Asian Journal of Chemistry Vol. 19, No. 6 (2007), 4151-4160

Statistical Concepts Utilized to Derive a QSAR and Their Application to Pharmaceutical Research

GAURAV VANDERKAR, DEEPTI JAIN and SURENDRA JAIN*

Shri Ravishankar College of Pharmacy, Bypass Road, Bhanpur, Bhopal-462 010, India

Tel: (91)(755)2737171; M: 09425011781; E-mail: jainsurendra@rediffmail.com

The combinatorial possibilities of hypothetical strategy for even simple systems can be explosive. For example, the number of compounds required to be synthesized in order to place 10 substituents on the four open positions of an asymmetrically disubstituted benzene ring system is *ca.* 10000. The alternative to this approach of compound optimization is to develop a theory that quantitatively relates variatioins in biological activity to changes in molecular descriptors that can easily be obtained for each compound. The present article explains the statistical concepts used to derive a QSAR and reviews the application of these techniques to pharmaceutical research.

Key Words: QSAR, Biological activity.

INTRODUCTION

Computational chemistry represents molecular structures as numerical models and simulates their behaviour with the equations of quantum and classical physics. Available programs easily generate and present molecular data including geometries, energies and associated properties (electronic, spectroscopic and bulk). The usual way of displaying and manipulating these data is a table in which compounds are defined by individual rows and molecular properties (descriptors) are defined by associated columns. QSAR attempts[1-10] to find consistent relationship 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.

Statistical concepts

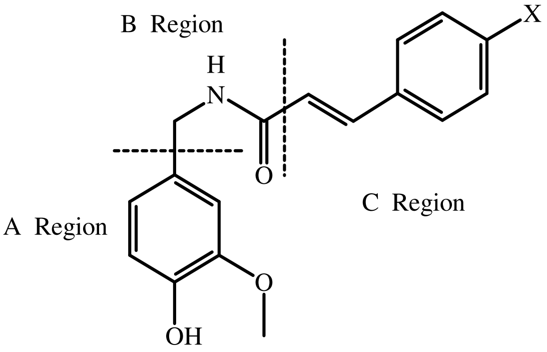
A QSAR gneerally takes the form of a linear equation

$$\text{Biological activity} = \text{Const} + (C_1 \bullet P_1) + (C_2 \bullet P_2) + (C_3 \bullet P_3) + \cdots\cdots$$

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 require the application of statistical techniques, a brief introduction to the principles behind the derivation of a QSAR is presented[11-28].

Taking example of novel analgesic agents[29] vanillylamides and vanillylthioureas (related to capsaicin) developed by the Sandoz Institute for Medical Research (biological activity tested by *in vitro* assay, which measured Ca^{2+} influx into dorsal root ganglia neurons). The data reported as the EC_{50} (μM) is shown in Table-1 (compound **6f** is the most active of the series).

TABLE-1
CAPSAICIN ANALOGS ACTIVITY DATA



Compound no.	Compound 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

In the absence of additional information, the only way to guess 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 authenticity of this guess cannot be guaranted.

The standard deviation of the data 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:

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

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

Rather than relying on this limited analysis, it is better to develop an understanding of the factors that influence activity within this series and use this understanding to predict activity for new compounds. The accomplish of this objective requires: (a) Binding data measured with sufficient precision to distinguish between compounds. (b) A set of parameters that can be easily obtained and which is likely to be related to receptor affinity. (c) A method for detecting a relationship between the parameters and binding data (the QSAR). (d) A method for validating the QSAR.

The QSAR equation is linear model that 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 syntheses, compounds that are structurally similar to the lead structure are prepared. The activity values for this series of compounds frequently span a limited range. In such cases, additional compounds have to be made and tested to fill out the training set.

The quality of any QSAR depends upon the quality of the data, which is used to derive the model. Thus, the dose-response curves need to be smooth, should contain enough points to assure accuracy and should span two or more orders of magnitude. Multiple readings for a given observations should be reproducible and must have relatively smaller errors. The important issue 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 case 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 these data will be questionable.

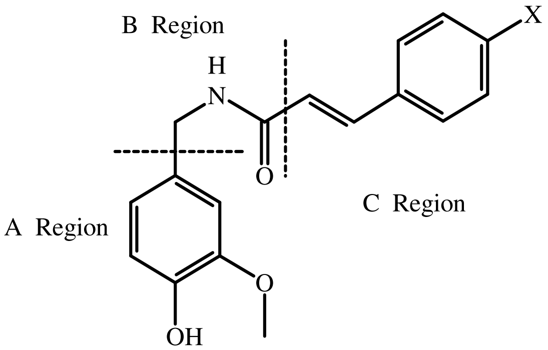
Biological data is often expressed in terms that cannot be used in a QSAR analysis. Since QSAR is based on the relationship of free energy to equilibrium constants the data for a QSAR study must be expressed in terms of the free energy changes that occur during the biological response.

While 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.

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

Since biological data are generally skewed, the log transformation moves the data to a nearly normal distribution. Thus, when measuring responses under equilibrium conditions, the most frequent transformed data for the capsaicin agonists is given in Table-2. The data points, projected onto the Y-axis, have become more uniformly distributed.

TABLE-2
CAPSAICIN ANALOGS TRANSFORMED DATA



Compd. no.	Compd. 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

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

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

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

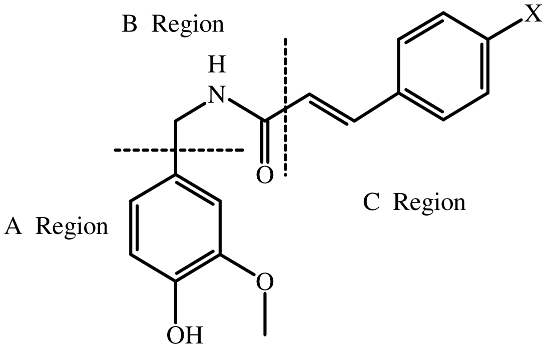
The example data sets are intended to show the general approach. The real data sets may have many more compounds and descriptors. Since the purpose of a QSAR is to highlight relationships between activity and structural features it is likely to find one or more structural features that relate the molecules and their associated activity. Additionaly, parameters that work consistently for all of the molecules in the series have to be found.

Several potential classes of parameters are used in QSAR studies. Substituent constant and other physico-chemical parameters *e.g.*, Hammett sigma constants (measures the electronic effect of a group on the molecule), Fragment counts (used to enumerate the presence of specific substructures), other parameters include, topological descriptors and values derived from quantum chemical calculations.

The selection of parameters is important in any QSAR study. Activity predictions are possible only if the association between the parameters(s) selected and activity is strong. Thus, for a given study, parameters that are relevant to the activity for the series of molecules under investigation and have values, which are obtained in a consistent manner, should be selected.

The analysis of capsaicin analogs can be divided into three regions: (A) region occupied by an aromatic ring, (B) region defined by an amide bond and (C) region occupied by a hydrophobic side-chain. The hypothesis for the (C) region assumed that a small, hydrophobic substituent would increase activity. Based on this assumption, the parameters that best define this characteristic being molar refractivity (size) and π (the hydrophobic substituent constant). Values are given in Table-3.

TABLE-3
CAPSAICIN ANALOGS PARAMETER VALUES



Compd. no.	Compd. name	X	log EC_{50}	MR
1	6a	H	1.07	1.03
2	6b	Cl	0.09	6.03
3	6d	NO_2	0.66	7.36
4	6e	CN	1.42	6.33
5	6f	C_6H_5	-0.62	25.36
6	6g	$N(CH_3)_2$	0.64	15.55
7	6h	I	-0.46	13.94

The data above can be analyzed graphically and statistically. The most visual approach with a limited number of variables being graphical. In this case, plot of activity *versus* molar refractivity or hydrophobicity gives some insight into the relationship between the parameters and activity.

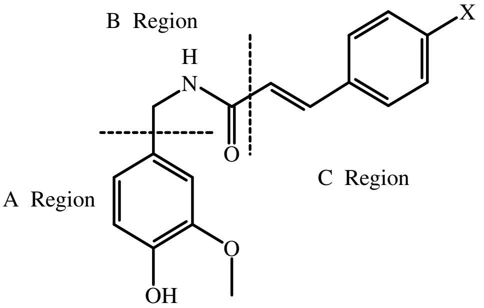
The ability of the graph to provide insight into the activity for compound **6i** and the estimate obtained regarding biological activity by knowing the value for either the hydrophobicity or molar refractivity is matter of concern.

In this example where only two values are examined, both the questions can be qualified but in more complex situations where multiple parameters are correlated to activity, statistics is used to derive an equation, which relates activity to the parameter set. The linear equation that defines the best model for this set of data being

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

The authenticity of the model can be answered by determining how well the equation predicts activities for known compounds in the series. The above equation estimates the average value for the EC_{50} based on the value of π; because assays vary, individual values differ from the regression estimate. This 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 are given in Table-4.

TABLE-4
CAPSAICIN ANALOGS CALCULATED VALUES



Compd. no.	Compd. name	X	log EC_{50}	Calculated log EC_{50}
1	6a	H	1.07	0.79
2	6b	Cl	0.09	0.21
3	6d	NO_2	0.66	1.02
4	6e	CN	1.42	1.26
5	6f	C_6H_5	-0.62	-0.81
6	6g	$N(CH_3)_2$	0.64	0.65
7	6h	I	-0.46	-0.12

The residuals can be used 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 can be calculated by taking the root-mean-squre of the residuals (in this calculation, the denominator shown as decremented by two to reflect the estimation of two parameters).

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

For being 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 previously calculated was 0.76 whereas the standard error from the QSAR model is 0.28. Clearly, the use of linear regression has improved the accuracy of analysis.

For a series of compounds several assumptions can be used in deriving a QSAR mode: (1) Parameters can be calculated (or measured in some cases) more accurately than activity can be measured. (2) Deviations from the best-fit line follow a normal (Gaussian) distribution. (3) Any variation in the line described by the QSAR equation is independent of the magnitude of both the activity and the parameters.

Considering 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:

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

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

The regression variance is defined as the original variance minus variance around the regression line. The original variance is the sum-of-the-squares distances of the original data from the mean.

Variance is calculated as follows:

Original variance = $(1.07 - 0.40)^2 + (0.09 - 0.40)^2 + \cdots\cdots$
Original variance = 3.49
Variance around the line = $(0.28)^2 + (-0.12)^2 + (-0.36)^2 + \cdots\cdots$
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

The 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 π, the hydrophobicity parameter.

While the data fitting to the regression line is excellent, it is important to decide whether if this correlation is purely based on chance. the higher the value for r^2 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 variables[1]. By considering 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 no relationship between activity and the π value is less than 1 % ($p = 0.01$).

Hydrophobicity values have found to corrleate well with biological activity. The effect of addition of a size parameter (MR) on the model should be analyzed, 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 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 indicates the relative contribution of the associated parameters to biological activity.

The two important caveats in applying multiple regression analysis are either based on the fact that for given enough parameters any data set can be fitted to a regression line. The consequence is, regression analysis generally requires significantly more compounds than parameters; a rule of thumb being three to six times the number of parameters under consideration. The difficulty being 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, following equation that relates hydrophobicity and molar refractivity to biological activity can be derived.

$$\log EC_{50} = 0.762 - (0.819)^{\pi} + (0.011)\ MR$$

$$s = 0.313,\ r^2 = 0.888$$

To judge the importance of a regression term, three following points need to be considered:

1. Statistical significance of the regression coefficient.
2. The magnitude of the typical effect b_ix_i (in this case, 0.011 • 25.36).
3. Any cross correlation with other terms.

As more terms are added to multiple linear regression, r^2 always gets larger. The previous calculations ($r^2 = 0.89$) carrying three significant figures have to be recomputed 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.

Conclusion

Developing a quantitative structure activity relationship is difficult. Molecules are typically flexible and it is possible to compute many possibly useful properties that might relate to activity. In early research program there are typically few compounds to model. Thus we have a few compounds in a very high dimensional descriptor space. Which are the important variables and how do we optimize them? It is clear that many experimental compounds need to span through the space and model fitting techniques need to address not only deriving a fit, but the predictive quality of the fit. While these methods have not discovered a new compound, they have aided scientists in examining the volumes of data generate in a research program. As the methods evolve, they will find broader application in areas such as combinatorial chemistry.

REFERENCES

1. J.G. Topliss, *J. Med. Chem.*, **15**, 1006 (1972).
2. Y.C. Martin, *J. Med. Chem.*, **24**, 229 (1981).
3. J.G. Topliss, Quantitative Structure-Activity Relationships of Drugs, Academic Press New York (1983).
4. R. Franke, Theoretical Drug Design Methods, Elsevier, Amsterdam (1984).
5. J.K. Seydel, QSAR and Strategies in the Design of Bioactive Compounds, VCH, Weinheim (1985).
6. Y.C. Martin, *Acc. Chem. Res.*, **19**, 392 (1986).
7. C. Hansch, *Acc. Chem. Res.*, **2**, 232 (1969).
8. C. Hansch and C. Silipo, *J. Am. Chem. Soc.*, **97**, 6849 (1975).
9. R.F. Gould, Biological Correlations - The Hansch Approach, Advances in Chemistry Series, No. 114, American Chemical Society, Washington, DC (1972).
10. Y.C. Martin, Quantitative Drug Design, Marcel Dekker, New York (1978).
11. S.M. Free and J.W. Wilson, *J. Med. Chem.*, **7**, 395 (1964).
12. S. Wold, *Patt. Recogn.*, **8**, 127 (1976).
13. A.J. Stuper, W.E. Brugger and P.C. Jurs, in ed.: B.R. Kowalski, Chemometrics: Theory and Application, American Chemical Society, Washington, DC (1977).
14. G. Klopman and M.I. Dimayuga, *J. Comput.-Aided Mol. Design*, **4**, 117 (1990).
15. J.W. McFarland and D.J. Gans, *J. Med. Chem.*, **30**, 46 (1987).
16. P.C. Jurs, in eds.: K.B. Lipkowitz and B.B. Boyd, Chemometrics and Multivariate Analysis in Analytical Chemistry, in Reviews in Computational Chemistry, VCH Publishers Inc., New York, Vol. 1 (1990).
17. P.C. Jurs, J.T. Chou and M. Yuan, *J. Med. Chem.*, **22**, 476 (1979).
18. L.B. Kier, Molecular Orbital Theory in Drug Resarch, Academic Press, New York (1971).
19. G.M. Grippen, *J. Med. Chem.*, **22**, 988 (1979).
20. M. Mabilia, R.A. Pearlstein and A.J. Hopfinger, in eds.: A.S.V. Burgen, G.C.K. Roberts and M.S. Tute, Computer Graphics in Molecular Shape Analysis, in Molecular Graphic and Drug Design, Elsevier Science Publishers, Amsterdam (1986).
21. H. Weinstein, in eds.: P. Politzer and D.G. Truhlar, Chemical Applications of Molecular Electrostatic Potentials, Plenum Press, New York (1981).
22. R. Potenzone, E. Cavicchi, H.J.R. Weintraub and A.J. Hopfinger, *Comput. Chem.*, **1**, 187 (1977).
23. G.R. Marshal, C.D. Barry, H.E. Bosshard, R.A. Dammkoehler and D.A. Dunn, in eds.: E.C. Olson and R.E. Christofferson, The Conformational Parameters in Drug Design: The Active Analog Approach, in Computer Assisted Drug Design, ACH Symposia, 112, American Chemical Society, Washington, DC (1979).
24. A.J. Hopfinger, *J. Am. Chem. Soc.*, **102**, 7196 (1980).
25. R.D. Cramer III, D.E. Patterson and J.D. Bunce, *J. Am. Chem. Soc.*, **110**, 5959 (1988).
26. V.E. Golender and A.B. Rozenblit, Logical Structural Approach to Computer Assisted Drug Design, in Drug Desin, Acadmic Press, Vol. 9 (1980).
27. V.E. Golender and A.B. Rozenblit, Logical and Combinational Algorithms for Drug Design, Resarch Studies Press, UK (1983).
28. V.E. Golender and E.R. Vorpagel, Computer Assisted Pharmacophore Identification, in 3D QSAR in Drug Design: Theory, Methods and applications, ESCOM Science Publishers, Netherlands (1993).
29. C.S.J. Walpole, R. Wrigglerworth, S. Bevan, E.A. Campbell, A. Dray, I.F. James, K.J. Masdin, M.N. Perkins and J. Winter, *J. Med. Chem.*, **36**, 2381 (1993).

(*Received*: 9 November 2005; *Accepted*: 28 March 2007) AJC-5535