Environmental Health Perspectives
Vol. 61, pp. 275-285, 1985

Parameter and Structure–Activity Data Bases: Management for Maximum Utility

by Albert Leo

Quantitative structive-activity relationships (QSAR) in the fields of medicinal chemistry, pesticide science, biochemistry and toxicology are being published at an ever increasing rate. In addition to these biological correlation equations, thousands of such equations have been published for all kinds of organic reactions. There is a great need to develop a computerized system to enable one to make comparisons and to draw generalizations about the effects of structure on chemical and biological activity. A proposal is made for a systematic approach to this problem based on the physicochemical properties of organic compounds.

Introduction

The general field of structure–activity relationships is currently suffering from an embarassing excess of "riches." In each of the major areas—pharmacokinetics, drug and pesticide design, and environmental hazard assessment—there is a massive outpouring of results, each interesting in its own right. But taken as a mighty flood, it prevents most investigators from reaching that magic "overview" point from which general principles can be discerned and formulated. The revolution in computers and in instrumentation has resulted in a nonlinear increase in output of each laboratory, and the number of laboratories is likewise increasing at some power function greater than one. Thus an incredible number of organic compounds, from the simple to the very complex, are being reacted with a huge variety of biological systems in every level of complexity from purified enzymes to whole animals or even ecosystems. The resulting publications inundate the researcher even if he is dedicated to keep abreast of the field.

Objectives

In structure–activity studies, the term "pattern recognition" is most often applied in a rather narrow sense to a certain mathematical treatment of a set of biological data. Of course it can be given a more fundamental "philosophical" definition where the patterns sought are far broader in extent—crossing the boundaries between scientific disciplines at times. It is "pattern recognition" in this more basic sense which can be expedited by a properly constructed bank of biological activities interfaced with a data base of suitable physicochemical parameters. Even though the early format of the data base maintained by the Pomona College Medicinal Chemistry Project needs a great deal of improvement, it still has been used to show a pattern of activity extending across disciplinary boundaries in physical chemistry, biochemistry, and animal and plant physiology. In these times, when organization within subspecialties leaves much to be desired, and integration between specialties is all too rare, any tool for organizing this kind of knowledge deserves close attention.

From the few examples given in the following section we hope to illustrate how the present structure of the data bases helped in the discernment of discipline-spanning patterns of activity and how this now leads us to propose improvements.

Nonspecific Hydrophobic Interactions

When silinized glass beads are placed in water they tend to aggregate, driven by the force which tends to minimize the hydrophilic/hydrophobic interface. This is the simplest sort of physicochemical model of biological membranes. Addition of alcohols to this system reduces the hydrophobic/hydrophilic difference and tends to disaggregate the beads according to the following relationship (*1*).

$$\log 1/C = 0.98 \log P - 0.80 \qquad (1)$$

with
$n = 4$
$s = 0.077$
$r = 0.995$

*Pomona College, Claremont, CA 91711.

where C = concentration of alcohol (in mole/L); n is the number of data points; s and r are the standard error and coefficient of regression, respectively, and P is the partition coefficient of the solute between octanol and water.

The membrane surrounding a red blood cell is much more complex than the silinized bead model above, yet the action of alcohols in rupturing this very important membrane can be expressed (*2*) in similar fashion:

$$\log 1/C = 0.96 \log P - 0.30 \quad (2)$$

with
$n = 6$
$s = 0.06$
$r = 0.999$

From this we see that the real biological membrane displays the same relative sensitivity to changes in hydrophobicity as does the model (same coefficient of the log P term) but its intrinsic sensitivity is greater. The latter characteristic is immediately evident if one eliminates the hydrophobic effect in both equations by letting log $P = 0$. The activity is then given directly by the intercept.

The resistance to an electrical impulse across the membrane in a nerve cell is of critical importance in transmission of nerve impulses. This resistance is affected by the presence of alcohols, and has been studied using synthetic "black lipid" membranes as models. In one such study (*1*) the data could be fit to the following equation:

$$\log 1/C = 1.16 \log P - 0.51 \quad (3)$$

with
$n = 7$
$s = 0.26$
$r = 0.985$

The effect of a series of miscellaneous neutral organic compounds on the blockage of a frog's nerve (*1*) is given by:

$$\log 1/C = 0.88 \log P + 0.63 \quad (4)$$

with
$n = 25$
$s = 0.297$
$r = 0.955$

Here again the real membrane of a nerve cell is intrinsically more susceptible to becoming depolarized (intercept + .63 vs. - 0.51), but its sensitivity to structural changes in the depolarizing agent is very similar to a physicochemical model (log P coefficient = 0.88 vs. 1.16).

Moving up the scale in complexity and crossing into the field of animal physiology, one can rationalize the narcotic action of alcohols on tadpoles (*1*).

$$\log 1/C = 0.90 \log P - 0.91 \quad (5)$$

with
$n = 57$
$s = 0.31$
$r = 0.962$

Again the hydrophobic effect, as modeled by octanol/water log P, can be seen as a common thread linking what may be various actions at membrane surfaces.

The significance of these relationships becomes more apparent when useful analgetics, anesthetics and narcotics are also fit into these equations. Even when a particular functional group imparts an unusually high level of activity and thus deserves the label of "pharmacophore," its analogs usually fit closely to an equation with a unit slope in log P but with a larger intercept.

In discussing the interaction of a wide variety of chemicals with biological systems, the importance of the use of generalized descriptors can hardly be overemphasized. This point can be illustrated by a study of penicillins (*3*).

$$\log 1/C = -0.45\pi + 5.67 \quad (6)$$

with
$n = 20$
$s = 0.191$
$r = 0.909$

where C is the concentration needed to cure mice of a *S. aureus* infection and π is a measure of substituent hydrophobicity (4) using the octanol/water model. The negative coefficient for the π term was unexpected, and it inspired a study to see if hydrophobicity could promote binding to a site of loss to a greater extent than it promoted bacterial toxicity. This study (*5*) produced the following equation:

$$\log (B/F) = 0.50\pi - 0.67 \quad (7)$$

with
$n = 79$
$s = 0.255$
$r = 0.924$

where B/F represents the ratio of penicillin bound to human serum albumin to that in the free state.

Penicillins are not unique in binding thus to serum albumin, as was shown by the study of a set of miscellaneous neutral organic solutes from which Eq. (8) was derived (*6*):

$$\log 1/C = 0.75 \log P + 2.30 \quad (8)$$

with
$n = 42$
$s = 0.16$
$r = 0.960$

Binding to hemoglobin can also reduce the effective concentration reaching the target (*7*). This is shown in Eq. (9):

The use of a direct measurement of activity as the dependent variable, instead of the concentration needed to reach a pre-set activity level, limits the usefulness of this equation, as will be discussed in detail in later sections. Direct comparisons of slopes and intercepts cannot be made unless the dependent variable is in the same form, usually log $1/C$.

$$\log 1/C = 0.71 \log P + 1.51 \quad (9)$$

with
$n = 17$
$s = 0.16$
$r = 0.95$

In Eq. (8) and (9), C refers to the molar concentration of ligand required for 1:1 binding of the solute to the bovine protein. It is evident that albumin has about 6.2 times the affinity of hemoglobin (antilog of 2.3–1.51), but the greater amount of the latter present in the bloodstream results in an important role for both proteins in determining how chemicals are distributed in animal bodies.

A broad spectrum of toxicity measurements appear to be nonspecific in nature and related simply to solute hydrophobicity. For a wide variety of alcohols (*1*) we have Eqs. (10)–(14).

50% Inhibition of bacterial luminescence

$$\log 1/C = 1.10 \log P + 0.21 \quad (10)$$

with
$n = 8$
$s = 0.103$
$r = 0.998$

50% Inhibition of tortoise heart

$$\log 1/C = 0.98 \log P + 0.52 \quad (11)$$

with
$n = 10$
$s = 0.124$
$r = 0.973$

50% Inhibition of oxygen consumption by guinea pig lung

$$\log 1/C = 0.84 \log P + 0.16 \quad (12)$$

with
$n = 7$
$s = 0.114$
$r = 0.994$

Toxity of vapor to tomato plants

$$\log 1/C = 0.68 \log P + 3.04 \quad (13)$$

with
$n = 14$
$s = 0.101$
$r = 0.972$

Inhibition of liver esterase

$$\log 1/C = 0.75 \log P + 3.70 \quad (14)$$

with
$n = 14$
$s = 0.322$
$r = 0.931$

The same sort of relationship often holds when the toxic reaction of alcohols is carried to lethality: LD_{-100} for cats (*1*):

$$\log 1/C = 1.06 \log P + 1.37 \quad (15)$$

with
$n = 8$
$s = 0.124$
$r = 0.986$

Sometimes outliers to a general relationship of this type give indication that a metabolite may contain a toxiphore not present in its precursor. It is interesting therefore to note that methanol is not an outlier in Eq. (15); i.e., the metabolite formaldehyde, which causes blindness, does not displace the failure of the autonomous nervous system as immediate cause of death in the time frame of the LD_{-100} test.

Even this simple pattern of hydrophobically dependent action has had important repercussions when its application to new areas was discovered. A number of elegant experiments have solidified the relationship between aquatic bioaccumulation and log P (o/w). The following relationship, which applies to chlorinated hydrocarbon pesticides in the waters of the Great Lakes, was taken from one such study (*8*):

$$\log \mathrm{BCF} = 0.791 \log P - 0.40 \quad (16)$$

with
$n = 122$
$r = 0.927$

Here BCF is the ratio of concentration of pesticide found in fish to the concentration in the waters in which they live. It is remarkable because it covers a 100,000-fold activity range and included 13 species studied in several laboratories. The usefulness of this expression to anyone responsible for the hazard assessment of untested compounds or yet-to-be synthesized structures hardly needs emphasizing, especially in view of improved methods of calculating log P (o/w) from structure (see later section).

One should no longer be surprised when relationships are discovered in the more esoteric specialties. In the field of forensic medicine, a relationship has been found between log P and the post-mortem concentration of barbiturates in human blood in cases of fatal poisoning (*9*):

$$\log 1/C = 0.44 \log P + 2.92 \quad (17)$$

with
$n = 5$
$r = 0.943$

The fact that the coefficient of the log P term is close to that found for barbiturate efficacy in mice is reassuring, but one still wonders why it is only half that for many other narcotics.

Early in their training, scientists learn the principal of Occam's razor, yet there is an understandable tendency for an investigator to consider the system he chooses to work on to be unique, and consequently he is very apt to concentrate on unique solutions to the problems it presents. As an example, someone working

on the inhibition of the enzyme, hydroxyindole-*O*-methyltransferase, by *N*-acyltryptamines might be excused for believing this reaction to be much more complex than any of those related in Eqs. (1)–(15). However, the equation for 50% inhibition of this enzyme can be expressed (*10*) as:

$$\log 1/C = 0.71 \log P + 1.51 \quad (18)$$

with
$n = 17$
$s = 0.16$
$r = 0.950$

This is so similar to the expression for binding to hemoglobin [Eq. (9)] or to a nonenzymic protein, serum albumin [Eq. (8)], that evidence for any uniqueness in the binding interaction must be found in other data.

A further example can be found in the 75% inhibition of influenza B virus by benzimidazoles for which the following expression holds (*10*):

$$\log 1/C = 0.58 \log P + 1.58 \quad (19)$$

with
$n = 15$
$s = 0.210$
$r = 0.903$

Again the nonspecific binding of solute to protein explains so much of the variance in this activity that nothing special about the structure of the viral protein is indicated by this data.

Specific Hydrophobic Interactions

If a biological effect does not involve a transport factor, or occasionally, if transport can be factored separately, then often there remains a hydrophobic effect limited to portions of the reactant structure. This is ascribed to a desolvation of only part of the substrate surface as it is bound to the active site in a specific orientation while the remainder is still exposed to solvent space. The biological activity bank contains many examples of enzyme inhibition where the overall log *P* of the inhibitor does not help the correlation but the π value of substituents in just one of the positions is clearly significant. One such example comes from a study of *S*-methylation of thiopurine by thiopurine methyltransferase (*11*). In a set of benzoic acids substituted at two or more of the 3, 4, and 5 positions, which acted as enzyme inhibitors, only π for the 3 position was significant. If both the 3 and 5 positions were substituted, then the substrate appeared to be oriented so that the more hydrophobic side contacted the enzyme and the π value of the other side did not matter. This single specific hydrophobic parameter accounted for only half of the variance in the data [Eq. (20)] but addition of an electronic term, which was not position restricted, resulted in considerable improvement [Eq. (21)].

$$pI^{50} = 1.54\,\pi^3 + 4.04 \quad (20)$$

with
$n = 12$
$s = 0.724$
$r = 0.746$

$$pI^{50} = 1.25\,\pi^3 + 2.2\Sigma\sigma + 4.04 \quad (21)$$

with
$n = 12$
$s = 0.457$
$r = 0.92$

In both the above equations, pI^{50} is the negative log of the concentration causing 50% inhibition. Further examples of the importance of electronic terms are given in the following section.

Electronic Effects

Even when the hydrophobic parameter plays an important role in toxicity there are many instances where an electronic parameter is needed to explain the activity of certain functional groups. The *in vitro* inhibition of the purified enzyme, alcohol dehydrogenase, by 4-substituted pyrazoles produced data (*12*) fitting this equation:

$$\log K_i = 1.22 \log P - 1.80\,\sigma_m + 4.87 \quad (22)$$

with
$n = 14$
$s = 0.32$
$r = 0.985$

where K_i is the inhibition constant. The Hammett sigma *meta* constant (*13*) is the obvious choice for the 4 position on a pyrazole ring, because it is meta to each of the two ring nitrogens.

The negative sign of the coefficient for the electronic parameter says that electron withdrawal from the ring decreases the binding strength. Substitution of pyrazoles with hydrophobic groups increases their binding to this enzyme, as indicated by the large positive coefficient of the log *P* parameter. This enzyme probably has limited hydrophobic binding space, and increasing solute log *P* past a certain point would not increase inhibition, but this point was not reached with the set studied.

When the same inhibitory action of the pyrazoles is studied in isolated liver cells, however, the processes of transport through the cell wall and absorption to plasma proteins can be expected to place an upper optimum on their hydrophobic nature. In Eq. (23) which correlates the whole cell data, a negative coefficient with the term (log *P*)2, results (*12*) in a parabolic relationship which fits the data rather well:

$$\log 1/K_i = 1.27 \log P - 0.20(\log P)^2 - 1.80\,\sigma_m + 4.75 \quad (23)$$

with
$n = 14$
$s = 0.320$
$r = 0.971$

Isolated liver cells carry out the dehydrogenation reaction with about the same efficiency as does the intact liver, and so it is gratifying to see that, in comparing *in vitro* and *in vivo* reactions, the electronic effect and the intrinsic activities remain essentially the same. The only additional term needed to correlate the *in vivo* results is one allowing for optimized transport to the active site.

Special Hammett sigma parameters apply to structures which allow the electronic effect to operate on a "through resonance" basis (*14*). In biological reactions, just as in ordinary solution chemistry, this parameter, called sigma minus, proves to be more effective whenever the reactant structure is appropriate for its use. For example, the ring-attached ester oxygen engages in "through resonance" in phenyl phosphates. When these compounds act as inhibitors of acetylcholinesterase, *in vitro*, Eq. (24) can be derived (*15*) for the 50% inhibition level:

$$\log 1/C = 2.37(\sigma') + 4.38 \qquad (24)$$

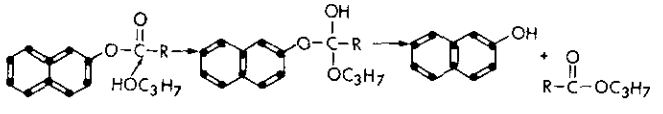
with
$n = 6$
$s = 0.297$
$r = 0.985$

For *in vivo* insecticidal activity, as is seen in the LD_{50} data of phenyl diethyl phosphates acting against houseflies (*15*), a hydrophobic term as well as an electronic term is justified:

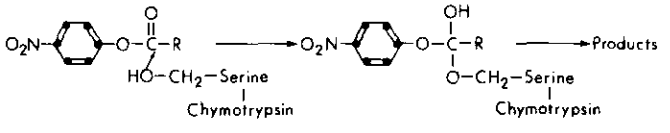
$$\log 1/C = 2.65(\sigma^-) + 0.36 \log P + 2.44 \qquad (25)$$

with
$n = 8$
$s = 0.206$
$r = 0.990$

In this set, none of the solutes approached the optimal hydrophobicity, and so a $(\log P)^2$ term or a bilinear equation was not justified.



(26)



(27)

Steric Effects

If the Hammett-Taft parameters with "extra-thermodynamic methodology" (*16*) is actually appropriate to detect patterns in reactions which extend across the boundaries of physical organic chemistry into biochemistry and animal and plant physiology, then one would expect to see some of the latter reactions to be dependent on the steric parameter E_s (*17*). An interesting example of such a reaction is a transesterification with and without the enzyme, chymotrypsin [Eqs. (26),(27)]. For the ester exchange [Eq. 26)], the data could be fit (*18*) into Eq. (18):

$$\log k = 1.67\, E_s - 1.13 \qquad (28)$$

with
$n = 6$
$s = 0.156$
$r = 0.989$

As the group R gets bulkier, its E_s value becomes more negative. Therefore, steric bulk hinders ester exchange as would be expected. What might not be expected, but which gives gratifying support to the use of Hammett-Taft parameters in biological QSAR, is based on data from hydrolysis of *p*-nitrophenyl esters by chymotrypsin (*18*):

$$\log k_2/k_m = 1.76\, E_s + 0.79\pi + 2.23 \qquad (29)$$

with
$n = 8$
$s = 0.201$
$r = 0.981$

The acylation step [Eq. (27)], which involves formation of a tetrahedral intermediate between substrate and the serine group of the enzyme, corresponds to the intermediate in the uncatalyzed reaction. The coefficient of the E_s term in Eq. (29) indicates that the reaction constant k_2 for this step, is also reduced by bulkiness of the R group to the same degree as in the *in vitro* reaction [Eq. (28)]. On the other hand, enzyme-binding, which favors the reaction, is enhanced by increased hydrophobicity.

Summary

This very brief overview of some patterns thus far recognized in biological QSAR at various levels of biocomplexity gives an idea of the types of information that we believe must be stored in a data base if it is to accomplish its purpose: expediting the discovery of more interesting patterns in the future. The Hammett-Taft parameters, which have proved so useful in solution chemistry, certainly must be made readily accessible, but automatic loading of preselected values should be optional. To refine the calculations from physicochemical data, which often is an order of magnitude more precise than that from biological systems, many modi-

fications of these parameters have been reported and discussed in the literature—so many in fact, that newcomers to the field may balk at the task of learning how to apply them all correctly. Some success has been made in the attempt to factor all of these electronic parameters into just two effects: field and resonance (*19*). Of course, it takes more data points to justify the two separate parameters where one sigma parameter served originally, but the further data gathering may be worth the effort if the separation of effects adds to overall understanding.

Currently it appears that there is so much "noise" in biological data that only about eight of the 35 types of electronic parameters are sufficiently different from the others that their use is justified in biological application. Nevertheless, there is an undisputed trend toward better precision in physiological data and especially in biochemical studies of purified enzymes. Therefore, the effort to keep these "unusual" parameters current and readily available in the database is justified, because at some later date they may provide insight into some subtle effects not discernable with techniques now in use. We cannot stress too strongly the fact the biological QSARs rest on a foundation of the corresponding QSARs in physical organic chemistry. Over the last 15 years we have carefully studied over 2200 physico-chemical data sets before adding them to the databank of the Pomona Medchem Project. These sets cover most of the reactions in solution chemistry that could be found relevant in biochemistry. The following sections will describe the databank as it is structured to accommodate biological information, since that is the more complex and poses greater problems for effective searching. It should be kept in mind that the physicochemical data are stored in a way that interfaces with the biological, a feature that encourages combined searches.

Database Structure and Content

From a maintenance aspect, it makes sense to separate the overall database into two parts: parameter information and activity information. When this work was initiated, it was uncommon for the majority of potential users to have ready access to a computer terminal on which searching and computation could be combined, and so the original design stressed effective methods of search and retrieval of information from hard copy (computer print or microfiche). One important facet of such use—searching for chemical structures by types of substructures—required the user to have at least a limited reading capability of Wiswesser line notation (WLN) (*20*). This could be acquired in a few hours, but many chemists resist its use. With the current availability of computer facilities and more "user-friendly" methods of structure searching, we have modified structure storage to take advantage of these advances, but we still maintain WLN where it can aid the user who by choice or of necessity searches manually. But of equal importance, WLN provides us with a more dependable and faster method of structure entry.

Parameter Databank

Parameters from Measured Physical Properties. Hydrophobicity is taken as the log of the partition coefficient between a nonpolar and an aqueous phase. Octanol is the preferred nonpolar phase, and only values for neutral solutes in this solvent are placed in the "select" category. While it is possible to automatically load these "select" values directly from the computer, rarely will all the desired values be found there. It is often more practical to use calculated log P (see below) or π constants, which are measures of substituent hydrophobicity. Most of the polar substituents in the data base have multiple entries of π constants, suffixed to indicate whether they apply to aliphatic, vinyl, or aromatic attachment, and, if the latter, if other polar groups present have altered their values by electronic interaction (*4,21*).

Electronic parameters include as many as 35 types of Hammett sigma constants which may be stored for any given substituent. The only ones designated "preferred" and available for automatic computer retrieval are: sigma *meta*, sigma *para*, sigma inductive, sigma star, sigma minus, and sigma plus. The orthogonalized parameters for the field and resonance effects, F and R (*14*), are also in the latter category.

Of steric parameters, the Taft E_s constant (*17*), derived from rates of acid hydrolysis of esters, is normally the preferred parameter, but the values containing Hancock's correction for hyperconjugation (*22*) can be retrieved if specified and available.

Molar refractivity is calculated via the Lorenz-Lorenz equation (14) from refractive index and density.

Parameters Calculated Directly from Structure. Log P (octanol/water) can be calculated by the fragment method (*4,23,24*) manually or by computer. The computer program, CLOGP-3, also retrieves a measured value from the "select" list for comparison, if one is available.† There are many advantages to using log P for the entire molecule rather than the substituent hydrophobic constant, π, as was commonly used in earlier work. Even when restricted to substituents on aromatic rings, different π value sets must be maintained for polar substituents depending on what other polar groups are present. The safest procedure is to calculate log P for all compounds including the "parent" if there is one. If it seems advantageous to study hydrophobicity in terms of substituent effect, then the log P of the parent is subtracted from each.

Provision for storage and retrieval of charge densities from molecular orbital calculations is being considered. To date, these measures of electronic effects have not been used as successfully in correlation analysis as the

†Part of software available through Pomona Medchem Project, Chemistry Dept., Claremont, CA 91711.

Hammett sigma values, and their inclusion would increase the size and complexity of the database manyfold.

Charton's steric parameters ν (*25*), calculated in part from van der Waal's radii, are stored for the most commonly encountered substituents, as are Verloop's sterimol parameters (*26*). This is especially important in view of the dearth of measured values for Taft E_s.

Molar refractivity, which reflects the overall bulkiness of the substituent together with a measure of London dispersion forces (*27*), can be calculated on the basis of an atom and bond additivity procedure programmed as CMR.† When calculated for a substituent rather than for an entire molecule, a conjugation correction may apply when attachment is on an aromatic or vinyl carbon atom.

In terms of the "mechanics" of parameter entry, the first problem to be dealt with is the method of handling chemical structures. At the outset, for solute log P values, WLN was chosen because structure entry was swift, storage space was minimal, and readily available alphanumeric sorting routines produced "hard copy" which could be efficiently searched manually. The original contracted WLN format was maintained, and structural isomerism was generally (but not always) indicated by the prescribed suffixes. Molformula (always) and CAS registry number (usually) accompanied each entry as additional means of access.

The use of WLN still has enough advantages to justify its continued use for entering molecular structure, but we have developed computer programs to circumvent most of its shortcoming as a searching tool. WISCT* converts WLN to a connection table, which makes more sophisticated searches possible. Also a new line-notation system called SMILES† has been implemented. It is very "user-friendly," taking perhaps ten minutes to learn. It dependably produces a unique notation which is more sparing of computer space than a connection table. And a program, WISSM, which converts WLN to SMILES completes the interconversion network.

For manual searching, files of substituent structures are ordered on "molformula," i.e., sum of atom types. Of course, substituent formula is often not a unique descriptor, and so each entry is accompanied by a typewritten structure (not standardized, but easily readable) and by the substituent WLN called SWLN. SWLN is in many ways easier to read and write than WLN, because the notation always begins at the attachment bond which is designated by an asterisk (*). SLWN is currently used for data entry and substituent searching, but provisions are being made to allow SMILES as an alternative.

We would like the primary parameter data base to be structured so that it will serve the needs of the most demanding specialist. To do so it must not only list a multiplicity of parameter types, but in many cases, quite a few values for each type even for the same structure. The justification of entering values that are nearly alike comes from the fact that they are usually from different secondary reaction sets, and this establishes their reliability to the primary set. We believe that very few of the "outlier" values present are the result of experimental error. The original reference often contains information about the reaction conditions, which could lead one to conclude that it might constitute a closer model to that under study.

The reference list, in addition to the usual journal designation, contains a short description of the method of parameter determination; e.g. "rate of solvolysis," "reaction rate with diphenyldiazomethane," "F-19 NMR," etc., which many times eliminates the need of searching for a journal that is not readily available.

There are alternative methods of determining parameter reliability other than accepting the judgment of those who assemble the data base. For instance, one can estimate the chemical reliability of a parameter for a given substituent of a specified type by observing the range of reported values in various model systems. Examples of lower reliability would be found in the comparatively large range of sigma para values for -NH- or -OH. Estimation of reliability in biological applications is a little more involved. However, it is not difficult to write a computer search of the structure–activity bank (see below) to determine how the deviation for a specified substituent compares with the standard deviation of the equation in which it appears and in which sigma para (for example) is found to be a significant parameter. One would not be surprised to find the predicted biological activity of p-NH- and p-OH more frequently at or beyond the "outlier boundary" (two times the standard deviation) because of the lower "chemical reliability." We plan to provide this confidence measure as a continually up-dated feature in the editions of the future data base.

Structure–Activity Banks: Biological and Physicochemical

The primary data are presently stored in the form of one, or at most two, regression equations. The equation stored is not arrived at by stepwise regression or a combination of the forward and reverse stepwise process. It is, instead, arrived at by examining all possible combinations of the most reasonable parameters as derived by a fast-permutation algorithm. The best 200 or 300 equations—out of perhaps millions possible if all combinations of eight or ten parameters are considered—are printed in ascending order of standard deviation and can be quickly scanned for those making the most sense both from a chemical and statistical standpoint.

Currently the structure activity databases contain no compounds whose measured activity is not part of a set of similarly acting substances. This is a weakness which

†Part of software available through Pomona Medchem Project, Chemistry Dept., Claremont, CA 91711.

is being remedied as quickly as funding permits. Strychnine serves as an example of the kind of important data gap which results. Its activity as a convulsant has become a standard against which anticonvulsants are often measured. Yet strychnine does not appear in the bank, because its convulsant activity has not been studied as part of a structurally related series.

In the regression equations making up the primary data, the dependent variable is the measure of activity. However, the preferred form for this variable is not the activity itself, such as percent hydrolysis in a chosen time period, but the concentration of reactant needed to produce a chosen activity level. The concentration should be expressed, whenever possible, as moles per liter, or moles per kilogram of test system, which is very similar. It is conventional to use the log of the reciprocal concentration, so that larger values denote higher activity. This also puts it on a free energy related basis. Any of the sets where the dependent variable is the log of the Michaelis constant K_m or the dissociation constant for an enzyme-inhibitor complex K_i can also be compared for slope and intercept in this manner.

It is, of course, much simpler to measure relative biological activity at a constant initial reactant level, and the bank contains a number of equations of this type, especially from the earlier literature. Some useful qualitative comparisons can be made with them, but they should be carefully excluded in any sweeping computer search. For these reasons, we are attempting to replace as many "direct-action" dependent variables as possible with comparable sets for which adequate dose/response data has been obtained.

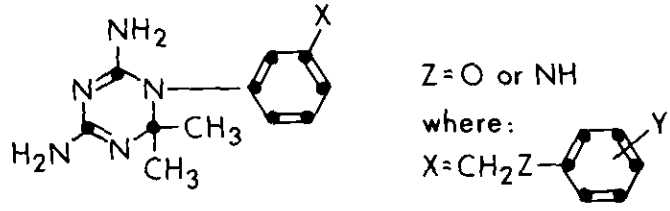
The independent variables in bank regression equations are generally the physicochemical parameters previously discussed. With the exception of molar refractivity (MR), and the calculated steric parameters (Charton's ν, and the Sterimol constants) they are on a log basis. MR has been divided by ten, and the others appropriately scaled to more nearly coincide with the other parameters.*

The bioactivity database contains a number of equations which use indicator variables instead of a true parameter. Like the "extrathermodynamic" method itself, this procedure has been criticized on theoretical grounds. In our use of indicator variables, a parametric method is combined with a nonparametric one, like the Free-Wilson (*28*). But even if purists are somewhat offended by this combination, one must face the reality of working with some structural variations that are best dealt with by the answer to the question: "present?" or "not present?" and so they are arbitrarily given the respective values of 1.0 or 0.0. The coefficient that appears in the resulting regression equation is a good measure of the relative importance of that feature in determining activity. As an alternate method, one could make separate sets for each of these structural variations, and the difference in intercepts would correspond to the indicator coefficient in the combined equation. The disadvantage to this second approach is that some of these sets might contain too few data points to establish slopes and intercepts with much confidence and the similarities might be overlooked. Of course, the real challenge is to devise a continuous variable that can be applied to that particular structural variation, making it a true parameter.

To make proper use of regression equation information, close attention must be given to the statistics accompanying it. The Pomona Medchem structure–activity data bases keep this statistical information readily accessible, for many searches could be misleading if it were not included. As a very simple example, one might search for all systems which, in the hydrophobic binding of reactants, appear to completely de-solvate them in a manner comparable to the octanol/water model. They should then show a coefficient with log P close to 1.0; say, 0.9 to 1.1. But to meet this criterion, the coefficient of each log P term can be expanded by the 95% confidence interval associated with it.

Whenever a squared parameter value is used as an independent variable and a parabolic relationship results, the optimal value of that parameter (usually hydrophobicity) is calculated together with its 95% confidence limits. Frequently there are not very many data points on the right side of the curve, and so the limits on the optimal value may include infinity. Obviously some searches should be made with these doubtful values excluded.

More and more frequently one finds data good enough to support a different slope for the upward and downward legs of a bilinear relationship. This is often the case with enzyme inhibition or cells in culture. An example is seen in Eq. (30) from a study (*30*) of growth inhibition of L5178 leukemia cells by triazenes having the structure:



a. b.

$$\log 1/C = 1.32\pi - 1.70 \log (\beta \cdot 10^{\pi} + 1) + 0.44I + 8.10 \quad (30)$$

with
$n = 37$
$s = 0.274$
$r = 0.929$
$\pi^{\circ} = 0.76$

*Scaling is just a matter of convenience for most applications, but if the parameters are used in cluster analysis, a more exact mathematical procedure must be followed (*29*).

Table 1. *S. typhimurium* TA92(04B)*1-(X-phenyl-3,3-dialkyltrizines* Mutation, 30/10⁸ cells, with S-9 liver fraction* J. Med. Chem., 22, 473 (1979)*

No.	X	R	WLN	log P	σ^+	Obsd	Calcd	Dev
1.	4-$CONH_2$[a]	*t*-Bu	ZVR DNUNN1&X	2.61	−0.30	3.83	6.26	2.43
2.	3,5-CN	CH_3	NCR CCN ENUNN1&1	2.18	1.12	3.46	3.50	0.04
3.	4-SO_2NH_2	CH_3	ZSWR DNUNN1&1	0.98	0.57	3.49	3.15	0.34
4.	3-$CONH_2$	CH_3	ZVR CNUNN1&1	1.21	0.28	3.51	3.86	0.35
5.	4-$CONH_2$	CH_3	ZVR DNUNN1&1	1.20	0.36	4.04	3.72	0.32
6.	4-$CONH_2$	allyl	ZVR DNUNN1&2U1	2.09	0.36	4.16	4.65	0.49
7.	3-$NHCONH_2$	CH_3	ZVMR CNUNN1&1	1.29	−0.03	4.19	4.45	0.26
8.	4-CN	CH_3	NCR DNUNN1&1	2.39	0.66	4.43	4.47	0.04
9.	4-$COCH_3$	CH_3	1VR DNUNN1&1	2.27	0.50	4.47	4.61	0.14
10.	H	CH_3	1N1&NUNR	2.59	0.00	5.32	5.75	0.43
11.	4-$CONH_2$	*n*-Bu	ZVR DNUNN4&1	2.46	0.36	5.41	5.03	0.38
12.	4-$NHCONH_2$	CH_3	ZVMR DNUNN1&1	1.25	−0.84[b]	5.59	5.72	0.13
13.	4-$NHCOCH_3$	CH_3	1VMR DNUNN1&1	1.54	−0.60	5.83	5.64	0.19
14.	4-CF_3	CH_3	FXFFR DNUNN1&1	3.70	0.61	5.99	5.91	0.08
15.	3-CH_3	CH_3	1N1&NUNR C	2.85	−0.07	6.44	6.14	0.30
16.	4-Cl	CH_3	GR DNUNN1&1	3.33	0.11	6.48	6.34	0.14
17.	4-CH_3	CH_3	1N1&NUNR D	2.93	−0.31	7.00	6.61	0.39
18.	4-C_6H_5	CH_3	1N1&NUNR DR	4.40	−0.18	7.67	7.93	0.26

[a]This data point not used in the derivation of Eq. (*1*).
[b]Estimated value.
$\log 1/C = 1.09\ (\pm 0.17) \log P - 1.63\ (\pm 0.35)\sigma^+ + 5.58\ (\pm 0.95)$; $n = 17$, $r = 0.974$, $s = 0.315$.

In this equation I indicates whether a bridge connects the *meta* substituent X with another ring, and β is a term that is responsible for the value at which π reaches an optimum. The slope of the "downward leg" in this relationship is given by the sum of the coefficients of the first two terms—in this case, - 0.52. β is calculated by an iterative procedure, and the "jacknife" procedure (*31*) provides confidence levels on the two slopes and $\log P_o$.

The methods for entering the structure of the test compounds and their storage and retrieval have been discussed above and apply to both biological and physicochemical data bases. Search screens (*32*) have been extracted from the WLNs, and, when applied at the set level, can expedite searches made with small computers. For most of the newer machines, computation time is so cheap that the flexibility of searching by SMILES or by connection table is easily justified.

As the biological data bank took shape, it quickly became apparent that the nomenclature used by the original investigators to describe biological systems was not sufficiently standardized to enable the usual "text searching" of the names as originally reported. Some common examples of the problem can be mentioned: "feline" is rarely used for cat, but "dog" and "canine" are both common; "mouse," "mice" and "murine," are all used referring to more than one test animal. We have always tried to use the simplest form of the plural word, but we also include a three-character class descriptor (*33*) beginning with the lowest numbers for the simplest systems. The six numbers currently in use are: 01, nonenzymatic macromolecules; 02, enzymes; 03, organelles; 04, single cell organisms; 05, isolated parts of organs; 06, large functioning organisms.

The third character is optional. It is a letter standing for a subdivision of that class, e.g., 04B refers to bacteria *in vitro*. The class descriptors can be combined. 04B, 06A refers to action against bacteria in a nonhuman animal.

The simplest of the possible systems involved, if it is clearly the target of the action, is given first billing. For instance, in a test of antibiotics against *Staphylococcus aureus* in mice, the system would be: *S. aureus*; mice; 04B, 06A. Inhibition of alcohol dehyrogenase in isolated liver cells would be entered as: dehydrogenase, alcohol; cells, liver; 02A,04C. The strain or other classification information is entered after the organism name, e.g., mice, nude; or monkey, rhesus.

Coping with the biological action descriptor is the most difficult problem in attempting to optimize the structure of this data bank. If they were entered just as reported, there would be perhaps one-third as many different actions as there are data sets. The problem becomes more acute as the test organism becomes more complex and very subtle behavior is being recorded as the end point. Narcosis, anesthesia, and analgesia are recorded as three separate actions, but distinction between them on the basis of the "tail flick" test seems to a nonexpert, to be some what arbitrary.* There is even some degree of ambiguity in simple bacterial tests. Some while others are reported as "kill." Obviously if the test were run long enough, bacteria showing 100% growth inhibition would be dead, unless they formed spores. But these details are not always clearly reported in the papers.

Our approach to this problem was to reduce the variety of "biological actions" by combining two or several

*After a rat is given an appropriate dose of the test compound, an intense spot of light is focused on its tail. If it feels pain, it flicks its tail away. Presumably, an analgesic can dull this pain with no noticeable loss in consciousness.

under one heading when this seemed appropriate. Then a thesarus was prepared of those remaining. Future entries are made after consulting this thesarus, and a reasonable effort is made to limit new designations.

Important test conditions are entered as secondary descriptors of the action. Thus "growth inhibition" might be followed by: "37 deg, 18 hr, Hartley broth + serum."

Table 1 shows the computer print of an example data set from the biological activity bank.

The "system," "compound," "action" and "reference" subfiles constitute the "Title information" for this set and are set apart by asterisks. They follow the format described above and thus need little comment, except to re-emphasize that, since the dependent variable is given as log $1/C$, C is the concentration of triazene needed to reach a mutation level of 30 mutants per 10^8 cells and that the presence of the S-9 liver microsome fraction is considered a modifier of "activity" rather than a modifier of the system. For each set, a fifth (optional) file is available for any notes describing any unusual steps in the computation—in this example, details of the steps taken to estimate the sigma plus value for p-$NHCONH_2$ would be entered in the fifth file, which is not computer-searchable, however.

Search Strategy Examples

A question which might often be asked by those concerned with environmental hazard assessment is the following: In which systems is this particular chemical substructure most toxic? For instance, N-monosubstituted-thiocarbamates?

The general search strategy could be set us as follows.

1. Dependent variable (i.e., activity) must be "log $1/C$," and > 3.0; AND
2. Compound SMILES must contain uniqued version of: *SC(=O)N* AND isolating carbons at asterisks can be of any type; THEN
3. Print systems in order of decreasing log $1/C$, with associated activity.

The majority of biological actions are toxic to some degree, but if a scan of the printout from the above search is too inclusive, and the output too voluminous to screen out by hand, then some attention must be given to the descriptors of the unwanted activities in order to add an appropriate restriction at step one.

As another example of a question which is easily answered with the present databank structure, one could ask: "How large is the range of ideal hydrophobicities of phenolic bactericides?"

1. System subfile must contain '4B'; AND
2. Compound structure file in SMILES must contain either: "-Oc"; OR "cO-" OR "c(O)" (where "c" is aromatic carbon); AND
3. Equation subfile contains a term in log P_o or π_o; THEN
4. Put log P_o and π_o in separate files; associate parent SMILES with each π_o; and print out in numerical order. To merge the π_o file with the log P_o, one must enter the parent SMILES into CLOGP-3 and add that value to each π_o. With prior knowledge of the bank contents, one can be sure that the majority of actions reported with bacterial systems are "kill," but to eliminate the few others, such as bacterial luminescence, one could add an appropriate restriction to the "action" subfile.

Conclusions

The use of substituent parameters, in linear regression equations or equations containing power terms, to rationalize either physicochemical or biochemical reactions cannot be justified on purely theoretical principles. While some of the parameters are from equilibrium data and thus have a basis in thermodynamics, others are from rate data and do not. Furthermore, in applying these values to any reaction system other than the one in which they were measured, one assumes either no entropy change or else an entropy/enthalpy cancellation. And yet, despite these seemingly valid objections, the "extrathermodynamic" methodology has been very successfully applied to physicochemical, biochemical and physiological interactions. Careful interpretations of the equations so far collected has given useful, if somewhat limited, insight into likely mechanisms of action. But more important, it has focused attention on certain similarities of action common to simple model systems as well as to complex whole animals and plants.

This methodology is not in competition with newer molecular graphics techniques which fit known substrates and prospective inhibitor structures to well-characterized active sites of enzymes. On the contrary, it has been shown to be a valuable adjunct to them (*34*), especially in the field of drug design. In the areas of predictive toxicology and environmental hazard assessment, however, such detailed knowledge of key enzyme structure is a seldom-enjoyed luxury, and regression analysis with readily available substituent parameters offers an "off-the-shelf" tool of proven utility.

This work has been supported by the National Institutes of Health under grant GM-30362

REFERENCES

1. Hansch, C., and Dunn, W. Linear relationships between lipophilic character and biological activity of drugs. J. Pharm. Sci. 61: 1–19 (1972).
2. Hansch, C., and Glave, W. Structure-activity relationships in membrane perturbing agents. Mol. Pharmacol. 7: 337–354 (1972).
3. Hansch, C., and Steward, A. The use of substituents constants in the analysis of the structure-activity relationships in penicillin derivatives. J. Med. Chem. 7: 691–916 (1964).
4. Fujita, T., Iwasa, J., and Hansch, C. A new substituents constant. *Pi*, J. Am. Chem. Soc. 86: 5175–5180 (1964).
5. Bird, A., and Marshall, A. Correlation of serum binding of penicillins with partition coefficients. Biochem. Pharmacol. 16: 2275–2290 (1967).
6. Helmer, F., Keihs, K., and Hansch, C. The linear free energy relationships between the partition coefficient and binding and

conformation of macromolecules by small organic compounds. Biochemistry 7: 2858–2863 (1968).
7. Kiehs, K., Hansch, C., and Moore, L. The role of hydrophobic bonding of organic compounds by bovine hemoglobin. Biochemistry 5: 2602–2605 (1966).
8. Veith, G., and Kosian, P. Estimating bioconcentration potential from O/W partition coefficients. In: Physical Behavior of PCBs in the Great Lakes (D. Mackay et al., Eds.), Ann Arbor Science, Ann Arbor, MI, 1983, pp. 269–282.
9. Moffat, A., and Sullivan, A. The use of quantitative structure-activity relationships as an aid to the interpretation of blood levels in cases of fatal barbiturate posioning. J. Forens. Soc. 21: 239–248 (1981).
10. Hansch, C. Quantitative approaches to pharmacological structure-activity relationships. In: International Encyclopedia of Pharmacology and Therapeutics: Structure-Activity Relationships, Section 5, Vol. 1 (C. J. Cavallito, Ed.) Pergamon Press, Oxford, 1973, pp. 75–166.
11. Woodson, L., Ames, M., Selassie, C., Hansch, C., and Weinshilboum, R. Thiopurine methyltransferase; Aromatic thiol substrates and inhibition by benzoic acid derivatives. Mol. Pharmacol. 24: 471–478 (1983).
12. Cornell, N., Hansch, C., Kim, K., and Henegar, K. The inhibition of alcohol dehydrogenase *in vitro* and in isolated hepatocytes by 4-substituted pyrazoles. Arch. Biochem. Biophys. 227: 81–90 (1983).
13. Hammett, L. Physical Organic Chemistry, McGraw-Hill, New York, 2nd Ed., 1970.
14. Hansch, C., and Leo, A. Substituent constants for correlation analysis in chemistry and biology. Wiley-Interscience, New York, 1979, pp. 2–3.
15. Fujita, T. Extrathermodynamic structure-activity correlations. In: Biological Correlations—The Hansch Approach (W. Van Valkenburg, Ed.), Advances in Chemistry Series 114, American Chemical Society, Washington, DC, 1972, pp. 1–19.
16. Leffler, J., and Grunwald, E. Rates and Equilibria of Organic Reactions. Wiley, New York, 1963.
17. Taft, R. Separation of Polar, Steric, and Resonance Effects in Reactivity. In: Steric Effects in Organic Chemistry (Melvin S. Newman, Ed.) Wiley, New York, 1956 pp. 556–675.
18. Hansch, C., and Coats, E. α-Chymotrypsin: a case study of substituent constants and regression analysis in enzymic structure-activity relationships. J. Pharm. Sci. 59: 731–743 (1970).
19. Hansch, C., Leo, A., Unger, S., Kim, K., Nikaitani, D., and Lien, E. "Aromatic" substituent constants for structure-activity correlations. J. Med. Chem. 16: 1207–1216 (1973).
20. Smith, E. The Wiswesser Line-Formula Chemical Notation. McGraw-Hill, New York, 1968.
21. Leo, A. The octanol-water partition coefficient of aromatic solutes: the effect of electronic interactions, alkyl chains, hydrogen bonds, and ortho-substitution. J. Chem. Soc. Perkin Trans. II : 825–838 (1983).
22. Hancock, C., and Falls, C. A Hammett-Taft polar-steric equation for the saponification rates in *m*- and *p*-substituted alkyl benzoates. J. Am. Chem. Soc. 83: 4214–4216 (1961).
23. Rekker, R. The Hydrophobic Fragmental Constant. Elsevier, Amsterdam, 1977.
24. Leo, A., Jow, P., Silipo, C., and Hansch, C. Calculation of hydrophobic constant (log *P*) from π and *f* constants. J. Med. Chem. 18: 865–868 (1975).
25. Charton, M. The prediction of chemical lability through substituent effects. In: Design of Biopharmacological Properties through Prodrugs and Analogs (E. Roche, Ed.), American Pharmaceutical Assoc., Washington, DC, 1977, pp. 228–280.
26. Verloop, A., Hoogenstraten, W., and Tipker, J. Development and application of new steric substituent parameters in drug design. In: Drug Design (E. Ariens, Ed.), Academic Press, New York, 1976, pp. 165–208.
27. Pauling, L., and Pressman, D. The serological properties of simple substances. IX. J. Am. Chem. Soc. 67: 1003–1012 (1945).
28. Free, S., and Wilson, J. A mathematical contribution to structure-activity studies. J. Med. Chem. 7: 395–399 (1964).
29. Hansch, C., Unger, S., and Forsythe, A. Strategy in drug design. Cluster analysis as an aid in the selection of substituents. J. Med. Chem. 16: 1217–1222 (1973).
30. Khwaja, T., Pentecost, S., Selassie, C., Guo, Z., and Hansch, C. Comparison of quantitative structure–activity relationships of the inhibition of leukemia cells in culture with the inhibition of dihydrofolate reductase from leukemia cells and other cell types. J. Med. Chem. 25: 153–156 (1982).
31. Dietrich, S., Dreyer, N., Hansch, C., and Bentley, D. Confidence interval estimators for parameters associated with quantitative structure-activity relationships. J. Med. Chem. 23: 1201–1205 (1980).
32. Granito, C., Becker, G., Roberts, S. Wiswesser, W., and Windlinx, K. Computer-generated substructure codes (bit screens). J. Chem. Doc. 11: 106–110 (1971).
33. Hansch, C., Leo, A., and Elkins, D. Computerized management of structure-activity data. I. Multivariate analysis of biological data. J. Chem. Doc. 14: 57–61 (1974).
34. Smith, R. N., Hansch, C., Kim, K., Omiya, B., Fukumura, G., Selassie, C., Jow, P., Blaney, J., and Langridge, R. The use of crystallography, graphics, and quantitative structure-activity relationships in the analysis of the papain hydrolysis of X-phenyl hippurates. Arch. Biochem. Biophys. 215: 319–328 (1982).