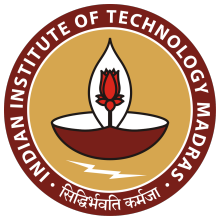
BT3041: Analysis and Interpretation of Biological Data

Term Project

Sumedh Sanjay Kangne (BE21B040)
Ramakrishna Paramahamsa Kompella (BE21B029)
Choppala Rahul (BE21B013)
Ibrahim Kaif (BE21B019)
Saggurthi Dinesh (BE21B032)
Suraj Manickam T (BE21B041)
Indian Institute of Technology, Madras



May 24, 2024

Abstract

This research delves into the role of small molecules in modulating the circadian rhythm, focusing on CRY1, a core clock protein. The study employs machine learning techniques to predict toxicity and period change in circadian rhythm. Unsupervised clustering using UMAP reveals two distinct chemotypes that bind to CRY1, with both clusters containing toxic and non-toxic molecules. This suggests that toxicity is not implicated in the pathways that mediate binding in this context. Further analysis with specific hyperparameters shows a subcluster completely devoid of toxic molecules, indicating a specific non-toxic chemotype. These findings provide a foundation for identifying non-toxic molecules that can modulate the circadian period, contributing to the development of therapeutics for conditions associated with circadian rhythm dysregulation.

Keywords: Circadian Rhythm, CRY1, Machine Learning, Chemotypes, Therapeutics

Motivation

What is circadian rhythm and why is it necessary to oblige by the circadian rhythm? The term 'Circadian Rhythm' refers to the natural operating rhythm of your body and is often called the 'body's internal clock'. It is a biochemical oscillator that modulates various physiological functions such as heart rate, alertness, memory, blood pressure, immune response and many more. Thus, a disruption in this rhythm can take a toll on the metabolic activities and may lead to mood and sleep disorders.

The significance of having a strong circadian clock for overall health is becoming more widely acknowledged. As a result, discovering molecules that can influence circadian rhythms has become a highly popular area of research. Use of classification methods on structure based drug discovery pipeline can help us separate toxic and inactive molecules, and further use of unsupervised learning can help us predict whether a particular molecule has additional uses beyond period lengthening.

Data Methodology: Identifying Candidates

The dataset was intended for the structure-based design of small molecules targeting CRY1, a core clock protein involved in regulating circadian rhythm. CRY1 variants have been associated with various diseases, making it an attractive target for drug development.

The crystal structure of CRY1 (PDB ID: 4K0R) was utilized for in-silico screening. Molecular dynamics (MD) simulations were performed to optimize the structure under physiological conditions, and the convergence of the simulation was monitored using root mean square deviation (RMSD) analysis.

A commercially available small molecule library containing approximately 8 million compounds was screened using Autodock Vina. The library was filtered to exclude irrelevant molecules, resulting in approximately 1 million compounds suitable for docking to the primary and secondary pockets of CRY1.

The top-ranking molecules were selected based on their Vina binding energies. Additionally, Pan Assay Interference compounds (PAINS) remover was employed to eliminate potential false positives. A total of 139 molecules designed for the primary pocket and 32 molecules designed for the secondary pocket of CRY1 were tested for toxicity.

Primary pocket is a FAD-binding domain and Secondary pocket is an alpha-beta domain.

The screening process aimed to identify molecules capable of modulating the circadian rhythm by interacting with the FAD-binding and secondary pockets of CRY1. These molecules hold potential for further development as therapeutics for conditions associated with circadian rhythm dysregulation, such as depression, mood disorders, hypertension, and sleep disorders.

Toxicity Dataset

The dataset includes 171 molecules designed for functional domains of a core clock protein, CRY1, responsible for generating circadian rhythm. 56 of these molecules are toxic and the rest are non-toxic.

The data consists of a complete set of 1538 molecular descriptors and needs feature selection before classification since some of the features are redundant.

The process of drug discovery using in-silico methods often produces datasets with a very large number of attributes (fields) per instance (record). Automated classification of such data on properties such as toxicity provides significant benefits for drug design but must cope effectively with the large number of attributes and the relatively small number of instances. By identifying a suitable small subset of the attributes that are effective for this classification task, experimental results indicate accuracies that compare very favorably with prior work on the same data.

In this study, in-silico screening using CRY1 crystal structure (ID: 4K0R) was performed to find molecules that regulate circadian rhythm in the human osteosarcoma (U2OS) cell line. 171 molecules were experimentally tested in terms of toxicity and activity. In vitro MTT toxicity assay is done to measure cytotoxicity of the small molecules in the cell. Out of these 56 molecules were found to be toxic, and 115 molecules were found to be nontoxic.

We will use machine learning to identify the features that are most important to predict the toxicity of molecules, as well as the descriptors that explain the period change in circadian rhythm. Results can be used in QSAR studies for identification of nontoxic and circadian period lengthener molecules using big libraries.

A classifier with a good fit on the training set may produce poor results on the test dataset. To prevent overfitting, it is necessary to select the best set of molecular descriptors and eliminate the redundant features. For this process of feature selection, we used Recursive Feature Elimination (RFE) together with Decision Tree Classifier (DTC) to get the best set of molecular descriptors for DTC. Subsetted data with 13 features is included as a supplementary file.

Analysis

Before we look at how these small molecules affect the period of the circadian cycle, it is important to look at their toxicity levels. Different molecules exhibit different levels of toxicity. 2.5µM was used as the threshold and any molecule with a relative cell viability <85% at this concentration was considered toxic.

Now, since this data exists in the form of molecules it is necessary to extract descriptors which define this data in an analytical form, which can be used for further ML analysis. To get these molecular descriptors, these molecules were fed into the ChemDes web-server. This gave us a total of 1538 PaDEL

molecular descriptors. However, 334 of them had the same value for all molecules and were thus omitted. The remaining 1203 features thus define our molecular data and were used for training the dataset.

These descriptors can be used to train a classifier and thus help separate toxic molecules from non-toxic ones. But since the number of descriptors is quite large and to avoid overfitting of the data it is necessary to eliminate the redundant features. This can be done by Feature Selection.

Number of features were calculated by calculating the accuracy using 10 fold cross validation and selecting the one with highest accuracy.

Number of Features	Mean CV-Score
10	0.7542
11	0.7484
12	0.6549
13	0.7719
14	0.7307

Table 1.1 - Mean CV-scores obtained for various number of features selected.

It was seen that the CV-score was highest for 13 features, and thus it seems appropriate to proceed with 13 as the number of features.

Recursive Feature Elimination (RFE) was used with Decision Tree Classifier (DTC) as the estimator to select the best 13 features. The parameters used for tuning the DTC were based on grid search. The following are the tuning parameters used for the DTC.

max_depth	10
max_features	8
min_samples_leaf	1
min_samples_split	3

Table 1.2 - Hyperparameters used for the estimator

Shown below are the 13 features obtained from RFE.

1. MDEC-23
2. MATS2v
3. ATSC8s
4. CrippenMR
5. SpMax7_Bhe
6. SpMin1_Bhs
7. C1SP2
8. BIC2
9. GATS8e
10. GATS8s
11. SpMax5_Bhp
12. VE3_Dzi
13. VPC-4

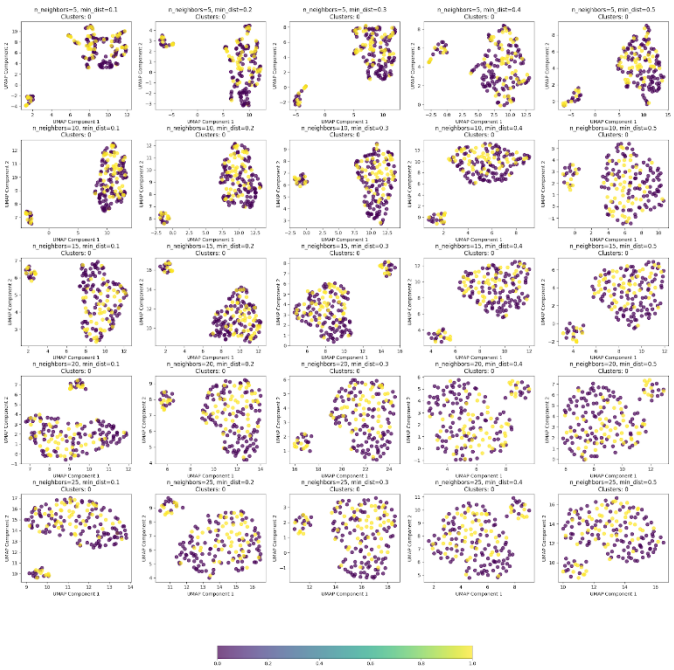
Unsupervised learning

Unsupervised clustering using UMAP reveals two clusters in the 2D Graph. Both clusters contain toxic as well as nontoxic molecules. We may hypothesize the following:

1. There exist two distinct chemotypes that bind to CRY1.
2. The two chemotypes may bind to the two different binding sites present in CRY1.
3. The existence of toxic and nontoxic molecules in both clusters indicates that this differential is not implicated in pathways(s) that mediate toxicity in this context.

For hyperparameters, n_neighbors =5 , min_dist = 0.1, We find that the larger cluster splits into further two clusters. The smaller subcluster is completely devoid of toxic molecules.

1. There is a specific chemotype which is not toxic
2. There could be elements which could be make otherwise toxic molecules non-toxic or vice versa.



These hypotheses merit an experimental study to confirm them.

Period Lengthening Dataset

The circadian rhythm period lengthening was observed and tested through a series of methodical steps involving real-time bioluminescence monitoring of U2OS cells, both in normal conditions and with CRY1 knockout. By analyzing the luminescence data over time, any changes in the circadian period were identified. This was observed as a shift in the periodicity of the bioluminescence oscillations recorded from the cells. The use of BioDare2 for data analysis allowed for precise determination of period length changes, and statistical analysis confirmed the significance of these observations.

Looking at the impact of 85 non-toxic molecules on the circadian rhythm of U2OS cells expressing a destabilized luciferase under the Bmal1 promoter (U2OS Bmal1-dLuc). Cells were treated with these molecules, and circadian period changes were analyzed using BioDare2. Twenty-one molecules significantly lengthened the circadian period, while one molecule, N8, shortened it and was excluded from further studies.

To confirm CRY1 dependency, CRY1 knockout U2OS cells (CRY1-/- Bmal1-dLuc) were generated using CRISPR/Cas9. These knockout cells exhibited a shorter circadian period, consistent with previous findings. When treated with the potent period-lengthening molecules, no changes were observed in the circadian period of CRY1 knockout cells, indicating CRY1 dependence.

Analysis

Now, that the toxicity classification is done, we can proceed with the period lengthening analysis. To start we choose all the non-toxic molecules which are CRY1 primary site targeting molecules. This gives us a total of 90 molecules that were used, out of which: 27 are period lengthening and 63 don't affect the period length.

We followed an approach similar to the toxicity dataset for the classification of the set of period-lengthening molecules. All of these molecules were fed into the ChemDes web-server to obtain a total of 1538 molecular descriptors. However, 360 of them had the same value for all molecules and were thus omitted. The remaining 1177 features were used for training the dataset.

Similar to the toxicity dataset, it is necessary to reduce the number of features as a higher number of features with respect to the number of molecules might lead to overfitting. This can be done by Feature Selection.

Number of features were calculated by calculating the accuracy using 10 fold cross validation and selecting the one with highest accuracy.

Number of Features	Mean CV-Score
5	0.8000
6	0.8000
7	0.7667
8	0.7778
9	0.7111

Table 2.1 - Mean CV-scores obtained for various number of features selected.

It was seen that the CV-score was highest for 5 features, and thus it seems appropriate to proceed with 5 as the number of features.

Recursive Feature Elimination (RFE) was used with Decision Tree Classifier (DTC) as the estimator to select the best 5 features. The parameters used for tuning the DTC were based on grid search. The following are the tuning parameters used for the DTC.

max_depth	6
max_features	4
min_samples_leaf	1
min_samples_split	8

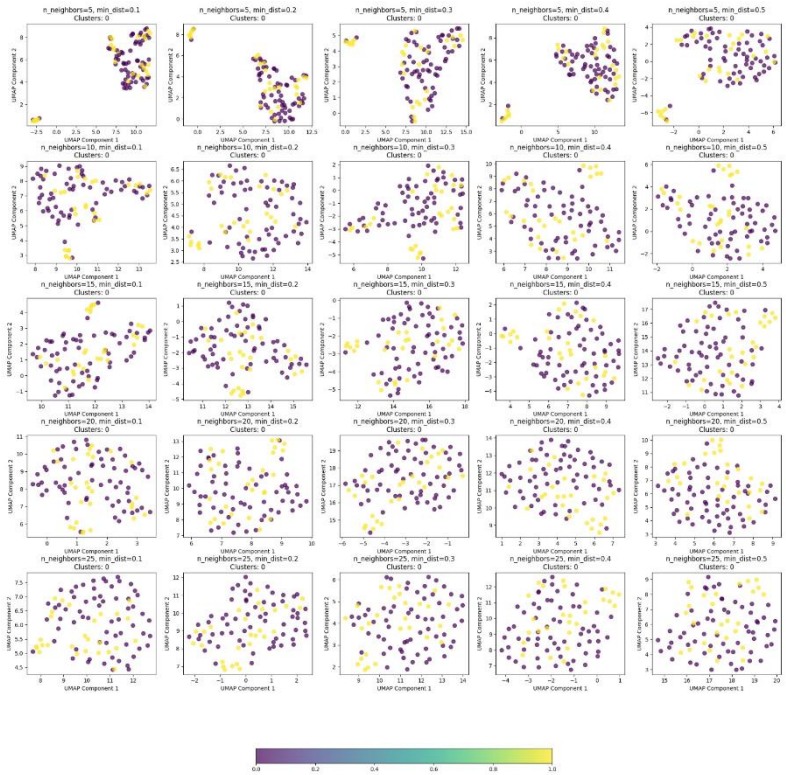
Table 2.2 - Hyperparameters used for the estimator

Shown below are the 5 features obtained from RFE.

1. MDEC-23
2. VR2_Dzs
3. ATSC3v
4. AATSC5c
5. VE2_Dzp

Unsupervised learning

Similar to the toxicity dataset, we find two clusters. Suggesting two chemotypes.



These hypotheses merit an experimental study to confirm them.

Conclusion

The study provides valuable insights into the potential of small molecules in modulating the circadian rhythm. It highlights the importance of machine learning in predicting toxicity and period change in circadian rhythm. The discovery of distinct chemotypes that bind to CRY1 and the identification of a non-toxic chemotype open new avenues for therapeutic development. These findings underscore the potential of in-silico methods in accelerating drug discovery and development for conditions associated with circadian rhythm dysregulation. Future work should focus on experimentally validating these hypotheses and exploring the therapeutic potential of these non-toxic molecules.

Code

https://colab.research.google.com/drive/1_89B5uojsxL6lrQpj_vgTaXrxkSwO2GD?usp=sharing

The End