Predicting Transfection Rates of Poly(β-amino ester) Compounds via Machine Learning Methods

Authors: Albert Lee[1], Susan Garwood[2], Brandon Walker[3], Aaron Tasset[3], Pengyu Ren[3], Huiliang Wang[3]

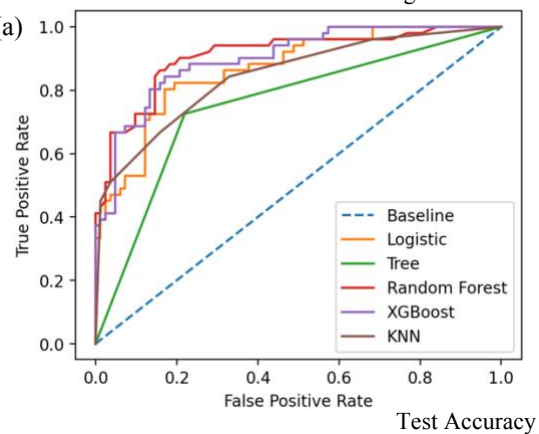
[1]Dept of Biomedical Engineering, Johns Hopkins University, Baltimore MD. [2]Dept of Electrical Engineering, Mississippi State University, Starkville MS. [3]Dept of Biomedical Engineering, University of Texas, Austin TX.

Introduction: Genetic defects occur in nearly 10 percent of all adults. In many cases, these defects become diseases such as inherited disorders, viral infections, and cancer. Gene therapies have emerged to combat these human genetic diseases. These gene therapies are a method to insert an adjustable gene into the host's genome to cure genetic diseases. To perform gene therapy, there are two primary gene delivery systems – viral and nonviral. Although viral methods are more effective, they pose production and toxicity concerns. As such, developing nonviral gene delivery systems that can effectively transfect into the cell is a frontier of gene therapies today. *In silico* models can help curb the high costs of *in vitro* experiments and provide insight into the transfection efficacy of novel poly(β-amino ester) compounds. This project, thus, aims to inform the development of a promising nonviral gene delivery system – poly(β-amino ester) compounds – using machine learning (ML) models.

Materials and Methods: Anderson et al. (2005) gathered transfection data on 443 poly(β-amino ester) compounds. We recreated each polymer compound *in silico* via SMARTS reactions. Once an *in silico* representation of all of the poly(β-amino ester) compounds had been generated, we extracted 175 chemical descriptors from each compound producing a 443 by 175 dataset. This dataset was used to predict a binary classifier with a '1' representing strong transfection rates and a '0' representing poor transfection rates. Then, five machine learning models were created: logistic regression, decision tree, random forest, XGBoost, and k-nearest neighbors. The resulting accuracies and chemical descriptors were ranked for effectiveness and informativity.

Results and Discussion:

(a) ROC Curve of Machine Learning Models



(c) Ranked Chemical Descriptors

Rank	Descriptor
1	'PEOEVSA11'
2	'S28'
3	'Smax45'
4	'PEOEVSA12'
5	'slogPVSA11'
6	'S53'

(b) Test Accuracy of Machine Learning Models

Classifier	Logistic Regression	Decision Tree	Random Forest	XGBoost	K-Nearest Neighbors
Accuracy	80.45%	72.93%	84.21%	81.95%	77.44%

Figure 1. Machine Learning Model Performance Metrics.

a Receiver operating characteristic (ROC) curve showing model performance via true positive and false positive rates

b Accuracy scores for each machine learning model

c Ranking the most informative chemical descriptors in predicting transfection rates for the best model

In predicting transfection efficacy from chemical descriptors of poly(β-amino ester) compounds, the random forest classifier proved to be the best ML model with an accuracy of 84.21% (Figure 1b). Within the random forest classifier model, the most informative chemical descriptors were 'PEOEVSA11', 'S28', and 'Smax45' which correspond to measures of partial charges, surface area, and electro-topological state index (Figure 1c).

Conclusions: This work demonstrates that the best performing random forest classifier can predict whether a novel poly(β-amino ester) compound will be effective 84.21% of the time. Further, the descriptors of interest to optimize are partial charges, surface area, and electro-topological states to achieve highest transfection rates. Future steps would involve more ML models and more data on other literature-scraped polymeric compounds.

References: Anderson, D. G., Akinc, A., Hossain, N., & Langer, R. (2005). Structure/property studies of polymeric gene delivery using a library of poly (β-amino esters). Molecular Therapy, 11(3), 426-434.