REVIEW ARTICLE

STATISTICAL THINKING AND TECHNIQUE FOR QSAR AND RELATED STUDIES. PART I: GENERAL THEORY

MERVYN STONE
Department of Statistical Science, University College London, Gower St., London WC1E 6BT, U.K.

AND

PHILIP JONATHAN
Shell Research Ltd., Sittingbourne, Kent ME9 8AG, U.K.

SUMMARY

The two parts of this paper form a critique of a variety of statistical techniques of actual or potential use in quantitative structure–activity relationship (QSAR) studies and related fields. Part I explores the statistical thinking that is needed to underpin those techniques. Emphasis is placed on (a) the role of 'exchangeability' as an alternative to unrealistic statistical modelling and (b) the use of cross-validation to limit self-deception in the use of any particular technique. The problem of the almost unlimited range of molecular descriptors is seriously addressed. (Part II provides a concise critical review of methods – some well-established and some new.)

KEY WORDS Biological activity Cross-validation Exchangeability Molecular descriptors Prediction Relationship Structure

1. INTRODUCTION

The largely empirical QSAR techniques may be seen as desperate responses to a pressing need: the prediction of activity of new compounds either not yet in existence or, if they do exist, not yet tested in the laboratory or field. The desperation arises for the following reasons.

(i) Any activity of interest must be lawfully related to molecular structure.
(ii) In many studies the relevant lawful relationship has been successfully approximated by empirically established QSARs.
(iii) High costs of synthesis and/or testing put a premium on predictive capability.
(iv) Ignorance of the processes relating structure to activity, or the complexity of those processes, obliges prediction to go beyond the remit of hard scientific theory.

Most of the more widely used QSAR techniques have been straightforward (and not so straightforward!) borrowings from mathematical statistics and the battery of methods associated with that academic discipline. The exceptions have in the main come from computationally oriented methodology in areas such as pattern recognition and neural

0886–9383/93/060455–21$15.50

Received June 1992
Accepted May 1993

networks. The QSAR literature is in fact refreshingly catholic: it almost seems that no corner of quantitative thinking has gone unscrutinized in the search for the key to success. The philosophy behind all these borrowings has been openly opportunistic: a technique is to be judged either by its performance in predicting the measured activity of newly synthesized or newly considered compounds or else by the untested estimate of such performance based on its application to a single series of existing compounds.

Given that 'statistical thinking' is broadly definable as being concerned with the quantification of uncertain inference or prediction, it is generally agreed that such thinking has an important role to play in QSARs – in the design of the database, in the choice of analysis and, above all, in the realistic evaluation of results in terms that allow comparison of different techniques. There are two special features of QSAR studies that distinguish them from most applications of statistical methods: (i) the specification of the series of existing compounds on which the study is based and (ii) the choice of descriptors of those compounds on which the predicting formula is based. The choice of descriptors has to be related to the series, but the series specification calls for a nice marriage of chemical insight and skill in 'experimental' design; the choice of descriptors calls for a serendipitous union of molecular science and a statistical legerdemain that will cope with an essentially unlimited number of descriptors without self-deception. Sadly, there are no reliable prescriptions for success in these unions: clairvoyance would certainly help.

The present paper modestly aims to serve two masters:

(a) the quantitatively minded chemist interested in deepening his or her understanding of those strands of statistical thinking that appear to be necessary for a proper assessment of the typical QSAR study

(b) the statistician interested in extending his or her appreciation of the particularities of the QSAR subject area and their importance for the choice of statistical method.

2. THE PREDICTION PROBLEM

One practical demand of a QSAR is that it should be able to answer a question such as:

> *If the particular compound C, perhaps yet unsynthesized, were introduced into the particular environment $\mathscr{E}$, what is the current prediction of the value of the specified activity measure y of that compound in $\mathscr{E}$?*

The ultimate need to predict does not preclude QSAR studies whose main purpose may be to gain understanding[1] of which features of a compound determine the particular activity. For such studies, the ability to predict and thereby perhaps to optimize the design of a new compound would play a deferred role. A good example is the recent study[2] of twelve compounds defined by variations in the substituents R^1 and R^2 of the 'rotenoid core structure' in Figure 1.

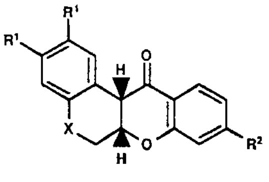

Figure 1. Rotenoid series[2]

Most QSAR studies start, like the rotenoid study, as an attempt to pinpoint the causes of the observed variation in activity in some series of congeneric compounds. They may then identify candidate compounds for synthesis in the light of the study findings.

3. DESCRIPTORS

In order to model chemical activity in some environment with any success, appropriate numerical descriptors of chemical structure and fundamental properties are essential. Three types of descriptor are in use today:

(i) *chemical descriptors* based on the three-dimensional molecular graph
(ii) *physicochemical descriptors* based on properties of the molecule in simple, specified environments
(iii) *high-dimensional descriptors* based on the representations of the molecule drawn from 'computational chemistry' and spectroscopy.

These three types are discussed in greater detail in the Appendix.

A range of different approaches, both experimental and theoretical, are available to generate numerical values for many of the above descriptors. For example, a physicochemical descriptor could be quantified using any one of

(a) experimental measurement in the physical chemistry laboratory
(b) a surrogate technique such as chromatography or spectroscopy
(c) an empirically established mathematical algorithm utilizing a database of descriptors for chemical fragments
(d) computational chemistry at the appropriate level of sophistication.

4. HIERARCHY OF ACTIVITY

There is clearly a hierarchy in the different activities that may be measured for a given compound, from the physicochemical 'properties' of the compound in specific experimental environments all the way to measures of its interaction with complex biological systems. From this viewpoint, some QSARs may be regarded as attempts to model or even explain the variation in higher-level activities in terms of a selection of elementary, scientifically comprehensible activities at lower levels. The latter may be easily measurable or adequately computable from knowledge of the molecular structure.

The activity of a compound C in a complex biochemical environment such as those found in living organisms is undoubtedly a correspondingly complex function of basic properties of the molecule that control the variety and rates of component processes, e.g. the transport of the compound to and from its main sites of chemical action or the compound's interaction with receptors at those sites. All of this is a statement of the obvious: that the task imposed on QSARs is both scientifically defensible and, at the same time, something of a tall order. QSARs attempt to bypass the hard science of their concern in the interests of expeditious discoveries—and must be allowed to use any 'soft science' device in their aid.

We believe that the concept of a hierarchy of activity may have been neglected as the basis of one such device: that, in predicting the measure y of the biological activity in which we are principally interested, it may be advantageous first to predict or measure a number, $z_1, \ldots, z_a$, of ancillary activities or properties, whose relevance might be argued on general scientific grounds, and then to predict y from the conjunction of $z_1, \ldots, z_a$ (now in hand) with other features of the structure $\mathscr{S}$ of the compound. For the case where $z_1, \ldots, z_a$ are not measured but are entirely predicted from knowledge of $\mathscr{S}$, this may appear to be a distinction without a difference, since a general function of $z_1, \ldots, z_a$ and $\mathscr{S}$ is then just another general function of $\mathscr{S}$. The potential value of the device is, however, apparent when it is considered that the zs may be relatively well-predictable functions of $\mathscr{S}$ whereas y itself may contain a large element of uncontrollable experimental variation. The latter would render difficult the direct determination of the best predictor of y from $\mathscr{S}$, because it would obscure the important role played by the zs, which are themselves well-predictable and need to be predicted reliably before being used in the second stage of the prediction of y. When the zs are themselves relatively easily measured activities, the potential predictive value of measurements of activities associated with parts of the process leading to y is obvious. For example, a z-variable might be a relevant biological activity measured *in vitro*, while y may be the complex and costly biological endpoint measurement *in vivo*. Yet another scenario is where z is the value of a physicochemical descriptor such as $\log P$ (see Appendix), predicted for compound C by a structural formula that has been validated by the congeneric series.

5. THE CONSTRUCTION SET OF COMPOUNDS

One possible framework for the prediction of the activity y of a compound C is as follows.

(i) C is a variant of a standard compound C_S and is defined by replacement of parts of C_S by alternative *substituents*.
(ii) There is a set $\mathscr{C}$ of so-called *congeneric* compounds in existence whose activities have been measured and which may all be regarded as related to C_S by different substitutions.
(iii) It is believed on reasonable scientific grounds that it may be possible to find an empirical relationship between the structure of the compounds in $\mathscr{C}$ and their measured values of y that would then provide a *statistical prediction* of the as yet unmeasured value of y for C.

A more realistic framework might be one in which the compound C is targeted, from the class of possible further variants of C_S, as of special interest only in the light of an empirical QSAR fitted to the compounds in $\mathscr{C}$.

The 'congeneric series' of compounds $\mathscr{C}$ is called the *construction set* for the prediction of y.

Note that, for cases of hierarchical prediction where the intermediate prediction of ancillary quantities z (Section 4) is to be part of the prediction procedure, the construction set for z may with advantage be larger than that for y.

6. STATISTICAL PREDICTION AND ITS ASSESSMENT

6.1. Generalities

The problem of statistical prediction of a complex activity of a compound C from knowledge of its structure alone is akin to that of forecasting the degree of marital harmony of a particular young man who has not yet met his partner.

The general approach that has been tried up to now has been

(i) a bold choice of potentially predictive descriptors
(ii) their deployment in a scientifically plausible formula

(iii) a rough evaluation of its predictive value based on its degree of statistical fit to the known values of y in the construction set $\mathscr{C}$.

Examples of the approach are given by the use of indicator variables for substituents in the Free–Wilson[3] method or by the use of the sum of tabulated values of log P for the substructures making up the compound. Even for the first of these techniques, the statistical problems associated with too many descriptors may be present. When the description of structure moves towards the quantum mechanical, with its potentially vast quantities of information, the choice of predictor variables undergoes a combinatorial explosion. Moreover, there is increasing difficulty in saying what is meant by 'the same descriptor' in the variety of quantum mechanical descriptions represented in $\mathscr{C}$. In some cases, variants of C_S introduce branches in the set $\mathscr{C}$ in which descriptors arise which are not definable for all compounds in $\mathscr{C}$; we then need the concept of *contingent* variables. Together with that idea goes the idea of a *sequential* tree-like analysis, in contrast to the balanced type of analysis of multiple regression.

When it comes to close examination of particular techniques of statistical prediction and its assessment, it has to be conceded that these are far from being standardized—even when there is no dispute about the probability model that would adequately represent the science behind the data.

For example, suppose that, in screening five congeneric compounds for their biological activity, we got the following LD50 activity estimates in log$_{10}$[1/(Molar concentration)] units:

Compound	1	2	3	4	5
Activity	5·27	6·85	4·94	5·01	4·62

Suppose for the sake of argument that it is known and agreed that for any one of these compounds the measurement of activity has a perfectly normal distribution about the true value with a standard deviation of 0·7 and that the five measurements are independently distributed. (The model here does not pretend to be realistic in the QSAR context. Indeed, we will make no further use of such formal models.) Suppose that a group of statisticians with a variety of backgrounds were individually asked:

> What is your prediction (estimate) of the true value of compound 2, selected because it has the largest measured activity? What is your assessment of the accuracy of your prediction?

Almost certainly, at least two differing answers would be obtained whose difference might be large enough to be of practical concern. One prediction might be the quite reasonable value of 6·85 with an assessment given by the experimental standard deviation of 0·7. Alternatively, the fact that the compounds do not have significantly different activities ($P = 0{\cdot}18$) suggests the prediction of 5·34 (the average of all five measurements) with a standard deviation of $0{\cdot}7/5^{1/2} = 0{\cdot}3$. These two predictions differ by a factor of more than 30 on the concentration scale.

If a disagreement of this magnitude can arise so readily for a simple, completely specified problem, it is clear that consensus cannot be expected in cases where the background science is only fragmentally developed and where, *a fortiori*, there can be no agreed probability model for the data. It is inevitable that in the absence of hard science, different investigators, driven by the pressure to predict, will arrive at very different predictions even when presented with exactly the same data. Accepting this diversity, the main emphasis has to be placed on attempting to get a realistic assessment of the statistical uncertainties of the different predictions.

We are therefore led to the consideration of general statistical methods for the assessment of statistical predictions. The first point that must be made is that, especially in an area like QSAR, it is much easier to identify potentially misleading techniques than it is to devise realistic ones. For example, there was the early recognition[4,5] of the phenomenon of chance correlations arising from a rich candidacy of descriptors and of the associated poor performance of the selected predicting formula when applied to new compounds. Such identification of misleading assessments requires, of course, that the reporting scientist will have given an honest account of the full choice of potential predicting formulae (predictors) that was available and of the way in which the variables used in the final predictor were actually chosen.

These necessary caveats should not be used, however, as a counsel of despair leading to the wholesale rejection of statistical techniques or to the insistence that the only reliable 'method' is the retrospective assessment that may be made on new compounds whose measured activities may be compared with the predictions made for them.

Indeed, the simplest simulation of retrospective assessment, known as the 'split sample technique', provides an alternative to despair that is sometimes used. In this method the construction set $\mathscr{C}$ is divided into a *training set* of compounds (from which the predictor is freely manufactured) and the remaining *validation set* (on which its performance for compounds not involved in its manufacture is assessed).

A refinement of the split sample method, known as cross-validatory assessment,[6] is available for those cases where the method of predictor construction can be specified with enough precision to be put on the computer. This method will be described in some detail after we have discussed some of the standard assessment techniques and listed some of the limitations on their use.

We will start the discussion at a somewhat unrealistic level with an example in which there are no unspecified parameters in the predictor that need to be estimated from the construction data. This example provides a baseline from which the increasingly questionable aspects of the more complex procedures may be more clearly appreciated.

6.2. Inference for a prespecified predictor

Suppose it is proposed to use the completely specified predictor $\hat{y} = f_o(\mathscr{S})$ in which the suffix 'o' is a reminder that there is nothing more to be specified or estimated in f_o.

Suppose $\mathscr{S}_1, \mathscr{S}_2, \ldots, \mathscr{S}_n$ are the structures of the n compounds in the construction set $\mathscr{C}$ and $\hat{y}_1, \ldots, \hat{y}_n$ are the corresponding 'predictions'. Then, for a QSAR study in which the only measurements for the construction set are the n values of y, the available raw material for a statistical assessment of the prediction $\hat{y}_C = f_o(\mathscr{S}_C)$ (for the unsynthesized or newly considered compound C with structure $\mathscr{S}_C$) is the n pairs of values $(\hat{y}_i, y_i)$, $i = 1, \ldots, n$. Suppose the deviations $d_i = y_i - \hat{y}_i$, $i = 1, \ldots, n$, between measured and 'predicted' values do not show any lawful dependence on $\mathscr{S}$ but appear to vary randomly over the compounds in $\mathscr{C}$ as if they were a normal random sample with zero mean (!). The reliability of $\hat{y}_C$ as predictor of y_C could then be assessed by means of a symmetric 95% confidence interval for y_C of the form $\hat{y}_C \pm t_n$ RMS, where

$$\mathrm{RMS} = [(d_1^2 + \cdots + d_n^2)/n]^{1/2}$$

and t_n is the upper 2·5% point of the t-distribution with n degrees of freedom.

However, it is unnecessary and arguably inappropriate in the QSAR context to make any assumption of normality for the inference. The $n+1$ deviations $d_1, \ldots, d_n$ and $d_C = y_C - \hat{y}_C$

are differences between the formula $f_o(\mathscr{S})$, presumably generated by some scientific theory, and the actual or potential values of y for the $n+1$ compounds. Such differences are not likely to fall obligingly into order as if from a normal distribution! Instead, the n values that are known to the investigator may well be unrelatable in any satisfactory way to the structure $\mathscr{S}$ – in other words they may look like a random sample of deviations. If this idea of randomness is extended to all $n+1$ values, including the value d_C that has not yet been revealed, on the grounds that there is nothing very special about the structure $\mathscr{S}_C$ in the context of $\mathscr{S}_1, \ldots, \mathscr{S}_n$, then the set $d_1, \ldots, d_n, d_C$ may be regarded as *weakly exchangeable* in the following sense: if these $n+1$ deviations are all different and we were somehow told the values of the deviations but not which of them was d_C, we would assign a probability $1/(n+1)$ of d_C being the smallest, a probability $1/(n+1)$ of it being the next smallest, and so on. It follows that if $d_{(1)} < d_{(2)} < \cdots < d_{(n)}$ are $d_1, \ldots, d_n$ placed in increasing order, the probability, on this view of the variability, that the unmeasured value of d_C lies between the ith smallest and the jth smallest of the values $d_1, \ldots, d_n$ is $(j-i)/(n+1)$. Another way of putting this is to say that the interval from $d_{(i)}$ to $d_{(j)}$ is a confidence interval for d_C with a confidence level equal to $(j-i)/(n+1)$. This becomes a confidence interval $(\hat{y}_C + d_{(i)}, \hat{y}_C + d_{(j)})$ for the quantity of interest, y_C.

Exchangeability also implies that $E(d_C^2) = E[(d_1^2 + \cdots + d_n^2)/n] = E(\text{RMS}^2)$, so that RMS^2 is a justified estimate of the mean-square error of prediction $E(d_C^2) = E[(y_C - \hat{y}_C)^2]$.

6.3. Multiple regression and exchangeability

An important class of incompletely specified predictors includes those that are linear in unspecified parameters. The reader is no doubt over-familiar with the simple case $\hat{y} = a + bf_o(\mathscr{S})$, so we will go straight to the case in which

$$\hat{y} = a + b_1 f_o^{(0)}(\mathscr{S}) + \cdots + b_p f_o^{(p)}(\mathscr{S})$$

This 'multiple-regression' predictor looks less intimidating written as

$$\hat{y} = a + b_1 x(1) + \cdots + b_p x(p) \tag{1}$$

which is called a general *linear* predictor because it is linear in the unspecified constants $a, b_1, \ldots, b_p$. The *predictor variables* $x(1), \ldots, x(p)$ may be simple descriptors or complex functions of descriptors.

We have gone for the special case where the predictor has an added constant a, in accordance with usual practice. Until further notice, we will also make two simplifying suppositions about the size and design of the construction set $\mathscr{C}$.

(a) There is no exact 'linear' relationship between the values of $x(1), \ldots, x(p)$ for the n compounds in $\mathscr{C}$ (e.g. we do not have $x(1) + \cdots + x(p) \equiv 100$ as would be the case if $x(1), \ldots, x(p)$ were percentages adding to 100).

(b) The number n of compounds in $\mathscr{C}$ exceeds the number $p+1$ of unspecified constants in $\hat{y}$.

These suppositions mean that there are no strictly mathematical complications in the ordinary least-squares fitting of the predictor formula to the construction data. (In Part II we look at some methods designed to tolerate the breakdown of these conditions.)

If $\hat{y}$ is then fitted by ordinary (unweighted) least squares to the construction data, what may be said in assessment of the resulting predictor $\hat{y}_C$? This depends on the level of acceptable assumption. Strengthening the definition of the previous section, suppose that there are unknown 'exchangeability values', a_e, b_{e1}, ..., b_{ep} say, of a, b_1, ..., b_p such that the associated deviations d_1, ..., d_n, d_C (between the values y_1, ..., y_n, y_C and the respective values of the formula $\hat{y}$ evaluated at the exchangeability values) are *(fully) exchangeable* in the following sense: if we were somehow told the values of the $n+1$ deviations but not which compounds had which values, we would regard all $(n+1)!$ possible allocations as equally probable. It may be shown that this assumption justifies the usual estimate of the root-mean-square prediction error of $\hat{y}_C$ as predictor of y_C at $x(1)$, ..., $x(p)$:

$$s(1 + 1/n + \mathbf{g}^T\mathbf{S}^{-1}\mathbf{g})^{1/2} \tag{2}$$

where (i) the *root-mean-square error* s is the square root of (residual sum of squares)/$(n-p-1)$, (ii) $\mathbf{S}$ is the sum-of-squares-and-products matrix for $x(1)$, ..., $x(p)$ and (iii) g_j, the jth component of $\mathbf{g}$, is the deviation of the value of $x(j)$ for compound C from its average value in the construction set. (Adding the assumption of normality to that of exchangeability would deliver a confidence interval for y_C centred at $\hat{y}_C$ and with a halfwidth given by the product of (2) and the appropriate value of Student's t.)

It is important to note that there are severe conditions on the applicability of formula (2) and the like. Exchangeability is a non-trivial assumption which would be invalidated if the compound C had been selected just because the values of its descriptors $x(1)$, ..., $x(p)$ gave a high value of $\hat{y}_C$. (Consider just $p=1$, a very noisy least-squares fit, and C chosen from two competing compounds.)

6.4. Testing and assessment

All prediction is a gamble and so is commentary on prediction in any area of 'soft' science such as QSARs. There is no statistical technique that eliminates the need for human judgement of the reasonableness of any assessment of the uncertainty of a particular prediction. (This is in contrast to a game of chance with known and agreed probabilities: the outcome may be a gamble but there is nothing uncertain about its prior assessment, provided that the mathematics is done correctly.)

The supposition that there is exchangeability of $n+1$ particular quantities, one of which involves the prediction error, requires such judgement. Given the complexity of the structure $\mathscr{S}$ and the environment $\mathscr{E}$ for even the simplest QSAR, it would be unrealistic ever to claim that such a supposition is a necessary consequence of the available information. Rather, in practice, the investigator may be able to make some limited tests of the exchangeability hypothesis before accepting the associated assessment as part of the decision making about synthesis of C. These tests may be only weakly informative, so that the investigator will be left with a large element of uncertainty about the quality of the decision. The only consolation from the statistical side may be that the decision is likely to be better when it is thus informed, with a necessarily long-run validation of the term 'better' over a number of independent assessments, i.e. gambles.

Without the extra assumption of normality (whose testing would add to an already difficult task), the only easily exploitable implications of the exchangeability assumption deployed in Section 6.3 are in the mean and variance–covariance structure of the least-squares residuals $r_i = y_i - \hat{y}_i$, $i = 1$, ..., n. It may be shown that a test statistic of the form $T = c_1r_1 + \cdots + c_nr_n$ with $c_1 + \cdots + c_n = 0$ is approximately normal with zero mean and standard deviation estimated as

$$s(c_1^2 + \cdots + c_n^2 - \mathbf{g}_c^T\mathbf{S}^{-1}\mathbf{g}_c)^{1/2}$$

where $g_{cj} = c_1[x_1(j) - \bar{x}(j)] + \cdots + c_n[x_n(j) - \bar{x}(j)]$. The quality of the approximation improves with the number n of compounds provided that T receives appreciable contributions from a large enough proportion of the residuals. With the test scene thus prepared, a significantly large value of T would cast doubt on the exchangeability supposition.

We can illustrate all this by a reanalysis of the data in Table 7 of Reference 7, for which $n = 11$, y is the natural logarithm of an 'apparent equilibrium constant' and $x(1)$ is π_N, a lipophilicity descriptor associated with substituents at site R in the 'QNB' series of Figure 2. Taking $x(2) = \pi_N^2$, the least-squares quadratic of y on π_N is

$$\hat{y} = -7\cdot64 + 7\cdot39\pi_N - 1\cdot55\pi_N^2 \tag{3}$$

in which the t-value of 3·7 for the quadratic coefficient is formally highly significant. (The t-test corresponds to an application of our exchangeability test of the residuals for a straight-line fit, using a standardized T with coefficients $c_1, \ldots, c_n$ themselves chosen as the residuals in a straight-line fit of π_N^2 on π_N.) Although theoretically reasonable, the fit of (3) to the data is not close enough (with $s = 1\cdot04$ on the natural logarithm scale) for the residuals to be accepted without further tests: in particular, theory suggests a look at their possible dependence on the steric descriptor E_S. The plot in Figure 3 is informative enough to reject exchangeability without formal testing. The final fit including E_S is

$$\hat{y} = -8\cdot35 + 8\cdot27\pi_N - 1\cdot68\pi_N^2 - 1\cdot45E_S \tag{4}$$

for which the residuals (with $s = 0\cdot56$) now seem experimentally reasonable. (The coefficients in (4) are in minor disagreement with the calculations of Mager and Rothe.[7] We cannot accept the logic of their arguments for rejecting such a simple analysis. They see (4) as an 'artefact' associated with 'nonsense correlations'.).

In the example just considered, if the tests of exchangeability had been arrived at after searching for a 'significant' refinement among a large number of possible candidates, then the possibility that 'significance' represents just a chance correlation[4,5] would have so reduced the impact of the finding that the investigator might stick with his or her initial choice of predictor formula. The weighing of this possibility raises difficult statistical questions whose

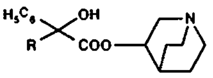

Figure 2. 3-Quinuclidinyl benzylate (QNB) series[7]

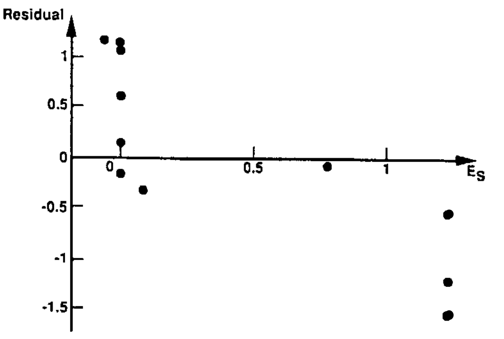

Figure 3. Residual plot for equation (3)

accommodation requires extensions to the framework of our discussion so far. In effect, in contrast to the usual role of significance tests, the testing of exchangeability may then be more helpful when it supports the chosen form of predictor than when it makes that choice questionable.

Exchangeability and subjectivity

The 'probability' involved in the definition of exchangeability has to be seen as a variety of subjective or betting probability, which is a necessary expression of the difficulty in conceiving any realistic chance mechanism as part of the underlying science. Those averse to the ideas of 'subjective' or 'betting' probability might wish to adopt some epithet such as 'judgemental' or 'quasi-random'.

Assessment

Suppose that a broad assessment of the predictive potential of the predictor $\hat{y}$ is wanted, without specification of any particular compound as candidate for synthesis. We list three assessment criteria that are more dedicated to this objective than is the mere quotation of the multiple-correlation coefficient (which is just the Pearson correlation coefficient between $\hat{y}_i$ and y_i) or the root-mean-square error s, useful though these two statistics are for general purposes. All three criteria assume that the n existing compounds in $\mathscr{C}$ are adequately representative of new compounds that might be proposed and that (interpolative) 'prediction' for the compounds in $\mathscr{C}$ will provide the required broad assessment.

(a) The 's^+ criterion' This is simply definable as an upward adjustment of s:

$$s^+ = s[1 + (p+1)/n]^{1/2}$$

Its derivation and interpretation rest on the hypothetical conception of n new compounds, different from those already in the construction set $\mathscr{C}$ but conveniently similar in that their positions in the p-dimensional space of $\mathbf{x} = (x(1), \ldots, x(p))^{\mathrm{T}}$ reproduce the n positions already occupied by the existing set $\mathscr{C}$ (given by the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$). We are also required to conceive of the set of unobserved activities $y_{n+1}, \ldots, y_{2n}$, say, of these artificially conceived compounds. The final conception is that for $i = 1, \ldots, n$ the values $y_1, \ldots, y_n, y_{n+i}$ satisfy the exchangeability condition that was imposed on $y_1, \ldots, y_n, y_C$ in order to justify formula (2) of Section 6.3.

With all this in mind, an estimate of the conceptual mean value of the square of the prediction error in $\hat{y}_{n+i}$ (the least-squares prediction of y_{n+i} at $\mathbf{x}_i$, say) is $s^2[1 + 1/n + (\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})]$ by application of formula (2). Averaging this over the index i gives us an estimate of the *mean value of the average of the squares of the prediction errors in* $\hat{y}_{n+1}, \ldots, \hat{y}_{2n}$:

$$s^2\left[1 + 1/n + \operatorname{trace}\left(\mathbf{S}^{-1}\sum_i(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}}\right)\Big/n\right] = s^2(1 + 1/n + \operatorname{trace}\,\mathbf{I}_{p\times p}/n)$$

$$= s^2[1 + (p+1)/n]$$

The latter expression has the uninformative label J_n in the statistical literature. The label s^+ for its square root is both informative and memorable, while the square root puts the criterion on the activity scale itself.

(b) An alternative to s^+ This is definable by

$$s^{++} = s[1-(p+1)/n]^{-1/2}$$

It is equivalent to the 'generalized cross-validation' criterion of Wahba[8] and is claimed to be an advance on RMSPE (see below). Its derivation is rather technical and it is not clear whether s^{++} is superior to either s^+ or RMSPE for the purpose of prediction assessment. Its adjustment of s always exceeds that resulting from the use of s^+ and becomes quite drastic as $(p+1)/n$ approaches unity, as the following numbers show:

	Multiplier of s									
When $(p+1)/n$ is	0·1	0·2	0·3	0·4	0·5	0·6	0·7	0·8	0·9	1·0
the multiplier in s^+ is	1·05	1·10	1·14	1·18	1·22	1·26	1·30	1·34	1·38	1·41
the multiplier in s^{++} is	1·05	1·12	1·20	1·29	1·41	1·58	1·83	2·24	3·16	∞

(c) The cross-validatory criterion RMSPE This is based on the simple idea that in turn, each of the compounds in $\mathscr{C}$ can simulate both a new compound prior to synthesis and the same compound after it has been synthesized and had its activity measured. For example, leaving out C_1, we can fit formula (1) of Section 6.3 to the $n-1$ compounds $C_2, \ldots, C_n$ and use the resulting prediction formula to give a prediction $\hat{y}_{-1}$, say, for C_1. The error $e_1 = y_1 - \hat{y}_{-1}$ in this prediction is clearly of a different character from the residual $r_1 = y_1 - \hat{y}_1$. Whereas r_1 is the result of a calculation that is adaptive to the value of y_1, the prediction error e_1 genuinely reflects the true hazards of prediction, embracing both bias and variance (even though it is necessarily based on a slightly reduced construction set). In fact, e_1 is always larger than r_1, the relationship between them being $e_1 = r_1/(1-h_1)$, where $h_1 = (\mathbf{x}_1 - \bar{\mathbf{x}})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x}_1 - \bar{\mathbf{x}})$. The sum $e_1^2 + \cdots + e_n^2$ is the well-known statistic PRESS (prediction error sum of squares). Practitioners may prefer, as we do, the equivalent root-mean-square prediction error criterion

$$\text{RMSPE} = [(e_1^2 + \cdots + e_n^2)/n]^{1/2}$$

since this criterion is both on a 'per compound' basis and is directly assessable on the activity scale.

Craig[9] has rightly remarked that such cross-validatory assessment is based on the measured activities of the compounds $C_1, \ldots, C_n$ and, of necessity, does not refer to new compounds. However, this remark also applies to any data-based assessment and we cannot agree with Craig's emphatic rejection of the leave-one-out procedure. (He refers to a paper by Snapinn and Knoke,[10] but this does not appear to relate to the valid point he makes.)

6.5. Problems of choice and 'chance correlation'

Up to now we have considered the assessment of predictors of prespecified form, involving a number of parameters that have to be determined by fitting the formula to the construction data. This type of problem is too restrictive to be in wide use: prespecification is too binding on the investigator, who would usually wish to be free to adapt the form of the predictor to the message of the construction data, prior to parameter determination in the chosen form.

In most QSAR problems there is a profusion of available descriptors whose value for prediction cannot be gainsaid in advance. The dimensionality of choice is usually very high and so therefore is the associated freedom to find a 'predictor' that fits the construction data unrealistically well, inducing a false sense of predictive competence in the user.

This statistical problem has long been recognized[11] as particularly serious in areas of soft science with a profusion of explanatory variables. However, the necessary statistical techniques are still relatively undeveloped, even primitive. (The option is not available in soft science to impose a hard probability model in which the necessary mathematics could be developed.) The problem of 'choice' from a wide variety of formulae has to be distinguished from the problem of 'overfitting' within the ambit of a specified but overparametrized formula. The problem of choice is an order of magnitude more difficult than that of overfitting. For the latter there is often a prior ordering of the variables in the order in which they would be added to the fitting by least squares and the multiple-correlation coefficient r may be made as close to unity as we like (reaching the value of unity when the formula fits all the y-values in $\mathscr{C}$ perfectly). However, the root-mean-square residual s will not in general show any parallel trend towards zero. This is because the 'degrees-of-freedom' divisor of the residual sum of squares makes an automatic adjustment for the overfitting. Moreover, criteria such as s^+ or RMSPE will reveal the point at which the overfitting starts and thereby provide the user with a safeguarding signal. (These criteria may be thought of as doing this by achieving a balance between the 'bias' from fitting an inadequately adaptive formula and the 'variance' in the fitted parameters due to the limited size n of the construction set.)

However, if the choice of successive variables was not predetermined but was made with an eye to the enhancement of the correlation with y (as, for example, in stepwise regression), unrealistically small values of s may be easily generated. (Cramer[12] has suggested that this may have happened in the analysis of Cheney *et al.*[13])

The most difficult problem in assessing the consequences of choice may, however, stem from the selection of a compound, for synthesis or testing, that has the maximum predicted activity in the chosen and fitted relationship. Especially when the dimensionality p is high, there is likely to be a large positive bias in the observed maximum as an estimate of the activity that would actually be found for the selected compound. The assessment techniques of Section 6.4 and the next section are essentially just averages of honest assessments of the values of $\mathbf{x} = (x(1), \ldots, x(p))^{\mathrm{T}}$ corresponding to $\mathrm{C}_1, \ldots, \mathrm{C}_n$. They do not tell us anything directly about the bias in any selected maximum. Unfortunately, there seem to be no general methods for correcting the bias in maximum predicted activity that do not depend on strong and probably insupportable assumptions about the underlying relationship.

6.6. Cross-validatory control

We will have nothing to say about the assessment of methods involving undocumented, perhaps subjective, choices of technique. (The only reasonable course of action for these would be documented reanalysis.) Rather, we consider now those methods for handling a multiplicity of descriptors that are completely specified apart from a small number (preferably one or two) of *control parameters* that have to be chosen before the prediction for any compound C can be calculated.

For example, *nearest-neighbour techniques* are controlled by the number of neighbouring compounds involved, while *partial least squares* is controlled by the number of construction stages used.

Cross-validatory choice of control parameters proceeds ideally by the 'leave-one-out' procedure. For each point of a suitably located grid covering the space of control parameters, each compound is left out in turn and a cross-validatory assessment is made of the corresponding 'prediction' for the omitted compound. The point of optimal aggregate assessment then determines the predictor to be used.

In the interests of computational economy, approximations to this ideal have been used in practice, such as the random or systematic division of $\mathscr{C}$ into training and validation sets (several such divisions being used) and the aggregation of the associated split-half assessments. Given the low cost of computing and the high cost of obtaining data, we recommend that efforts should always be made to achieve the symmetrical leave-one-out procedure, not forgetting the possible role of algebra in cutting computational comers. The method of cross-validatory choice is itself adaptive to the construction data and therefore runs the risk of 'overfitting'. However, when the number of control parameters is small compared with the number of degrees of freedom after fitting the construction data, the value of the cross-validatory assessment criterion at the optimized choice should be a reasonably realistic measure of the predictive performance of the method as a whole.

In some cases it is feasible to carry out what may be termed a 'two-deep' procedure, in which the whole calculation, 'cross-validatory choice of control parameters + optimized prediction', is itself subjected to a leave-one-out cross-validatory assessment! This clearly requires the leaving out of pairs of compounds: hence the 'two-deep'. It is feasible only for simple procedures such as those involving least-squares matrix algebra, but the need for it or for some approximation to it should be recognized in those cases where the number of control parameters is not small.

6.7. Cause and effect

With few caveats, the logic of causal inference is applicable to the simplest of QSAR comparisons of existing compounds. Suppose the activities y_1 and y_2 of compounds C_1 and C_2 have been established, effectively error-free, by valid unbiased experimental procedures. Any difference between y_1 and y_2 must then be attributed to the difference between C_1 and C_2. For example, if C_1 and C_2 differ only in their substituents, A and B respectively, at a particular location in the molecule, then we may say that the change of substituent A → B is the *cause* of the *effect* $y_1 \rightarrow y_2$. The logic is less straightforwardly applicable either when more complex comparisons are made than between just two compounds with one site of difference between them or when, for the design of a new molecule, we wish to predict a (necessarily causal) 'effect' for that molecule. The difference between C_1 and C_2 may be expressed as a trivial Free–Wilson QSAR

$$y = y_1 I_A + y_2 I_B$$

for the congeneric class $\{C_1, C_2\}$, where $I_A = 1$ or 0 according to whether A is present or not (ditto for I_B). In this simple case the Free–Wilson relationship encodes the 'cause and effect'.

When C_1 and C_2 differ in a more complex fashion, e.g. in their substituents at two locations, causal inference requires more data. The minimal case invokes a third compound C_3 with measured activity (also supposed for simplicity to be effectively error-free). The set-up for four substituents A, B, α and β could then be as follows:

Compound	Location		Activity
	1	2	
C_1	A	α	y_1
C_2	B	β	y_2
C_3	A	β	y_3

If y_1, y_2 and y_3 are all different, then it can be truly said that (i) in the presence of β at location 2 the change A → B causes the effect $y_3 \rightarrow y_2$, and (ii) in the presence of A at location 1 the change $\alpha \rightarrow \beta$ causes the effect $y_1 \rightarrow y_3$. However, no certain prediction can be made about the effect of joint changes A → B and $\beta \rightarrow \alpha$ in the compound C_3 that would give a new compound C with B and α at locations 1 and 2 respectively and with unknown activity y_c. At best we might *hypothesize* that, for example, the joint changes will have an additive effect, so that we would predict

$$y_c = y_3 + (y_2 - y_3) + (y_1 - y_3) = y_1 + y_2 - y_3$$

However, any faith in this prediction is not evidentially supported by the construction data in $\mathscr{C} = \{C_1, C_2, C_3\}$: its strength is no more than that of the extraneous hypothesis of additivity. The associated Free–Wilson QSAR

$$y = y_1 + (y_2 - y_3) I_B + (y_3 - y_1) I_\beta$$

is also evidentially neutral.

Alternatively, consider the possibility of a good fit of y_1, y_2 and y_3 by the two-parameter Hansch model $y = a + bx$, where $x = H_1 + H_2$ is the sum of tabulated hydrophobic fragmental constants[15] of the substituents at locations 1 and 2. A correlation coefficient of 0·9999 would give strong support from the data for the hypothesis that $y = a + bx$ is a causally interpretable QSAR that can predict the activity of the new compound with the variable x playing a fully causal role.

Similar support for causal interpretation can be given to more complex QSARs provided that their derivation has avoided the pitfalls of choice discussed in Sections 6.5 and 6.6. Although the QSAR area is a soft science, it does not have the degree of softness that in some areas, such as economics or sociology, renders causal interpretation of discovered relationships extremely hazardous. A chemical compound is a good model of an isolatable, fully describable agency. Moreover, the introduction of one compound rather than another into an environment in the determination of $y_1, \ldots, y_n$ is usually under the control of the experimenter. If one finds a strong, cross-validated QSAR in a rationally chosen set $\mathscr{C}$ of congeneric compounds, the expectation is reasonable that the relationship should extend to any new compound that would have been classed with $\mathscr{C}$ had it existed at the time. To deny this expectation by the counter-argument of 'correlation not causation' would be to invoke Murphy's law, i.e. to suppose that some other latent influential causal variable has misleadingly correlated with the QSAR for the compounds in $\mathscr{C}$ only to strike against it when the new compound has been synthesized.

The reassurance just offered does not remove the need for adequate 'congenericity' between the new compound C and the construction set $\mathscr{C}$. At the more technical statistical level the expectation of predictability has already been interpreted as an assumption of exchangeability. The prediction for C should not be so 'extrapolative' that doubt may reasonably be cast on the applicability of the predicting formula. (The terms 'congenericity' and 'extrapolative' are put in quotes for lack of precise definitions.)

6.8. Collinearity

The QSAR literature abounds with warnings[16–21] about the dangers of *collinearity* or *near-collinearity* of the predictor variables constituting the vector $\mathbf{x} = (x(1), \ldots, x(p))^T$. Such warnings serve a useful purpose and the following alternative formulation may serve to bring out the key point more clearly.

Suppose that the values $H_1, \ldots, H_n$ of some natural (i.e. not specially contrived) function $H(x(1), \ldots, x(p))$ for $C_1, \ldots, C_n$ respectively happen to be approximately equal (to their mean $\bar{H}$). If a new compound has a value of H, H_C say, appreciably different from $\bar{H}$, then either the prediction of its activity should be treated with caution on the grounds that the discrepant value H_C may be associated with lack of congenericity or extrapolability, or the prediction will have high assessed standard deviation (if extrapolation is carried out).

For the case where $H = c_1x(1) + \cdots + c_px(p)$ (which corresponds to the case of near-collinearity), the mathematical expression of the higher variability of least-squares prediction may be obtained from Section 6.3. For this it may be shown that in formula (2)

$$\mathbf{g}^T\mathbf{S}^{-1}\mathbf{g} > (H_C - \bar{H})^2 \Big/ \sum_1^n (H_i - \bar{H})^2$$

which will be large under the suppositions made.

We think that some confusion may have arisen from the use of the term 'independent variables' for what we have called the predictor variables. The 'independent' may originally have been intended in the sense that the variables could be varied independently of each other at the whim of the experimenter or, when the framework is observational rather than experimental, as the result of chance influences acting on each variable separately. The usage has since weakened to mean simply any variables that are used in an explanatory role with respect to a variable that is 'dependent' on them in a functionally specified relationship. The 'independence' in the term 'statistical independence' has a different technical meaning which happens to imply uncorrelatedness. The chain of association here might suggest that independent variables should ideally be uncorrelated and that caution is required (in some unspecified way) when they are not. (The term 'uncorrelated' is otherwise expressed as 'orthogonal' in the language of vector spaces.) It is the case, of course, that for efficient least-squares estimation and prediction, the 'design' of the points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ for $C_1, \ldots, C_n$ respectively should be well spread out, spanning their p-dimensional space, and that this property is often associated with zero or small correlations. However, designs with high correlation can be more efficient than orthogonal designs. For example, with $n = 4$ and $p = 2$ the design labelled '○' in Figure 4 has high correlation but would be more efficient (for an additive predictor) than the '×' design. Furthermore, there are no greater difficulties in interpretation of the fitted coefficients for the '○' design.

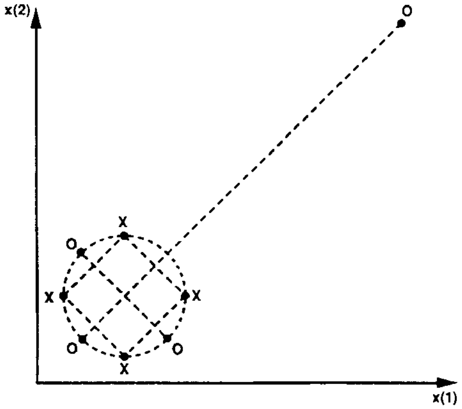

Figure 4. Comparison of an uncorrelated design (×) with a correlated design (○).

It may therefore be that warnings about correlation have been overdone. A possible example of their subtle influence may be seen in the analysis of Rowberg and Hopfinger[22] with $n = 22$ and unspecified p. The authors found a correlation between the indicator variables for pOH and mOH, corresponding to the imbalance in numbers of compounds in the 2×2 table

		$I_{p\mathrm{OH}}$	
		1	0
$I_{m\mathrm{OH}}$	1	7	0
	0	7	8

which 'shows the specific hydroxyphenyl substituent combination lacking in the database which is needed in order to determine the importance of the C-4′ and C-3′ hydroxy positions for inhibition potency'. They found that $I_{m\mathrm{OH}}$ was not statistically significant in a Free–Wilson analysis with just $I_{p\mathrm{OH}}$. However, when the model is extended to include the three indicator variables THP, PYR and $I_{N\text{-actyl}}$ (all in Rowberg and Hopfinger's best-fitting Free–Wilson equation), we find that $I_{m\mathrm{OH}}$ is significant at $P = 0{\cdot}03$, with a coefficient corresponding to a multiplicative factor of 2·7 on the $1/I_{50}$ scale, consistent with the other findings gently questioned by the authors. (Duewer[23] has raised a number of statistically contentious issues for Free–Wilson modelling, but these are without import for the present analysis.)

6.9. Scale standardization

Many standard statistical techniques have the property that their output is unaffected by the choice of units of measurement ('scale') of the input variables. An influential example is that of prediction in the non-singular case with $n \geqslant p + 1$ (the 'multiple regression' of Section 6.3): with $a, b_1, \ldots, b_p$ fitted by least squares, the actual value of the predictor (1) is unchanged by any rescaling of the descriptors $x(1), \ldots, x(p)$. (The least-squares coefficients $b_1, \ldots, b_p$ automatically compensate to keep $\hat{y}$ the same.)

By contrast, principal component regression[14] and partial least squares[14] are examples of techniques that are seriously affected by changes in the descriptor scales. For example, a useful predictor found with descriptors recorded in the old cgs units might have gone undiscovered had the newer SI units been used – or vice versa.

It is commonly held that the output of *any* statistical procedure for QSARs should not depend on what the units of measurement of the individual descriptors happen to be. This view has led to the usual practice of *scale standardization* (or *autoscaling*) whereby the variables (whether 'centred' to have zero means or not) are divided by their individual standard deviations over the construction set. There is clearly much to be said for this practice where the individual descriptors are all of very different characters, e.g. 'number of carbon atoms' and 'molar refractivity', and where prior knowledge about their relative importance is not available. In such cases autoscaling serves the interest of *scientific* standardization, reducing the risk of subjective, deceptive scaling. However, standardization may have undesirable consequences. Suppose, for example, that the set of descriptors includes those making up a digitized NIR spectrum. Typically,[24] the informative wavelengths are those of high variance over the construction set. For a number of techniques this information would be degraded if we were to autoscale each individual wavelength. (Of course, this does not exclude other forms of spectrum preprocessing.[25]) In this context, what might be termed *group autoscaling* should be considered, in which subgroups of similar descriptors (e.g. the NIR wavelengths) are jointly autoscaled as a group in a way that does not destroy the relative informativeness of individual descriptors within groups.

Technical footnotes

(i) As a non-linear operation, autoscaling has the disadvantage that it stops algebraically slick cross-validation in least-squares multiple regression based on downdating formulae for matrix inversion: valid cross-validation has to use the *n*-fold leave-one-out procedure.

(ii) The use of logarithms of positively valued descriptors should always be considered as an alternative to scale standardization. Their use gives independence of measurement units too and its value depends on the informativeness of percentage variations.

ACKNOWLEDGEMENTS

We thank Simon Belsten of Shell Research for the artwork and Mac McCarthy of Shell Research for his generous support.

APPENDIX: CHEMICAL DESCRIPTORS FOR QSAR

A1. The role of computational chemistry

The possibility of establishing QSARs based on theoretical descriptors offers the opportunity to predict the likely activity of a hypothetical chemical prior to synthesis. For this reason, considerable effort has been devoted to the development of chemical descriptors using computational chemistry. From the quantum mechanical perspective a complete description of any molecular system could *in principle* be obtained by solution of the appropriate Schrödinger equation.[26] In practice, this solution is unobtainable except for the hydrogen atom! (For the most elementary molecular ion, H_2^+, we are forced to adopt the Born–Oppenheimer approximation, which permits a decoupling of the Schrödinger equation into nuclear and electronic equations.[27]) In general, approximate solution of the Schrödinger equation requires a number of further simplifying assumptions. The quality of the solution obtained is dependent on many factors, notably

(i) the completeness of the Hamiltonian (energy) function used
(ii) the approximate form of the wavefunction assumed
(iii) the method used for solution of the Schrödinger equation.

Three levels of computational methods are in current use[28] to generate descriptors for QSARs. Two of these, the *ab initio*[29] and the *semi-empirical*,[30] provide approximate quantum mechanical solutions; the third approach, known as *molecular mechanics*,[31,32] utilizes a classical Newtonian 'ball and spring' method parametrized by experiment or high-level computation. Molecular mechanics provides a means for rapid calculation of minimum energy 3D molecular conformations essential for many structure-based QSAR descriptors.[33] Molecular mechanics also permits the modelling of molecular interactions[34,35] and hence the estimation of free energies of binding[36] and time-averaged properties[37] (via molecular dynamics[38]) for example.

A2. Chemical descriptors based on 3D molecular structure

Useful chemical descriptors have been devised using nothing more than a knowledge of the molecular constituent atoms and their connections. In addition, some approaches make use of the precise 3D molecular geometry (incorporating bond lengths, angles, etc.), the determination of which involves either computational or experimental (e.g. crystallographic[39]) work. These descriptors are discussed here.

The Free–Wilson descriptors

Consider a congeneric set of chemicals whose molecular structures differ only in the presence or absence of certain substituents on a common parent structure. In Free–Wilson analysis[3] the variation in molecular structure is encoded as a set of *binary indicator variables* X_i ($i = 1, 2, \ldots$), each referring to the presence ($X_i = 1$) or absence ($X_i = 0$) of some substituent at some location.

Minimal topological difference[40]

This can be considered as an extension of Free–Wilson analysis. The 3D structures for a set of chemicals are superimposed. As the first stage of the analysis, the molecular structure of each chemical is encoded as a set of binary indicator variables indicating the presence or absence of particular substituents at given locations in space.

Topological descriptors

Using the topology of a molecular structure, a number of chemical descriptors have been developed via graph theory.[41] In general terms, these descriptors describe the disposition of atoms in the molecule. For example, Wiener[42] proposed a measure of molecular branching based on the *molecular distance matrix*, whose elements are the numbers bonds between appropriate pairs of atoms in the molecule. Kier and Hall[43] introduced *connectivity indices*, which have been widely used for QSARs.

A3. Chemical descriptors based on elementary physicochemical properties

Historically, QSARs were attempts to relate chemical activity with simple measurable chemical properties.[44] One of the earliest reported QSARs involved correlation of the toxicity of simple organic compounds with their solubility in water.[45] Hansch *et al.*[46] showed that the octanol–water partition coefficient, a measure of hydrophobicity, could be correlated with the biological activity of certain plant growth regulators. Hammett[47] demonstrated that a measure of substituent electronic effect could be correlated with the reactivity of substituted benzenes. The comparative success of these studies has motivated the widespread development of chemical descriptors for QSARs based on one or more simple measures of hydrophobicity, electronic and steric properties. Originally, these descriptors were measured experimentally. In recent years, however, computational chemistry and particularly molecular-fragment-based algorithms such as CLOGP and CMR[48,49] have made reasonable theoretical approximation more attractive.

The following are noteworthy physicochemical properties useful in QSARs.

Hydrophobic descriptors

The distribution (or partition) of a solute between two immiscible liquid phases is clearly of considerable interest for QSARs, involving biological systems[50] for example. A typical measure of hydrophobicity is the molecular octanol–water partition coefficient introduced by Hansch *et al.*[46] Expressed as a logarithm, log P(oct/water) is a popular descriptor for both pharmaceutical[51] and agrochemical[52] QSARs. Initially, only experimental values for log P were obtainable. The popularity of this measure has resulted, however, the development of analogous descriptors using chromatography.[53] From the theoretical viewpoint, Rekker[15] and colleagues pioneered an empirical method for estimating log P using molecular fragment data; this culminated in the development of software such as CLOGP[49] which provides rapid approximation of log P values. Direct quantum mechanical approximation is also possible.[54]

Electronic descriptors

Hansch *et al.*[46] clearly established the importance of electronic effects in drug design following Hammett's work on substituted benzenes.[47] Refined versions of Hammett's approach[55] have been widely applied to QSARs, with varying success. Today, a wide variety of electronic properties are used for QSARs. These include reaction constants,[55] ionization constants,[56] dipole moments[57] and calculated energies for the highest occupied and lowest unoccupied molecular orbitals.[58]

Steric descriptors

Size and shape properties were first shown to be related to chemical activity by Meyer.[59] Since then, a number of physical and chemical descriptors of steric properties have been proposed; notable early descriptors were Taft's steric constants.[60] Many steric descriptors are calculated theoretically from a knowledge of the van der Waals radius of each atom in the molecule concerned. The van der Waals radius is defined as the distance at which the repulsion between the electron densities of two neighbouring atoms balances the attractive force between them. The value of this radius can be estimated from crystallography or computational chemistry.

Various descriptors based on the van der Waals radius have been used for QSARs. In general, however, it is unlikely that any one of these would be of use in isolation. For this reason, Verloop and Tipler[52] devised the STERIMOL method, which describes molecules or substituents in terms of five steric parameters. Knowledge of the van der Waals radii permits the computation of the molecular van der Waals volume,[61] namely the volume of intersection of atom-centred spheres whose radii are given by the corresponding atomic van der Waals radii.

Another popular descriptor of steric properties is the *molar refractivity* (MR),[62] related to refractive index and molar volume by the Lorenz–Lorentz equation. The precise meaning of MR as a QSAR descriptor is unclear; its close relation to molecular volume and molecular mass is obvious, however. A fragment-based method, CMR,[49] has been developed for rapid theoretical estimation of molar refractivities. Molecular volume and molecular mass[63] have also been used as steric parameters for QSARs.

A4. High-dimensional chemical descriptors

The advent of accessible computing facilities, the ensuing ease of data generation, acquisition, manipulation and application of multivariate statistical techniques have encouraged the use of high-dimensional descriptors for QSARs. These are drawn primarily from computational chemistry and spectroscopy. Weinstein *et al.*[64] initiated the use of electrostatic potential maps,[65] corresponding to the electrostatic potential of a molecule, calculated on a lattice of points surrounding the molecule. Cramer *et al.*[66] have used a similar approach in their comparative molecular field analysis (CoMFA); this technique yields steric and electrostatic maps quantifying the interaction between the molecule of interest and a chosen probe atom. The use of measured or simulated spectroscopic data (infrared, near-infrared and nuclear magnetic resonance spectra) for prediction of the constituent proportions in a complex mixture[67,68] has also prompted the use of these and similar data as chemical descriptors for QSARs.

REFERENCES

1. C. Hansch, in *Drug Design: Fact or Fantasy?*, ed. by G. Jolles and K. R. H. Wooldridge, Academic, New York (1984).
2. L. Crombie, J. L. Josephs, J. Cayley, J. Larkin and J. B. Weston, *Bio-org. Med. Chem. Lett.* **2**, 13 (1992).
3. S. M. Free and J. W. Wilson, *J. Med. Chem.* **7**, 395 (1964).
4. J. G. Topliss and R. J. Costello, *J. Med. Chem.* **15**, 1066 (1972).
5. J. G. Topliss and R. P. Edwards, *J. Med. Chem.* **22**, 1238 (1979).
6. M. Stone, *J. R. Stat. Soc. B*, **36**, 111 (1974); corrigendum **38**, 102 (1976).
7. P. P. Mager and H. Rothe, *Pharmazie*, **45**, 758 (1990).
8. G. Wahba, in *Applications of Statistics*, ed. by P. R. Krishnaiah, p. 507. North-Holland, Amsterdam (1976).
9. P. N. Craig, in *Comprehensive Medicinal Chemistry*, Vol. 4, ed. by C. Hansch, P. G. Sammes and J. B. Taylor, p. 664, Pergamon, Oxford (1990).
10. S. M. Snapinn and J. D. Knoke, *Technometrics*, **27**, 199 (1985).
11. R. A. Fisher, *Philos. Trans. R. Soc. Lond. B*, **213**, 89 (1929).
12. R. D. Cramer, in *Quantitative Structure Activity Relationships of Drugs*, ed. by J. G. Topliss, p. 253. Academic, New York (1983).
13. B. V. Cheney, J. B. Wright, C. M. Hall and H. G. Johnson *J. Med. Chem.* **21**, 936 (1978).
14. M. Stone and P. Jonathan, *J. Chemometrics*, **8**, in press (1994).
15. R. F. Rekker, *The Hydrophobic Fragmental Constant*, Elsevier, Amsterdam (1977).
16. C. J. Blankley, in *Quantitative Structure Activity Relationships of Drugs*, ed. by J. G. Topliss, p.1. Academic, New York (1983).
17. W. J. Dunn, *Chemometrics Intell. Lab. Syst.* **6**, 181 (1989).
18. P. Lewi, in *Drug Design*, Vol. X, ed. by E. J. Ariens, p. 307. Academic, New York (1980).
19. E. R. Marengo and M. Conterno, in *QSAR: Rational Approaches to the Design of Bioactive Compounds*, ed. by C. Silipo and A. Vittoria, p. 173. Elsevier, Amsterdam (1991).
20. M. Randic, *J. Mol. Struct. (Theochem.)*, **233**, 45 (1991).
21. J. K. Seydel and K.-J. Schafer, *Pharm. Ther.* **15**, 131 (1982).
22. K. L. Rowberg and A. J. Hopfinger, *Chemometrics Intell. Lab. Syst.* **8**, 183 (1990).
23. D. L. Duewer, *J. Chemometrics*, **4**, 2299 (1990).
24. M. Stone and R. J. Brooks, *J. R. Stat. Soc. B*, **52**, 237 (1990).
25. H. Martens and T. Naes, *Multivariate Calibration*, p. 314, Wiley, New York (1989).
26. P. W. Atkins, *Molecular Quantum Mechanics*, Oxford University Press, Oxford (1983).
27. M. Born and R. Oppenheimer, *Ann. Phys.* **84**, 457 (1927).
28. T. Clark, *A Handbook of Computational Chemistry*, Wiley, New York (1985).
29. W. J. Hehre, R. Ditchfield, R. F. Stewart and J. A. Pople, *J. Chem. Phys.* **52**, 2769 (1970).
30. M. J. S. Dewar and W. Thiel, *J. Am. Chem. Soc.* **99**, 4899, 4907, 5231 (1977).
31. N. L. Allinger, *J. Am. Chem. Soc.* **99**, 8127 (1977).
32. U. Burket and N. L. Allinger, *ACS Monographs*, **177**, 1 (1982).
33. T. Ryhanen, F. J. Bemejo, J. Santoro and M. Rico, *Comput. Chem.*, **11**, 13 (1987).
34. S. N. Rao, U. C. Singh P. A. Kollman, *J. Am. Chem. Soc.* **108**, 2058 (1986).
35. B. Pullman (ed.), *Intermolecular Interactions from Diatomics to Biopolymers*, Wiley, New York (1973).
36. T. A. Andrea, S. W. Dietrich, W. J. Murray, P. A. Kollman, E. C. Jorgensen and S. Rothenberg, *J. Med. Chem.* **22**, 221 (1979).
37. J. A. McCammon and S. C. Harvey, *Dynamics of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge (1987).
38. P. Dauber-Osguthorpe and D. J. Osguthorpe, *J. Am. Chem. Soc.* **112**, 7921 (1990).
39. A. S. Horn and C. J. de Ranter (eds), *X-ray Crystallography and Drug Actions*, Clarendon Press, Oxford (1984).

40. I. Motoc, in *Steric Effects in Drug Design*, by M. Charton and I. Motoc, Springer, New York (1983).
41. A. T. Balabon (ed.), *Chemical Applications of Graph Theory*, Academic, New York (1967).
42. H. Wiener, *J. Am. Chem. Soc.* **69**, 2636 (1947).
43. L. B. Kier and L. H. Hall, *Molecular Connectivity in Structure–Activity* Analysis, Research Studies Press, Letchworth (1986).
44. M. S. Tute, in *Comprehensive Medicinal Chemistry*, Vol. 4, ed. by C. Hansch, P. G. Sammes and J. B. Taylor, Pergamon, Oxford (1990).
45. C. Richet, *C. R. Sci. Soc. Biol. Ses. Fil.* **9**, 775 (1893).
46. C. Hansch, P. P. Maloney, T. Fujita and R. M. Muir, *Nature*, **194**, 178 (1962).
47. L. P. Hammett, *Physical Organic Chemistry*, McGraw-Hill, New York (1940).
48. C. Hansch and A. J. Leo, *Substituent Constants for Correlation Analysis*, Wiley, New York (1979).
49. Medchem Software, Pomona College Chemistry Lab., Claremont, CA 91711 U.S.A.
50. P. J. Taylor, in *Comprehensive Medicinal Chemistry*, Vol. 4, ed. by C. Hansch, P. G. Sammes and J. B. Taylor, p. 241. Pergamon, Oxford (1990).
51. C. Hansch, A. Leo, C. Schmidt, P. Y. C. Jow and J. A. Montgomery, *J. Med. Chem.* **23**, 1095 (1980).
52. A. Verloop and J. Tipler in *Pharmaco-chemistry Library: Biological Activity and Chemical Structure*, Vols 4 and 10, ed. by J. A. Keverling Buisman, Elsevier, Amsterdam (1977).
53. S. H. Unger, P. S. Cheung, G. H. Chiang and J. R. Cook, in *Partition Coefficient Determination and Estimation*, ed. by W. J. Dunn, J. H. Block and R. S. Pearlman, Pergamon, Oxford (1986).
54. G. Klopman and L. D. Iroff, *J. Comput. Chem.* **2**, 157 (1981).
55. K. Bowden in *Comprehensive Medicinal Chemistry*, Vol. 4, ed. by C. Hansch, P. G. Sammes and J. B. Taylor, p. 205. Pergamon, Oxford (1990).
56. A. Albert, *Selective Toxicity: The Physico-chemical Basis of Therapy*, Chapman and Hall, London (1973).
57. L. J. Lien, Z. R. Guo, R. L. Li and C. T. Su, *J. Pharm. Sci.* **71**, 641 (1982).
58. E. L. Mehler and J. Gerhards *Mol. Pharmacol.* **31**, 284 (1983).
59. V. Meyer, *Chem. Ber.* **27**, 510 (1894).
60. R. W. Taft, in *Steric Effects in Organic Chemistry*, ed. by M. S. Newman, Wiley, New York (1956).
61. R. S. Pearlman, in *Physical Chemical Properties of Drugs*, ed. by S. H. Yalkowsky, A. A. Sinkula and S. C. Valvani, Dekker, New York (1980).
62. M. Charton and B. I. Charton, *J. Org. Chem.* **44**, 2284 (1979).
63. S. C. Dash and G. B. Behera, *J. Indian Chem. Soc.* **57**, 542 (1980).
64. H. Weinstein, S. Maayani, S. Srebrenik, S. Cohen and M. Sokolovsky, *Mol. Pharmacol.* **9**, 820 (1973).
65. A. Pullman and B. Pullman, *Rev. Biophys.* **14**, 289 (1981).
66. R. D. Cramer, D. E. Patterson and J. D. Bunce *J. Am. Chem. Soc.* **110**, 5959 (1988).
67. W. J. Dunn, M. G. Koehler and S. L. Emery, *Chemometrics Intell. Lab. Syst.* **1**, 321 (1987).
68. P. A. Salamin, Y. Cornelis and H. Bartels, *Chemometrics Intell. Lab. Syst.* **3**, 329 (1988).