Investigation of a Charge-transfer Substituent Constant Using Computational Chemistry and Pattern Recognition Techniques

David J. Livingstone,* David A. Evans and Martin R. Saunders
SmithKline Beecham Pharmaceuticals, The Frythe, Welwyn, Herts AL6 9AR, UK

Using the techniques of computational chemistry, a set of 58 parameters were calculated for 43 monosubstituted benzenes. Substituent constant values, κ, derived from the measurement of formation constants of charge-transfer complexes are available for 38 of these compounds. Relationships between κ and the calculated descriptors have been investigated using a variety of multivariate statistical techniques. Unsupervised analysis by two methods showed groupings of similar substituents and a tendency to order the compounds according to their κ values. It has been shown that κ can be predicted using these parameters but the different multivariate methods yielded different results. Partial least squares and principal components regression both tended to underpredict κ values whereas simple linear regression equations under- or over-predicted depending on the number of terms included. Examination of the descriptors involved in the correlations has shown that electronic effects are relatively unimportant in these complexes and that bulk and hydrophobicity parameters are most useful for the description of κ.

A substituent constant, κ, has been reported[1,2] which was proposed to measure the influence of substituents on the formation of charge-transfer complexes. Perhaps such complexes are better termed Electron-Donor-Acceptor (EDA) complexes since a charge-transfer band is not always observed in their spectra, particularly for the weaker complexes, and since it has been argued that transfer of electronic charge makes only a small contribution to the forces which stabilize their ground state.[3] Indeed, in the original study and the work reported here it would appear that electronic factors make only a small contribution to the description of κ. This substituent constant was suggested as a measure of EDA effects since it was derived from NMR determination of the formation constants of weak EDA complexes. It was intended for use in the quantitative description of the biological effects of compounds using the approach known as Quantitative Structure–Activity Relationships (QSAR) pioneered by Hansch.[4] This involves the evaluation of a correlation equation, such as that shown below, in which substituent effects are represented by linear free energy related parameters describing hydrophobic[4] (π), electronic[5] (σ) and steric[6] (E_s) effects:

$$\log 1/C = a\pi + b\sigma + cE_s + d \qquad (1)$$

where C is the dose to produce a given effect and the coefficients a, b, c and d are usually estimated by a least-squares fit. This approach has been extended considerably over the last thirty years by the consideration of numerous other descriptors of molecular properties[7,8] and by the application of other methods of data analysis.[8,9]

More recently, computational chemistry, in the form of molecular mechanics and quantum mechanics, has been seen as an alternative to QSAR for drug design. This has been aided by the rapid evolution of hardware and the increasing availability of 'user friendly' software. However, most applications of computational chemistry have involved the use of a graphics system in order to display[10] and overlay structures.[11] Such examples might be properly described as Structure–Activity Relationships since they make use of the structure of the test molecules but tend not to use quantitative descriptions of molecular properties. Computational chemistry and QSAR should not be viewed as alternatives; they are, in fact, complementary.[12] For example, there are reports in which a regression equation has been rationalised in terms of the fit of an inhibitor to an enzyme as determined by X-ray crystallography.[13–15] There are also examples in which parameters derived from computational chemistry have been used in attempts to correlate biological activity. The choice of such descriptors has often been based on some mechanistic rationale but it is also possible to adopt a systematic approach in which all possible parameters for a data set are calculated.[16–18]

One of the advantages of this method is that it is possible to calculate a very large number of parameters which can be used to describe physicochemical properties. Indeed, in terms of charge properties alone it is possible to create several times as many descriptors as there are atoms in a molecule. Such data sets are expected to contain some very detailed information although they also contain a considerable degree of redundancy.[18,19] One of the problems with this approach is that the large data matrices call for methods of analysis other than multiple regression. Such techniques are readily available[9,12,18,20] although they are perhaps not as well known as the various regression methods. Another problem with the use of parameters calculated by computational chemistry is that there is no general consensus of opinion on which properties should be used to represent a molecule, and which are the best parameters to represent those properties. This problem also occurs to some extent when using the 'traditional' QSAR descriptors since although it is generally accepted that parameters are required which represent hydrophobic, steric and electronic effects, there are many descriptors which may be used to this end.[7,8] Perhaps the biggest danger in the use of large numbers of parameters lies in the possibility of chance correlations. As pointed out by Topliss and co-workers,[21] the larger the number of variables screened then the higher the possibility of a seemingly significant relationship arising by chance.

The utility of parameters based on chemical model systems has been demonstrated many times, both as a means of describing chemical reactions and also the more complex interactions involved in biological systems. If the descriptors calculated by computational chemistry methods are to be of value in QSAR studies then it ought to be possible to describe simple models of chemical interactions using them. There is, of course, nothing new in this. A number of studies have been reported of the use of semi-empirical methods to calculate acidity constants and hence Hammett-type substituent constants.[22–24] This approach has been discussed specifically from the point of view of QSAR studies[25] and molecular orbital calculations of proton transfer involving amines has been proposed as a model of the binding of opiates to their receptor.[26]

In the work reported here we have examined relationships between calculated descriptors and κ substituent constant values. An advantage of this over the examples mentioned above[22–26] is that the formation of weak intermolecular complexes between two small molecules might serve as a better model of a drug receptor complex than the simple interaction of a proton with an acidic or basic group. In addition, it is possible that the examination of a model system such as this using theoretical parameters may give some insight into the forces responsible for the stabilisation of such complexes. Finally, the nature of the data set generated by the molecular modelling calculations forces the use of a number of multivariate statistical methods. Thus, the results from these analyses allow the comparison of a variety of analytical procedures.

Computational Methods

The chemical model system involving EDA complexes has been described before[1,2] so it will only be briefly mentioned here. Formation constants were measured using an NMR technique for a set of mono-substituted benzene electron donors with a common electron acceptor, 1,3,5-trinitrobenzene, in carbon tetrachloride solution. Experimental values have been reported for 35 substituents, a further three are given here. A substituent constant, κ, was derived from the values of the formation constants in an analogous fashion to the Hammett equation, but without the reaction constant ρ, as shown below:

$$\kappa_X = \log_{10}K_X - \log_{10}K_H \qquad (2)$$

where K_X is the formation constant for an X-substituted donor and K_H is the formation constant for the unsubstituted parent (benzene). Further experimental details are given in the references cited.

All calculations were performed using the COSMIC suite of molecular modelling programs.[10,27] A comprehensive series of mono-substituted benzenes was built. Each was geometry optimised with molecular mechanics, and the MOPAC[28] hamiltonian was calculated. Information for the statistical treatment was collated using a module in COSMIC designed specifically for tasks of this type. The series of molecules is considered as a common 'core' of atoms (up to 32) and a series of substituents (up to 12). For each molecule, the program records simple geometric data such as maximum and minimum dimensions in the cartesian axes, calculated log P and molar refractivity using the MEDCHEM algorithms,[29] moments of inertia, and some wavefunction derived properties such as dipole moment and its components, energies of HOMO and LUMO obtained from CNDO, MOPAC or *ab initio* wavefunctions as available.

For each atom of the common core of the series, a set of wavefunction-derived parameters such as Mulliken charges, self atom polarisabilities and superdelocalisabilities are recorded from each of the available wavefunctions. For each of the substituents, some geometric information, the sum of Mulliken charges and the mean square Mulliken charge is recorded, along with (where available) a set of physicochemical parameters from a suitably formatted lookup table of published substituent constants.[7] We thus have a framework in place which can store information based on the whole molecule, information based on the substituents, or more generally on arbitrary sections of the molecule, and information based on the common core atoms, or more generally on individual atoms in the molecule. We are currently extending the programme to give us the ability to include properties at arbitrary positions around the molecule such as electrostatic potential and electric field gradients.

The program may be run on a single molecule or on a list of molecules. The user-defined common core is automatically perceived within the molecule, giving the user the opportunity to intervene if multiple matches exist, or to override the automatic selection. The matching of substituents is treated similarly. The various collected and calculated properties for each molecule are then stored in a disk file for subsequent processing. A companion module takes sets of the property files, ascertains which sets of properties are present in all of the files and checks for consistency between the files (same number of core atoms, same number of substituents). Having established the largest complete data set available for the series of molecules, the user is offered the option of leaving out sections of the dataset. The whole table of raw data is then written out either to an RSE (BBN Software products, UK Ltd., Staines, Middlesex, UK) table or as a file in the format of the desired statistical package. A number of statistical procedures are available for the examination of relationships between the variables, for example hierarchical cluster analysis, principal component analysis and the reduction of redundancy by the removal of highly correlated properties.[19] The reduced data set can be written to an RSE table or to a disk file. Arbitrary experimental information such as biological activity or category data for classification procedures can be added at this stage if desired, or by manually editing the RSE table.

Statistical calculations were carried out using RSE, the pattern recognition package ARTHUR (Infometrix, Inc, Seattle, WA 98121) and the general purpose statistical package GENSTAT (Numerical Algorithms Group, Ltd., Oxford, OX2 7DE, UK).

Results and Discussion

As a result of the molecular modelling calculations, a total of 58 descriptors were assembled as the raw data set; these are shown in Table 1. This preliminary set consisted of parameter values for 40 compounds, *i.e.* the same substituents as reported in ref. 2. Due to solubility problems, values of κ are available for only 35 of these compounds, the remainder can be used as a test set. Substituents and their κ values are shown in Table 2. The first step in the analysis of these data was to remove redundancy on the basis of pairwise correlations. This was carried out with the preliminary data analysis module of COSMIC using a correlation coefficient limit of 0.7 to leave a reduced data set of 31 parameters. The variables in this set were autoscaled to give 'new' variables with a mean of zero and a standard deviation of 1. This is achieved as shown below, eqn. (3), where X'_{ij} is the autoscaled value of variable j for compound

$$X'_{ij} = (X_{ij} - \bar{X}_j)/\sigma_j \qquad (3)$$

i, X_{ij} is the raw data value, σ_j and $\bar{X}_j$ are the standard deviation and mean, respectively, for variable j. Autoscaled data are less sensitive to outliers and have the advantage when used in variance related methods, such as principal components analysis, of contributing one unit of variance per variable. A two-dimensional display (not shown) of the data points from this set using a non-linear mapping routine[9,31] showed a tendency to group similar substituents together. Further examination of this plot also showed that the κ values of the substituents increased in one direction along the plot.

Principal components analysis (PCA) of these data produced 10 principal components with eigenvalues greater than 1, a commonly used criterion to test for the number of 'significant' components since, for autoscaled data, an eigenvalue of less

Table 1 The 58 parameters in the starting data set

Whole molecule	Substituent	Atom centred
Moments of inertia in the x, y and z directions (I_x, I_y and I_z)	Calculated log P^a	Charge [chge(atom No.)]
Principal ellipsoid axes (P1, P2 and P3)	Calculated molar refraction^a	Self atom polarizability [alp()]
Calculated log P	Minimum and maximum dimensions in the x, y and z directions (x_{min}, x_{max}, *etc.*)	Electrophilic [F_e()] and nucleophilic [F_n()] frontier orbital densities
Calculated molar refractivity		Electrophilic [S_e()] and nucleophilic [S_n()] superdelocalizabilities
Energy of the highest occupied (E_{HOMO}) and lowest unoccupied (E_{LUMO}) molecular orbitals		
Dipole moment (μ) and components of the dipole moment in the x, y and z directions (μ_x, μ_y and μ_z)		

[a] These two properties are perfectly correlated with the whole molecule quantities since there is no other substitution.

Table 2 κ values for monosubstituted benzenes

Substituent	κ	Substituent	κ
H	0.00	Br	0.00
Me	0.11	Cl	−0.01
Et	0.13	I	0.01
Pr	0.04	F	−0.16
$CH(Me)_2$	0.07	NO_2	0.26
$(CH_2)_3Me$	0.07	NH_2	0.66
$C(Me)_3$	−0.07	NHMe	0.73
C_6H_5	0.45	NHEt	0.79
CHO	0.32	CH_2CN	0.39
COMe	0.48	$N(Me)_2$	0.90
CO_2Me	0.48	$N(Et)_2$	0.81
CO_2Et	0.55	CH_2OH	0.59
OMe	0.44	$CON(Me)_2$	1.31
OEt	0.39	$CON(Et)_2$	1.31
OH	0.40	$SO_2N(Me)_2$	1.24
SMe	0.40	$SO_2N(Et)_2$	1.31
CF_3	−0.09	$SO_2N(Pr)_2$	1.33
CN	0.23		

Table 3 The 11 parameter SELECT set in the order chosen

Number	Quantity
1	CMR
2	Clog$P$
3	E_{HOMO}
4	P3
5	μ_x
6	$S_n(1)$
7	$S_n(2)$
8	P1
9	$F_e(4)$
10	μ
11	$S_n(3)$

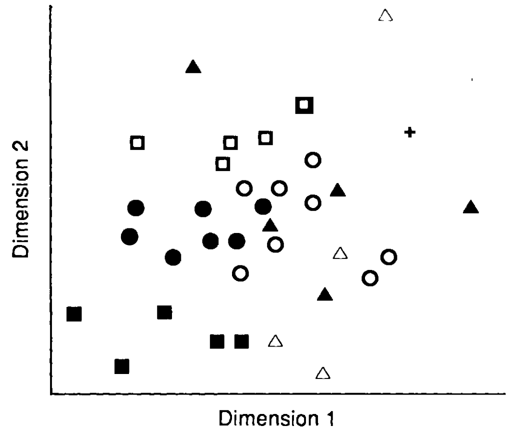

Fig. 1 Non-linear map derived from the 11 variable set, substituents classified as shown. +, hydrogen; ●, alkyl; ○, esters/ethers/carbonyl; ▲, halogen; □, amine; ■, sulfonamides/amides; △, no group.

than 1 represents the variance contribution of less than one of the original variables (however, see later). These components accounted for 88% of the total variance of the data set. A plot of the principal components scores (not shown) for the first two principal components also showed groupings of similar substituents and some classification of the compounds according to their κ values.

This unsupervised analysis was encouraging in that it appeared that the data did contain information which could be used to describe the model system. However, even after the preliminary removal of redundant variables the data set still contains a large number of descriptors compared with the number of data points. One means by which this number can be reduced without the loss of 'useful' information is to use the ARTHUR routine SELECT to choose parameters on the basis of their ability to predict κ values. This procedure is similar to forward stepping regression except that a decorrelation step is involved after the first and subsequent variables are chosen.[32] Although there is the potential danger of chance correlations[21] involved in this approach, it is expected that the use of a number of multivariate methods will help to identify chance effects. The SELECT procedure, using the 35 substituents with measured κ values, identified a set of 11 variables which are listed in Table 3 in descending order of selection. A non-linear map of this set (Fig. 1) shows a quite marked clustering of substituents according to their chemical classes with some ordering of κ values across the diagonal of the plot.

Analysis of this set by PCA resulted in the explanation of 80% of the variance using five principal components. The loadings (correlations) of the variables on these components are shown in Table 4 along with their eigenvalues (amount of variance explained). It can be seen from the table that the fifth component has an eigenvalue less than 1 and would thus not normally be considered to be important. However, Lukovits[33] demonstrated that a principal component, derived from quantum chemical data, with an eigenvalue smaller than 1 (0.52) was of importance in the explanation of pharmacological data. Examination of the dependence of κ on the principal components by forward stepping regression using the PC scores as independent variables gave the following equations, (4)–(6):

$$\kappa = 0.191\ \mathrm{PC1} + 0.453 \qquad (4)$$

$$R^2 = 0.5 \qquad F = 33.01 \qquad \mathrm{SE} = 0.32$$

Table 4 Variable loadings[a] for the first five principal components derived from the reduced data set of 11 variables

	Component (eigenvalue)				
	1 (2.73)	2 (2.19)	3 (1.78)	4 (1.23)	5 (0.95)
Variable	Loading				
CMR	0.48	−0.34			
$C\log P$		−0.41		−0.47	
E_{HOMO}		−0.36		0.49	
P3	0.41				0.33
μ_x		0.48		−0.37	
$S_n(1)$	−0.31				0.42
$S_n(2)$			−0.59		
P1				−0.41	0.60
$F_e(4)$	−0.39			−0.38	
μ	−0.40	0.40			0.38
$S_n(3)$			−0.60		

[a] For simplicity, only loadings above 0.3 are shown.

Table 5 Modelling κ by PLS

Dimension of PLS model	Percentage of κ variance explained using: 11 variables	31 variables
1	78.6	78.7
2	92.9	90.4
3	94.9	95.1

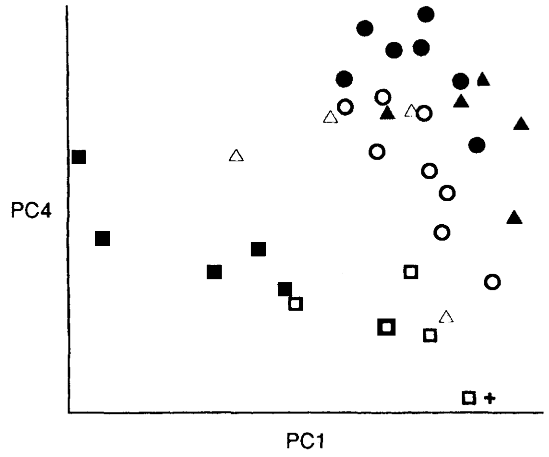

Fig. 2 Principal componen s plot calculated from the 11 parameter set. Symbols as in Fig. 1.

$$\kappa = 0.191\ \text{PC1} + 0.193\ \text{PC4} + 0.453 \qquad (5)$$
$$R^2 = 0.732 \qquad F = 43.77 \qquad \text{SE} = 0.24$$

$$\kappa = 0.191\ \text{PC1} + 0.193\ \text{PC4} + 0.130\ \text{PC5} + 0.453 \quad (6)$$
$$R^2 = 0.814 \qquad F = 45.22 \qquad \text{SE} = 0.20$$

$N = 35$ for all three equations, and the t statistics for individual regression coefficients are all significant at greater than the 1% level.

A plot of the principal component scores for the first two components included in the PC regression (PC1 and 4) shows a similar grouping of substituents (Fig. 2) to the non-linear map. It is interesting to note that the more 'important' principal components 2 and 3 are not involved in the two- or three-term regression equations. Although these components are required to explain the variance in the independent set, it is clear that this variance is not correlated with κ values. It is also of interest to see that the third term to be included in the regression is a component which has two 'bulk' terms (P1 and P3), nucleophilic superdelocalizability [Sn(1)] and the magnitude of the dipole moment.

A related technique to principal components regression is partial least squares (PLS) which combines the generation of principal components, known as latent variables, with regression on a dependent variable or set of variables.[34] The latent variables are calculated so as to maximise their correlation with the dependent variable with the provision that latent variables, like principal components, explain as much variance in the independent set as possible and are orthogonal to one another. Application of PLS to the selected data set of 11 variables resulted in the explanation of over 90% of the variance in the dependent data (κ) using a two-dimensional PLS model as shown in Table 5. This table shows that the addition of a third PLS dimension, equivalent to a third principal component, only improves the description of the κ data by 2% and thus would not be considered necessary. In this example, PLS has performed better than principal components regression although, of course, it should since the PLS latent variables are chosen so as to have high correlations with the dependent data. One of the features of the PLS technique is that it is claimed to be able to cope well with data sets which contain redundant information. A PLS analysis was also carried out on the de-correlated data set of 31 parameters giving very similar results, shown in Table 5, which demonstrates that this claim is justified, for this data set at least.

How do the PLS latent variables compare to the principal components ? Table 6 shows the loadings of the independent variables on the first three PLS latent variables (LV). It can be seen that the first LV has much similarity to the first principal component; the variables with high loadings on this LV correspond to the highly loaded variables on PC 1. They have coefficients with the same sign and mostly identical values, the major difference between this latent variable and the principal component is the addition of $C\log P$. Interestingly, $C\log P$ loads on to PC 2, which was not included in the regression on PCs, and PC 4 which was. The second latent variable corresponds most closely to the second principal component except that the sign of the $C\log P$ coefficient is changed. These similarities and differences highlight one of the major disadvantages of these 'latent variable'* techniques, the difficulty of interpretation. This is of little importance if all that is required is prediction, but

Table 6 Variable loadings[a] for the first three latent variables from PLS analysis of 11 variables

	Loading on latent variable		
Variable	1	2	3
CMR	0.48		−0.37
$C\log P$	−0.32	0.67	0.40
E_{HOMO}		−0.51	0.39
P3	0.42		
μ_x		0.36	−0.36
$S_n(1)$	−0.24		
$S_n(2)$			
P1			
$F_e(4)$	−0.34		−0.42
μ	−0.39	0.40	
$S_n(3)$			

[a] For simplicity, only loadings above 0.3 are shown apart from Sn(1) in LV 1 for comparison with PC 1.

* This includes principal components analysis which is also often referred to as a latent variable approach.

Table 7 Predicted and measured κ values for three new substituents

	κ values				
Substituent	Measured[a]	PLS[b]	PCR[c]	MLR[d]	MLR[e]
$OCH(Me)_2$	0.51	0.288	0.25	0.24	0.67
OC_6H_5	0.43	0.297	0.45	0.23	0.76
$OCH_2C_6H_5$	0.60	0.56	0.38	0.55	1.14

[a] From ref. 36. [b] From the two dimensional PLS model. [c] PC regression from eqn. (6). [d] Variable regression from eqn. (8). [e] Variable regression from eqn. (9).

if information concerning mechanism is desired then other methods may be more informative.

Application of stepwise regression to the 11 parameter set yielded the following relationships:

$$\kappa = 0.03\ \mathrm{CMR} - 0.82 \qquad (7)$$

$$R^2 = 0.43 \qquad F = 24.69 \qquad \mathrm{SE} = 0.34$$

$$\kappa = 0.04\ \mathrm{CMR} - 0.34\ C\log P - 0.47 \qquad (8)$$

$$R^2 = 0.91 \qquad F = 166.1 \qquad \mathrm{SE} = 0.14$$

$$\kappa = 0.05\ \mathrm{CMR} - 0.35\ C\log P + 0.09\ E_{\mathrm{HOMO}} + 0.53 \qquad (9)$$

$$R^2 = 0.95 \qquad F = 182.24 \qquad \mathrm{SE} = 0.11$$

Once again, $N = 35$ for all three equations and the t statistics for individual regression coefficients are significant at greater than the 1% level. A further term can be included in this forward stepping regression, P3, which lowers the standard error slightly to 0.098 and raises the multiple correlation coefficient to 0.96. Although this term is statistically 'significant' it appears to offer little improvement to the description of κ. These regression equations appear to offer a better fit to the κ data than the two- and three-term principal component regressions, eqns. (5) and (6), and, of course, are simpler to interpret.

Eqns. (8) and (9) are very similar, at least with respect to the CMR and $C\log P$ terms, to the correlation equations previously reported[2] for κ, as shown below, eqns. (10) and (11).

$$\kappa = 0.05\ \mathrm{MR} - 0.36\pi + 0.02 \qquad (10)$$

$$R^2 = 0.93 \qquad F = 77.41 \qquad \mathrm{SE} = 0.12$$

$$\kappa = 0.04\ \mathrm{MR} - 0.33\pi - 0.21\ \mathrm{R} + 0.0 \qquad (11)$$

$$R^2 = 0.95 \qquad F = 76.74 \qquad \mathrm{SE} = 0.10$$

These regression equations were based on tabulated values of substituent constants and R is the positionally weighted Swain and Lupton component of σ.[35] An interesting feature of these equations and eqns. (7)–(9) is that an electronic term does not become involved until after the bulk and partition coefficient based descriptors. Thus, it might appear that the popular view that charge-transfer complexes involve primarily electronic effects is not correct. However, this observation perhaps serves to illustrate the problem of the separation of steric, volume and 'bulk' effects. Whilst MR is correlated with other steric parameters,[37] it is also related to polarizability.[37] Perhaps the electronic terms included in the principal component regressions and the PLS treatment are also serving as measures of polarizability.

Encouragingly, these analytical methods all appear to be giving similar results in their description of the κ data, but how well do they predict? A further three substituent κ values have become available[36] since the original measurements were reported. These are shown in Table 7 together with predicted values derived from the principal components regression, eqn. (6), the two-dimensional PLS model and the multiple regressions on individual variables, eqns. (8) and (9). Predictions for the isopropoxy substituent are poor by all methods; the three-term variable regression equation does perhaps the best job but this gives overpredicted results for the other two substituents. The phenoxy substituent is best predicted by the principal components regression, eqn. (6), and the benzyloxy substituent is well predicted by both the PLS model and the two-term variable regression. From these results it is difficult to say that any one method is 'best'; principal components regression might be said to be giving the best overall results but the differences between the techniques are small. The fact that the PLS method does not make the best predictions is surprising since PLS latent variables are generated in such a way that correlations with the dependent variable are maximized. The PLS predictions, in fact, are little better than the variable regression predictions and the variable regressions are much easier to interpret. The results of the predictions from these two- and three-term regression equations demonstrate an important feature of how such models should be assessed. The statistics (R^2, F, SE *etc.*) for the three-term equation are all 'significant' and appear better than those of the two-term equation and yet the three-term predictions are worse. This may be a feature of this particular combination of compounds in the 'training' and 'test' sets, but this is a good example of a 'real' situation.

Conclusions

It has been shown that substituent constant values from a simple chemical model system may be described using parameters calculated by computational chemistry methods. This provides encouragement in the use of these techniques in attempts to model the effects of candidate drug molecules in biological systems. However, in agreement with an earlier analysis, the most important descriptors were shown to be calculated values of log P and molar refraction, two parameters which have been widely used in the 'traditional' QSAR approach. This perhaps serves to illustrate the fundamental importance of these properties or to indicate the utility of the mix of factors which make up these parameters. It also implies that the major contribution to the forces which stabilize these complexes is not electronic, contrary to popular opinion, unless it is the electronic components of log P and, more likely, MR which provide the correlation with κ. One way in which the computed descriptors might be improved would be to carry out the modelling calculations on complexes. This might serve as a better model of the EDA system but it is unlikely that this could be generally applied to biological systems since we often have little idea of the nature of the site of action of a drug.

Analysis of these data has provided an opportunity to compare several statistical methods. Unsupervised techniques such as PCA and non-linear mapping have shown that the physicochemical data can be used to group similar substituents together and, in part, to order substituents according to their κ values. PLS has been shown to describe the variance of the κ data in a smaller number of latent variables than principal components regression and comparisons between latent variables and principal components have been made. Multiple linear regression gave a better statistical fit than principal components regression and, of course, is easier to interpret. However, it is not possible clearly to identify a best technique since some of the predictions for 'test set' substituents are poor. This demonstrates that multivariate models should be evaluated by their performance in prediction as well as the statistics of their fit. Finally, the results show the complementary nature of these analytical methods.

References

1 R. Foster, R. M. Hyde and D. J. Livingstone, *J. Pharm. Sci.*, 1978, **67**, 1310.

2 D. J. Livingstone, R. M. Hyde and R. Foster, *Eur. J. Med. Chem.*, 1979, **14**, 393.

3 R. Foster, *Organic Charge-Transfer Complexes*, Academic Press, London and New York, 1969, pp. 2–4, 18–31.

4 C. Hansch, P. P. Maloney, T. Fujita and R. M. Muir, *Nature*, 1962, 178.

5 L. P. Hammett, *J. Am. Chem. Soc.*, 1937, **59**, 96.

6 R. W. Taft, in *Steric Effects in Organic Chemistry*, ed. M. S. Newman, Wiley, New York, 1956, p. 556.

7 H. Van de Waterbeemd, N. El Tayar, P. A. Carrupt and B. Testa, *J. Comput.-Aided Mol. Design*, 1989, **3**, 111.

8 D. J. Livingstone, in *Similarity Models in Organic Chemistry, Biochemistry and Related Fields*, ed. T. M. Krygowski, R. Zalewski and J. Shorter, Elsevier, Amsterdam, 1991, pp. 557–627.

9 D. J. Livingstone, in *Molecular Design and Modeling: Concepts and Applications*, vol. 203 of *Methods in Enzymology*, ed. J. J. Langone, Academic Press, San Diego, 1991, pp. 613–638.

10 J. G. Vinter, A. Davis and M. R. Saunders, *J. Comput.-Aided Mol. Design*, 1987, **1**, 31.

11 Y. Kato, A. Itai and Y. Iitaka, *Tetrahedron*, 1987, **43**, 5229.

12 R. M. Hyde and D. J. Livingstone, *J. Comput.-Aided Mol. Design*, 1988, **2**, 145.

13 M. Recanatini, T. Klein, C.-Z. Yang, J. McClarin, R. Langridge and C. Hansch, *Mol. Pharmacol.*, 1986, **29**, 436.

14 C. Hansch, T. Klein, J. McClarin, R. Langridge and N. W. Cornell, *J. Med. Chem.*, 1986, **29**, 615.

15 C. D. Selassie, Z.-X. Fang, R.-L. Li, C. Hansch, T. Klein, R. Langridge and B. T. Kaufman, *J. Med. Chem.*, 1986, **29**, 621.

16 O. Kikuchi, *Quant. Struct.-Act. Relat.*, 1987, **6**, 179.

17 R. C. Glen and V. S. Rose, *J. Mol. Graph.*, 1987, **5**, 79.

18 M. G. Ford and D. J. Livingstone, *Quant. Struct.-Act. Relat.*, 1990, **9**, 107.

19 D. J. Livingstone and E. Rahr, *Quant. Struct.-Act. Relat.*, 1989, **8**, 103.

20 D. J. Livingstone, *Pestic. Sci.*, 1989, **27**, 287.

21 J. G. Topliss and R. P. Edwards, *J. Med. Chem.*, 1979, **22**, 1238.

22 R. D. Gilliom, J.-P. Beck and W. P. Purcell, *J. Comput. Chem.*, 1985, **6**, 437.

23 R. Voets, J.-P. Francois, J. M. L. Martin, J. Mullens, J. Yperman and L. C. Van Poucke, *J. Comput. Chem.*, 1990, **11**, 269.

24 R. Karaman, J.-T. Huang and J. L. Fry, *J. Comput. Chem.*, 1990, **11**, 1009.

25 R. M. Hyde, *Prog. Clin. Biol. Res.*, 1989, **291**, 91.

26 L. K. Bennett and R. L. Beamer, *J. Pharm. Sci.*, 1986, **75**, 769.

27 S. D. Morley, R. J. Abraham, I. S. Haworth, D. E. Jackson, M. R. Saunders and J. G. Vinter, *J. Comput.-Aided Mol. Design*, 1991, **5**, 475.

28 J. J. P. Stewart, QCPE Program No. 455 (version 5), University of Indiana, Bloomington, IN, USA.

29 MEDCHEM software system, Daylight Chemical Information Systems, Inc, Irvine, CA 92713–7821, USA.

30 L. B. Kier, *Molecular Orbital Theory in Drug Research*, Academic Press, New York and London, 1971, ch. 4.

31 B. Hudson, D. J. Livingstone and E. Rahr, *J. Comput.-Aided Mol. Design*, 1989, **3**, 55.

32 B. R. Kowalski and C. F. Bender, *Pattern Recognition*, 1976, **8**, 1.

33 I. Lukovits, *J. Med. Chem.*, 1983, **26**, 1104.

34 W. G. Glen, W. J. Dunn and D. R. Scott, *Tetrahedron Comput., Methodol*, 1989, **2**, 349.

35 S. G. Williams and F. E. Norrington, *J. Am. Chem. Soc.*, 1976, **98**, 508.

36 J. A. Chudek, University of Dundee, personal communication.

37 M. Charton, in *Steric Effects in Drug Design*, ed. M. Charton and I. Motoc, Springer-Verlag, Berlin, 1983, pp. 107–118.

Paper 2/02231A
Received 29*th April* 1992
Accepted 2*nd June* 1992