Abstract

Quantitative structure activity relationship (QSAR) models are are a statistical solution to the problem of directly calculating physical and biological properties of molecules from their physical structure. The direct prediction of properties is in general not feasible either owing to lack of computing resources or lack of knowledge about the relationship between structure and property. The goal of a QSAR model is to extract information from a set of numerical descriptors characterizing molecular structure and use this information to develop inductively a relationship between structure and property. Two important questions arise during the modeling process. First, are the data used to build the model representative of the whole dataset and can the model be extended to predict properties for new molecules? Second, given that a model encodes information about the structures of molecules and relates this to their properties, can we extract and interpret the encoded information? The focus of the work reported in this thesis is on the validation and interpretation of QSAR models and presents both applications of interpretation techniques as well as the development of validation and interpretation methodologies.

The first study describes a technique to develop representative QSAR sets using a self-organizing map (SOM). The SOM was used to classify a dataset consisting of dihydrofolate reductase inhibitors with the help of an external set of global descriptors. The resultant classification was used to generate training, cross-validation and prediction sets (collectively known as QSAR sets) for QSAR modeling using the ADAPT methodology. The results were compared to those of QSAR models generated using sets created by activity binning and a sphere exclusion method. The results indicated that the SOM was able to generate QSAR sets that were representative of the composition of the overall dataset in terms of similarity. The resulting QSAR models were half the size of those published and had comparable RMS errors. Furthermore, the RMS errors of the QSAR sets were consistent, indicating good predictive capabilities as well as generalizability.

The determination of the validity of a QSAR model when applied to new compounds is an important concern in the field of QSAR modeling. Various scoring techniques can be applied to specific types of models to obtain measures of confidence in the predicted property for new compounds. The second study describes the development of a methodology which allows one to state whether a new compound will be well predicted

by a previously built QSAR model. The study focuses on linear regression models only, though the technique is general and can also be applied to other types of quantitative models. The technique is based on a classification method that divides regression residuals from a previously generated model into a good class and bad class and then builds a classifier based on this division. The trained classifier is then used to determine the class of the residual for a new compound. The performance of a variety of classifiers, both linear and nonlinear, was investigated. The technique was tested on two data sets from the literature and an artificial data set. The data sets selected covered both physical and biological properties and also presented the methodology with quantitative regression models of varying quality. The results indicate that this technique can determine whether a new compound will be well or poorly predicted with weighted success rates ranging from 73% to 94% for the best classifier.

The remaining studies focus on methods to interpret QSAR models. The third and fourth studies describe applications of a partial least squares (PLS) based method to interpret linear regression models. The third study developed QSAR models to predict the biological activity of 179 artemisinin analogues. The structures of the molecules were represented by numerical descriptors. Both linear (multiple linear regression) and nonlinear (computational neural network) models were developed to link the structures to their reported biological activity. The best linear model was subjected to a PLS analysis to provide model interpretability. While the best linear model did not perform as well as the nonlinear model in terms of predictive ability, the application of PLS analysis allows for a sound physical interpretation of the structure-activity trend captured by the model. On the other hand, the best nonlinear model was superior in terms of pure predictive ability, as characterized by low training and prediction set root mean square errors.

The fourth study consisted of the development and interpretation of QSAR models to predict the activity of a set of 79 piperazyinylquinazoline analogues which exhibited platelet derived growth factor (PDGFR) inhibition. Linear regression and nonlinear computational neural network models were developed. The linear regression model was developed with a focus on interpretative ability using the PLS technique. However, it also exhibited good predictive ability after outlier removal. The nonlinear CNN model had superior predictive ability compared to the linear model, having a prediction set root mean square errors nearly half that of the linear model. A random forest model was also developed to provide an alternate measure of descriptor importance. This approach ranks descriptors, and its results confirmed the importance of specific descriptors as characterized by the PLS technique.

Studies five and six describe the development of methods to provide interpretability to computational neural network (CNN) models. The fifth study focuses on a measure of relative importance of the descriptors present in a CNN model. The approach is based on a sensitivity analysis of the descriptors and is similar in concept to the descriptor importance measure for random forest models. The method was tested on three published data sets for which linear and CNN models were previously built. The original work reported interpretations for the linear models. This study compared the results of this method to the importance of descriptors in the linear models as described by the PLS technique. The results indicate that the proposed method is able to rank descriptors such that important descriptors in the CNN model correspond to the important descriptors in the linear model.

The sixth study presents the development of a method to provide a detailed interpretation of a CNN model. This methodology provides a means to analyze the correlation between specific input descriptor and the predicted output of the network, rather than simply providing a ranking of all descriptors. The method consists of two parts. First, the nonlinear transform for a given neuron is linearized, allowing us to determine how a given neuron affects the downstream output. Next, a ranking scheme for neurons in the hidden layer is developed. This scheme allows for the development of interpretations of a CNN model similar in manner to the PLS interpretation method for linear models. The method was tested on three datasets covering both physical and biological properties. The results of this interpretation method correspond well to PLS interpretations for linear models using the same descriptors as the CNN models.