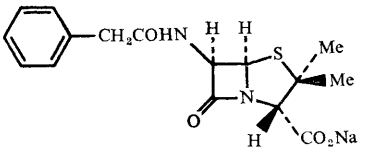
Quantitative Drug Design

By G. Redl, R. D. Cramer tert., and C. E. Berkoff
TECHNOLOGY ASSESSMENT, SMITH KLINE & FRENCH LABORATORIES, PHILADELPHIA, PENNSYLVANIA 19101, U.S.A.

1 Introduction

The discovery and clinical availability of unique drug therapy are increasingly governed by a number of severe competitive and regulatory constraints.[1] To develop a potentially useful therapeutic agent to the point of U.S. regulatory approval typically demands an investment of some $11.5 million in support of an 8—10 year programme. By 1977 the necessary investment is expected to rise to $40 million.[2] With little understanding at the molecular level of how one designs a new agent in any field of clinical need, only 1 in perhaps 15000 compounds emerges as a commercial product. Any rational synthetic strategy for improving these odds must appeal to those concerned with the practical aspects of discovery and development of new drug therapy.

Although still an emerging discipline, quantitative drug design can, in very practical terms, contribute both to the discovery of new therapeutic agents and to the progress of biomedical research in general. While, to our knowledge, there is no example of any molecule having found its way into the physician's armamentarium *via* quantitative structure–activity analysis, developing techniques are becoming increasingly more capable of directing synthetic effort from compounds that have a low probability of success to structural variations often overlooked by the most experienced and imaginative medicinal chemist. These powerful tools become even more relevant as modern organic chemistry makes vast numbers of compounds synthetically accessible. For example, following the discovery of penicillin (1) as an effective and useful antibiotic in



Sodium Penicillin G
(1)

[1] F. A. Robinson, *Chem. in Britain*, 1974, **10**, 129.
[2] L. H. Sarett, *Research Management*, 1974, 18.

the treatment of bacterial disease, many laboratories set out to modify the structure of the molecule to improve its efficacy while reducing adverse side-effects. Penicillin thus became a 'lead' structure – the point of departure for further synthesis and biological testing. In the typical quest for a better drug the lead structure is modified in stepwise fashion, changing both substituent groups and parent ring system. It is a sobering thought, however, that in the case of penicillin, for 20 substituents (including hydrogen) attached to 3 of the 5 available positions of the phenyl ring, the number of possible analogues is 36537. Coupling these variations with different amide side-chains, structural changes at other positions on the thia-azabicycloheptane nucleus, and variation of the heterocyclic ring system *per se* (including stereochemical possibilities at extant asymmetric centres), the medicinal chemist is rapidly confronted with astronomic numbers of possible structures; many of these may be arguably reasonable in terms of their biological potential, their synthetic (or semi-synthetic) accessibility, and, in very commercial terms, their patentability. Indeed, enormous effort has already been expended in the synthesis of literally thousands of penicillins and the related cephalosporins, very few of which have in fact emerged as successful therapeutic agents. The discovery of quantitative structure–activity relationships can reduce such problems to practical dimensions and increase the chances of a synthetic programme successfully meeting its objectives.

The ultimate goal of drug design is to enable the chemist to design compounds with a prescribed biological profile. The achievement of this aim still appears to be far in the future and will require breakthrough advances – especially in our understanding of biological and disease processes at the molecular level. Two approaches to the problem have emerged and both have been of significant (albeit indirect) value in the discovery and development of new drug therapy. The biochemical approach is to design new drugs on the basis of established and/or hypothesized models and mechanisms of drug action at the molecular level. The other, and the focus of this Review, is primarily a statistical approach in which a variety of computational techniques are applied to establish correlations between the level of a given biological activity and other measurable (or calculable) properties associated with the chemical structure of the molecule. The separation of the two approaches is, of course, artificial; they must coalesce if the ultimate goal of drug design is to be achieved. It is, in fact, an indication of the state of the art that the two approaches can be pursued with only very small degrees of overlap or cross-impact.

Virtually all of the published literature in drug design has dealt with the analysis of series of chemically closely related compounds, where the primary objective is the design of the optimally active compound within a lead series, *i.e.* 'lead optimization'. In practice, drug-design problems cannot be restricted to a selected compound series. Although very few attempts have been described dealing with the problem of analysis of structurally diverse compounds, its importance and challenge have persuaded us to give strong emphasis in our own studies to this particular aspect of the overall problem. This we call 'lead generation'.

Drug design has been the subject of numerous articles and monographs.[3] In the review that follows we attempt to describe the state of the art in very practical terms to a general chemical readership. In this context, it is worth noting that much of the developing methodology lends itself to the correlative analysis of chemical structure with any measurable or calculable property of a molecule, whether biological, chemical, or physical. Thus, appropriate quantitative drug-design techniques can, for example, also be applied to an understanding of chemical reactivity, interpretations of spectral behaviour, and to the design of other new chemical products, such as paints, plastics, and pesticides.

2 General Considerations

A. Biological.—We define a drug as a chemical substance that exerts a reproducibly observable effect on a biological system. The effect may be called biological activity or biological response, and is usually dose-dependent. The dose is the quantity or concentration of the administered compound; the response is the observed effect, which can be as objective as the percent inhibition of an enzymatic reaction or a change in heart rate, or as subjective as a change in the mood or posture of a test animal. In view of the complexity of factors that affect the *in vivo* activity of a drug (absorption, transport, tissue distribution, metabolism, *etc.*), it is not surprising that structure–activity correlation has been most successful with data obtained from less complex biological systems such as isolated organs, cell cultures, or purified enzymes.

To reduce the dimensions of dose and response to a single parameter, biological activity is usually expressed as $\log(1/c)$, where c is the minimum concentration required to produce a specific response. Its evaluation requires interpolation of experimental results obtained at several concentrations; in practice this is carried out only for the relatively few compounds of special interest. In most primary test systems (screens), where compounds are initially tested at a single dose only, the activity may be expressed as a percentage (P) of some maximal response. For this type of data the logit transformation, $\ln[P/(100-P)]$, can be useful.[4]

Experimental variability of biological measurements is much larger than that of most physical or chemical measurements. For meaningful correlations, the need for reliable biological data with well-defined confidence limits is as evident as it is frequently ignored. The drug designer should satisfy himself that the experimental design of the test system, in particular the nature and reproducibility of the observed effects, provides data which are appropriate for analysis.

[3] (*a*) E. J. Ariens, 'Drug Design', Academic Press, New York and London, 1971; (*b*) W. P. Purcell, G. E. Bass, and J. M. Clayton, 'Strategy of Drug Design', John Wiley & Sons, New York, 1973; (*c*) A. Burger, 'Medicinal Chemistry', Part I, Wiley–Interscience, New York, 1970, pp. 25—245.

[4] J. Thomas, C. E. Berkoff, W. Flagg, J. J. Gallo, R. F. Haff, C. A. Pinto, C. Pellerano, and L. Savini (manuscript submitted for publication).

B. Computational.—In its most general sense, quantitative drug design embraces all attempts to relate biological activities mathematically to other properties of molecular structure. Inasmuch as the mathematically simplest relationships among several properties are linear equations, the majority of quantitative drug-design methods are essentially attempts to derive a linear equation of the form (1), where the x_i are structural properties and the coefficients a_i emerge

$$\text{biological activity} = a_0 + \sum_{i=1}^{m} a_i x_i \qquad (1)$$

from the analysis. This equation allows prediction of the biological activity of any compound for which the x_i are known.

Procedures for obtaining the m coefficients in equation (1) require an experimentally determined biological activity and a group of m structural properties for each of the n compounds in a series. Because of the relatively low precision of biological data and the uncertainty that a linear correlation model is applicable to any specific structure–activity relationship, good practice in drug design requires that n be considerably larger than m, preferably by a factor of five or more. The number of compounds in excess of the minimum ($m + 1$) required for an analysis is called the number of degrees of freedom.

The set of coefficients a_i that constitutes the solution for equation (1) is usually obtained from the data by the least-squares method using multiple-regression computer programs. Statisticians have developed a number of criteria[5] to evaluate the appropriateness of a regression equation such as (1) for correlating the data and for extrapolating to new results. The most important of these are: the multiple correlation coefficient (R), where R^2 is the proportion of variation within the observations that is explained by the equation; and the F-test, an assessment of the probability (p) that the relationship derived is actually a chance occurrence. It should be recognized that over-reliance on statistical criteria to the neglect of common sense is a dangerous and all too frequent abuse.[6]

Though undoubtedly an oversimplification, the assumption of a linear relation among biological activity and a few structural parameters has proved useful among series of related compounds where the biological activity can be quantified. To handle more complex drug-design problems, for example the analysis of data from the testing of structurally diverse compounds, the various methods of pattern recognition may ultimately prove useful.

The recognition of patterns involving two or three variables is readily achieved by the scientist, without computer aid, using spatial representations of the data. For example, the existence of any relationship, linear or otherwise, between activity and a single function of structure is easily detected by graphical means.

[5] G. W. Snedecor and W. G. Cochran, 'Statistical Methods', Iowa State University Press, Ames, 1967; see also ref. 36, p. 27.
[6] S. H. Unger and C. Hansch, *J. Medicin. Chem.*, 1973, **16**, 745.

A correlation of activity with two functions of structure can still be recognized by plotting the two functions on Cartesian axes with an activity classification indicated for each compound/point. For the analysis of multivariate data exceeding three dimensions, however, human cognitive processes are poorly suited. While there are nearly as many computational methods for pattern recognition as workers in the field, those recommended for chemical problems[7] and employed in drug design[8–11] include various forms of cluster analysis, discriminant analysis, and linear learning machines. Cluster analysis,[12] a mathematical technique for classification, seeks similar sets of data; similarity is defined in terms of a 'distance' between points representing the objects (*i.e.*, compounds) in multidimensional variable space. Different clustering procedures can yield quite different classifications of the same sets of data, depending on the precise definition of similarity and the heuristic methods used to simplify the computational burdens. Discriminant analysis,[13] a technique firmly grounded in classical statistical theory, seeks a linear equation which can be used to place an unknown object into either of two classes (*i.e.*, active or inactive). In geometric terms, it is obvious that discriminant analysis defines a hyperplane which optimally bisects a multidimensional data space. Linear learning machines,[14] heuristic methods originating in the field of artificial intelligence, are useful in deriving linear equations.

The power of any of these methods is enhanced by preprocessing the data in a variety of ways which may weigh all features equally, give weight to the features expected to have particular importance, or completely remap the features. One objective of such remapping is to reduce the data to two or three dimensions while retaining as much as possible of the original information. This kind of two- or three-dimensional representation can then be displayed by the computer, allowing visual pattern recognition by the scientist.[7]

3 Lead-optimizing Techniques

Lead optimization is the phase of the drug development process in which the principal goal is to improve the biological profile of a lead compound, typically by increasing the separation between a dose that produces desirable activity and a dose that produces undesirable side-effects. A lead-optimizing programme, which focuses on synthesizing and testing structural modifications of the lead,

[7] B. R. Kowalski and C. F. Bender, *J. Amer. Chem. Soc.*, 1972, **94**, 5632; 1973, **95**, 686.
[8] Y. C. Martin, J. B. Holland, C. H. Jarboe, and N. Plotnikoff, *J. Medicin. Chem.*, 1974, **17**, 409.
[9] B. R. Kowalski and C. F. Bender, *J. Amer. Chem. Soc.*, 1974, **96**, 916.
[10] K. C. Chu, R. J. Feldmann, M. Shapiro, G. F. Hazard, jun., C. L. Chang, and R. Geran, *Abstracts of 167th American Chemical Society Meeting*, April 1974, CHLT 24; K. C. Chu, *Analyt. Chem.*, 1974, in the press.
[11] K. H. Ting, R. C. Lee, G. W. A. Milne, M. Shapiro, and A. M. Guarino, *Science*, 1973, **180**, 417.
[12] R. M. Cormack, *J. Royal Statistical Assoc.*, 1971, **134**, 321.
[13] T. W. Anderson, 'An Introduction to Multivariate Statistical Analysis', Wiley, New York, 1958.
[14] N. J. Nilsson, 'Learning Machines', McGraw-Hill, New York, 1965.

requires the identification of a particular structural moiety which is associable with the observed biological activity of a compound. This problem is by no means trivial, as is well illustrated by the belated discovery of the cephalosporin antibiotics. The principal structural difference between the penicillins and the cephalosporins is that the thiazolidine ring of the former has been expanded to the corresponding thiazine ring. Nevertheless, some fifteen years of synthetic effort failed to uncover the worth of this relatively minor structural modification; the exciting antimicrobial utility of the cephalosporins was obliged to await the testing of soil microflora samples for its discovery. Although there clearly can be no final answer to the question of which part of a structure is responsible for the observed biological effects of a molecule, in practical terms the problem can be tentatively resolved. Choice of the structural moiety defining the scope of a lead-optimizing synthetic programme usually rests on objective and very practical considerations such as synthetic accessibility and potential patentability; unquestionably it also embraces more subjective issues such as a scientist's propensity for one particular type of chemistry over another.

A. The Physicochemical Model.—Introduction of different substituents into a lead molecule alters its chemical and consequently its biological properties in ways which can often be related linearly to the physicochemical properties of the substituents themselves. If such a relation can be found, knowledge of the physicochemical properties of unexplored substituents will permit prediction of the activities of the unsynthesized members of a lead series. These considerations form the basis of the physicochemical model which underlies the development of the 'multiple parameter' or 'linear free energy' approach to drug design. The wealth of publications devoted to the physicochemical approach, generally associated with the name of Hansch, suggests it to be by far the most popular of quantitative drug-design methods.

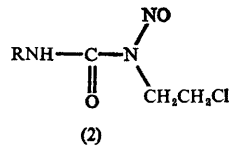
The physicochemical properties associated with a substituent may be loosely classified as electronic, steric, or solvent partitioning. However, it is not clear which laboratory measurements or calculated parameters best define a class of substituent properties.[15] For example, very different steric effects can be estimated from solution kinetics, crystallography, molecular models, quantum mechanics, and polarizability data. Understandably perhaps, the number of physicochemical substituent properties that have been tried in correlation studies has now reached 32;[16] many of these, however, are highly intercorrelated.[17] The vast majority of published studies have been based on the Hansch π, often augmented by the Hammett σ, and occasionally by one or more other properties.

Although the effect of the oil/water distribution ratio on drug action had been recognized and even quantified in the nineteenth century,[18] it was Hansch

[15] J. Shorter, *Quart. Rev.*, 1970, **24**, 433.
[16] Reference 3*b*, pp. 43—45.
[17] A. Leo, C. Hansch, and C. Church, *J. Medicin Chem.*, 1969, **12**, 766.
[18] E. Overton, *Z. physiol. Chem.*, 1897, **22**, 189.

who constructed a theoretical rationale for the effect,[19] developed a standard reference system for its measurement,[20] and demonstrated its general relevance with numerous correlations.[21] The π value of a substituent is defined as $\log(P/P_0)$, where P is the partition coefficient between octanol and water for the substituted compound and P_0 the coefficient for the unsubstituted compound. This value not only is essentially independent of compound series, but is also well approximated for an unknown substituent by summing the π values of the substituent fragments.[22] The additive property of π values is extremely useful when synthesis of a compound containing a new substituent is being considered. The classical Hammett σ, the second most widely used substituent property, is an expression of the electronic effect of a substituent.

Regression equation (2), typical of correlation results involving physicochemical substituent properties, was obtained for a series of nitroso-ureas (2) tested for their ability to delay the growth of a solid tumour, the Lewis lung carcinoma, in mice.[23]



(2)

$$\log(1/c) = -0.08(\log P)^2 + 0.14(\log P) + 1.23 \qquad (2)$$

[Statistics: $n = 13$; $R^2 = 0.585$; $F_{2,10} = 7.1$ ($p < 0.025$)]

The log P values in equation (2) are the logarithms of experimentally determined octanol–water partition coefficients for the whole molecule; $\log(1/c)$ and the statistical indices are as explained above (*cf.* Section 2). While the relatively low value of R^2 indicates that a substantial proportion of the variation in the observed biological data has yet to be explained, the results of the F-test and the large structural variation among the substituent groups R (from adamantyl to carboxycyclohexyl) leave little doubt that the structure–activity relationship implied by equation (2) is real.

As is often the case, equation (2) contains a term in $(\log P)^2$ and thus describes a parabolic rather than a linear relationship between hydrophobicity and biological activity. The negative value of the parabolic term in (2) indicates further that there is a particular value of the partition coefficient for which

[19] C. Hansch, *Accounts Chem. Res.*, 1969, **2**, 232.
[20] C. Hansch and T. Fujita, *J. Amer. Chem. Soc.*, 1964, **86**, 1616.
[21] C. Hansch and W. J. Dunn, *J. Pharm. Sci.*, 1972, **61**, 1.
[22] A. Leo, C. Hansch, and D. Elkins, *Chem. Rev.*, 1971, **71**, 525; G. G. Nys and R. F. Rekker, *Chimie Therapeutique*, 1973, **5**, 521.
[23] J. A. Montgomery, J. G. Mayo, and C. Hansch, *J. Medicin. Chem.*, 1974, **17**, 477.

biological activity will be maximized. Several theoretical attempts to classify series of drugs according to their optimal partition coefficient have appeared.[19,21,24]

In principle, physicochemically based structure–activity correlation equations should be useful in lead development programmes by allowing predictions of the activity of unsynthesized compounds. Resulting data would then of course be incorporated into refined analyses. In practice, reliable equations often either fail to appear or emerge only after interest in further development of a series has waned. An important reason can be that the properties of the initial members of a series are poorly suited for analysis. Thus recent efforts to present physicochemical data in a form that would be useful at an early stage of lead development should be welcome. For example, Craig has advocated the use of the σ *versus* π scatter diagram shown in Figure 1.[25] Selection of representative substituents from each of the four quadrants of the graph in Figure 1, in the planning

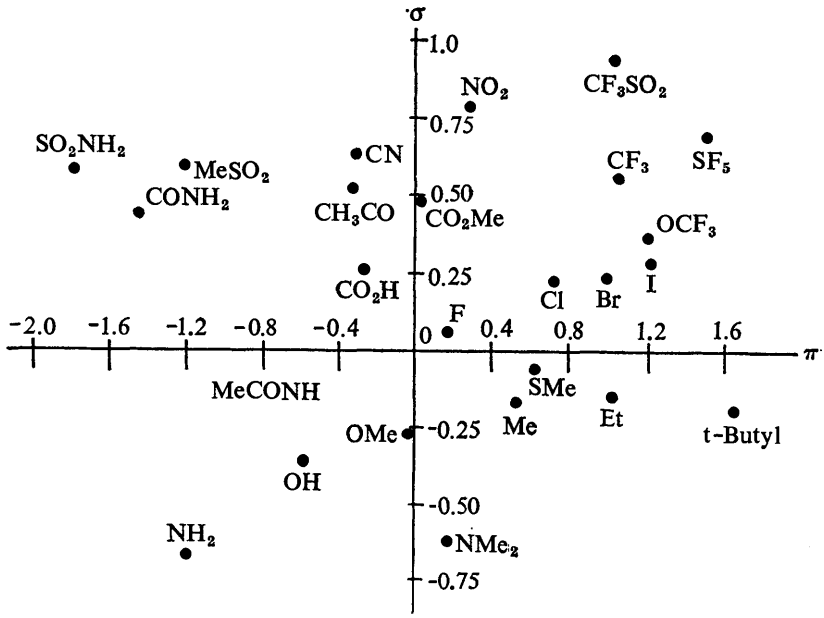


Figure 1 *Relationship between the Hammett* σ *and Hansch* π *values of some commonly used para-substituents.*
(Reproduced by permission from *J. Medicin. Chem.*, 1971, **14**, 682.)

[24] R. Franke and W. Schmidt, *Acta Biol. Med. Germ.*, 1973, **31**, 273; T. Higuchi and S. S. Davis, *J. Pharm. Sci.*, 1970, **59**, 1376.
[25] P. N. Craig, *J. Medicin. Chem.*, 1971, **14**, 680.

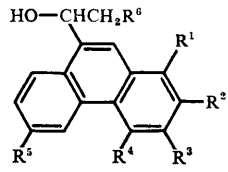
stages of a programme, increases the chances of early discovery of the σ/π region of maximum activity. Generalizing this approach to five important physicochemical parameters, Hansch *et al.*[26] used a hierarchical form of cluster analysis to classify substituents according to their overall similarity and dissimilarity. Most important from a practical viewpoint, Topliss[27] has developed a set of physicochemically based decision rules whose use during a synthetic programme requires no computers or statistics. In six retrospective cases cited,[27,28] use of the Topliss decision rules to guide a lead-optimizing synthetic programme would have identified the most active compound with considerably less chemical effort than that actually expended.

B. The Additive Model.—One of the simplest models that can form the basis for structure–activity correlation among series of related compounds is one that assumes that biological activity is an additive property of the substituents that vary within the series. Analyses based on this model have been able to account for a substantial part of the variation of the biological activity in numerous series of compounds.

Several methods based on the additive model have been described[29] but it was Free and Wilson who in 1964 developed the technique[30] into an elegant and now generally accepted form. In their mathematical formulation every substituent is assigned a substituent constant which represents the contribution of that substituent to the overall biological activity of the molecule in which it is present. These substituent constants are evaluated by the least-squares solution of a set of linear equations of the form (3), one for each of the molecules in the series.

$$\begin{matrix}\text{biological}\\ \text{activity}\end{matrix} = \begin{matrix}\text{mean biological}\\ \text{activity}\end{matrix} + \begin{matrix}\text{sum of substituent}\\ \text{contributions}\end{matrix} \qquad (3)$$

A recent analysis[31] of the antimalarial activity of a series of phenanthrene-carbinols (3) illustrates both the usefulness and the problems of applying the Free–Wilson methodology.



(3)

[26] C. Hansch, S. H. Unger, and A. B. Forsythe, *J. Medicin Chem*, 1973, **16**, 1217.
[27] J. G. Topliss, *J. Medicin. Chem.*, 1972, **15**, 1006.
[28] Y. C. Martin, *J. Medicin. Chem.*, 1973, **16**, 578.
[29] T. C. Bruice, N. Kharasch, and R. J. Winzler, *Arch. Biochem. Biophys.*, 1956, **62**, 305; J. Kopecky, K. Bocek, D. Vlachova, and M. Krivucova, *Experientia*, 1964, **20**, 667.
[30] S. M. Free, jun., and J. W. Wilson, *J. Medicin. Chem.*, 1964, **7**, 395.
[31] P. N. Craig and C. H. Hansch, *J. Medicin. Chem.*, 1973, **16**, 661.

Possible combination of the substituents represented within the 43 compounds available for analysis is $3\times3\times3\times6\times6\times3 = 2916$. The minimum number required to solve the equations is $1+(2+2+2+5+5+2) = 19$. Thus the number of compounds available in excess of this minimum is $43 - 19 = 24$ (the number of degrees of freedom), which is satisfactory from a statistical point of view. However, as seen in Table 1, 6 of the 24 substituents occur only once in the series of compounds and the distribution of the other substituents is relatively uneven. This, of course, is undesirable but is not atypical of series available for retrospective analysis. Such problems could be avoided by application of the Free–Wilson methodology in the planning stages of a synthetic lead-optimization programme.

The rank order of the substituent constants at a given position parallels the substituent contributions to the biological activity. To estimate the significance of the differences between values the student *t* test is commonly applied.[5] The tempting conclusion that the optimal compound is one with substituents having the highest constant in each position would be valid only if additivity were perfect and biological variability of the test system negligibly small, conditions seldom if ever satisfied.

The range of the substituent constants (Table 1) at the different positions on the molecule varies substantially (0.093—1.021); this reflects the relative sensitivity of the biological activity to substitution of the molecule. The larger the range the more important is that position for optimizing the biological activity. A small range suggests the position to be relatively unimportant, but it may also be that appropriate substituents have not been explored. Thus the data presented in Table 1 suggest that the substituents explored at position R^6 have minimal influence on the observed biological activity; in contrast, position R^5 appears to be most sensitive to substitution.

Analyses based on regression techniques allow the comparison of calculated and experimentally determined biological activities. A useful way of examining such data is to construct a plot of calculated *versus* experimental values; if most of the compounds fall into a zone bisecting the axes with a zone-width comparable to a specified confidence limit (*e.g.* 95%) of the experimental values then the additive model applied can be deemed appropriate. Compounds represented by points clearly outside this zone should be retested biologically and examined for possible error in structural assignment. If neither is responsible for the deviation, specific interaction between substituents is suggested. The identification of such hidden synergisms between groups is an important point of departure for further research.

Further useful information can emerge by recognizing correlations between substituent constants and the physicochemical parameters (electronic, partitioning, steric) of the substituents. The relative importance of the latter can only be established by exploring substituents covering a sufficiently wide range of parameter values; in this regard, reference to substituent scatter diagrams of the type shown in Figure 1 can be of considerable value. If the correlation proves significant after simple or multiple regression analysis, the synthesis of com-

Table 1 *Free–Wilson analysis of antimalarial phenanthrenecarbinols* (3)*: summary of results*

Position	*Group*	*No. of examples*	*Substituent constant*	*Range*
R^1	Cl	3	0.130	0.468
R^1	H	39	−0.001	
R^1	Br	1	−0.338	
R^2	Cl	7	0.301	0.370
R^2	CF_3	1	0.292	
R^2	H	35	−0.069	
R^3	CF_3	7	0.384	0.578
R^3	Br	1	0.296	
R^3	Cl	10	0.155	
R^3	I	1	0.129	
R^3	F	2	−0.193	
R^3	H	22	−0.194	
R^4	Cl	1	0.273	0.280
R^4	CF_3	1	0.043	
R^4	H	41	−0.008	
R^5	CF_3	18	0.451	1.021
R^5	Br	2	0.363	
R^5	Cl	6	−0.187	
R^5	F	2	−0.196	
R^5	H	13	−0.477	
R^5	OCH_3	2	−0.570	
R^6	2-piperidyl	13	0.037	0.093
R^6	dibutylamino	13	0.0142	
R^6	diheptylamino	17	−0.056	

Statistics: $n = 43$; $R^2 = 0.853$; $F = 7.82$ ($p < 0.01$)

pounds containing previously untested substituents might be indicated. By extending the predictive potential of the analysis beyond new combinations of 'old' substituents, this approach constitutes a very desirable coupling of the Hansch and Free–Wilson techniques.

C. **Quantum Chemical Methods.**—As early as 1945, the pioneering application of quantum chemical reasoning by Pullman led to recognition of the role of so-called *K* and *L* regions in the carcinogenic activity of fused aromatic hydro-

carbons.[32] The valence-bond method used originally gradually yielded to molecular orbital (MO) methods; the latter are now used almost exclusively when quantum chemistry is applied to problems of structure–activity correlation. In early studies the classical Hückel MO methods were employed; these were restricted to π electrons and thus to planar molecules or structural fragments. In the past decade, MO theory and computational techniques have advanced rapidly, and convenient programme packages are now available for many different all-valence-electron MO methods. Particularly extensive use has been made of the iterative, semi-empirical approaches based on Hoffman's Extended Hückel Theory (EHT)[33] and Pople's Complete Neglect of Differential Overlap (CNDO).[34]

Besides providing numerical values for molecular electronic parameters, the all-valence-electron MO methods allow the calculation of conformational energy profiles. Preferred (minimum energy) conformations of agonist molecules have been assumed to be those required for biological activity. For example, on the basis of comparing the minimum energy conformations of acetylcholine, muscarine, and muscarone, the muscarinic pharmacophore shown in Figure 2 was proposed.[35,36] Extensive studies of this kind on nicotinic, adrenergic, histaminic, and other agonist and antagonist molecules have been reviewed by Kier.[36]

Different calculations on the same molecules can yield different conformational energy profiles depending on the molecular parameters (*e.g.*, bond lengths and bond angles) and the MO method used. There is no consensus and apparently no clear answer as to which of the semi-empirical methods is most reliable. *Ab initio* methods are presumably more accurate, and recent refinements based on the molecular fragment approach[37] have made these methods suitable for molecules as complex as the antibiotic lincomycin ($C_{18}H_{34}N_2O_6S$).[38]

More important than the problem of relative accuracy of the various methods is whether the preferred conformation of an isolated molecule is likely to be involved in its interaction with a receptor. Differences between conformational energy minima are often only a few kilocalories and can be more than compensated by the energy of interaction between agonist and receptor. Thus delineation of the nature and magnitude of the conformational barriers might be more important than detailed knowledge of the configurations corresponding to energy minima. For example, in a recent study of conformational energy profiles of histamine and some of its methyl-substituted derivatives, Ganellin[39] presented good evidence to suggest that the conformation of histamine for inter-

[32] A. Pullman, *Compt. rend. Soc. Biol.*, 1945, **139**, 1956.
[33] R. Hoffmann, *J. Chem. Phys.*, 1963, **39**, 1397.
[34] J. A. Pople, D. P. Santry, and G. A. Segal, *J. Chem. Phys.*, 1965, **43**, S129.
[35] L. B. Kier, *Mol. Pharmacol.*, 1973, **9**, 820.
[36] L. B. Kier in 'Advances in Chemistry Series', No. 114, American Chemical Society, Washington, 1972, see also J. P. Green, C. L. Johnson, and S. Kang, *Ann. Rev. Pharm.*, 1974, **14**, 319.
[37] R. E. Christofferson, *Adv. Quantum Chem.*, 1972, **6**, 333.
[38] L. L. Shipman, R. E. Christofferson, and B. V. Cheney, *J. Medicin. Chem.*, 1974, **17**, 583.
[39] C. R. Ganellin, *J. Medicin. Chem.*, 1973, **16**, 620.

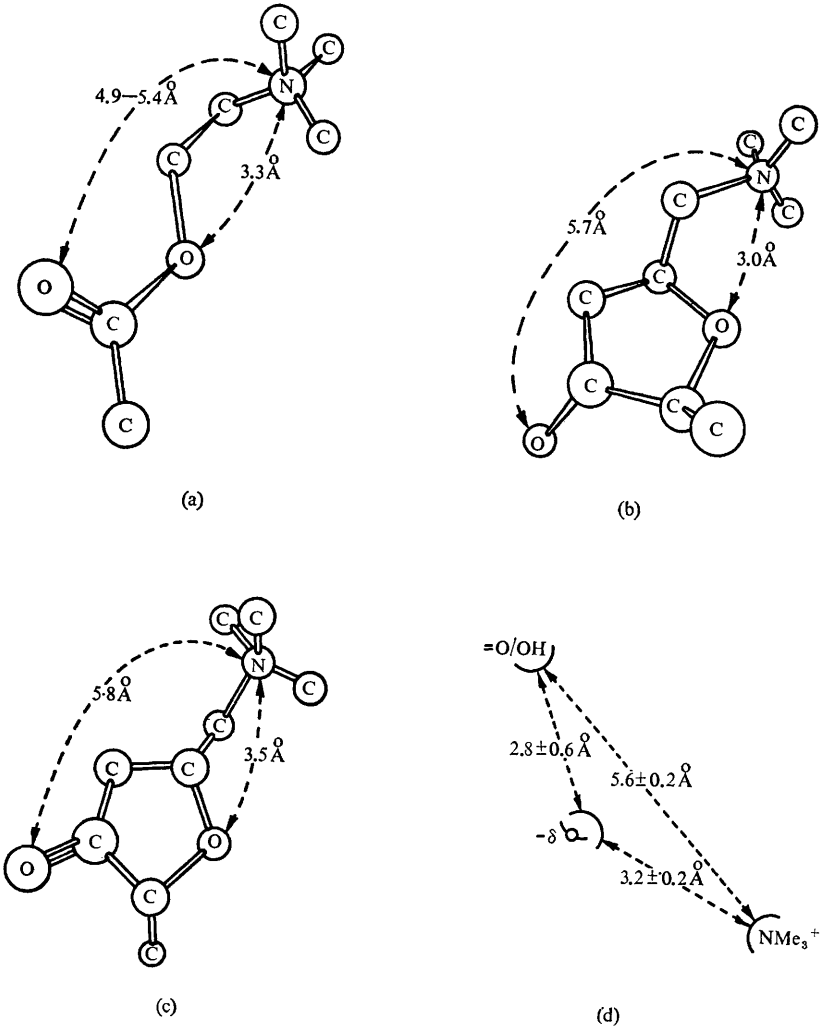


Figure 2 *Predicted conformations of acetylcholine* (a), *muscarine* (b), *muscarone* (c), *and proposed muscarinic pharmacophore.*
(Reproduced by permission from Advances in Chemistry Series, No. 114, 1972, p. 120)

action with one of its two recognized receptors is in the region of a local maximum in the energy profile of the molecule.

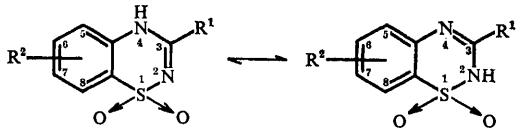
Although the computer time required for the MO calculation of a medium size (30—60 atom) drug molecule in a single conformation is moderate, a detailed

conformational analysis involving several simultaneously variable bond parameters can increase the demand for machine time to the limits of practicability. Increasing use of the computationally very much faster PCILO (Perturbation Configuration Interaction using Localized Orbitals) MO method[40] has now greatly reduced this problem, and even the consideration of solvation effects has become possible.[41]

Many of the structure–activity correlations using MO calculations are based on computed indices reflecting the electronic structure of the molecules.[42] Net atomic charges, frontier electron and orbital densities, superdelocalizabilities, energies of the highest occupied (HOMO) and lowest empty (LEMO) molecular orbitals are some of the most frequently used indices. A comprehensive list of MO indices used in structure–activity studies has recently been compiled.[43] Similar to, and sometimes in conjunction with, experimentally determined physicochemical parameters (*cf.* Section 3A), MO indices can be used in multiparameter regression analyses. While there is considerable question regarding the reliability of the absolute values of the calculated indices, it is generally assumed that errors due to the approximate nature of the MO methods, the lack of experimentally determined bond lengths and bond angles, the neglect of solvation effects, *etc.*, are largely parallel within structurally closely related series. Thus, trends and differences among the indices might be meaningful even though the absolute values are not.

If the molecules in a series under study are conformationally flexible, choice of the conformer for evaluation of the indices poses an additional problem. Typically, conformational energy investigation is carried out for one or two representative members, and preferred conformations are assumed to apply to the rest of the series. This is clearly an oversimplification, but separate energy minimization for every member of a series is rarely feasible.

A study[44] of the correlation of antihypertensive potencies with EHT MO indices in a series of benzothiadiazines clearly illustrates these problems. In this



(4) (5)

[40] S. Diner, J. P. Malrieu, F. Jordan, and M. Gilbert, *Theor. Chim. Acta*, 1969, **15**, 100.
[41] B. Pullman, Ph. Courrière, and H. Berthod, *J. Medicin. Chem.*, 1974, **17**, 439
[42] L B. Kier, 'Molecular Orbital Theory in Drug Research', Academic Press, New York and London, 1971; A. Cammarata in 'Molecular Orbital Studies in Chemical Pharmacology', ed. L. B. Kier, Springer Verlag, New York, 1970.
[43] Reference 3*b*, pp. 45—46.
[44] A. J. Wohl, *Mol. Pharmacol.*, 1970, **6**, 195.

example, the further complicating possibility of tautomerism was resolved by a separate preliminary study[45] leading to the conclusion that the equilibrium strongly favours the 4*H*-tautomer (4). It should be borne in mind, however, that in the macrobiological milieu activity may reside in an unfavoured tautomer even if only a minute fraction of the molecules are in that form at equilibrium. Of the various regression equations examined, equation (4) was found to be most satisfactory.

$$pA_2 = 64.12 - 5.16\,E_{HOMO} + 55.09\,S_5^N + 115.78\,S_6^N + 5.16q(3R) \quad (4)$$
$$[\text{Statistics: } n = 23;\ R^2 = 0.96;\ F = 62.31\ (p < 0.0005)]$$

[where pA_2 = *in vitro* potency based on competitive antagonism of Ca^{2+}; E_{HOMO} = energy of HOMO in $-$eV; S_5^N = nucleophilic superdelocalizability on atom 5; S_6^N = nucleophilic superdelocalizability on atom 6; and $q(3R)$ = summed regional charge over all atoms in the 3-R group].

The statistical significance of the equation is impressive, although 5-substituted derivatives were excluded in its derivation and the predicted values for these molecules were consistently high. The activity of the 5-substituted derivatives could be satisfactorily accounted for by an alternative regression equation which, however, required 8 indices, rather too many for a series containing only 25 compounds.

In general, many of the MO indices are calculated for individual atoms, and thus the parameter pool available for correlation analysis is large. Selection from this pool should ideally be based on biochemical reasoning since reliance on stepwise regression methods, which automatically select the statistically significant indices, can lead to an often overlooked pitfall. As Topliss has convincingly demonstrated,[46] the likelihood of obtaining chance correlations increases considerably with the number of parameters tried. The common practice of quoting only best regression equations without mentioning the size of the parameter pool can be very misleading.

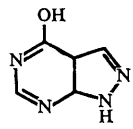
Reliance on the somewhat arbitrary and artificial atom-by-atom MO indices makes less than optimal use of the informational content of the MO calculations. However, few constructive alternative approaches have yet been suggested. A recent structure–activity study[47] of some anticholinergic phencyclidine derivatives using CNDO-generated electrostatic potential maps appears to be a promising new departure.

4 Lead-generating Techniques

A research programme whose objective is the optimization of a lead represented by an existing drug product is unlikely to produce a truly novel therapeutic

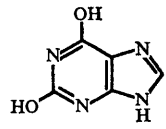
[45] A. J. Wohl, *Mol. Pharmacol.*, 1970, **6**, 189.
[46] J. G. Topliss and R. J. Costello, *J. Medicin. Chem.*, 1972, **15**, 1066.
[47] H. Weinstein, S. Maayani, S. Srebrenik, S. Cohen, and M. Sokolovsky, *Mol. Pharmacol.*, 1973, **9**, 820.

agent. New therapy is more reasonably found either by utilizing established testing systems to search for new lead compounds, or by devising new test systems where a lead structure may not exist. In either of these cases, a new lead might be suggested by consideration of a biochemical model or hypothesis. For example, if the objective were to inhibit a particular enzyme, one logical strategy would be to give provisional lead status to compounds that resemble the natural enzyme substrate, either in the ground or transition state. The discovery of allopurinol (6), a xanthine oxidase inhibitor used in the treatment of gout, exemplifies this approach.[48]



allopurinol

(6)



xanthine

(7)

Unfortunately, however, available biochemical hypotheses are often inadequate to identify meaningful leads, especially in the areas of most pressing clinical need. In the absence of any hypothesis, biochemical or otherwise, lead identification must depend primarily upon the screening of structurally diverse compounds. By virtue of the empirical and all but random character of their compound requirements, such screening programmes will usually have a relatively high throughput capacity. Screens capable of processing a thousand compounds per year are common, and the U.S. National Cancer Institute recently increased the target for its primary screen to a thousand compounds per week.[49] The analysis of data generated at this rate poses challenging problems – and opportunities – for the drug designer:

(i) Based on accumulated experience, how might priorities be established for the selection and testing of new compounds?

(ii) In which of several possible competing screens should the (probably limited) supply of a novel compound be consumed?

(iii) Can a large body of test data, frequently generated over long periods of time, be made to yield hitherto unidentified lead structures?

While the volume of the data and the structural variations represented within high-capacity screening programmes suggest the potential for computer assistance in addressing these problems, useful techniques are, surprisingly, only beginning to emerge.

[48] G. H. Hitchings, 'Progress in Drug Research', P.M.A. Research Symposium, Washington, D.C., March 6, 1969.

[49] T. H. Maugh, jun., *Science*, 1974, **184**, 970.

Recent independent experiments[9-11, 50-54] aimed at developing systems capable of meeting the challenges of lead generation have been based on substructural features (functionality, rings, chains, hetero-atoms, and combinations thereof).[55] The substructural approach has intuitive appeal. In fact, the lead concept underlying much of drug design might be paraphrased as 'a complex substructure whose incorporation tends to confer activity on a molecule'. Although the term 'substructural analysis' has been used for one of the techniques,[51,52] the phrase seems more appropriate as a general name encompassing all of these approaches.

Most substructural analyses can be described formally as attempts to devise linear equations in which the likelihood of activity of a compound is related to the sum of contributions from its constituent substructures; each substructural contribution is in turn computed from past testing experience with compounds containing the substructure. Differences among the analyses result from the ways that past testing experience is allowed to influence substructural contributions, and in the coding of the substructures themselves.

The most common procedure is to form a set of linear equations much like those used in the Free–Wilson additive model, except that, for computational tractability, the contribution of a substructure must be assumed to be independent of its molecular environment. However, inasmuch as the variety of substructures coded among a series of unrelated compounds of moderate complexity usually greatly exceeds the number of compounds, there initially are far too few degrees of freedom for confident regression solution of the equations. Therefore, Kowalski and Bender,[9] Chu *et al.*,[10] and Hiller *et al.*,[50] have used pattern-recognition techniques to extract only the substructures which seem most influential in determining activity. Cramer *et al.*[51,52] use all coded substructures, avoiding the problem of degrees of freedom by initially assigning a value to the coefficient of each substructural contribution, *e.g.* the proportion of actives among tested compounds containing the substructure. Other pattern-recognition approaches have also been used in substructural analysis, in particular the *k*-nearest-neighbour technique and other methods of cluster analysis.[9-11,53]

For reasons of expedience, the substructures employed in these analyses have generally been drawn from existing data-retrieval systems, a less than ideal situation since these systems were designed for different purposes. Nevertheless, innovative substructural descriptors have been used. For example, instead of allowing the substructural parameter simply to be the number of occurrences of

[50] S. A. Hiller, V. E. Golender, A. B. Rosenblit, L. A. Rastrigin, and A. B. Glaz, *Computers and Biomedical Research*, 1973, **6**, 411.
[51] R. D. Cramer tert., G. Redl, and C. E. Berkoff, *Abstracts of 167th American Chemical Society Meeting*, April 1974, CHLT 3.
[52] R. D. Cramer tert., G. Redl, and C. E. Berkoff, *J. Medicin. Chem.*, 1974, **17**, 533; G. Redl, R. D. Cramer tert., and C. E. Berkoff, 'Proceedings of the Conference on Chemical Structure – Biological Activity Relationships', Prague, June 1973 (in the press).
[53] P. J. Harrison, *J. Appl. Statistics*, 1968, **17**, 226.
[54] G. W. Adamson and J. A. Bush, *Nature*, 1974, **248**, 406.
[55] C. E. Granito and E. Garfield, *Naturwiss.*, 1973, **60**, 189.

a substructure, Kowalski and Bender have defined substructural parameters having continuous properties, such as the number of sulphur atoms per carbon atom.[9] Another experiment was based on the use of mass spectral fragments as substructures.[11]

Since general statistical criteria for analyses carried out by pattern-recognition techniques have not been developed, the validation of a substructural analysis requires empirical yardsticks. The degree of success in predicting activity within a group of compounds not used in the analysis is compared with some benchmark success rate. Ideally this prediction set will be completely distinct from the training set used in the analysis. When the compounds are too few to permit a permanent division, 'leave-one-out'[56] or 'leave-*n*-out'[52] techniques may be employed. Using these techniques, a compound or set of n compounds is excluded from the training set and the likelihood of its (their) activity is recomputed. The leave-out procedure is repeated until all the compounds have been made members of a small prediction set, and the overall results are again compared with the benchmark rate. When making these comparisons, consideration should be given to existing statistical criteria for distributions, such as the χ^2 test.[5] We note parenthetically that leave-out tests could also be used to evaluate retrospective analyses which employ regression techniques (*cf.* Section 3).

Some very recent work[51] exemplifies the broad scope and statistically significant, but limited, predictive powers of current approaches. Among 540 structurally diverse compounds screened for their ability to inhibit passive cutaneous anaphylaxis in the rat (a measure of their potential anti-allergic effect), 259 showed some degree of activity. For each of the 401 substructures among the tested compounds, a substructure activity score was computed as $[A_i - T_i(259/540)]$, where A_i is the number of active compounds and T_i the number of tested compounds containing the ith substructure. A leave-3-out technique was employed. After exclusion of the testing results for a three-compound prediction set, the three compounds were ranked by descending values of molecular activity score (defined as the sum of the substructure activity scores). This procedure was repeated until all 540 compounds had been included in a three-compound prediction set, while a record was kept of the number of times that the first-place, second-place, and third-place compounds were actually active.

As can be seen in Table 2, there is a relationship between total activity score

Table 2 *The distribution of activity among compounds ranked by descending values of the molecular activity score*

Rank	*Proportion of active compounds*	*Average molecular activity score*
1	100/180 = 0.56	40.6
2	95/180 = 0.53	0.5
3	64/180 = 0.36	−31.3

[56] B. R. Kowalski and C. F. Bender, *Analyt. Chem.*, 1972, **44**, 1405.

and the likelihood of activity, since compounds of higher rank were active more often. The probability of obtaining a distribution this skewed by chance, were there no relationship between molecular activity score and biological activity, is less than 2% according to a χ^2 test.

Close examination of the results from our substructural analysis of 771 compounds tested for anti-arthritic activity[52] suggests that many compounds fall into a category that could be regarded as 'anti-lead'. These compounds can be predicted to have a significantly *lower* than average chance of being active, since a preponderant number of their substructures have occurred mostly in inactive compounds. Exclusion of this type of compound from testing consideration would enhance the lead-generating efficiency of the screen.

To date, two of the three published substructural analyses involving related series of compounds[9,11] have been criticized for their apparent triviality; the conclusions drawn from the analyses have appeared obvious upon re-examination of the data.[57,58] The third example, an impressive correlation obtained in a regression study of the substructures involved in penicillin binding to proteins,[54] seems less remarkable to us considering the strong dependence of drug–protein binding on hydrophobicity[59] and the known additive properties of partition coefficients.[22]

In view of the severe approximations involved in substructural analysis, the ability to obtain any correlation is encouraging. However, the utility of the methods in terms of the problems of lead generation raised above remains to be conclusively demonstrated. The use of substructural analysis to establish screening priorities will, of course, depend upon its reliability, and on the relative costs of computer and biological testing.[50] The more exciting challenge of designing new lead structures (perhaps by combination of substructural fragments into more complex moieties) is a more distant but nonetheless realistic goal as sophisticated structural representations become available.[60]

5 Conclusions

'Is quantitative drug design of any practical use?' – the provocative question often put by the disbeliever. Despite the limited number of successful predictive analyses,[61] we believe that drug design methodologies should be of great value in many of the problems faced by the medicinal chemist. At today's unfavourable odds against any particular compound becoming a drug product, the traditional measure of 'success' appears to be an unrealistic challenge, rather similar to

[57] S. H. Unger, *Cancer Chemotherapy Reports*, 1974, in the press.
[58] C. L. Perrin, *Science*, 1974, **183**, 551.
[59] W. Scholtan, *Arzneim.-Forsch.*, 1968, **18**, 505.
[60] W. T. Wipke, S. R. Heller, R. J. Feldmann, and R. Hyde, 'Computer Representation and Manipulation of Structural Information', John Wiley and Sons, New York, 1974.
[61] R. W. Fuller, M. M. Marsh, and J. Mills, *J. Medicin. Chem.*, 1968, **11**, 397; J. G. Beasley and W. P. Purcell, *Biochim. Biophys. Acta*, 1969, **178**, 175; Y. C. Martin, T. M. Bustard, and K. R. Lynn, *J. Medicin. Chem.*, 1973, **16**, 1089; P. J. Goodford, F. E. Norrington, W. H. G. Richards, and L. P. Walls, *Brit. J. Pharmacol.*, 1973, **48**, 650; H. Cousse, G. Mouzin, and L. Dussourd d'Hinterland, *Chimie Therapeutique*, 1973, **4**, 466.

expecting a professional golfer to demonstrate his superior techniques by shooting a hole-in-one.

To the question put by the less disbelieving, 'Which of the many methodologies in current use is the best?' we respond that there is no direct answer, save that it is the wrong question to ask. Different methods require different types of data and answer different questions; all approaches must be considered when the analysis of a new problem is being planned.

The application of lead-optimizing regression techniques requires series of active compounds and is restricted to relatively narrow structural classes. Identification of a lead is therefore a prerequisite. Nevertheless, by identifying the physicochemical properties that most influence biological activity in a given series, multiparameter analysis may help elucidate the biological mechanisms of action and thus contribute to the discovery of new leads as well as to the optimization of existing ones. This underscores the somewhat arbitrary nature of the distinction between lead-optimizing and lead-generating techniques.

Until recently, quantitative drug design has not been applied to the problems of generating new structural leads. Substructural analysis now offers great promise, in particular because of its capacity to accommodate qualitative data on large numbers of diverse structures.

Drug design is undeniably still in its infancy, and quantal improvements are needed in virtually all aspects of available methodologies. To realize its full potential, readjustment of existing attitudes towards application of the techniques is mandatory at critical points in a research programme. For example, its impact should be anticipated in the planning stages of any chemical programme devoted to the synthesis of an optimally active compound. Further, in the absence of rationally founded chemistry it is tempting, even wise, to be guided by the most tenuous of predictions based on structure–activity correlations. However, the original tenuousness of the predictions must be remembered, especially if negative data appear. Only with better integration into the overall research process can quantitative drug design assume its proper place and emerge as a mature technology.

We thank Dr. A. D. Bender for stimulating discussion and continuing encouragement.