PROCEEDINGS OF THE 30th European Symposium on Computer Aided Process Engineering (ESCAPE30), May 24-27, 2020, Milano, Italy

Data-Driven Model Development for Cardiomyocyte Production Experimental Failure Prediction

Bianca Williams[a], Caroline Halloin[b], Wiebke Löbel[b],Ferdous Finklea[a], Elizabeth Lipke[a], Robert Zweigerdt[b], Selen Cremaschi[a]*

[a]*Auburn University, Department of Chemical Engineering, Auburn, AL, USA*
[b]*Leibniz Research Laboratories for Biotechnology and Artificial Organs (LEBAO), Department of Cardiac, Thoracic, Transplantation, and Vascular Surgery, Hannover Medical School, 30625 Hannover, Germany*
selen-cremaschi@auburn.edu

Abstract

Cardiovascular diseases (CVD) are the leading cause of death worldwide. Engineered heart tissue produced by differentiation of human induced pluripotent stem cells may provide an encompassing treatment for heart failure due to CVD. However, considerable difficulties exist in producing the large number of cardiomyocytes needed for therapeutic purposes through differentiation protocols. Data-driven modeling with machine learning techniques has the potential to identify factors that significantly affect the outcomes of these differentiation experiments. Using data from previous cardiac differentiation experiments, we have developed data-driven modeling methods for determining which experimental conditions are most influential on the final cardiomyocyte content of a differentiation experiment. With those identified conditions, we were able to build classification models that can predict whether an experiment will have a sufficient cardiomyocyte content to continue with the experiment on the seventh (out of 10) day of the differentiation with a 90% accuracy. This early failure prediction will provide cost and time savings, as each day the differentiation continues requires significant resources.

Keywords: cardiac differentiation, machine learning, random forests

1. Introduction

Cardiovascular diseases (CVD) are the leading cause of death worldwide, meaning there are more deaths annually due to CVD than any other cause. These diseases can lead to heart attacks, which can result in the loss of more than one billion heart cells, leading to congestive heart failure (Kempf, Andree, et al., 2016). Patients who suffer from advanced stages of heart failure have a poor prognosis for survival, and the large disparity between numbers of donors and recipients leaves few viable treatments. Artificial prosthetic hearts and heart assist devices have demonstrated some success in prolonging the lives of patients receiving treatment, but their development is slow and clinical trials have seen limited. Due to the nature of heart transplants and the stigma surrounding artificial organs, engineered heart tissue may provide an encompassing treatment for heart failure (Kempf, Andree, et al., 2016).

Mature cardiomyocytes, the contracting cells in the heart, are some of the least regenerative cells in the body. This characteristic carries over into the laboratory environment and thus limits in vitro expansion capabilities of cardiomyocytes.

Difficulties in direct culture of cardiomyocytes can be overcome by differentiation from human pluripotent stem cells (hiPSCs) (Kempf, Andree, et al., 2016). The indefinite turnover potential of pluripotent cells allows for the expansion of large quantities for differentiation into therapeutic engineered tissues. However, the differentiation of hiPSCs into specific cell types is a highly complex and costly process that is sensitive to the impact of a high number of factors (Gaspari et al., 2018), and significant difficulties exist in reliably and consistently producing the large number of cardiomyocytes needed for therapeutic purposes (Kempf, Andree, et al., 2016).

Data-driven modeling with machine learning techniques has the potential to identify factors and patterns that most significantly affect the outcomes of these differentiation experiments. Previously, machine learning techniques have successfully been used to identify key factors and assist in optimization for production of several proteins and cell lines (Sokolov et al., 2017; Zhou et al., 2018). The goal of this work is to use machine learning techniques to identify key process parameters to be used in predictive modeling of bioreactor cardiac differentiation outcomes. The high number of experimental factors that influence the differentiation results in a large set of possible inputs to be considered for modeling. This high data dimensionality, in addition to the low number of data points due to the time-consuming nature of these experiments, represent significant challenges for modeling the differentiation process. Our aim is to use machine learning models to predict whether or not the cardiomyocyte content at the end of differentiation process will be sufficiently high. We define insufficient production as having a cardiomyocyte content on the tenth day of differentiation (dd10) that is less than 90%, meaning less than 90% of the cells produced at the end of the differentiation are cardiomyocytes. Predicting if the cardiomyocycte content will be insufficient before the end of the differentiation will provide cost and time savings, as each day the differentiation continues requires significant resources.

Using existing data from bioreactor experiments, we have applied feature selection techniques, including correlations, principal component analysis, and built-in feature selection in machine learning models, to identify the conditions in the bioreactor, which we define as bioreactor features, are the most influential on and predictive of the cardiomyocyte content. Bioreactor features considered include values related to the cell concentration, size of cell aggregates, pH, dissolved oxygen concentration, and concentrations and timings of certain nutrients, such as glucose, and small molecules known to direct the differentiation. We then used the identified features as inputs to build models to classify the resulting cardiomyocyte content of a particular bioreactor run as being sufficient or insufficient to justify continuing with the differentiation.

2. Machine Learning Techniques for Cardiomyocyte Content Prediction

2.1. Multivariate Adaptive Regression Splines (MARS)

Multivariate adaptive regression spline (MARS) models are made up of a linear summation of basis functions. The three types of possible basis functions are a constant, a hinge function (or “spline”), or a product of two or more hinge functions. The training of a MARS model starts with an initial model that is a constant value equal to the mean of the data outputs. On its initial training pass, the model is overfit to the data using a greedy algorithm, adding basis functions to reduce the sum of the squared errors (SSE) between the given and predicted outputs. Then, a backward, pruning pass is performed to remove terms that have little effect on the SSE until the best model is identified based on generalized cross validation (GCV) criteria (Friedman, 1991). In order to make

cardiomyocyte content classifications, MARS models were trained to predict the value of the cardiomyocyte content using the selected bioreactor features as inputs, and a classification was assigned based on the predicted value.

2.2. Random Forests (RF)

Random forest (RF) models are machine learning models that make output predictions by combining outcomes from a sequence of regression decision trees. Each tree is constructed independently and depends on a random vector sampled from the input data, with all the trees in the forest having the same distribution. The predictions from the forests are averaged using bootstrap aggregation and random feature selection. RF models have been demonstrated to be robust predictors for both small sample sizes and high dimensional data (Biau & Scornet, 2016). RF classification models were constructed that directly classified bioreactor runs as having sufficient or insufficient cardiomyocyte content.

2.3. Gaussian Process Regression (GPR)

Gaussian process regression (GPR) is a method of interpolation where interpolated values are modeled by a Gaussian process governed by prior covariances. Under suitable assumptions on the priors, GPR gives the best linear unbiased prediction of the intermediate values (Rasmussen & Williams, 2005). It uses a kernel function as measure of similarity between points to predict the value for an unseen point from the training data. This method has been successfully used with small dataset sizes. In order to make cardiomyocyte content classifications, GPR models trained were similarly to the MARS models.

3. Data Collection and Feature Selection Methods

3.1. Experimental Data

Experimental data was generated and collected from 58 cardiac differentiation experiments performed by (Halloin et al., 2019). The differentiation experiments were carried out in chemically defined conditions in stirred tank bioreactors. Details of the experiments are described in Halloin et al. (2019). The set of independent variables include experimental conditions such as the rotation speed in the bioreactor and measurements such as differentiation day dependent cell densities and aggregate sizes, and continuous time measurements of dissolved oxygen (DO) concentration and pH. The set of independent variables measured from the experiments was expanded to include engineered features such as estimated gradients in cell densities and DO concentrations, resulting in a total of 101 variables, which we refer to as bioreactor features. The dependent variable is the percentage of the cells in the bioreactor that have differentiated into cardiomyocytes, or the cardiomyocyte content, on the last day of the differentiation experiment, dd10. Data from 42 of the experiments was designated as training data and used for feature selection and classification model construction. The remaining experiments were reserved as test data for testing the classification models.

3.2. Feature Selection Methods

We performed feature selection using the training data set in order to discover which of the bioreactor features were most influential on the cardiomyocyte content. The set of features considered consists of all the collected bioreactor features measured up until the seventh day of differentiation (dd7).

3.2.1. Correlations

The Pearson and Spearman correlations (Bonett & Wright, 2000) between the collected bioreactor features and the cardiomyocyte content were calculated. The Pearson

correlation measures the strength of the linear relationship between two variables. It has a value between -1 to 1, with a value of -1 meaning a total negative linear correlation, 0 being no correlation, and +1 meaning a total positive correlation. The Spearman correlation measures the strength of a monotonic relationship between two variables with the same scaling as the Pearson correlation.

3.2.2. Principal Component Analysis

Principal component analysis (PCA) converts a set of possibly correlated variables into a set of linearly uncorrelated ones through an orthogonal transformation (Hotelling, 1933). The resulting principal components (PCs) are linear combinations of the original set of variables.

3.2.3. Machine Learning Technique Built-In Feature Selection

Each of the machine learning techniques applied has its own method for selecting features and ranking their predictive importance. During the MARS model construction, a pruning pass is performed over the model that removes terms and features based on the level of their effect on GCV criteria. For RF models, features are selected based on how well they improve the separation of the data at each decision node. GPR selects features using its built-in automatic relevance determination method.

4. Classification Performance Metrics

The metrics used to evaluate the performance of the classification models (i.e., the classification of insufficient/sufficient cardiomyocyte content) are accuracy, precision, (Sokolova & Lapalme, 2009), and the Matthews correlation coefficient (MCC) (Matthews, 1975). The accuracy is the proportion of the classifications made by the models that were correct. Given that the classification model predicts an insufficient cardiomyocyte content for a bioreactor run, precision is the probability that the cardiomyocyte content of that run will actually be insufficient. The MCC is the correlation between actual and predicted classifications. It has the same range and scale of the Pearson and Spearman correlations. Figure 1 depicts the workflow of the process taken to construct the models and calculate the performance metrics.

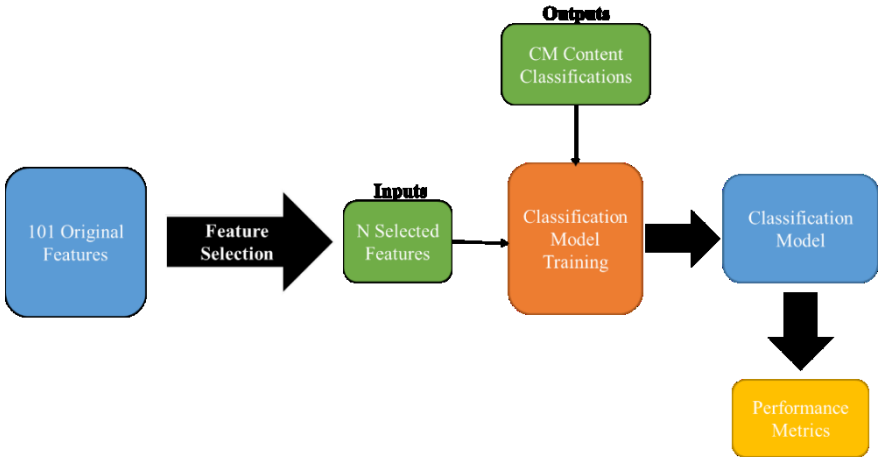

Figure 1-Feature selection and model training process (CM = cardiomyocyte)

5. Results and Discussion

5.1. Feature Selection Results

PCA of the collected feature set yielded five principal components that explained 94% of the variance in the input data. Correlations and PCA did not yield any results for significant features, with the strongest linear correlation between a feature and the cardiomyocyte content being -0.51, with the time that the IWP2 molecule remained in the bioreactor. The strongest linear correlation between the PCs and the cardiomyocyte content was 0.16. However, we had success in reducing the feature set using the built-in feature selection methods of each of the machine learning approaches investigated. From the original 101 features, MARS, RF and GPR identified 12, 12, and 7 significant features, respectively. Common features that were selected as significant include the cell densities and their gradients during the first two days of the differentiation protocol (dd0 and dd1). This selection agrees with previous experimental studies concluding that cell density during early differentiation influences differentiation into specific cell lineages (Kempf, Olmer, et al., 2016).

5.2. Classification Model Results

Results for classification model performance are summarized in Tables 1 and 2. The performance metrics in Table 1 were calculated using the leave one out (LOO) cross validation (Wong, 2015) on the training data. Two classification models were trained for each method. One model utilized the bioreactor features selected by the built-in feature selection as the inputs, and the other employed the PCs obtained from PCA as the inputs.

Table 1 – Performance of classification models on training data evaluated using LOO cross validation

	MARS		**RF**		**GPR**	
	Features	**PCA**	**Features**	**PCA**	**Features**	**PCA**
Accuracy	0.74	0.64	**0.90**	0.74	**0.90**	0.67
Precision	0.81	0.66	**0.90**	0.74	**0.93**	0.67
MCC	0.55	-0.11	**0.78**	0.36	**0.79**	0

For all of the machine learning techniques tested, the classification models using the model-selected features yielded better performance (Table 1). This suggests that while the principal components successfully explain the variance in the data, they fail to accurately characterize the relationship between the features and the cardiomyocyte content. RF models and GPR had similar performance with an accuracy and precision both of about 90%, while MARS models did not perform as accurately.

Table 2 – Performance of classification models on test data

	RF	**GPR**
Accuracy	0.89	0.89
Precision	0.92	0.87
MCC	0.72	0.72

The performances of the RF and GPR classification models trained using the model-selected features were evaluated on the test data (Table 2). Both classification models performed comparably for the test data with an accuracy of 89%, precisions near 90%, and MCC values of 0.72. The results obtained for the test data are comparable to those obtained from LOO cross validation on the training data, indicating that the models

accurately captured the relationship between the features and the cardiomyocyte content, while avoiding overfitting.

6. Conclusions and Future Directions

Using existing data from previously conducted cardiac differentiation experiments, we were able to identify on dd7 if an experiment would have an insufficient final cardiomyocyte content of less than 90% with accuracy and precision of about 90% with both RF and GPR models. We were able to make these predictions using less than 16% of the collected features. Future work with this data will focus on predicting the experimental outcomes at earlier timepoints in the differentiation. This modeling will enable the early interruption of failing experiments, providing cost and time savings.

7. Acknowledgements

This work was partially funded by Department of Education GAANN grant #P200A150075 and NSF grant #1743445.

8. References

Biau, G., & Scornet, E. (2016). Rejoinder on: A random forest guided tour. Test, 25, 264-268.

Bonett, D. G., & Wright, T. A. (2000). Sample size requirements for estimating Pearson, Kendall and Spearman correlations. Psychometrika, 65, 23-28.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines - Rejoinder. Annals of Statistics, 19, 123-141.

Gaspari, E., Franke, A., Robles-Diaz, D., Zweigerdt, R., Roeder, I., Zerjatke, T., & Kempf, H. (2018). Paracrine mechanisms in early differentiation of human pluripotent stem cells: Insights from a mathematical model. Stem Cell Res, 32, 1-7.

Halloin, C., Schwanke, K., Lobel, W., Franke, A., Szepes, M., Biswanath, S., Wunderlich, S., Merkert, S., Weber, N., Osten, F., de la Roche, J., Polten, F., Wollert, K., Kraft, T., Fischer, M., Martin, U., Gruh, I., Kempf, H., & Zweigerdt, R. (2019). Continuous WNT Control Enables Advanced hPSC Cardiac Processing and Prognostic Surface Marker Identification in Chemically Defined Suspension Culture. Stem Cell Reports.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24, 417-441.

Kempf, H., Andree, B., & Zweigerdt, R. (2016). Large-scale production of human pluripotent stem cell derived cardiomyocytes. Advanced Drug Delivery Reviews, 96, 18-30.

Kempf, H., Olmer, R., Haase, A., Franke, A., Bolesani, E., Schwanke, K., Robles-Diaz, D., Coffee, M., Gohring, G., Drager, G., Potz, O., Joos, T., Martinez-Hackert, E., Haverich, A., Buettner, F. F. R., Martin, U., & Zweigerdt, R. (2016). Bulk cell density and Wnt/TGFbeta signalling regulate mesendodermal patterning of human pluripotent stem cells. Nature Communications, 7.

Matthews, B. W. (1975). Comparison of Predicted and Observed Secondary Structure of T4 Phage Lysozyme. Biochimica Et Biophysica Acta, 405, 442-451.

Rasmussen, C. E., & Williams, C. K. I. (2005). Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning, 1-247.

Sokolov, M., Ritscher, J., MacKinnon, N., Souquet, J., Broly, H., Morbidelli, M., & Butte, A. (2017). Enhanced process understanding and multivariate prediction of the relationship between cell culture process and monoclonal antibody quality. Biotechnol Prog, 33, 1368-1380.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45, 427-437.

Wong, T. T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recognition, 48, 2839-2846.

Zhou, Y., Li, G., Dong, J., Xing, X. H., Dai, J., & Zhang, C. (2018). MiYA, an efficient machine-learning workflow in conjunction with the YeastFab assembly strategy for combinatorial optimization of heterologous metabolic pathways in Saccharomyces cerevisiae. Metab Eng, 47, 294-302.