**Capstone Project Proposal**

**Introduction**

Discovery and development of a new drug can include the following stages: 1) target identification 2) lead finding and optimization 3) pre-clinical studies 4) clinical studies.1 Discovery of new drug candidates is becoming increasingly difficult and laborious. This process can take between 12-20 years and sometimes fraught with complications that at any step, the drug can fail clearance. Interestingly, high-throughput screening is becoming more used in the exploratory phase of the drug-discovery process. In this period, thousands of small molecule compounds are tested for inhibitory capacity in order to distinguish drug-like and non-drug-like molecules. Various statistical calculations can be applied to the screening data to glean information on the patterns of drug behavior. Statistical machine learning methods have been used in drug discovery studies for classification purposes, such as discriminant, tree-based, kernel-based and other algorithms.

**Data Source**

One of the projects I worked on in my dissertationwas to find inhibitors of the Dengue virus targeting the non-structural 3 (NS3) helicase of the Dengue genome. Dengue virus (DENV) is one of the most significant human viral pathogens transmitted by mosquitoes. It causes 50 million or more cases of infection worldwide each year, resulting in around 24,000 deaths. It is primarily a tropical disease. About 80% of people infected with dengue virus are asymptomatic or only have mild symptoms such as fever and severe joint pain. However, more serious syndromes, such as dengue hemorrhagic fever or dengue shock syndrome, can manifest following dengue infection.2 Researchers currently believe that the deadly dengue hemorrhagic disease is caused when a person is infected with one subtype, and then infected later by a second subtype. Control of dengue virus through the use of vaccination or a direct-acting antiviral has proven to be elusive.

The data I’ll be using came from the high-throughput screening I performed as part of my dissertation. In the study, I used an optimized and automated malachite green-based colorimetric assay to detect compounds that directly inhibit helicase-catalyzed ATP hydrolysis of the Dengue virus. My hypothesis is predicated upon the NS3 helicase protein being vital in the creation of the Dengue viral genome. Therefore, by inhibiting the function of the NS3 helicase, the growth of the Dengue virus is also inhibited. In this study, I screened over 6,000 samples containing small organic molecules.

**Data Wrangling**

The initial data was converted to an excel spreadsheet file by the microplate reader machine used for screening. I further merged it with the molecule file that came from the vendor which contains molecular information such as SMILES (1D chemical representation of the molecules). I then imported it to Instant Jchem which is free to use for academic users. This software calculated molecular descriptors based on the chemical structure, such as LogP, LogD, # of H bond donors, # of H bond acceptors, # of Rotatable bonds, and whether it satisfies Lipinski’s Rule of 5.

***Initial Table Columns:***

Lab ID

SMILES String

Molecular Formula

Formula Weight Registry Number

Lot Number

Purity

Daughter Plate

Plate Map

DENV ATPase %Inhibition

**Approach to Data Analysis and Visualization**

I would like to find out if certain patterns of behavior of the molecule correlate to its structure. For example:

1. Do bigger-sized molecules more/less potent inhibitors?
2. Does the number of benzene rings increase/decrease inhibitory capability?
3. Does the octanol/water index affect ligand activity?

Besides classification algorithms such as decision tree classification (DTC), random forest (RF), I’d also like to create heat maps and dendogram for visual inspection of the molecules. I would also like to see if I can use principal component and hierarchical cluster analysis.

**Bibliography**

1. Lock, E. F., Abdo, N., Huang, R., Xia, M., Kosyk, O., O'Shea, S. H., ... Rusyn, I. (2012). Quantitative high-throughput screening for chemical toxicity in a population-based in vitro model. Toxicological Sciences, 126(2), 578-588. DOI: [10.1093/toxsci/kfs023](http://dx.doi.org/10.1093/toxsci/kfs023)
2. Guzman M. G. *et al*. Dengue: A continuing global threat. *Nature Reviews Microbiol*ogy **8**, S7–S16 (2010).