Inverse QSAR Analysis for Improving Predictions of Chemical Toxicity

Shawn Martin[1] (smartin@sandia.gov), John Kenneke[2], W. Michael Brown[1], and Jean-Loup Faulon[3]

[1]Sandia National Laboratories, Albuquerque, NM; [2]U.S. Environmental Protection Agency, National Exposure Research Laboratory, Ecosystems Research Division, Athens, GA; [3]Sandia National Laboratories, Livermore, CA.

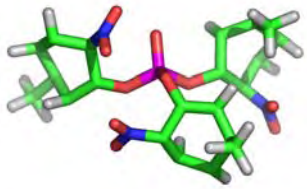
Motivation

The toxic outcomes associated with environmental contaminants are often not due to the chemical form that was originally introduced into the environment, but rather to the chemical having undergone a transformation prior to reaching the vulnerable species. More importantly, the chemical is often transformed (or metabolized) to the toxic form inside the species of interest. This situation is so common that any tool for accurately predicting toxicity must include a module that accurately predicts metabolism. In response to this need, NERL/ERD-Athens is developing a metabolic simulator in support of ORD's Computational Toxicology Program. In a joint project between EPA and Sandia National Laboratories, inverse quantitative structure activity relationships (QSARs) are being used to elucidate structural motifs that lead to activated, and potentially harmful, metabolites. Utilization of these QSARs in the design and development of the metabolic simulator will result in greater accuracy than current chemoinformatic methods, provide a source for a universal descriptor from which other descriptors could be computed, provide a means to control descriptor degeneracy, and be used to generate molecular structures (i.e., new chemicals), which will then be used to test the metabolic simulator and target areas requiring further research or data.

Inverse QSAR

Step 1: Gather Data

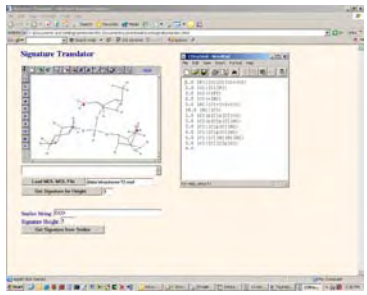
We first gather datasets containing molecules related to the problem at hand. As a first step we are using two datasets: a 27 conazole fungicide dataset with Fish ChV activities obtained from the PBT Profiler and the EPA's Fathead minnow dataset, containing 707 molecules and associated LC_{50} values. The compound below is from the Fathead Minnow database.



TRIS(5-METHYL-2-NITROPHENYL) PHOSPHATE

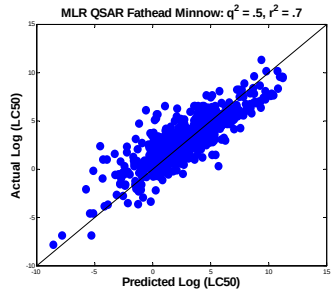
Step 2: Encode Structure using Signature

We next encode the compounds in a dataset using the *signature* molecular descriptor. Signature is a fragmental descriptor which allows reconstruction of structures and motifs similar to compounds already in the dataset. Shown below is an example of the previous compound encoded in signature format.



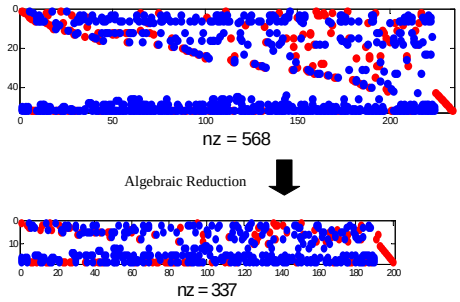
Step 3: Compute QSAR

After the compounds are encoded using signature, we using multi-linear regression (MLR) or Support Vector Regression (SVR) to correlate structure with activity. Shown below is the QSAR for the Fathead minnow dataset.



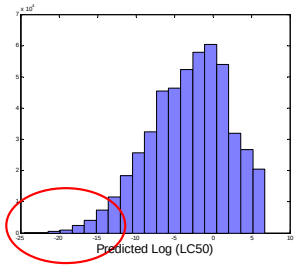
Step 4: Enumerate Motifs

Using the signature encoding of the compounds in the dataset, we can also calculated constraints which can be solved to enumerate similar structures & motifs (small substructures). These constraints are Diophantine equations that consist of integer coefficients and positive integer solutions. They are *very* difficult to solve.



Step 5: Screen Motifs

Finally, we screen the motifs enumerated in the Step 4 using the QSAR computed in Step 3. The motifs with high toxicity must now be analyzed for similarities.



Toxic Motifs

Results

We are in the process of developing the inverse QSAR method. As a proof-of-concept, we have used tested our algorithms on two datasets: a 27 conazole fungicide dataset with corresponding fish ChV values computed using EPA's PBT profiler as well as EPA's Fathead minnow database. We have trained QSARs using forward stepping MLR and signature. The final QSARs have achieved q^2 values of .65 and .5 respectively, and r^2 values of .91 and .7 respectively, in both cases indicating predictive ability. To invert the QSARs, we derived 29 equations with 91 variables for the conazoles and 52 equations with 230 variables for the Fathead minnow. Using a newly developed method, we have algebraically reduced the conazole equations 15 equations and 77 variables and the Fathead minnow equations to 18 equations and 200 variables. Although we have been unable to enumerate completely the solutions to these equations using our standard method (a Contejean-Devie solver), we have nevertheless obtained 155,340 solutions for the conazoles using a partial basis and 499,312 solutions for the Fathead minnow. Finally, we have reconstructed motifs from these solutions by screening with the QSARs. Our next step will be the application of our method to a database oriented towards metabolites.