Chapter 10

Summary

The focus of this thesis has been the validation and interpretation of QSAR models. The preceding chapters cover examples of the application of interpretation method applied to a variety of models as well as the development of interpretation and validation methods.

Chapters 2 and 3 provide a brief introduction to the modeling techniques used in this work and the general methodology used to develop predictive QSAR models using the ADAPT software package. Though the descriptions have focused on quantitative models, the principles underlying the development of these types of models are also applicable to the development of classification models. As described in Chapter 3, the model building procedure involves a number of steps. Aspects such as feature selection have been studied extensively in the statistical literature, and developments in that field have yielded important insights into the methods by which good subsets of descriptors are selected. The model building process is also fundamentally a statistical problem and advances in the fields of data mining and pattern recognition have allowed the field of QSAR modeling to build models with increasing predictive ability and reliability.

Chapters 4 and 5 focus on two specific steps of the QSAR modeling process. Chapter 4 describes a method that was developed to create representative QSAR sets. One of the initial steps in building a QSAR model is to divide the dataset into a number of subsets, collectively termed QSAR sets. The training set is used to build the model, and the prediction set is used to test the final model for its predictive ability. In the case of a neural network an additional subset, namely the cross-validation set, is also created which is used to prevent overfitting. A number of methods exist to create these sets but one requirement is that each set be representative of the composition of the whole dataset. That is, features that are present in the overall dataset should also be present in these subsets, otherwise different phases of the model development process may be biased. This aspect of set creation is not addressed at all by the random selection method. The activity binning method does solve the problem to some extent but mainly concentrates on representing the range of activities in the whole dataset in the individual subsets. The

method described in Chapter 4 utilizes a self-organizing map (SOM), together with a set of holistic descriptors, to classify the dataset into two classes. The QSAR sets are then created such that the distribution of molecules in the two classes of the original dataset is maintained in each of the subsets. The assumption underlying this procedure is that models built with these QSAR sets will be able to capture the representative features of the dataset during training and will also exhibit better predictive ability compared to models built with QSAR sets in which the features of the dataset are not evenly distributed. The method was tested on a dataset of dihydrofolate reductase inhibitors which had been studied by Mattioni et al.[1] The best CNN model developed using this method had a simpler architecture and exhibited statistics similar to the previously reported model. An interesting observation regarding this model was that the statistics for the individual QSAR sets were consistent. That is, the RMSE and R^2 values were similar to each other for the training, cross-validation and prediction sets. This was not the case for the original model developed by Mattioni.[1] We believe that this indicates that the model developed using representative QSAR sets was better trained and had better generalizability than the reported model. The SOM method was also compared to the sphere exclusion method of generating QSAR sets using the diversity index as the metric. The results indicated that the latter method did not lead to better QSAR models than the SOM based method. Finally, we investigated a number of holistic descriptor combinations for use in the classification stage and employed a diversity metric to study the diversity of the QSAR sets created using the SOM and different descriptor types.

After a QSAR model has been built it must be validated. The goal of validation is to test whether a model has been overfit or is due to chance factors. In addition, the validation step provides a measure of the model's predictive ability. The traditional QSAR methodology uses scrambling tests and the prediction set to perform validation. However, these methods are not able to answer the question of a how a given model will perform when faced with data that it has not seen before. Though the prediction set statistics provide some indication of the model's generalizability, it is restricted, due to the fact that it has features in common with the dataset it was taken from. Furthermore, the prediction set statistics do not provide us with a measure of confidence for predictions made for molecules that were not present in the original dataset. The technique described in Chapter 5 allows us to understand whether a new molecule will be well predicted by a model or not and thus is more general than the statistics of the prediction set. In addition the method also provides a quantitative measure of the models predictive ability for a new molecule. That is, it provides a measure of confidence

in the models's prediction. Though confidence measures can be evaluated for specific types of models (confidence bounds for linear models, frequency scores for random forest models and so on), the method described in this work is quite general so that it can be applied to any type of regression model. We first investigated an approach which tried to correlate similarity to model quality, since one would expect that a new molecule that is similar to the training set should be predicted well. However, this did not yield any conclusive results. An alternative method was based on a classification approach. The residuals from a regression model were divided into two classes, good and bad, by choosing a cutoff value. Once the residuals were divided into two classes a classifier was trained, using these class assignments. We investigated both linear and nonlinear classifiers, and a nonlinear neural network classifier exhibited the best results. The input to the neural network was the descriptor set that was used to build the original model. Once the neural network was trained, it was then used to predict the class of the residual for a new compound. We tested the method on three datasets covering both biological activities and physical properties. The results indicated that the neural network classifier was able to correctly predict the class of the residual for a new molecule 80% to 90% of the time. The neural network classifier was also able to provide a probability of class membership, which can be viewed as a confidence measure. Plots of probability of membership versus residual were created to visualize the results of the technique for the training sets. Given that similarity should play an important role in the prediction of new compounds, we also attempted to include similarity measures in the classification model. However, the results indicated that their inclusion did not significantly improve results.

Chapters 6 and 7 focused on applications of the QSAR methodology and interpretation of linear models. The first study developed linear regression and neural network models to predict and interpret the anti-malarial activity of a set of artemisinin analogs. The dataset had been studied by Avery et al.,[2] who developed a set of models using the CoMFA technique. This method is a 3-D QSAR method and is dependent on accurate alignments. We attempted to build predictive models using the 2-D ADAPT methodology. The results indicated that the models we developed compared reasonably with those reported in the original work. The R^2 for the training and prediction sets were 0.68 and 0.77 respectively, compared to R^2 values ranging from 0.82 to 0.88 for the original models. However, the neural network model that was developed showed significantly better performance with R^2 values of 0.96, 0.94 and 0.88 for the training, cross-validation and prediction sets, respectively. The linear regression model was subsequently interpreted

using the PLS technique which was able to explain how each descriptor in the model correlated to the predicted activity. The conclusions from the interpretation corresponded well with the established mode of action of the artemisinin group of anti-malarials.

The second study (Chapter 7) focused on modeling the inhibitory activity of a set of PDGFR inhibitors. This class of compounds have been studied for their role in cell signal transduction pathways. The dataset considered in this work was obtained from Pandey et al.[3] who investigated the biological activity a set of 79 piperazinylquinazolines using a phosphorylation assay. The original dataset was studied in the presence and absence of human plasma. The latter data were modeled by Khadikar et al.,[4] who restricted themselves to the use of topological descriptors and linear regression models. The study described in Chapter 7 used the data from the assays carried out in the presence of human plasma and built linear regression and neural network models using the full suite of ADAPT descriptors. The best linear model exhibited a R^2 value of 0.84 and a RMSE of 0.24. The statistics for the neural network were significantly better than for the linear model. The R^2 for the training, cross-validation and prediction sets was 0.94, 0.90 and 0.61, respectively. The study also provided an interpretation of the structure-activity trends characterized by the linear model. The main conclusions that were drawn, namely, the presence of bulky hydrophobic groups and nitrogen centers increase inhibitory activity, matched closely to observations made by Pandey and other workers for this class of compounds. The study also developed a random forest model to determine descriptor importance. The ranking generated by the random forest model corresponded closely to the descriptors present in the best linear and nonlinear models, exhibiting the ability of the feature selection algorithms to select information rich descriptor subsets.

The last two chapters described approaches to the interpretation of neural network models. Chapter 8 described a method to provide a broad interpretation of a neural network model similar in manner to the descriptor importance measures for random forest models. The method described in this work was essentially a sensitivity analysis of the network and consisted of scrambling individual descriptors and making new predictions for the training set. The result of scrambling a descriptor was to increase the RMSE of the predictions. Furthermore, the more important a descriptor was to the model's predictive ability, the larger the difference between the RMSE of the original predictions and the RMSE for the predictions obtained after scrambling that descriptor. This procedure allowed us to rank the descriptors in the model in order of importance. By analogy with the descriptor importance plots that can be created for

random forest models, the method was also able to create importance plots which were used to visualize the relative importance of descriptors in a neural network model. This method was applied to neural network models built for three datasets using the ADAPT methodology. Linear models had been previously developed for these datasets. Each linear model was interpreted using the PLS technique. The results indicated that the descriptors that were highly ranked in the neural network models were very similar in nature (and in some cases identical) to the descriptors deemed the most important from the PLS analysis of the linear models. Assuming that both types of models captured similar structure-property trends in each dataset, these results indicate that the broad interpretation method correctly identifies the descriptors that play an important role in the neural network model's predictive ability.

Though this method provides some insight into the working of a neural network model, it does not help us to understand the role played by a specific descriptor in the model. In other words, how does the network model the relationship between a given descriptor and the predicted output? This problem was addressed in Chapter 9 which described a method to provide a detailed interpretation of a neural network model, similar to the type of interpretation that is possible for linear regression models using the PLS technique. The method used only the weights and biases of the trained network and did not require the training set to develop an interpretation. The core of the method involved the linearization of the network by defining effective weights. A second feature was to order the hidden neurons using the effective weights. By considering the hidden neurons as latent variables, the interpretation used the effective weights to provide a detailed breakdown of the roles played by each descriptor in the model's predictive ability. By evaluating the contribution of each hidden neuron to the output neuron, the method allowed the user to focus on the most important hidden neurons and the most important descriptors within them. The effective weights were also used, in conjunction with the training set data, to develop score plots, by analogy with the PLS interpretation technique. These plots allowed for the easy visualization of the behavior of each hidden neuron and resulted in a very detailed, compound-wise interpretation of the structure-property trends present in the dataset, as encoded by the neural network. The technique was tested on three datasets covering biological and physical properties. First, linear models were developed for each dataset using the ADAPT methodology and then interpreted using the PLS method. Next, the descriptor subsets from the linear models were used to build neural network models. The assumption was, that given the same datasets and descriptor subsets, both linear and nonlinear models would capture the same

structure-property trends. The results indicated that the neural network interpretation matched very closely (and in some case was identical to) the interpretation of the linear models. This indicates that the interpretation method was able to extract, in a detailed fashion, the structure-property trends encoded by the neural network models for the datasets studied.

In summary, this thesis has focused on specific steps in the QSAR model development process, some of which have not been considered in detail previously. The SOM method illustrates a way to create QSAR models utilizing as much information as is available to the modeler. The problem of model performance when faced with new compounds has not been studied in detail in the cheminformatics or QSAR literature and the method described in this work represents a generalized, quantitative approach to this problem. As described in this thesis, interpretation of QSAR models greatly improves their usability. For the case of linear models, the PLS approach allows us to understand, in detail, the various structure-property trends encoded in the model. The two studies presented in this work exemplify the ease with which linear models can be dissected using this technique. In the case of neural network QSAR models, interpretability has been lacking, resulting in their reputation as black box models. The two methods discussed in this work have have been successful in alleviating this problem. The broad interpretation method has been shown to be a useful method to quickly summarize the roles played by individual descriptors in a neural network model. For a more in-depth view of the relationship between input descriptors and network output, the detailed interpretation method has been shown to be able to provide a comprehensive view of the structure-property trends encoded in the model. Together with score plots, this method allows for a detailed, compound specific analysis of the model. The method thus places neural network models on par with linear regression models in terms of interpretability. As a result of this method, 2-D QSAR studies involving both linear regression and neural network models are expected to be more comprehensive, leading to better understanding of structure-property relationships. The studies presented in this thesis have been shown to improve the quality, reliability and interpretability of the QSAR modeling process.

References

[1] Mattioni, B. E.; Jurs, P. C. Prediction of Dihydrofolate Reductase Inhibition and Selectivity Using Computational Neural Networks and Linear Discriminant Analysis. *J. Mol. Graph. Model.* **2003,** *21,* 391–419.

[2] Avery, M.; Gao, F.; Wesley, C.; Mehrotra, S.; Milhous, W. Structure-Activity Relationships of the Antimalarial Agent Artemisinin. 1. Synthesis and Comparative Molecular Field Analysis of C-9 Analogs of Artemisinin and 10-Deoxoartemisinin. *J. Med. Chem.* **1993,** *36,* 4264–4275.

[3] Pandey, A.; Volkots, D. L.; Seroogy, J. M.; Rose, J. W.; Yu, J.-C.; Lambing, J. L.; Hutchaleelaha, A.; Hollenbach, S. J.; Abe, K.; Giese, N. A.; Scarborough, R. M. Identification of Orally Active, Potent, and Selective 4-Piperazinylquinazolines as Antagonists of the Platelet-Derived Growth Factor Receptor Tyrosine Kinase Family. *J. Med. Chem.* **2002,** *45,* 3772–3793.

[4] Khadikar, P. V.; Shrivastava, A.; Agrawal, V. K.; Srivastava, S. Topological Designing of 4-Perazinylquinazolines as Anatagonists of PDGFR Tyrosine Kinase Family. *Bioorg. Med. Chem. Lett.* **2003,** *13,* 3009–3014.