

## QSAR – Quantitative Structure-Activity Relationships and Drug Discovery

### References:

Chapter 16 in *Chemoinformatics in Drug Discovery* edited by Oprea

Chapter 11 in *Modern Methods in Drug Discovery* by Hillisch and Hilgenfeld

“An Introduction to QSAR Methodology” by A. B. Richon and S. S. Young  
<http://www.netsci.org/Science/Compchem/feature19.html>

“QSAR and Drug Design” by D. R. Bevan  
<http://www.netsci.org/Science/Compchem/feature12.html>

### What is QSAR?

The goal of quantitative structure-activity relationship (QSAR) studies are to relate the structural, chemical, physical and other properties of a compound to his biological activity. These physiochemical properties include hydrophobicity, topology, electrical properties, steric effects as well as others. The properties are described quantitatively so that a scoring function can be derived to judge the suitability of a compound as a potential therapeutic agent.

### History of QSAR

Early QSAR related studies date back over 100 years. In 1863, A. F. A. Cros at the University of Strausbourg that alcohol toxicity was inversely correlated with the solubility of the alcohol in water. In 1899, H. H. Meyer published a paper that described that the toxicity of organic compounds depended on their lipophilicity. This suggested a relation between solvent partitioning and biological activity.

L. Hammet correlated the electronic properties of organic acids and bases with their equilibrium properties. He found that a linear relationship resulted when substitutions of different groups were made to aromatic compounds.

$$\log \frac{K}{K_0} = \rho \log \frac{K'}{K'_0} = \rho \sigma$$

$K_0$  and  $K'_0$  are equilibrium constants for unsubstituted compounds and  $K$  and  $K'$  are the equilibrium constants for substituted compounds. He used benzoic acid as reference compound yielding the  $\sigma$  seen at the far right. To interpret this equation, if the linear relation defines  $\rho > 1$ , then the effect of the substitutions is greater than making the same substitutions on benzoic acid. The  $\sigma$  describes the properties of the substitution groups. If  $\sigma$  is positive, the group is electron withdrawing. If  $\sigma$  is negative, the group is electron donating. The magnitude of  $\sigma$  indicates the degree of these effects.

The modern field of QSAR was established by Hansch, Fujita and co-workers in the 1960's. In their approach presented in a publication in 1964, whole molecule properties such as 1-octanol/water partitioning coefficients, molar refractivity (a measure of molar volume), shape, and topology indices for groups of related compounds were correlated with biological activity measurements using statistical techniques.

Hammett's equation can be used to predict the inhibition of bacterial growth by sulfonamides. Assume that various substitutions were made on one of the benzene rings and the minimal amount of compound that inhibited *E. Coli* growth was observed (C). A QSAR can be made based on the  $\sigma$  values of the substitution groups

$$\log\left(\frac{1}{C}\right) = 1.05\sigma - 1.28$$

This relation indicates that electron-withdrawing group substitution groups inhibit growth. It is important to note that the addition of substitution groups is not necessarily an additive phenomenon when it comes to biological activity. For example, addition of 4-chloro or a 3-fluor might each increase activity, but addition of both might not be more potent than either single addition.

Unfortunately, this approach does not lead to good drug candidates as it does not account for the delivery of the drug to a target. It turns out the lipophilicity is an important indicator of bioavailability as suggested by Hansch. The compound must be able to pass through cell membranes, however, it must not get stuck there. Hence the lipophilicity cannot be too low or too high. A better equation might be

$$\log\left(\frac{1}{C}\right) = k_1\sigma + k_2\pi + k_3$$

$$\pi = \log\left(\frac{P_X}{P_H}\right)$$

Where  $P_X$  is the partition coefficient of the compound and  $P_H$  is the coefficient for a reference compound.

Another example of QSAR involves a series of 1-(X-phenyl)-3,3-dialkyl triazenes (R-N=N-Ph-4,X) that were of interest as anti-tumor drugs. A drawback was that they were also mutagenic. The goal of QSAR in this case is to understand how the mutagenicity can be reduced without reducing the anti-tumor activity. The mutagenic activity was observed by the Ames test in which C is the molar concentration of the drug needed to cause 30 revertants in  $10^8$  bacteria. This yields the following equation"

$$\log\left(\frac{1}{C}\right) = 1.09 \log P - 1.63\sigma^* + 3.06$$

where  $\sigma^*$  is the “through resonance”, an electronic parameter and P is the partition coefficient. Studies of the anti-tumor activity against L1210 leukemia cells in mice yielded the following equation:

$$\log\left(\frac{1}{C}\right) = 0.10\log P - 0.042(\log P)^2 - 0.31\sigma^* - 0.18MR + 0.39E_sR + 4.12$$

Where C is the molar concentration producing a 40% increase in the mice lifespan, MR is molar refractivity, and  $E_sR$  is a steric parameter for the R group. The results are summarized in the table

4-X substitution group	Mutagenicity (1/C)	Antitumorigenicity (1/C)
-H	5.75	3.58
-SO <sub>2</sub> HN <sub>2</sub>	3.15	4.48

The results indicated that the sulfonamide group reduced antitumor activity only 1.2 fold while it reduced the tumorigenicity 400 fold.

The question arises – How do they choose compounds. In cases like this, many compounds would be tested by QSAR and the ones that satisfied the criteria would be studied further. Drug companies maintain databases with many of the parameters used in QSAR stored for the different compounds. Hence, a large number of compounds might be screened by QSAR quite rapidly.

### Modern QSAR

Up until now we have discussed traditional QSAR, now termed 2D QSAR. The QSAR relations were determined using logic about important parameters and had to rely on statistical correlations of structural descriptors with biological activities. There were some significant limitations with using the QSAR relations developed to predict better drugs. Today, 3D QSAR uses model computational methods and technologies such as statistical correlation, machine learning, and 3D visualization, yielding rational or computer assisted drug design. There are software products that enable the design and construction of the QSAR models in an optimal fashion.

### 3D QSAR

In most 3D QSAR methods a molecule of the compound of interest is put in a cage of test atoms at predefined grid points. A map value is calculated at each test atom. This map value consists of a quantitative measure the interaction of the test atom (for a given property) with each of the atoms of the compound of interest. For example, three possible properties used by the comparative molecular field analysis (CoMFA), an early commonly used and accepted method, are the steric ( $S_i$ ), electrostatic ( $E_i$ ), and lipophilic ( $L_i$ ) potentials given as

$$S_t = \sum_i \varepsilon_{it} [1.0 / p_{it}^{12} - 2.0 / p_{it}^6], p_{it} = r / (R_i + R_t)$$

$$E_t = 332.17 \sum_i Q_i Q_t / D_{it} r$$

$$L_t = \sum_i a_i S_i e^{-r}$$

where  $\varepsilon_{it}$  is the van der Waals constant,  $r$  is the distance between the test atom (t) and the molecular atom (i),  $R_i$  and  $R_t$  are the van der Waals radii, 332.17 is a unit conversion constant,  $Q_i$  and  $Q_t$  are the partial atomic charges,  $D_{it}$  is the dielectric function for i and t,  $a_i$  is the hydrophobic atom constant and  $S_i$  is the solvent accessible surface area for i. The “field” around the molecule are more important than the molecular structure, so grid points within the van der Waals structure of the molecule are omitted. This is important because if an atom in the molecule and gridpoint are in the same location, there would be a divide by zero. Also, as we are looking at molecular binding (receptor or binding site recognition) the three properties modeled are more important than the positions of individual atoms.

There are other ways of implementing these types of calculations in different methods such as the HASL or CoMSIA (Comparative Molecular Similarity Indices Analysis). In HASL, each grid point within the van der Waals surface of the molecule are evaluated noting hydrophobes, hydrogen bond donors, and hydrogen bond acceptors as these are important in docking. In CoMSIA, the steric, electrostatic, hydrophobic, hydrogen bond donor, and hydrogen bond acceptor fields are calculated.

These methods just described generate an array of these descriptors. These data are assumed to be linear so that statistical techniques such as multilinear regression, partial least squares, genetic algorithms, etc can be used to characterize the descriptor data and relate it to the biological activity. This is referred to as the “alignment problem”. Once this is done it becomes possible to compare the different molecules.

### **The alignment problem**

One of the purposes of performing QSAR is so that different compounds can be compared with respect to a desired biological activity. The most important step is the alignment of the molecule in space. The chosen alignment of the molecule will effect how the molecule interacts with the test atoms. Hence, it is extremely important to be consistent about how the different molecules are aligned as this will greatly affect the predictive power of the method. For that reason, the approximation of a good structure for the molecule is essential. This is possible from X-ray crystal structure, NMR-based structure prediction, as well as computational methods such as secondary structure prediction and molecular dynamics. This is easier when two molecules have a similar core structure. However, if the core structures are significantly different a different approach must be used. It then is necessary to develop a “pharmacore hypothesis” i.e. identify a set of features common to the different molecules. One possibility is to look at common interaction points within the receptor, which would be defined by the presence

of hydrogen bond donors or acceptors, positions of hydrophobic residues, the presence or absence of heteroatoms, etc. For the methods to work each molecule has to have a similar set of descriptors. Typically a set of differing pharmacore hypotheses will be developed and implemented in a study.

### **Example: cannabanoids**

The compound  $\Delta$ -9-tetrahydrocannabinoid ( $\Delta$ -9-THC) is the principal hallucinogenic component of cannabis. Cannabanoids were studied by a Pfizer research group in the mid 1980's as an analgesic (pain killer) that might not have the adverse side effects of opiates. CP55,940 is a cannabanoid developed by this group. Cannabanoids are currently being used as an anti-emetic in chemotherapy and as an appetite stimulant for AIDS patients. There is a possibility that they also might help as a neuroprotective agent in multiple sclerosis. The problem with cannabanoids, depending upon your perspective, is the hallucinogenic effects. It was possible to develop a QSAR model using the pharmacore hypothesis for testing these compounds.

### **QSAR Model Development: Validation with Experimental Studies**

QSAR model development has three steps: 1) data preparation, 2) data analysis, and 3) model validation. Data preparation includes the selection of the dataset to be used. This may include any additional experimental studies needed or simply the extraction of data from a database. This also includes the calculation of molecular descriptors described above.

The second step is data analysis. It includes the selection of the techniques for statistical analysis and correlation techniques. The correlation techniques are linear such as partial least squares analysis or non-linear such as artificial neural network models. There are many different software packages as mentioned above. For this part, a set of compounds whose biological activity is known is needed to train the method.

The final step is the model validation. This involves the evaluation of the predictive power of the model. This can be done by using the model to predict the biological activity of an independent set of compounds whose biological activity is known. By independent, we mean that the set is different from the training set. Often the set of compounds whose biological activity is known is small. In this case, it might be necessary to divide the data into a series of training and test sets with which to test the model.

Often for the model validation procedure, a validation correlation coefficient  $q^2$  is calculated

$$q^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$$

where  $y_i$ ,  $\hat{y}_i$ , and  $\bar{y}_i$  are the actual, estimated and average activities, respectively. These calculations are performed over the training set. When a high correlation is obtained it suggests the predictive power. Many researchers stop here. However, to properly validate the model, it should be used on the test set.