Molecular Modeling for Predicting the Toxicity of Dioxins Furans and PolyChloro Biphenyl (PCB) Compounds

Pluton Pullumbi and Ivan Milosavljevic

Air Liquide, Centre de Recherche Claude Delorme,
1 chemin de la Porte des Loges, B.P. 126, Les Loges-en-Josas, 78354 Jouy-en-Josas, France

Introduction

In the present study we have used the QSAR (Quantitative Structure/Activity Relationships) approach to characterize the relative toxicity of a number of halogen substituted dibenzo-p-dioxins, dibenzo-furans and PolyChloro Biphenyls (PCB). These compounds range in toxicity from virtually none to very high as, e.g., that of 2,3,7,8 tetrachlorodibenzo-p-dioxin (tcdd). For many classes of molecules, it is now possible to carry out studies aimed at researching the relations between the molecule's structure and its biological activity. QSAR (Quantitative Structure/Activity Relationships) studies take into consideration the interaction between the molecule under examination and its corresponding biological receptor. The ultimate aim is to connect toxicity and pharmacodynamics, consequences of receptor-binding, to the molecular structure. In practice, the method attempts to demonstrate that molecule **A** is toxic because it has a given structure, that molecule **B** is less dangerous because it has a slightly different structure, while molecule **C** is the least dangerous because it has yet another structure. In the present work we focus on establishing reliable correlations between structural features or molecular properties and the toxicity of this class of compounds.

QSAR-based Method

Theoretical descriptors for each molecule in the database are calculated using the computer program of MSI as a first step. We have used 47 descriptors based on quantum calculations, topological, information-theoretic and graph theoretic analysis. The Principal Component Analysis (PCA) has been then used to assess the intrinsic dimensionality of the problem and extract the components (linear combinations of the descriptors) that explain most of the variance in the original data. Properties dependent on topology, stereochemical configuration, and charge distribution (such as lipophilicity, dipole moment and hydrophobic moment) have been added (as appropriate) to the descriptor list . Cross-validation is performed by dividing the input dataset into several distinct training and test subsets such that each training set covers all of the substituent positions and represents the structural diversity in the original dataset.

Information-rich structure descriptors are key to meaningful QSAR models. In models for predicting toxicity solely from molecular structure, an effective numerical representation of molecular structure is extremely important. From the analysis of the processes leading to a toxic response, it can be rationalized that the structure descriptors should be able to quantify transport, bulk, and electronic attributes of molecular structure. A number of theoretically calculated and experimentally measured property values have been employed to numerically encode these structural features. The graph theoretic and information theoretic indices used in our analysis are derived from the adjacency matrix and distance matrix of a chemical graph . The Genetic Function

Approximation (GFA) algorithm has been used to build the structure-activity models. GFA automates the search for QSAR models by combining a genetic algorithm with statistical modeling tools. Thousands of candidate models are created and tested during evolution; only the superior models survive which are then used as "parents" for the creation of the next generation of candidate models. GFA is the procedure of choice when the data set contains many more descriptors than samples, when it is desired to select among competing correlated descriptors, or when it is suspected that there may be nonlinear relationships in the data. In these cases GFA can rapidly point out the most information-rich combinations of features, and can expose patterns in the data set that may otherwise remain hidden.

In principle, a QSAR-based model is a quantitative relationship between a numerical measure of toxicity and structure descriptors, i.e.,

$$\mathbf{T = f\,(S)}$$

where **T** is a measure of toxicity and **S** is a set of numerical quantities representing different structural attributes (**f** being a mathematical function). The structure may be quantified at any level of complexity ranging from a mere count of certain atoms or groups to sophisticated quantum mechanical indices, and a variety of methods ranging from linear multiple regression analysis to neural networks are available to determine the explicit form of the function **f**. These structure-toxicity relationships are generally called quantitative structure-toxicity relationship (QSTR) models or equations, because by knowing **f** and providing the values of **S** for any compound one could estimate **T**.

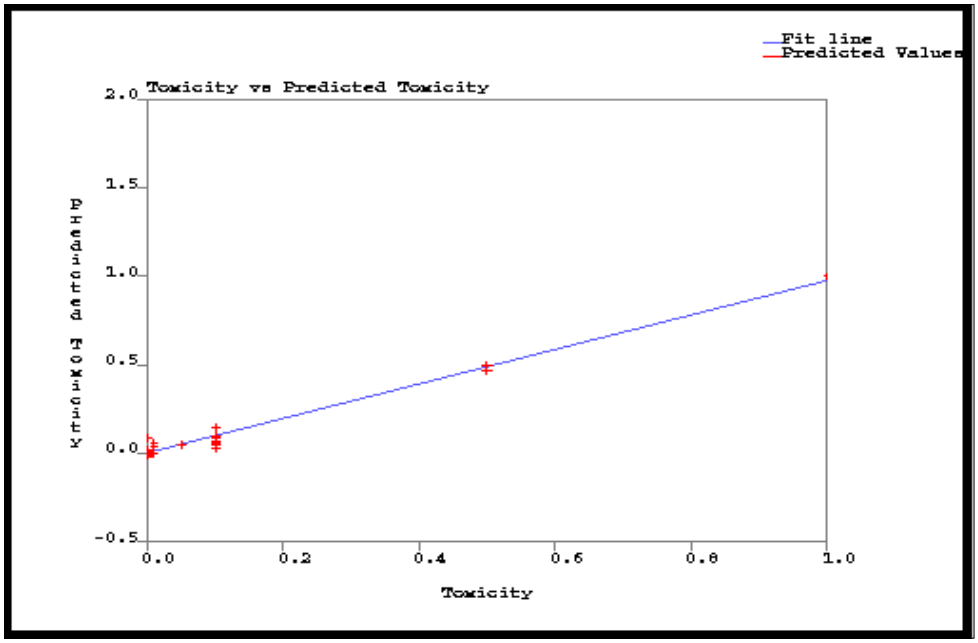

Figure 1: Toxicity vs Predicted Toxicity (Equation 1)

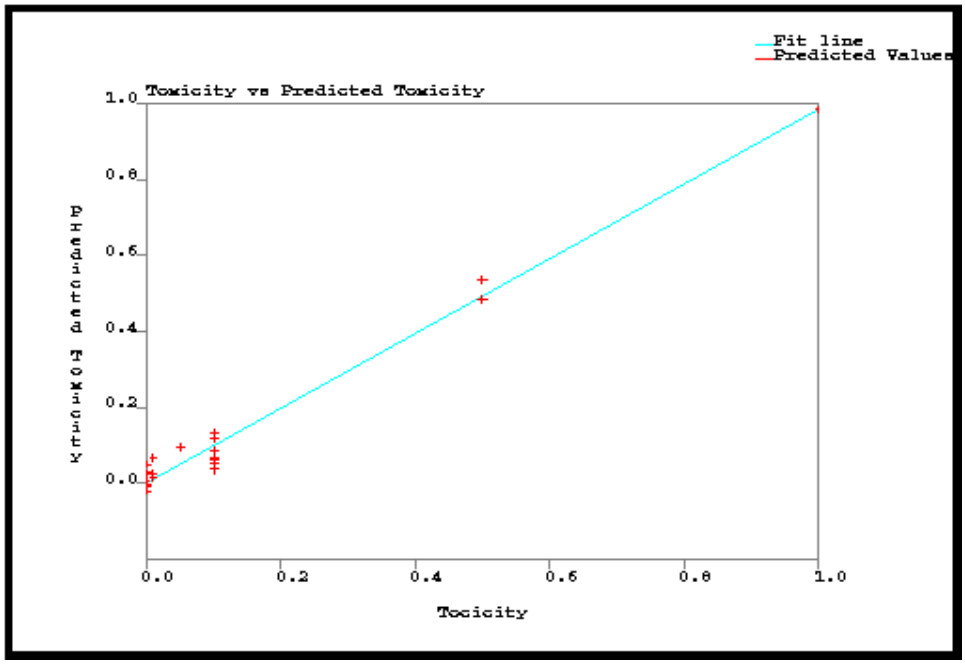

Figure 2: Toxicity vs Predicted Toxicity (Equation 2)

In figures 1 and 2 are reported two of the correlation examples established from a set of about 100 GFA-generated equations. In both figures we report the predicted toxicity as a function of the experimentally derived toxicity are reported. They were obtained by using the following equations:

$$\text{Toxicity} = 1.14482 + 3.60596(\text{HOMO\_MOP} + 9.09366) - 0.013296\text{MolRef} + 10.3424(\text{JURS\_RPCS} - 0.722871)^2 - 0.034864[(\text{CHI\_2}) - 9.59104]^2 \quad (1)$$

$$\text{Toxicity} = -0.016623 + 12.6668(\text{JURS\_RPCS} - 0.722871)^2 + 4.05871(\text{HOMO\_MOP} + 9.09366) + 0.038791\text{Sr} \quad (2)$$

We have carried out internal validation of the data set from which the model was derived and checked for internal consistency. The procedure uses a reduced set of structural data to test the predicted correlations. The new model is used to predict the toxicities of the molecules that were not included in the reduced-model set. This was repeated until all compounds have been deleted and predicted once. It is well known that the internal validation is less rigorous than the external one. As more experimental toxicity data for other compounds of this series was not available, we used the bromine substitution of one of the chlorine atoms of the training set for an preliminary external validation by supposing that such substitutions would not dramatically affect the toxicity of the conformers. The error of data fit for our regression model can be estimated by the r^2 values (square of the correlation coefficient) $r^2 = 0.995$ and 0.993 respectively for two relations. An analysis of the descriptors that appear in these relations indicates that the toxicity is related to the

topology of the compound but also to its electronic properties. These descriptors are detailed in the following.

HOMO_MOP: The value of the Highest Occupied Molecular Orbital calculated using the MOPAC program which is based on a semi-empirical quantum method. This descriptor is related to the ability of the molecule to be an electron donor.

JURS_RPCS: This set of JURS descriptors (Stanton and Jurs 1990) combines shape and electronic information to characterize the molecules. The descriptors are calculated by mapping atomic partial charges on solvent-accessible surface areas of individual atoms. A total of 30 different descriptors are included in the set. The Relative Positive Charge Surface (RPCS) descriptor calculates the solvent-accessible surface area of the most positive atom divided by descriptor.

MolRef: The molar refractivity is a molecular descriptor that can be used to relate chemical structure to observed chemical behavior. It is a combined measure of the size of a group and its polarizability.

CHI_N descriptors: The molecular connectivity index of order *n* corresponding to subgraph type *s* is denoted by ${}^{n}\chi_{s}$. Given an order *n* and a subgraph type *s* one considers all connected subgraphs of type *s* consisting of *n* edges. This is a structural-topological type of descriptor.

Sr (Superdelocalizability): One could interpret this as measure of the delocalizability of the electrons of the system: inner electrons that are tightly held are not very delocalizable. On the other hand, the upper occupied states (especially HOMO), i.e. the electrons in the higher-energy orbitals, are less tightly bound, which means that they are relatively delocalizable. Therefore the upper energy levels will dominate the Superdelocalizability term. The sum of *Sr* for all atomic positions of a molecule gives a metric of electrophilicity, which may be used to predict relative reactivity in a series of molecules.

Conclusions

Toxicity of dioxins and furans has been correlated to structural-electronic features of these compounds using the GFA algorithm within the QSAR approach. The training set has been enlarged with PCB's whose e toxicity is known experimentally.

Basic correlation was established. It was shown that the correlation is less satisfactory for very low toxicities. It is suggested that the correlations can be improved by using more detailed experimental data as well as refining the computation. These improvements will be the subject of future communication.