Development of Machine Learning-Based Prediction Models for Chemical Modulators of the Glucocorticoid Receptor Signaling Pathway Using Public-Domain Bioactivity Data

[1]Shreya Singh, [1]Sunghwan Kim, [1]Evan Bolton, [1]James Ostell

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health

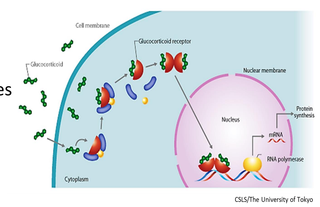
Introduction

Executive Summary

- **Our goal was to predict glucocorticoid receptor (GR) activity based on the structure of its small molecule chemical modulators, via a machine learning approach.**
- **PubChem** is the world's largest freely accessible chemistry database, maintained by the NIH.
- The model was trained on **Tox21 bioassay data,** stored on PubChem.
- We used **small molecule substructure fragments** to train our models and predict **GR activity.**
- We created **structure alerts** via identifying the substructures whose presence was statistically likely to indicate a specific activity.
- **The best models were Random Forest, Support Vector Machines, and Neural Networks** with an **AUC of 0.96 –** performance varying slightly with fingerprint type.
- The calculated **applicability domain** excluded a majority of external data, indicating **Tox21 data is not representative of all PubChem compounds**.
- **This exploration shows how machine learning methods can be used on PubChem's open-source bioassay data to generate useful predictive insights.**

NR3C1 – Glucocorticoid Receptor

- Class of **Nuclear Receptor** --in cytoplasm
- **Binds to Glucocorticoids** -- Class of adrenal hormones *(ex. Cortisol)*
- Upon binding, travels to the nucleus and **acts as a transcription factor**
- **Signaling pathways vary:**
  - **Agonist/ Antagonist**
- Clinical Significance
  - Inflammatory responses
  - Cellular proliferation
  - Target tissue differentiation



Data Sources

Fingerprint Type	String Type	Length
MACCS Keys	Bit String	166
Topological Fingerprint	Bit String	2048
Morgan Circular Fingerprint (ECFP)	Bit String	1024
Morgan Circular Fingerprint (FCFP)	Bit String	1024
PubChem Fingerprint	Hexacode	881

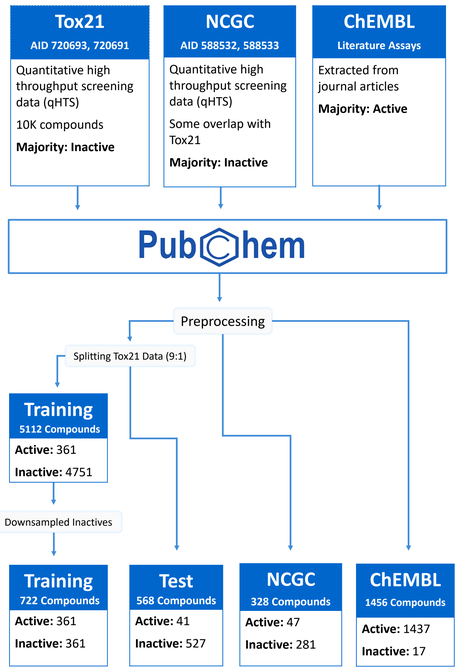
- All fingerprints were downloaded and decoded using **the RDKit Cheminformatics Python Package**

-- Does molecule have ... -- Does molecule have 4 Hydrogen? -- Does molecule have 1 Lithium? -- Does molecule have any ring size 3? -- Does molecule have any nitrogen-containing rings? -- Does molecule have a Li-Cl bond? -- Does molecule have ...

1 0 1 1 0 1 0

Methods

Data Sources and Preprocessing Pipeline



Conflict Resolution

1. Replaced mixtures and salts with parent compound.
2. Removed duplicate compounds.
3. Removed compounds with conflicting bioactivity.

Molecular Descriptors/PUG-REST

1. Downloaded Molecular Properties were from PubChem via the PUG-REST URL-based request system.
   - *Molecular Weight, Hydrogen Bond Donor/Acceptor Count, Rotatable Bond Count, XLogP, Topological Surface Area, Heavy Atom Count, Complexity*

**Data Imbalance Correction (only Tox21)

1. Set aside 10% of the Tox21 data as a test set.
2. Normalized the Tox21 training data, along with the downloaded molecular properties, via Principal Component Analysis
3. Balanced the Tox21 training data via down-sampling inactive compounds .
   - Kept all 361 active compounds
   - Used k-nearest-neighbors clustering to identify 361 inactive points. Removed remaining inactive points.

**For External Datasets

1. Removed compounds that existed in both the Tox21 data and an external set from the external set.

Machine Learning Models

- **X-Features:**
  - **Molecular Fingerprint Bits**
- **Y-Target:**
  - **GR Receptor Activity**

Model Type	Hyperparameters
Naïve Bayes	alpha: α (10^{-10} ~ 1)
Decision Tree	max_depth_range: (3 ~ 7) min_samples_split_range: (3 ~ 7) min_samples_leaf_range: (2 ~ 6)
Random Forest	n_estimators: (10 ~ 200)
K-Nearest Neighbors	weights: (uniform, minkowski, jaccard), n_neighbors: (1 ~ 25)
Support Vector Machine	c_range: (2^{-10} ~ 2^{10}) gamma range (γ): (2^{-10} ~ 2^{10})
Neural Network	solver: (lbfgs, adam) alpha: α (10^{-7} ~ 10^{7})

Evaluation Metrics

$$\textbf{\textit{Sensitivity}} = \frac{TP}{TP + FN} \qquad \textbf{\textit{Specificity}} = \frac{TN}{TN + FP}$$

$$\textbf{\textit{Balanced Accuracy (BACC)}} = \frac{Sensitivity + Specificity}{2}$$

$\textbf{\textit{Area under the Curve (AUC)}}$ = *used for hyperparameter optimization*

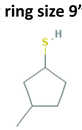
Results

Structure Alerts

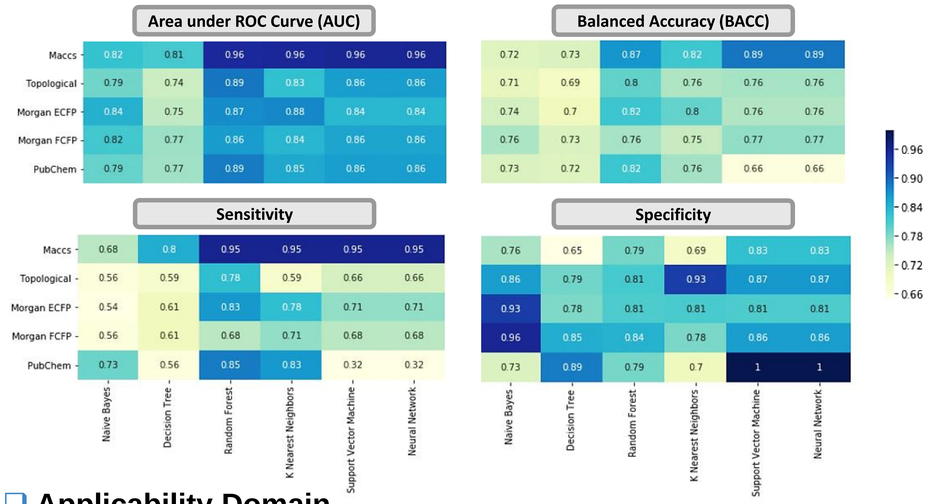
- Analyzed the occurrence of each substructure in the data
  - Used PubChem FP – best documented definition
- Compared that to the overall active/inactive data distribution
- Identified the bits/structure fragments that varied significantly from the activity distribution
  - Chi-Squared Statistic
  - Central Limit Theorem (n > 30) $\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$
- Determined that the presence of those significant fragments could be used to predict GR receptor activity
  - 95% degree of confidence

4 significant fragments identified:

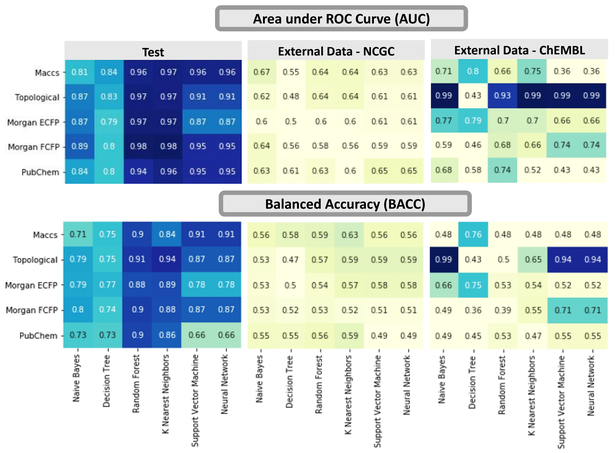
- **"1 saturated or aromatic nitrogen-containing ring size 9"** (#243)
  - Present in 131 actives, 59 Inactives
- **3-methylcyclopentane-1-thiol (#841)**
  - "CC1CC(S)CC1"
  - Present in 112 Actives and 31 Inactives
- **2-methylcyclopentane-thiol** (#862)
  - "CC1C(S)CCC1"
  - Present in 112 actives, 32 Inactives
- **2-sylfanylcyclopentane-1-O** (#867)
  - "OC1C(S)CCC1"
  - Present in 14 actives, 1 inactive



Model Performance



Applicability Domain



Distance Threshold

- Eliminated points whose distances were over half a standard deviation away from the average distance to the training set.
- Re-ran models with remaining points.

Model performance increased slightly for the Test set, and **increased significantly for NCGC and ChEMBL external data.**

	Tox21	NCGC	ChEMBL
Data Type	qHTS	qHTS	Literature Extracted
Chemical Domains	Environmental Chemicals	Environmental Chemicals	Medicinal Chemistry/ Natural Product
Compounds within applicability domain (Range over all 5 FPs)	568 initial compounds **~(69 – 90) %** remaining	328 initial compounds **~(58 – 84) %** remaining	1456 initial compounds **~(10 – 14) %** remaining

Conclusions

Model/Fingerprints

- **Random Forest, Support Vector Machines, and Neural Networks were the best models. (AUC = 0.96)**
- The MACCS fingerprints (fewest number of bits) performed best.
- Longer fingerprints (Topological, PubChem) suffered slightly from overfitting.

Structure Alerts

- Cyclopentanes appear to be highly related to GR activity.
  - ¾ of significant fragments
- Identifying activity-related substructures allows for a greater understanding of GR signaling pathway dynamics. This information can be used in predictive analytics, cell-behavior modeling and drug discovery.

Applicability Domain

- Distance analysis shows that NCGC data is more similar to Tox21 data then ChEMBL data, likely due to their chemical domains.
- The model's increased performance indicates it performs well on data similar to the training data.
- Tox21 compounds are not indicative of all PubChem compounds.

Future Directions

- Removing correlated features to continue addressing overfitting.
- Testing against more external GR data.
- Engineering a ternary classifier to distinguish between active agonist and active antagonist binding behavior.
- Testing the model's performance against additional receptors with available PubChem bioactivity data.
  - Further investigate cyclopentanes.
  - See if structure alerts are GR receptor specific.

Acknowledgements

Thank you to Dr. Sunghwan Kim, Ph.D. for his direct supervision and guidance. This research was supported by the College Summer Opportunities to Advanced Research Program of the Office of Intramural Training and Education at the National Institutes of Health (NIH) as well as the Intramural Research Program of the National Library of Medicine at NIH.