Statistical concepts in QSAR.

Computational chemistry represents molecular structures as a numerical models and simulates their behavior with the equations of quantum and classical physics. Available programs enable scientists to easily generate and present molecular data including geometries, energies and associated properties (electronic, spectroscopic and bulk). The usual paradigm for displaying and manipulating these data is a table in which compounds are defined by individual rows and molecular properties (or descriptors) are defined by the associated columns. A QSAR attempts to find consistent relationships between the variations in the values of molecular properties and the biological activity for a series of compounds so that these "rules" can be used to evaluate new chemical entities.

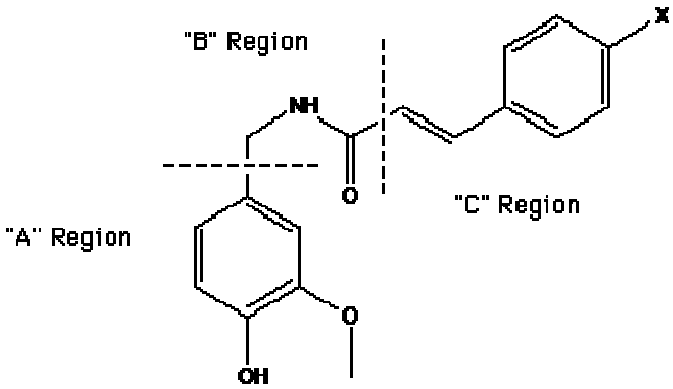
A QSAR generally takes the form of a linear equation:

$$\text{Biological Activity} = \text{Const} + (c_1 \times P_1) + (c_2 \times P_2) + (c_3 \times P_3) + \ldots$$

where the parameters P_1 through P_n are computed for each molecule in the series and the coefficients c_1 through c_n are calculated by fitting variations in the parameters and the biological activity. Since these relationships are generally discovered through the application of statistical techniques, a brief introduction to the principles behind the derivation of a QSAR follows.

The work reported from The Sandoz Institute for Medical Research on the development of novel analgesic agents can be used as an example of a simple QSAR. In this study, vanillylamides and vanillylthioureas related to capsaicin were prepared and their activity was tested in an in vitro assay which measured $^{45}Ca^{2+}$ influx into dorsal root ganglia neurons. The data, which was reported as the EC_{50} (μM), is shown in Table 1 (note that compound 6f is the most active of the series).

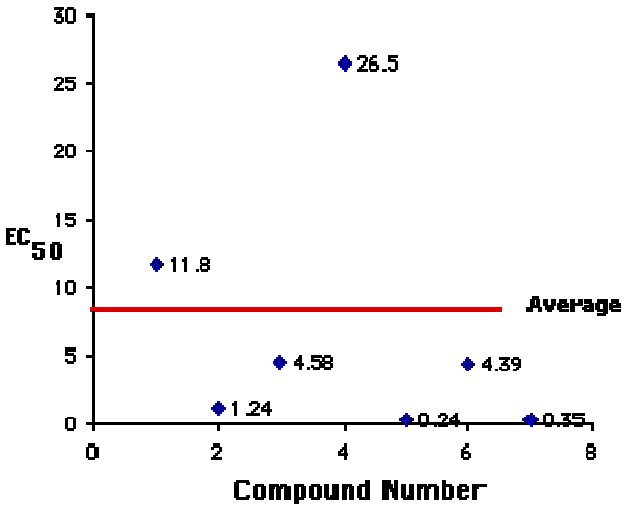
TABLE 1: Capsaicin Analogs Activity Data



Cmpd Number	Cmpd Name	X	EC_{50}(μM)
1	6a	H	11.80 ± 1.90
2	6b	Cl	1.24 ± 0.11
3	6d	NO_2	4.58 ± 0.29
4	6e	CN	26.50 ± 5.87
5	6f	C_6H_5	0.24 ± 0.30
6	6g	$N(CH_3)_2$	4.39 ± 0.67
7	6h	I	0.35 ± 0.05
8	6i	NHCHO	???

In the absence of additional information, the only way to derive a best "guess" for the activity of 6i is to calculate the average of the values for the current compounds in the series. The average, 7.24, provides a guess for the value of compound 8 but, how good is this guess? The graphical presentation of the data points is shown in Graph 1.

GRAPH 1: Capsaicin Analogs Activity Data.



The standard deviation of the data, s, shows how far the activity values are spread about their average. This value provides an indication of the quality of the guess by showing the amount of variability inherent in the data. The standard deviation is calculated as shown below.

$$s = \sqrt{\frac{(11.8 - 7.24)^2 + (1.24 - 7.24)^2 + (\ldots)^2}{7 - 1}}$$

$$s = \sqrt{\frac{539.41}{6}} = 9.48$$

Rather than relying on this limited analysis, one would like to develop an understanding of the factors that influence activity within this series and use this understanding to predict activity for new compounds. In order to accomplish this objective, one needs:

- binding data measured with sufficient precision to distinguish between compounds;
- a set of parameters which can be easily obtained and which are likely to be related to receptor affinity;
- a method for detecting a relationship between the parameters and binding data (the QSAR) and
- a method for validating the QSAR.

The QSAR equation is a linear model which relates variations in biological activity to variations in the values of computed (or measured) properties for a series of molecules. For the method to work

efficiently, the compounds selected to describe the "chemical space" of the experiments (the training set) should be diverse. In many synthesis campaigns, compounds are prepared which are structurally similar to the lead structure. Not surprisingly, the activity values for this series of compounds will frequently span a limited range as well. In these cases, additional compounds must be made and tested to fill out the training set.
The quality of any QSAR will only be as good as the quality of the data which is used to derive the model. Dose-response curves need to be smooth, contain enough points to assure accuracy and should span two or more orders of magnitude. Multiple readings for a given observation should be reproducible and have relatively smaller errors. The issue being addressed is the signal-to-noise ratio.
The variation of the readings obtained by repeatedly testing the same compound should be much smaller than the variation over the series. In cases where the data collected from biological experiments do not follow these guidelines, other methods of data analysis should be utilized since the QSAR models derived from the data will be questionable.
Once biological data has been collected, it is often found that the data is expressed in terms which cannot be used in a QSAR analysis. Since QSAR is based on the relationship of free energy to equilibrium constants, the data for a QSAR study must be expressed in terms of the free energy changes that occur during the biological response. When examining the potency of a drug (the dosage required to produce a biological effect), the change in free energy can be calculated to be proportional to the inverse logarithm of the concentration of the compound.

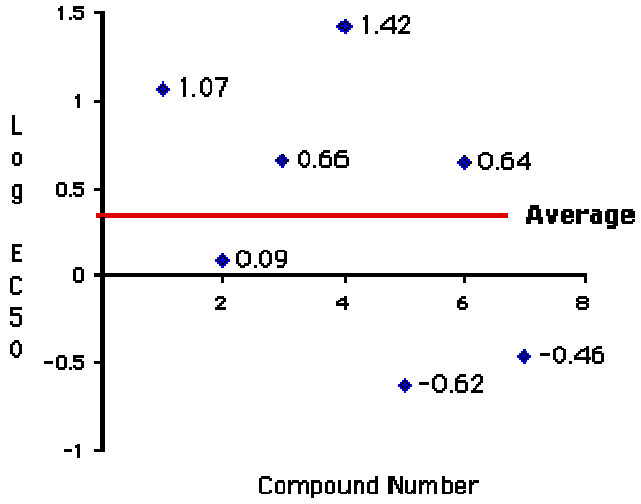
$$\Delta G_0 = - 2.3RT\log K = \log 1/[S]$$

Further, since biological data are generally found to be skewed, the log transformation moves the data to a nearly normal distribution. Thus, when measuring responses under equilibrium conditions, the most frequent transformation used is to express concentration values (such as IC50, EC50, etc.) as log[C] or log 1/[C]. The transformed data for the capsaicin agonists are shown in Table 2.

TABLE 2: Capsaicin Analogs Transformed Data

Cmpd Number	Cmpd Name	X	EC_{50}	Log EC_{50}	Log 1/EC_{50}
1	6a	H	11.80 ± 1.90	1.07	-1.07
2	6b	Cl	1.24 ± 0.11	0.09	-0.09
3	6d	NO_2	4.58 ± 0.29	0.66	-0.66
4	6e	CN	26.50 ± 5.87	1.42	-1.42
5	6f	C_6H_5	0.24 ± 0.30	- 0.62	0.62
6	6g	$N(CH_3)_2$	4.39 ± 0.67	0.64	-0.64
7	6h	I	0.35 ± 0.05	- 0.46	0.46
8	6i	NHCHO	?? ± ??	??	??

GRAPH 2: Capsaicin Analogs Transformed Data



Given the transformed data, our best guess for the activity of 6i is still the average of the data set (or 0.40). As before, the error associated with this guess is calculated as the square root of the average of the squares of the deviations from the average.

$$s = \sqrt{\frac{(1.07 - 0.40)^2 + (0.09 - 0.40)^2 + (\ldots)^2}{7-1}}$$

$$s = \sqrt{\frac{3.4906}{6}} = 0.76$$

This is an example data set intended to show the general approach; real data sets would have many more compounds and descriptors. Since the purpose of a QSAR is to highlight relationships between activity and structural features, we would like to find one or more structural features which relate these molecules and their associated activity. Additionally, we would like to find a parameter that works consistently for all of the molecules in the series.

There are several potential classes of parameters used in QSAR studies. Substituent constants and other physico-chemical parameters (such as Hammett sigma constants) measure the electronic effects of a group on the molecule. Fragment counts are used to enumerate the presence of specific substructures. Other parameters can include topological descriptors and values derived from quantum chemical calculations.

The selection of parameters is an important first step in any QSAR study. If the association between the parameter(s) selected and activity is strong, then activity predictions will be possible. If there is only weak association, knowing the value of the parameter(s) will not help in predicting activity. Thus, for a given study, parameters should be selected which are relevant to the activity for the series of molecules under investigation and these parameters should have values which are obtained in a consistent manner.

The Sandoz group divided their analysis of capsaicin analogs into three regions: the A-region which was occupied by an aromatic ring; the B-region which was defined by an amide bond; and the C-region which was occupied by a hydrophobic side-chain (See figure in Table 1). The hypothesis for the C-region assumed that a small, hydrophobic substituent

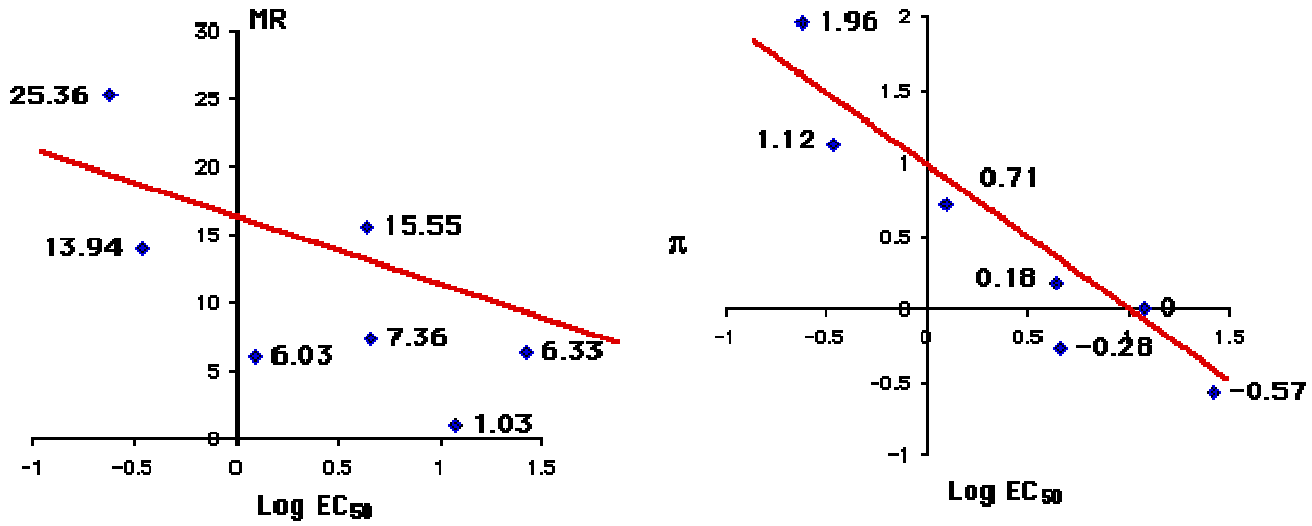
would increase activity. Given this assumption, the parameters selected to best define this characteristic were molar refractivity (size) and , the hydrophobic substituent constant. These values are given in Table 3.

TABLE 3: Capsaicin Analogs Parameter Values

Cmpd Number	Cmpd Name	X	Log EC_{50}	π	MR
1	6a	H	1.07	0.00	1.03
2	6b	Cl	0.09	0.71	6.03
3	6d	NO_2	0.66	- 0.28	7.36
4	6e	CN	1.42	- 0.57	6.33
5	6f	C_6H_5	- 0.62	1.96	25.36
6	6g	$N(CH_3)_2$	0.64	0.18	15.55
7	6h	I	- 0.46	1.12	13.94
8	6i	NHCHO	??	??	??

The data above can be analyzed for relationships by two means: graphically and statistically. The most visual approach to a problem with a limited number of variables is graphical. In this case, a plot of activity versus either molar refractivity or hydrophobicity gives some insight into the relationship between the parameters and activity. The plots derived by the Sandoz group are reproduced in Graph 3.

GRAPH 3: Capsaicin Analogs Parameter Values



Does the graph provide insight into the activity for compound 6i? Does knowing the value for either the hydrophobicity or molar refractivity parameters for this compound provide a good estimate for activity?

Since this is a simple example where only two values are examined, the answers to these questions are a qualified yes. In more complex situations however, where multiple parameters are correlated to activity, statistics is used to derive an equation which relates activity to the parameter set. The linear equation which defines the best model for this set of data is

$$\text{Log } EC_{50} = 0.764 - (0.817)\pi$$

How much confidence should we place in this model? The first step to answering this question is to determine how well the equation predicts activities for known compounds in the series. The equation above estimates the average value for the EC_{50} based on the value for; because assays vary, it is not surprising that individual values will differ from the regression estimate. The difference between the calculated values and the actual (or measured) values for each compound is termed the residual from the model. The calculated values for activity and their residuals (or the errors of the estimate for individual values) are shown in Table 4.

TABLE 4: Capsaicin Analogs Calculated Values

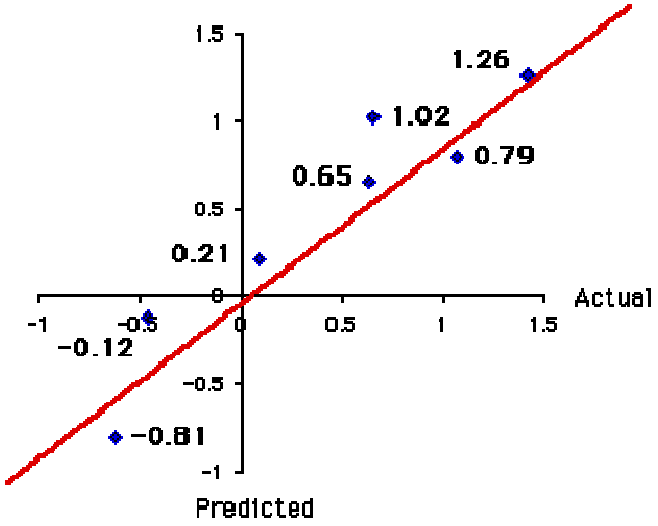
Cmpd Number	Cmpd Name	X	Log EC_{50}	π	Calculated Log EC_{50}	Residual
1	6a	H	1.07	0.00	0.79	0.28
2	6b	Cl	0.09	0.71	0.21	- 0.12
3	6d	NO_2	0.66	- 0.28	1.02	- 0.36
4	6e	CN	1.42	- 0.57	1.26	0.16
5	6f	C_6H_5	- 0.62	1.96	- 0.81	0.19
6	6g	$N(CH_3)_2$	0.64	0.18	0.65	- 0.01
7	6h	I	- 0.46	1.12	- 0.12	- 0.34
8	6i	NHCHO	??	- 0.98	1.60	??

The residuals are one way to quantify the error in the estimate for individual values calculated by the regression equation for this data set. The standard error for the residuals is calculated by taking the root-mean-square of the residuals (in this calculation, the denominator shown as decremented by two to reflect the estimation of two parameters).

$$S = \sqrt{\frac{0.28^2 + (-0.12)^2 + (-0.36)^2 + \ldots + (-0.34)^2}{(7-2)}} = 0.28$$

In order to be an improved model, the standard deviation of the residuals calculated from the model should be smaller than the standard deviation of the original data. The standard error about the mean was previously calculated to be 0.76 whereas the standard error from the QSAR model is 0.28. Clearly, the use of linear regression has improved the accuracy of our analysis. The plot of measured values versus calculated is shown in Graph 4 with a 45° line.

GRAPH 4 Capsaicin Analogs Predicted Versus Actual EC50 Values



There are several assumptions inherent in deriving a QSAR model for a series of compounds. First, it is assumed that parameters can be calculated (or measured in some cases) more accurately and cheaply than activity can be measured. Second, it is assumed that deviations from the best fit line follow a normal (Gaussian) distribution. Finally, it is assumed that any variation in the line described by the QSAR equation is independent of the magnitude of both the activity and the parameters. Given these assumptions, the quality of the model can be gauged using a variety of techniques.

Variation in the data is quantified by the correlation coefficient, *r*, which measures how closely the observed data tracks the fitted regression line. Errors in either the model or in the data will lead to a bad fit. This indicator of fit to the regression line is calculated as:

$$r^2 = \frac{\text{Sum-of-Squares of the deviations from the regression line}}{\text{Sum-of-Squares of the deviations from the mean}}$$

$$r^2 = \frac{\text{Regression Variance}}{\text{Original Variance}}$$

where the "Regression Variance" is defined as the "Original Variance" minus the Variance around the regression line. The Original Variance is the sum-of-the-squares distances of the original data from the mean. This can be viewed graphically as shown in Graph 5.

The calculation is carried out as follows:

Original Variance = $(1.07 - 0.40)^2 + (0.09 - 0.40)^2 + \ldots$
Original Variance = 3.49

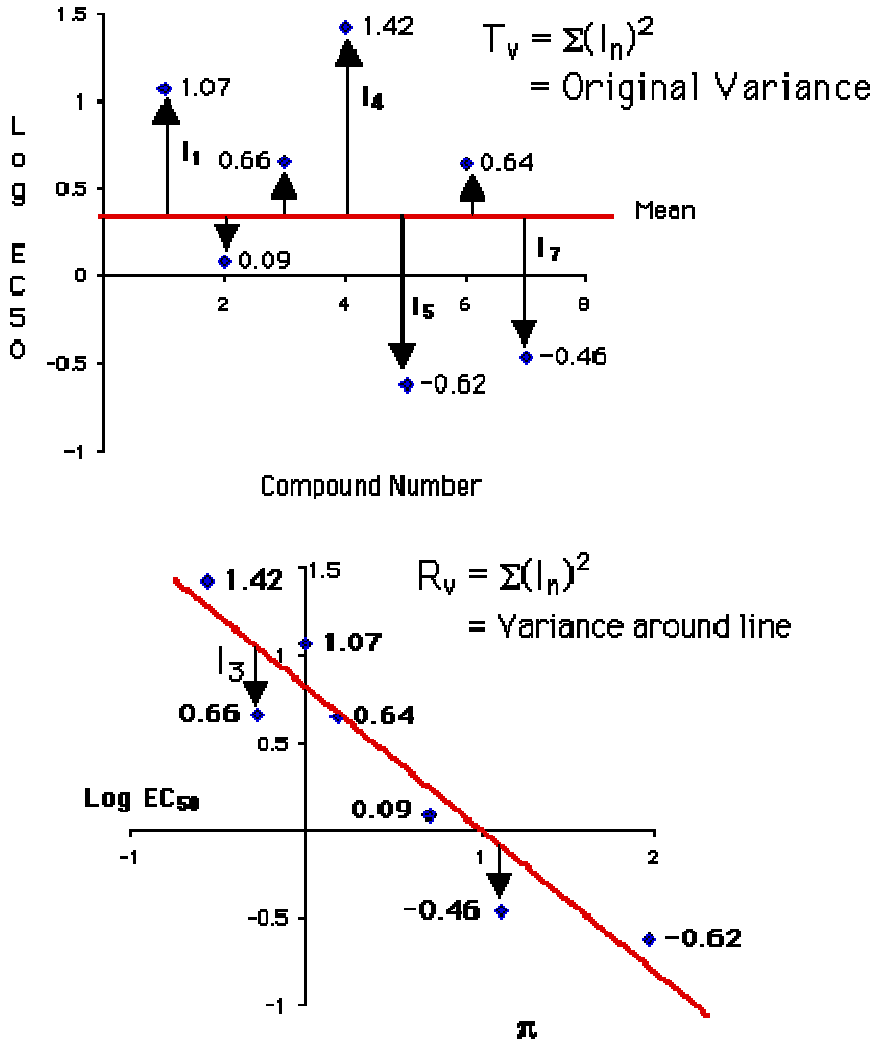
Variance around the line = $(0.28)^2 + (-0.12)^2 + (-0.36)^2 + ...$
Variance around the line = 0.40

Regression Variance = Original Variance - Variance around the line
Regression Variance = 3.49 - 0.40 = 3.09

r^2 = Regression Variance/Original Variance
r^2 = 3.09/3.49
r^2 = 0.89

Possible values reported for r^2 fall between 0 and 1. An r^2 of 0 means that there is no relationship between activity and the parameter(s) selected for the study. An r^2 of 1 means there is perfect correlation. The interpretation of the r^2 value for the capsaicin analogs is that 89% of the variation in the value of the Log EC_{50} is explained by variation in the value of , the hydrophobicity parameter.

GRAPH 5 Capsaicin Analogs Derivation of r2 values



While the fit of the data to the regression line is excellent, how can one decide if this correlation is based purely on chance? The higher the value for r2 the less likely that the relationship is due to chance. If many explanatory variables are used in a regression

equation, it is possible to get a good fit to the data due to the flexibility of the fitting process; a line will fit two points perfectly, a quadratic curve will fit three, multiple linear regression will fit the observed data if there are enough explanatory variables2. Given the assumption that the data has a Gaussian distribution, the F statistic below assesses the statistical significance of the regression equation.

The F statistic is calculated from r^2 and the number of data points (or degrees of freedom) in the data set. The F ratio for the capsaicin analogs is calculated as:

$$F_{1,n} = (n-2)\frac{r^2}{1 - r^2} = (7-2)\frac{0.89}{1 - 0.89} = 40.46$$

This value often appears as standard output from statistical programs or it can be checked in statistical tables to determine the significance of the regression equation. In this case, the probability that there is no relationship between activity and the value is less than 1% (p=0.01).

We have found that hydrophobicity values correlate well with biological activity. Does the addition of a size parameter (MR) improve our model? In order to analyze a relationship which is possibly influenced by several variables (or properties), it is useful to assess the contribution of each variable. π and MR appear to be somewhat correlated in this data set so the order of fitting can influence how much the second variable helps the first. Multiple linear regression is used to determine the relative importance of multiple variables to the overall fit of the data.

Multiple linear regression attempts to maximize the fit of the data to a regression equation (minimize the squared deviations from the regression equation) for the biological activity (maximize the r^2 value) by adjusting each of the available parameters up or down. Regression programs often approach this task in a stepwise fashion. That is, successive regression equations will be derived in which parameters will be either added or removed until the r^2 and s values are optimized. The magnitude of the coefficients derived in this manner indicate the relative contribution of the associated parameter to biological activity.

There are two important caveats in applying multiple regression analysis. The first is based on the fact that, given enough parameters any data set can be fitted to a regression line. The consequence of this is that regression analysis generally requires significantly more compounds than parameters; a useful rule of thumb is three to six times the number of parameters under consideration. The difficulty is that regression analysis is most effective for interpolation and it is extrapolation that is most useful in a synthesis campaign (i.e., the region of experimental space described by the regression analysis has been explained, but projecting to a new, unanalyzed region can be problematic).

Using multiple regression for the capsaicin analogs, one can derive the following equation which relates hydrophobicity and molar refractivity to biological activity.

$$\text{Log } EC_{50} = 0.762 - (0.819)\pi + (0.011)MR$$
$$s = 0.313,\ r^2 = 0.888$$

To judge the importance of a regression term, three items need to be considered.

1. Statistical significance of the regression coefficient.
2. The magnitude of the typical effect "$b_i x_i$" (in this case, 0.011x25.36).
3. Any cross-correlation with other terms.

As more terms are added to multiple linear regression, r^2 always gets larger. We recomputed the previous calculations (r^2 = 0.89) carrying three significant figures so that rounding does not lead to confusion.

These results of this analysis indicate that, within this series, steric bulk is not an important factor in activity. The influence of the hydrophobicity constant confirms the presence of a hydrophobic binding site. Given the limited number of substituents in this analysis, it is unlikely that more can be learned from further analysis.

This section has developed the fundamental mathematics of QSAR studies. Several authors have published reviews of QSAR and have discussed various aspects of the methods. Each of the examples to follow uses these techniques to derive information about the chemical factors which are important for activity.