Y. Lee, J. Thomson and T. Kamp. Human Embryonic Stem Cells Develop into Multiple Types of Cardiac Myocytes: Action Potential Characterization *Circulation research*, vol. 93, p.32-39, 2003.

[5] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf and A. Smola . A Kernel Method for the Two-Sample-Problem. Advances in neural information processing systems, pp. 513-520, 2007.

[6] T. O'Hara, L. Virág, A. Varró, and Y. Rudy. Simulation of the undiseased human cardiac ventricular action potential: model formulation and experimental validation. *PLoS Computational Biology*, vol. 7, no. 5, p. e1002061, 2011.

[7] A. Nygren, C. Fiset, L. Firek, J. Clark, D. Lindblad, R. Clark, and W. Giles. Mathematical model of an adult human atrial cell: the role of k+ currents in repolarization. *Circulation Research*, vol. 82, no. 1, pp. 63-81, 1998.