Capstone Project Proposal

Logit Data Science, 8/19/2016

Arjun Narayanan and Philip Chong

**Background:**

The treatment and eradication of malaria is currently receiving a lot of attention in the popular media. Recently, the Bill and Melinda Gates Foundation identified malaria as a “top priority” and committed nearly US $2 billion in grants to combat malaria1. Bill Gates has declared that “*we’re now in a position to eradicate malaria—that is, wipe it out completely in every country—within a generation*”2. Part of this effort involves the development of novel therapeutics that capable of clearing all malaria parasites from the body. While there are a lot of academic research groups and pharmaceutical companies conducting research on this topic, current drug discovery methodologies involve the use of “brute-force” methods, namely, high-throughput screening and iteratively improving the biological activity (as measured by specific assays) of small molecules through synthetic organic methods. There is a lot of scope for the *intelligent* design of models that can accurately predict the biological activity of new molecules based on data acquired for structural homologs. This represents an area in which Machine Learning algorithms (both supervised and unsupervised) can be applied for the prediction of biological activity and discovery of other actionable insights.

**Dataset and Methods:**

The dataset under investigation in this study is the UW Kinase screening hits dataset available on ChemBL3. This is available under ChemBL’s Neglected Tropical Diseases category4, and was uploaded for public access on March 5, 2016. 13,452 molecules are screened against 5 protein kinases previously identified to be promising anti-malaria targets, and the assay data is included for all compounds, as well as other useful chemical data (polar surface area, number of rotatable bonds, whether the molecule obeys Lipinski’s “Rule of Five”, etc.). Thus, this dataset can be analyzed both with supervised and unsupervised Machine Learning methods – one can try to build a regression model to predict inhibitory action (a continuous numeric variable) from other features in the dataset. Potential bioactive molecules can also be clustered using unsupervised methods. Molecular similarity can be measured using the Tanimoto similarity metric, and this is available in the *rdkit5* Python package for parsing chemical structure data.

**References:**

1. <http://www.gatesfoundation.org/What-We-Do/Global-Health/Malaria>
2. <https://www.gatesnotes.com/Health/Eradicating-Malaria-in-a-Generation>
3. <https://www.ebi.ac.uk/chemblntd/download>
4. <https://www.ebi.ac.uk/chemblntd/>
5. <http://www.rdkit.org/docs/GettingStartedInPython.html#fingerprinting-and-molecular-similarity>