DEVELOPING A QUALITY ASSESSMENT METRIC FOR DRUG DISCOVERY DATASETS AND PREPARING PROTEIN DATASETS FOR QSAR APPLICATIONS

By

Margaret Rose McLellan

An Abstract of a Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subject: CHEMISTRY

The original of the complete thesis is on file
in the Rensselaer Polytechnic Institute Library

Examining Committee:

Professor Curt Breneman, Thesis Adviser
Professor Chris Bystroff, Member
Professor Steven Cramer, Member
Dr. Doug Kitchen, Member
Professor Mark Wentland, Member

Rensselaer Polytechnic Institute
Troy, New York

October 2011
(For Graduation December 2011)

ABSTRACT

Quantitative Structure-Activity Relationship models are often mis-applied in drug discovery. This has motivated the creation of a number of model validation techniques, but none of these directly determine the stability of model predictions over changes in the model. Often in drug discovery, the numerical value of a molecular prediction is less important than the prediction of rank orders. Evaluating the stability of rank order, or Rank Order Entropy, therefore requires the use of a rank order metric. Models were evaluated using Kendall Tau as a rank order metric. Data Truncation Analysis (DTA) was created to evaluate the predictive power of a model over decreasing information in the training set by iteratively and randomly reducing the information in the training set. The Shannon entropy of Kendall Tau was calculated over the truncations as a measure of stability.

The ROE metric was applied to combinations of 71 data sets of different sizes and 6 sets of descriptors, and was found to reveal more information about the behavior of the models than traditional metrics alone. In the end, ROE metrics suggested that some QSAR models that are typically used should be discarded, and some QSAR models that might be typically discarded could be used. ROE evaluation helps to discern which combinations of data set, descriptor set, and modeling methods lead to usable models in prioritization schemes, and provides confidence in the use of a particular model within a specific domain of applicability.