Prediction of Hydroxylation Rate Constants from Molecular Structure by Cheminformatics Methods and Computational Neural Networks

Darin Vaughan

Chemistry and Mathematics

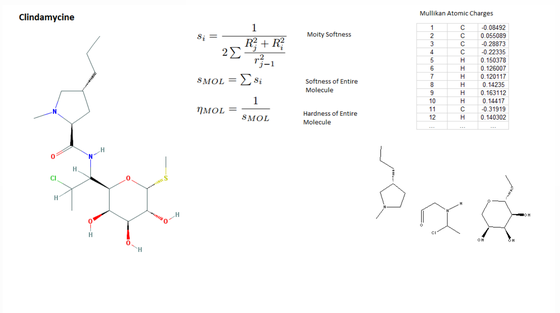
Morehead State University, Morehead, KY

Introduction and Motivation for Cheminformatics Methods

Pharmaceutical drug discovery and design has always been a complicated and time consuming process, especially for purely experimental chemists. This study presents a highly interdisciplinary approach to solving difficult problems in drug discovery through cheminformatics.

Quantitative structure-activity relationship models are a multiple linear regression technique for aiding in predicting certain properties of compounds, such as the relationship between the chemical structure of an enzyme inhibitor and its biological activity. To explore the strength and utility of QSAR methods, a model was developed relating four classes of structure descriptors for a set of unsaturated hydrocarbons which undergo hydroxylation reactions to accurately predict k[OH] rate constants. To aid in the advancement of novel and promising technology, computational neural networks (commonly named Artificial Neural Networks) and genetic algorithm optimization were employed for producing a predictive learning model which resulted in exceptionally low error.

This technique provides a strong measure of confidence for scalability into more complex molecular relationships such classification of compounds based on antibiotic activity. The time saving benefit of QSAR and cheminformatics in general is that enzyme assays and extensive reaction spectroscopy are not necessary to determine rate constants. Reaction dynamics may be simulated entirely *in silico* using such a multidisciplinary approach as this. Combining relatively accessible methods from computer and data science, quantum computational chemistry, applied mathematics, and organic and pharmaceutical chemistry, cheminformatics provides the opportunity for people from any discipline to enter an exciting field.



Example hybrid descriptor for clindamycin shows calculation of moiety electronegativity with respect to atomic radii using Gaussian 03

Experimental Methods

Four sets of chemical descriptors were calculated on 310 unsaturated alkenes using the R programming language package RCPI after significant data pre-processing. The set of descriptors include inductive, hybrid, topological, and 3D electronic structure based calculations. Preprocessing of the datasets included generating SMILES formatting for the compound set using COSMOquick, removing descriptors for which the column was primarily zero or duplicate values for more than half the column. Additionally, pairwise correlation provided a smaller set by removing dependent variables.

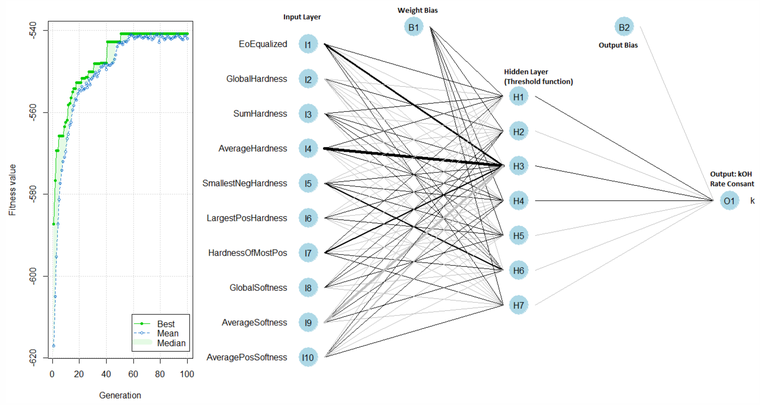
Finally, descriptor sets were optimized with a genetic algorithm to select the most efficient predictor variables within each set before being fit to a linear model with multiple least squares regression and computational neural network.

This process was repeated for a second study for classification of a set of 163 antibiotic compounds with the goal of creating a linear model which accurately assigns an antibiotic activity level from 0 to 1.

Descriptor sets considered here are:

- Topological, calculations based on the chemical graphs such as weight paths and Kier shape indices.
- Inductive, which are based on the Linear Free Energy Relationship principle.
- 3D Electronic include HOMO and LUMO values and other electrostatic potentials
- Hybrid descriptors are combinations between sets, such as electronegativity with respect to geometric indices

Genetic Algorithm and Computational Neural Network Using 'Inductive Descriptor' Set



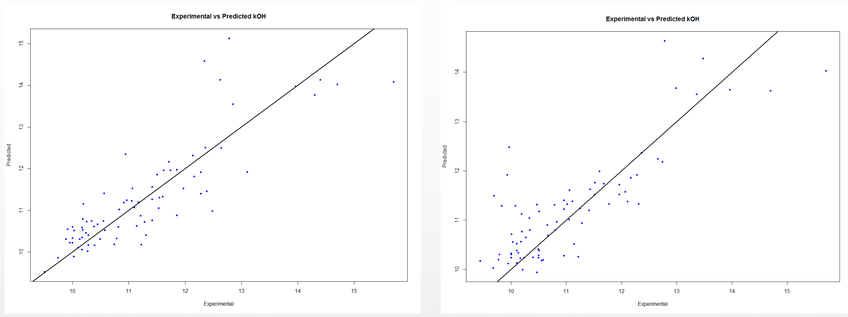
GA shows minimization convergence to 31 predictor variables from original 55. Neural network in image is scaled to 30% of experimental number of networks with weights removed for clarity

The above figures show the process of optimizing descriptor variables using a genetic algorithm and fitting a linear model with a computational neural network. A GA iteratively selects the best fitting variables over many generations of fitness tests until converging at a set of predictor variables that most efficiently estimate the regression relation of the predictor descriptors and the response variable, k[OH].

The computational neural network shown is a scaled example of the network for fitting 'inductive' descriptors and though typically used as a 'black box' method, a regression formula is obtainable just as with standard regression.

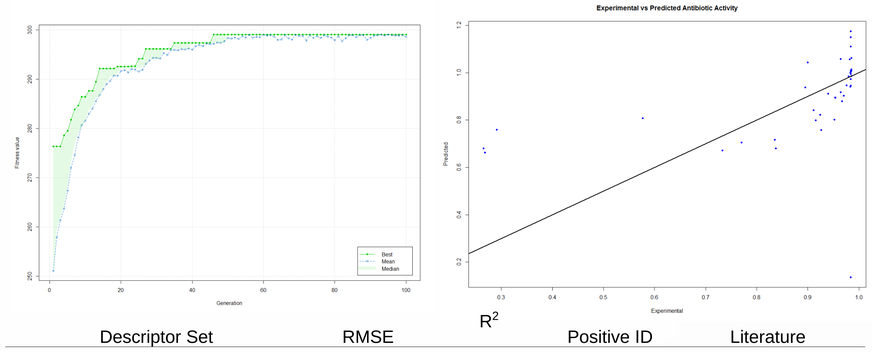
Below is an unoptimized linear model from the neural network output compared with the optimized response obtained from the GA. It is notable that some small fitting is lost when variables are cut since the removed predictors did provide information about the regression relation. This loss in fitting is acceptable, only approximately 0.02 on R^2 was typically lost.

177 Variable Unoptimized and 77 Variable Optimized k[OH] Rate Prediction Models



Unoptimized (left) shows excellent fit while the optimized model sacrifices a 0.02 points on R^2 fitting for nearly half the number of variables required to form the model

Genetic Algorithm Optimization and Linear Model of Antibiotic Activitv Classification



Descriptor Set	RMSE	R^2	Positive ID	Literature
Antibiotics				
Induction	0.091044	0.7127	97.50%	93% [2]
Topological	0.088211	0.9445	42.50%	-
3D Electronic	0.130584	0.4825	100%	-
Hybrid	0.117367	0.5374	94.90%	-
Hydroxylation				**(RMS log units)**
Induction	0.556991	0.7907		
Topological	0.443955	0.9066	-	0.139 [1]
3D Electronic	0.798652	0.5342	-	0.281 [1]
Hybrid	0.66018	0.6667	-	0.187 [1]
Optimized (Topological)				
Antibiotic (99)	0.07505	0.8768	90%	-
Hydroxylation (77)	0.434714	0.8705	-	-

Results of QSAR using each descriptor set for Antibiotic classification and Hydroxylation k[OH] rate constant prediction. Empty cells indicate values for calculations the literature studies did not consider

Results and Discussion

Linear models from the neural networks were cross checked with multiple linear regression models in both R and STATA to ensure the positive results were not due to coding errors. At least three of the descriptor sets were rank-deficient, meaning variable independence was not fully satisfied.

Topological descriptors provided the best fit and lowest error in both sets from both optimized and unoptimized data. Evolutionary optimization with the genetic algorithm successfully reduced variable size from 177 to 99 and 77 for Antibiotics and Hydroxylation studies respectively. This is important since it reduces the required calculations and amount of work by over half. RMSE error calculated was acceptable for this study.

Antibiotic activity classification successfully predicted high percentages except for the topological descriptors. This may be due to the very wide range of chemical shape and structure and the topology of a compound may not be as relevant to it's classification. Error in the data sets could also explain this since the RMSE was exceptionally low.

Ultimately, it is very clear that QSAR is an extremely valuable time and cost saving technique for *in silico* computational physical chemistry methods as well as providing excellent work for interdisciplinary chemists interested in computational drug design

Acknowledgements

- Dr. Nathan Coker, Morehead State University, Dept. Chemistry – For his continued mentorship and support
- Corey Broadwell, Morehead State University, Mathematics, Computer Science undergraduate – Programming collaboration

Resources

- [1] Gregory A. Bakken, Peter C. Jurs, **Prediction of Hydroxyl Radical Rate Constants from Molecular Structure,** J. Chem. Inf. Comput. Sci. 1999, 39, 1064-1075
- [2] Artem Cherkasov, **Inductive QSAR Descriptors. Distinguishing Compounds with Antibacterial Activity by Artificial Neural Networks,** Int. J. Mol. Sci. 2005, 6, 63-86