Explorations of Machine Learning Methodologies to Enhance the Design of RNA-based Dopamine Biosensors

James Craven and Matthew Lindsey

Friday, June 28th 2024

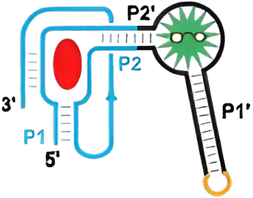
Benefits of a Dopamine Sensor

Dopamine is a neurotransmitter and plays a role in a variety of functions such as memory, learning, and reward systems. Detecting dopamine levels could help with diagnosing:

- addiction
- mental illness
- neurodegenerative disorders

Building a Synthetic Ribosensor

Research Question: How can we leverage machine learning to inform the necessary nucleotide sequence for the designed ribosensor in order for the ribosensor to emit a strong fluorescence in the presence of dopamine?



. . . GUCCA GCUGC GGAAGAAACUGUGGCACUUCGGUGCCAG GCAGC UUGU. . .

Data Understanding for ML

In order to begin building such a model, we first need to understand:

- is our data labeled?
- how much data do we have?

Data Understanding for ML

In order to begin building such a model, we first need to understand:

- is our data labeled?
- how much data do we have?

The Fernandez lab supplied us with roughly 15 labeled instances of ribosensor data, along with labeled data from a literature review of similarly built ribosensors, for total a dataset of 103 instances.

Data Understanding for ML

In order to begin building such a model, we first need to understand:

- is our data labeled?
- how much data do we have?

The Fernandez lab supplied us with roughly 15 labeled instances of ribosensor data, along with labeled data from a literature review of similarly built ribosensors, for total a dataset of 103 instances.

While this seems like a healthy amount of experimental data, many machine learning models need 10-100x that amount to discover relationships between the features and target of a dataset.

Previous Work

In an effort to better understand how machine learning is currently used in synthetic biology, we came across the work of Angenent-Mari et al. [1]. Results from this work include:

Previous Work

In an effort to better understand how machine learning is currently used in synthetic biology, we came across the work of Angenent-Mari et al. [1]. Results from this work include:

- creation of a dataset of over 90,000 toehold switches

Previous Work

In an effort to better understand how machine learning is currently used in synthetic biology, we came across the work of Angenent-Mari et al. [1]. Results from this work include:

- creation of a dataset of over 90,000 toehold switches
- training of a multilayer perceptron (MLP) to adequately predict fluorescence for the toehold switch data

Previous Work

In an effort to better understand how machine learning is currently used in synthetic biology, we came across the work of Angenent-Mari et al. [1]. Results from this work include:

- creation of a dataset of over 90,000 toehold switches
- training of a multilayer perceptron (MLP) to adequately predict fluorescence for the toehold switch data
- the MLP outperforms other regression-based models in predicting fluorescence

Previous Work

In an effort to better understand how machine learning is currently used in synthetic biology, we came across the work of Angenent-Mari et al. [1]. Results from this work include:

- creation of a dataset of over 90,000 toehold switches
- training of a multilayer perceptron (MLP) to adequately predict fluorescence for the toehold switch data
- the MLP outperforms other regression-based models in predicting fluorescence
- all models tested performed better when trained on nucleotide sequences instead of derived thermodynamic parameters

Research Questions

Since some of the Fernandez lab ribosensors are riboswitch-like in design, our research questions for this work are two-fold:

Research Questions

Since some of the Fernandez lab ribosensors are riboswitch-like in design, our research questions for this work are two-fold:

1. Using both the toehold switch dataset and the Fernandez lab data, do models perform better when trained on nucleotide sequences than thermodynamic parameters? Additionally, does the MLP outperform other regression-based models for both datasets?

Research Questions

Since some of the Fernandez lab ribosensors are riboswitch-like in design, our research questions for this work are two-fold:

1. Using both the toehold switch dataset and the Fernandez lab data, do models perform better when trained on nucleotide sequences than thermodynamic parameters? Additionally, does the MLP outperform other regression-based models for both datasets?
2. Can we leverage the 90,000+ labeled toehold switch dataset to train a ML model and make accurate predictions on the Fernandez ribosensor data?

Handling Sequences

Most machine learning models require the input data to have the same dimensionality. An option to handle sequences of variable length is to pad them to the same length.

Handling Sequences

Most machine learning models require the input data to have the same dimensionality. An option to handle sequences of variable length is to pad them to the same length.

. . . GUAGAGUGUGAGCUCCGUAACUAGUCGCGUC

Handling Sequences

Most machine learning models require the input data to have the same dimensionality. An option to handle sequences of variable length is to pad them to the same length.

. . . GUAGAGUGUGAGCUCCGUAACUAGUCGCGUC

↓

. . . GUAGAGUGUGAGCUCCGUAACUAGUCGCGUCAUAUAUAUAUAUAUAU

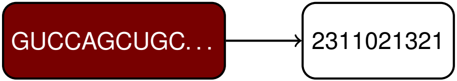
Handling Sequences

In addition, sequence data has to be adequately formatted before being used to train a machine learning model. Plain-text characters can't be used directly.

GUCCAGCUGC. . .

Handling Sequences

In addition, sequence data has to be adequately formatted before being used to train a machine learning model. Plain-text characters can't be used directly.

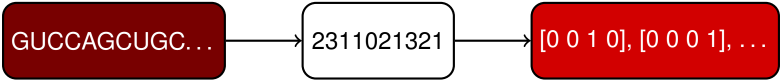


GUCCAGCUGC. . .

2311021321

Handling Sequences

In addition, sequence data has to be adequately formatted before being used to train a machine learning model. Plain-text characters can't be used directly.



GUCCAGCUGC. . . → 2311021321 → [0 0 1 0], [0 0 0 1], . . .

Data Preparation for Model Building

- The two data sets also have two distinct target values, namely fold increase and on/off ratio.

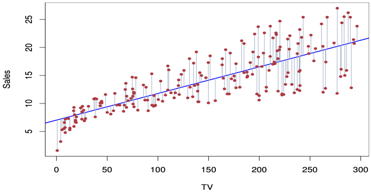
Data Preparation for Model Building

- The two data sets also have two distinct target values, namely fold increase and on/off ratio.
- To properly compare the results of the data sets, these two values were normalized using range-based normalization.

Data Preparation for Model Building

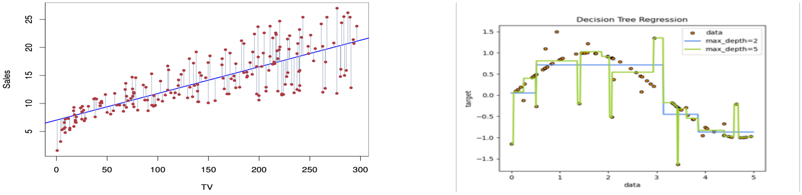
- The two data sets also have two distinct target values, namely fold increase and on/off ratio.
- To properly compare the results of the data sets, these two values were normalized using range-based normalization.
- With the data preprocessed, we trained and cross-validated seven regression-based models on the nucleotide sequence and thermodynamic parameters separately for both datasets.

Constructing Regression Models



(a) Linear Regression

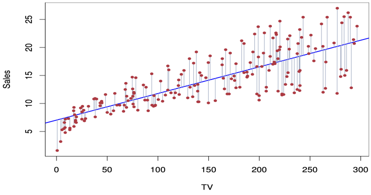
Constructing Regression Models



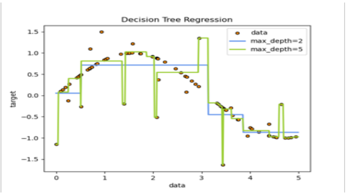
(a) Linear Regression

(b) Decision Tree

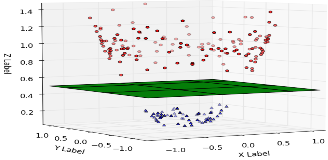
Constructing Regression Models



(a) Linear Regression

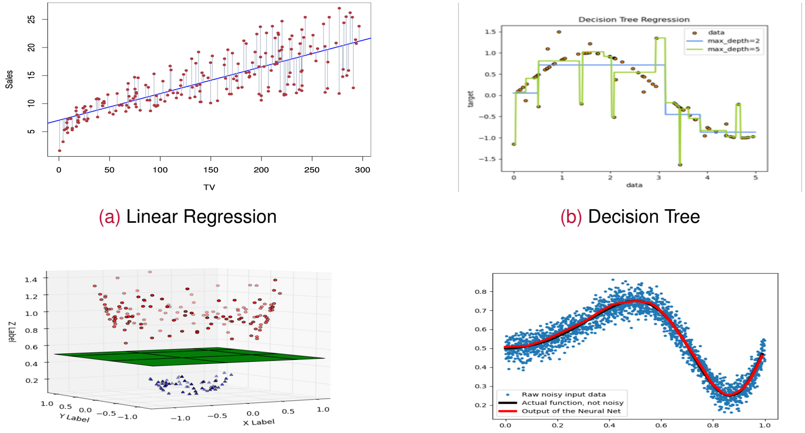


(b) Decision Tree



(c) Support Vector Machine

Constructing Regression Models



(a) Linear Regression

(b) Decision Tree

(c) Support Vector Machine

(d) Neural Network

A Metric for Comparing Regression Models

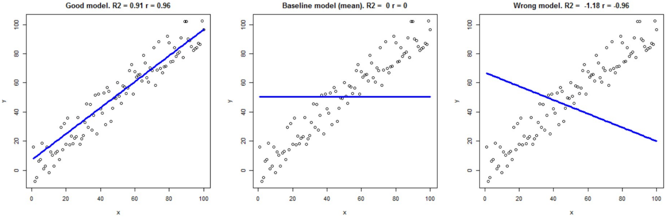
We utilize scikit-learn's implementation of the coefficient of determination:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

A Metric for Comparing Regression Models

We utilize scikit-learn's implementation of the coefficient of determination:

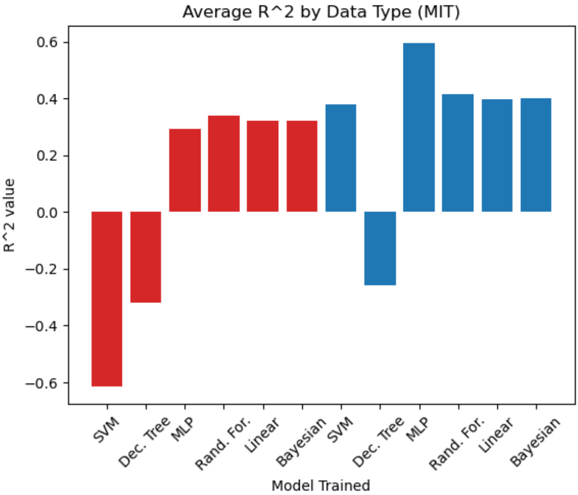
$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$



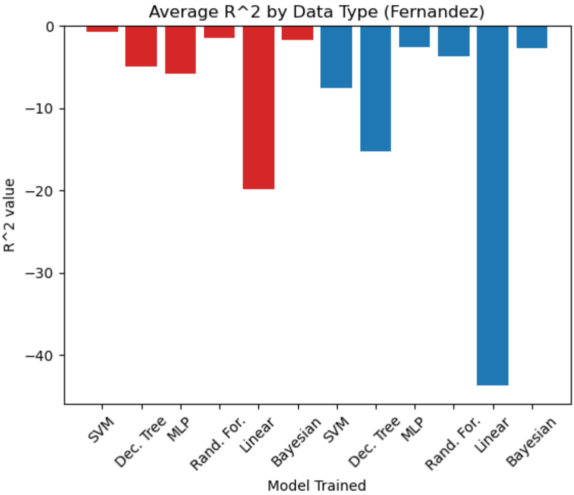
Research Question 1

- Using both the toehold switch dataset and the Fernandez lab data, do models perform better when trained on nucleotide sequences than thermodynamic parameters?
- Additionally, does the MLP outperform other regression-based models for both datasets?

Results for Toehold Switches



Results for Ribosensor Data



Research Question 2

Can we leverage the 90,000+ labeled toehold switch dataset to train a ML model and make accurate predictions on the Fernandez ribosensor data?

Research Question 2

Can we leverage the 90,000+ labeled toehold switch dataset to train a ML model and make accurate predictions on the Fernandez ribosensor data?

Results: We trained a MLP on the full 90,000+ toehold switch dataset and tested on the ribosensor data. In doing so, we obtained an R^2 score of -0.458.

Key Accomplishments and Summary of Results

Key Accomplishments:

- Developed framework for one-hot encoding nucleotide sequences to train ML models

Key Accomplishments and Summary of Results

Key Accomplishments:

- Developed framework for one-hot encoding nucleotide sequences to train ML models
- Created data preprocessing pipeline to pad sequences and normalize output values

Key Accomplishments and Summary of Results

Key Accomplishments:

- Developed framework for one-hot encoding nucleotide sequences to train ML models
- Created data preprocessing pipeline to pad sequences and normalize output values
- Constructed neural net model architecture for training on both sequence and parameter data

Summary of Results:

- Sequence data appears to be better for model training due to potential information loss in calculating thermodynamic parameters

Key Accomplishments and Summary of Results

Key Accomplishments:

- Developed framework for one-hot encoding nucleotide sequences to train ML models
- Created data preprocessing pipeline to pad sequences and normalize output values
- Constructed neural net model architecture for training on both sequence and parameter data

Summary of Results:

- Sequence data appears to be better for model training due to potential information loss in calculating thermodynamic parameters
- Neural Networks are effective models for this type of problem

Key Accomplishments and Summary of Results

Key Accomplishments:

- Developed framework for one-hot encoding nucleotide sequences to train ML models
- Created data preprocessing pipeline to pad sequences and normalize output values
- Constructed neural net model architecture for training on both sequence and parameter data

Summary of Results:

- Sequence data appears to be better for model training due to potential information loss in calculating thermodynamic parameters
- Neural Networks are effective models for this type of problem
- There is either not enough similarity between the datasets or there is not enough data, in general, to do cross-training between the two

Future Directions of the Project

- Pre-training the neural net on Harvard data and then unfreezing layers and retraining on a portion of Fernandez data

Future Directions of the Project

- Pre-training the neural net on Harvard data and then unfreezing layers and retraining on a portion of Fernandez data
- Semi-supervised learning by generating millions of sensor sequences and using Large Language Models (LLM) to learn general RNA patterns that can be applied to the smaller labeled dataset

Future Directions of the Project

- Pre-training the neural net on Harvard data and then unfreezing layers and retraining on a portion of Fernandez data
- Semi-supervised learning by generating millions of sensor sequences and using Large Language Models (LLM) to learn general RNA patterns that can be applied to the smaller labeled dataset
- Revisiting the literature to see if other data sources are closer in design to our dopamine sensors

References

Nicolaas M Angenent-Mari, Alexander S Garruss, Luis R Soenksen, George Church, and James J Collins.
A deep learning approach to programmable rna switches.
Nature communications, 11(1):5057, 2020.

Ann-Christin Groher, Sven Jager, Christopher Schneider, Florian Groher, Kay Hamacher, and Beatrix Suess.
Tuning the performance of synthetic riboswitches using machine learning.
ACS synthetic biology, 8(1):34–44, 2018.

Jacqueline A Valeri, Katherine M Collins, Pradeep Ramesh, Miguel A Alcantar, Bianca A Lepe, Timothy K Lu, and Diogo M Camacho.
Sequence-to-function deep learning frameworks for engineered riboregulators.
Nature communications, 11(1):5058, 2020.

Acknowledgements

Thanks! Any questions?

GitHub

This work was supported primarily by the National Science Foundation EPSCoR Program under NSF Award #OIA-2242812. Any Opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the National Science Foundation.