Title: Predicting the Oral Toxicity of a Molecule Using a Decision Tree Approach: A Machine Learning Application to Toxicology

Objectives/Goals
Toxicity assessment of potential drug molecules is an important step in drug discovery and can be improved using computational models for toxicity evaluation. The efficiency of creating toxicity profiles for new chemicals that are introduced in increasing numbers each year can be improved using computational models for toxicity prediction. These models for toxicity evaluation will also reduce the need for animal testing. This project uses decision tree approach to accurately predict the level of the oral toxicity of any organic molecule.
Methods/Materials
The classification scheme for molecules that are orally active consists of three classes correspond to low, moderate, and high levels of toxicity. First, the model determines whether a molecule is orally active or inactive using the Lipinski's Rule of Five. If the molecule is considered orally active, the model predicts whether the molecule's oral toxicity is of a low, moderate, or high level using the molecule's chemical structure. I used a machine learning algorithm to train the decision tree using the structural properties of the molecules in the training set. Through the use of machine learning and the repeated modification of the set of properties used to train the decision tree, the efficiency of the model was improved. In the algorithm the best split was chosen depending on the predictive accuracy of the potential splits. Overfitting became evident in some instances when the some decisions made in the trained decision tree applied only to the training set. Therefore, I limited the depth of the tree to ensure the efficiency of the model.
Results
The actual toxicity class of the molecule in the test set was decided based on the classification scheme. The accuracy of the prediction made by decision tree was measured by comparing to the toxicity class outputted by the decision tree to the actual toxicity class.
The algorithm was able to predict the toxicity class of eighty percent of the forty molecules in test set accurately. The null hypothesis was that the accurate prediction of toxicity of the molecules in the test set was due to chance alone. Therefore, I validated the model using a t-test in which the null hypothesis was assumed to be true. The t-test produced a p-value of 0.004 which means that the null hypothesis should be rejected.
Conclusion
The model provides the potential to improve the models for toxicity assessment in drug discovery and the evaluation of chemicals. The model contributes to toxicity prediction because it demonstrates the efficiency using a trained decision tree to associate certain properties with certain levels of toxicity. The model also validates the approach of the building toxicity profiles for molecules based their chemical structure. This computational model also reduces the need for animal testing.
Conclusion
The decision tree model created provides an efficient approach to classify organic molecules based on their toxicity.