(34) R. S. Mulliken, *J. Chem. Phys.*, **23**, 1833 (1955); see also D. B. Boyd and W. N. Lipscomb, *J. Theor. Biol.*, **25**, 403 (1969), for the dependence of overlap populations on bond lengths and, hence, on bond strengths.
(35) A definition of leavability is given in ref 15. Further, related connotations of the term are apparent in the writings of S. L. Vail and H. Petersen, *Ind. Eng. Chem., Prod. Res. Dev.*, **14**, 50 (1975); W. N. Olmstead and J. I. Brauman, *J. Am. Chem. Soc.*, **99**, 4219 (1977); and E. S. Gould, "Mechanism and Structure in Organic Chemistry", Holt, Rinehart, and Winston, New York, N.Y., 1959, p 261.
(36) J. M. Indelicato, T. T. Norvilas, R. R. Pfeiffer, W. J. Wheeler, and W. L. Wilham, *J. Med. Chem.*, **17**, 523 (1974).
(37) M. Charton, *J. Org. Chem.*, **29**, 1222 (1964); C. D. Ritchie and W. F. Sager, *Prog. Phys. Org. Chem.*, **2**, 323 (1964); E. M. Kosower, "An Introduction to Physical Organic Chemistry", Wiley, New York, N.Y., 1968, p 49; M. S. Newman, "Steric Effects in Organic Chemistry", Wiley. New York, N.Y., 1956, p 592.
(38) Numerous series of related cephalosporins are known; see, e.g., J. A. Webber and J. L. Ott in "Structure–Activity Relationships among the Semisynthetic Antibiotics", D. Perlman, Ed., Academic Press, New York, N.Y., 1977, p 161.
(39) R. M. Sweet, in ref 8, p 280.
(40) D. B. Boyd, *J. Phys. Chem.*, **78**, 2604 (1974).
(41) The conjugated N_5=C_4=C_3=C_{18} portion of **2**–**4** can be *rudimentarily* modeled by *trans*-1,3-butadiene. CNDO/2 calculations were done on butadiene with CO_2 near the β face of C_1 and OH^- near either the α or β face of C_4. CO_2 (representing the COOH part of **2**–**4**) was placed in the plane bisecting the H–C_1–H bond angle such that the carbon of CO_2 is only 2.0 Å from C_1 and C--C_1=C_2 equals 135° and O=C--C_1 equals 90°. The OH^- (representing OAc^-) was placed in the plane bisecting the H–C_4–H bond angle such that C_4--O is 3.5 Å and C_3=C_4--O equals 109.4712°. C_4 of butadiene was taken to be either planar or pyramidal. With this simplified model, the essential features of the cephem system are isolated for scrutiny and computer times are reduced. Significantly, the most stable arrangement has CO_2 and OH^- on opposite sides of the C_1=C_2–C_3=C_4 plane. This configuration corresponds to the cephem situation where OAc^- is departing from the α face of C_{18} and, hence, anti to C_8 near N_5. Electron-density maps of the butadiene complexes and its substructures were obtained from CNDO/2D wave functions (ref 32 and 40). Difference density maps were calculated by subtracting from the valence-electron density of each complex the sum of the valence-electron densities of the noninteracting substructures using the same atomic coordinates as in that complex. These plots show that perturbation of the electron density in the C_3=C_4 π region due to CO_2 is relatively small but that due to OH^- is large. Interelectron repulsion is the dominant effect when OH^- is present: electron density is significantly reduced on whichever face of C_4 is closest to OH^-. The reduction in the C_4--OH^- region is a little greater when OH^- is on the α face, which would be anti to the CO_2 on the β face of C_1. Thus, electron density between OH^- and butadiene is lower when CO_2 and OH^- are on opposite faces. Lowered electron repulsions are consistent with the anti configuration being more stable.
(42) G. Stork and W. N. White, *J. Am. Chem. Soc.*, **78**, 4609 (1956); G. Stork and A. F. Kreft III, *J. Am. Chem. Soc.*, **99**, 3850, 3851 (1977), and references cited therein.
(43) J. M. Dereppe, J. P. Declercq, G. Germain, and M. Van Meerssche, *Acta Crystallogr., Sect. B*, **33**, 290 (1977).
(44) M. Gorman and C. W. Ryan, in ref 8, p 532.
(45) D. B. Boyd, *J. Chem. Educ.*, **53**, 583 (1976).
(46) D. B. Boyd, *J. Med. Chem.*, **16**, 1195 (1973).
(47) J. R. Knox, P. E. Zorsky, and N. S. Murthy, *J. Mol. Biol.*, 79, 597 (1973); J. R. Knox, J. A. Kelly, P. C. Moews, and N. S. Murthy, *J. Mol. Biol.*, 104, 865 (1976); J. R. Knox, M. L. DeLucia, N. S. Murthy, J. A. Kelly, P. C. Moews, J.-M. Frère, and J.-M. Ghuysen, *J. Mol. Biol.*, **127**, 217 (1979); R. Aschaffenburg, D. C. Phillips, B. J. Sutton, G. Baldwin, P. A. Kiener, and S. G. Waley, *J. Mol. Biol.*, **120**, 447 (1978), and references cited therein.

Mathematical Considerations in Series Design

Yvonne Connolly Martin* and Helen Nehrich Panas

Abbott Laboratories, North Chicago, Illinois 60064. *Received January 15, 1979*

It is proposed that a series designed to explore the potential of a "lead" should have the following characteristics: (1) the analogues should be synthetically feasible, (2) the series should contain enough variation in the properties which may influence potency, (3) these properties should be varied independently of each other, and (4) the series should be the minimum acceptable size, i.e., each analogue should contribute unique information. Point 2 is evaluated by a consideration of the definition of R^2. As a rule of thumb, the standard deviation of a property should usually be ≥1.0. Point 3 is evaluated by analyzing the correlation matrix of properties. If it has fewer significant eigenvalues than properties, then factor analysis reveals which properties are artificially correlated. Point 4 is evaluated by distance between analogues in property space. In order to be certain that the proposed molecular descriptors are independent, a large data set of possible substituents was analyzed. Factor analysis of the physicochemical properties of 78 aromatic substituents revealed that π, S, P, and MR are orthogonal descriptors. The proposed criteria have been applied to series designed by cluster analysis, multidimensional nonlinear mapping, Topliss batch methods, and to two Abbott series. The other mathematical methods of series design suffer from their lack of attention to all four points simultaneously.

If one accepts the premise that the biological properties of organic molecules are a direct consequence of their chemical and physical properties, then it becomes possible to propose strategies to make the process of drug discovery more efficient. This report suggests a method of evaluating the suitability of a set of analogues which have been proposed to follow up a lead. Synthesis and testing of this set of analogues will adequately explore the effect of variations in those properties used in its design. Hence, if it is found that this series does not possess or suggest analogues of sufficient potency, one can confidently decide not to explore further variations of these properties. Thus, the method is useful in both initial series design and in the decision to terminate further synthesis.

The proposed criteria for a suitable set are as follows: (1) It will be the smallest size consistent with the objectives of the synthetic program. (2) Those chemical and physical properties which are hypothesized to determine biological potency or profile will be varied widely enough that it is theoretically possible to find a useful relationship. (3) The series will exhibit variation in these molecular properties in such a way that the variation of each property is independent of the variation in all other properties. The two latter characteristics assure that the data space has been adequately explored. (4) It will contain only analogues which are not too difficult to synthesize. This point will

0022-2623/79/1822-0784$01.00/0

be established by the synthetic chemists of the team.

Various strategies to produce a suitable set of planned analogues have been advanced[1-3] but, as will be shown below, each has serious shortcomings. We suggest that once a series has been proposed, it can be evaluated and perhaps improved by the use of standard statistical methods of factor and cluster analysis.

The basic steps of this method of series design and improvement are as follows: (1) The key decisions discussed below are made. (2) The initial series is proposed. This initial series may be chosen from published[1] or internally generated clusters and/or from considerations of synthetic feasibility plus intuition. (3) The series is evaluated by factor and cluster analysis. Factor analysis programs provide measures of relatedness or independence of each property. Although the most common use of factor analysis (or its close relative, principal component analysis) in chemistry has been to assess the relatedness of properties, the method is equally appropriate for the analysis of independence of properties. It is explained in the Appendix. Cluster analysis provides measures of the uniqueness of each compound. A discussion of these measures of suitability as well as examples of evaluated series follow. (4) Depending on the outcome of the evaluation, as many as are required of propose–evaluate cycles are done until all four criteria have been satisfied.

Steps in the Evaluation of a Series

Key Initial Decisions. (1) Choice of Physical Properties to be Used to Describe Substituent Effects. A critical decision in the evaluation or design of a series is the choice of structural properties on which the optimization is to be accomplished. For example, one could consider indicator variables which denote the presence or absence of certain specific substructures or pharmacophores, conformational properties such as distances between key atoms, and/or physical and chemical properties of the molecules as a whole or of a substructural feature of the molecule. The large number of possible descriptors presents a problem. On the one hand, only if all molecular properties which influence biological activity are considered will the lead be adequately explored. On the other hand, as the number of molecular descriptors is increased, so is the number of analogues necessary to examine them.

Because of their known utility, and in spite of their obvious limitations, the series considered in this report are described almost exclusively by those properties traditionally considered in the linear free-energy analysis of structure–activity relationships. These properties are divided into the hydrophobic, electronic, and steric effects of the variable substituent on the (constant) parent compound.[5] If such substituent effects are truly independent in the series, then each represents a different dimension of physical property space; that is, each property is totally uncorrelated (i.e., orthogonal) with every other property. This section deals with the selection of a set of orthogonal physicochemical parameters.

(a) Hydrophobic Substituent Effects. Hydrophobicity is parameterized in this report by the substituent effect on the logarithm of the octanol–water partition coefficient, the Hansch π value[6] or by the log P of the total molecule.

(b) Electronic Substituent Effects. The electronic influences of substituents have been commonly assumed to be attributable to their separate orthogonal inductive-field and resonance effects. Various methods of calculating these two effects have been proposed. From the standpoint of maximizing the number of substituents for which values would be available, it is preferable to consider those parameters which can be calculated from the common Hammett m and p values. On this basis, the Swain–Lupton $\mathcal{F}$ and $\mathcal{R}$ values[7] and the Unger–Swain S and P values[8] are attractive. We used factor analysis to analyze these two pairs of inductive-field and resonance values for independence between the two electronic effects and to possibly make a descision as to which pair is more orthogonal. This analysis was performed on the data set of substituents[6] for which both σ_m and σ_p values are available, 125 in all. The derived parameters were calculated from the following equations derived from the original references:

$$\mathcal{F} = 1.369\sigma_m - 0.373\sigma_p - 0.009$$

$$\mathcal{R} = \sigma_p - 0.921\mathcal{F}$$

$$P = 4.60\sigma_p - 3.81\sigma_m + 0.085$$

$$S = (\sigma_m - 0.178P + 0.002)/0.419$$

The correlation matrix of the data (σ_m, σ_p, $\mathcal{F}$, $\mathcal{R}$, S, and P) matrix has only two significant eigenvalues, 4.22 and 1.78. This confirms the chemical postulate that there are two and only two electronic substituent effects, inductive field and resonance.

Table I. Rotated Factor Pattern of the Electronic Variables

variable	factor 1	factor 2
σ_m	0.57	0.82
σ_p	0.86	0.52
$\mathcal{F}$	0.33	0.94
$\mathcal{R}$	0.99	0
S	0	0.99
P	0.99	0

The rotated factor pattern is shown in Table I. The results clearly show the orthogonality of S and P and the less satisfactory performance of $\mathcal{F}$ and $\mathcal{R}$.

(c) Steric Substituent Effects. The selection of parameters for the steric effects of substituents is a very complex problem and many different choices have been suggested. For drug analogue studies, the most commonly used steric parameters are the Taft E_s value, the Verloop–Tipker Sterimol parameters, and molar refractivity or a related value. The traditional steric parameter E_s is an experimental measure. It is based on the presumed steric component of the effect of substituents on the rate of ester hydrolysis.[9] The Sterimol parameter L is the calculated van der Waal's length (of the presumed minimum energy conformation) of the substituent along the line of the bond between the parent and the substituent. B_1 is the smallest dimension perpendicular to L, and B_4 is the largest dimension perpendicular to L.[10] Finally, the molar refractivity of a substituent is an additive constitutive property which denotes the effect of that substituent on the refractive index of the compound. It is also the parameter which measures the dispersion bonding potential of a substituent.[6]

How many steric parameters are necessary? What is the relationship between the various parameters? Again, factor analysis was used to investigate this problem. There are 34 substituents for which the Sterimol parameters, E_s values, and molar refractivity are available. The eigenvalues of the correlation matrix of this data set are 3.14, 1.19, 0.38, 0.20, and 0.08.

The rotated factor pattern of this data is shown in Table II. The previously noted[10] relationship between E_s and

Table II. Rotated Factor Pattern of Steric Variables

	factor		
var	1	2	3
MR	0.80	0.35	0.38
E_s	−0.50	−0.81	0
L	0.41	0	0.90
B_1	0	0.97	0
B_4	0.92	0	0.33

Table III. Rotated Factor Pattern of Hydrophobic and Steric Parameters

	factor				
var	1	2	3	4	5
π	0	0.99	0	0	0
MR	0.32	0	0.27	0.29	−0.86
L	0.92	0	0	0.24	−0.26
B_1	0	0	0.98	0	0
B_4	0.23	0	0	0.94	−0.23

Table IV. Means, Standard Deviations, Squared Multiple Correlations, and Rotated Factor Pattern of Physical Properties in the Total Data Set of $N = 78$

			sq mult	factor			
var	mean	SD	correlat	1	2	3	4
π	0.21	1.11	0.14	0	0	0	0.98
MR	18.06	12.04	0.07	0	0	−0.99	0
S	0.57	0.41	0.10	0.99	0	0	0
P	−0.21	0.99	0.04	0	1.00	0	0

B_1 is seen in factor 2. Note, however, the B_4 and E_s are also related. The most interesting result is the observation that MR is related to each of the Sterimol parameters. On the basis of these results it looked as if, for preliminary investigations at least, MR might be a good compromise steric parameter.

However, because MR is essentially a measure of dispersion bonding, it was important to verify in a larger set of substituents that π and MR are orthogonal and the MR is a measure of steric effects only. Thus, we performed a factor analysis on those 71 substituents for which π, MR, L, B_1, and B_4 are available. The results are shown in Table III. This analysis verified both points.

Hence, our recommended set of parameters to be used to characterize a substituent is π, S, P, and MR. Table IV summarizes the factor analysis of 78 substituents for which these four parameters are available. The eigenvalues of the correlation matrix are 1.42, 1.08, 0.89, and 0.61. The significance of each eigenvalue indicates four separate properties. The identification of each eigenvector with a parameter is confirmed by the factor pattern.

If the series to be designed or evaluated is substituted in one position only and if a suitable σ value unambiguously applies to the series, then one may choose to use that σ value rather than S and P. Although the series optimization may be done on MR, it would seem sensible to choose substituents for which E_s or Sterimol parameters are also available since at later stages in the evaluation of a series regression analysis using these variables may be desirable.

(2) Number of Compounds to be Synthesized. The minimum number of compounds necessary in a series is governed by the number of properties to be examined and by the results of Topliss and Edwards.[4] They studied the relationship between the number of analogues, the number of variables examined, and the possibility of chance but statistically significant correlations. For example, they showed that in order to investigate five variables it is necessary to have 14 observations if the risk of a chance correlation is to be to less than 0.01. The similar figure for ten variables is 21 observations. Hence, the decision of the number of analogues to synthesize involves an assessment of one's willingness to accept the risk that an untrue relationship would be accepted as true vs. the risk that a true relationship would be discarded as a chance correlation. The weightings assigned to these risks, in turn, depend on the intended use of the information.

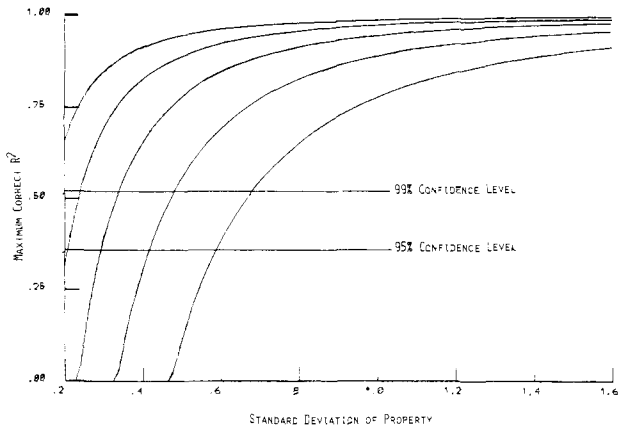


Figure 1. A plot of the maximum R^2 possible without overfitting the data vs. the standard deviation of the predictor variable. The curves are drawn for the following coefficients of the predictor variable in the regression equation (in order from left to right): 4.0, 2.0, 1.0, 0.5, and 0.25.

It has been our experience that if a series contains sufficient variation in the properties of interest and if these properties are not highly correlated, then there are enough analogues that the risk of chance correlation is acceptably low.

Measures of Suitability of a Series. (1) Variability in Physical Properties. One measure of variability in a property is its standard deviation. The question then is: "What standard deviation of a property is consistent with the reasonable likelihood that I will have varied that property sufficiently to see its effect on log potency?" For the case of log potency correlated with one property only, a tentative answer may be found by manipulation of the equation for R^2, the fraction of the variance in the data which is explained by the correlation equation:[6]

$$R^2 = \frac{SS_{reg}}{SS_{mean}} = \frac{s_m^2(n-1) - s_r^2(n-2)}{s_m^2(n-1)} = \frac{as_p^2(n-1) - s_r^2(n-2)}{as_p^2(n-1)}$$

in which SS_{mean} is the sum of squares about the mean, SS_{reg} is the sum of squares due to regression, n is the number of analogues, s_r is the standard deviation of log potency from the regression line, s_m is the standard deviation of log potency from the mean, s_p is the standard deviation of the physical property in question from its mean, and a is the coefficient of the regression equation between log potency and the physical property.

Figure 1 is a plot of R^2 vs. s_p when $n = 10$ and $s_r = 0.25$. In the case in which all of the real variance in the data is fit by the equation, s_r is equal to the standard deviation between replicate determinations of the log potency of the same compound; 0.25 has been selected as a typical value. Note that the lower the coefficient of a parameter in a regression equation, the larger the standard deviation required to reach the same R^2. Although statistical significance is reached at lower values, the usual rule of

Table V. Univariate Statistics of Properties in the Whole Set of Possible Substituents

var	n	mean ± SD
L	89	4.89 ± 1.47
B_1	89	1.69 ± 0.32
B_4	89	37.6 ± 1.59
MR	137	23.38 ± 13.36
E_s	44	−0.75 ± 1.20
π	79	0.21 ± 1.10
$\mathcal{F}$	125	0.23 ± 0.22
$\mathcal{R}$	125	−0.08 ± 0.26
S	125	0.56 ± 0.45
P	125	−0.12 ± 0.97
σ_m	125	0.21 ± 0.25
σ_p	125	0.12 ± 0.38
σ^*	125	1.07 ± 1.06

thumb for a useful equation is one for which R^2 is at least 0.80. On this basis, the standard deviation of a property must be approximately 0.75 if its coefficient in the regression equation can be as low as 0.50. If the regression coefficient could be even lower and/or if a higher R^2 is required, then an even larger standard deviation of the property will be necessary.

Table V lists the mean and standard deviation of typical variables used in quantitative structure–activity analysis. Note that in the whole data set the standard deviation of σ_m is 0.25 and that of σ_p is 0.38. This suggests that it may be difficult to design a series with sufficient variation in electronic properties if only monosubstituted analogues are considered.

(2) Independence of Physical Properties. It is suggested that factor analysis of the physical property corrrelation matrix of the proposed series be used to determine the independence of the physical properties. The first clues to a lack of independence are small eigenvalues and large squared multiple correlations. Identification of the problem relationships may be accomplished by examination of the rotated factor pattern. (Factor analysis is described further in the Appendix.)

(3) Uniqueness of Each Compound. A series designed with efficiency in mind contains no analogues of essentially duplicate properties. The appropriate measure of duplication is the distance between analogues in property space. In the absence of compelling arguments for other measures, we have chosen to measure this distance by the Euclidean distance in standardized variable space. This is simply the square root of the sum of the squares of the distance between two analogues in each dimension. Each variable was standardized by subtracting the mean and dividing by the appropriate standard deviation listed in Table V. The distance was calculated with the BMDP2M cluster analysis program.[11]

The Euclidean distance is easily visualized and has the advantage that the distance between analogues which are different in only one dimension is not exaggerated. However, it is dependent on the number of dimensions. For example, a difference of 0.5 (standard deviation) unit in each of two dimensions yields a distance of $(0.5^2 + 0.5^2)^{1/2} = 0.71$, whereas in three dimensions the distance is $(0.5^2 + 0.5^2 + 0.5^2)^{1/2} = 0.87$ and in four dimensions it is 1.00. The discussions of distance in this paper will use the criterion of a distance of 0.5 standard deviations in each direction as the cutoff between close and distant.

Examples: Evaluation of Selected Series

Examples of the use of factor and cluster analysis to evaluate series follow. The variables chosen in some examples are based on those used in the source of information.

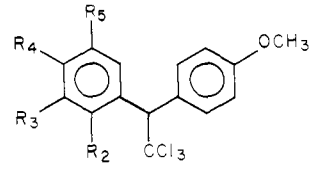
Table VI. Means, Standard Deviations, and Squared Multiple Correlations of Physical Properties of a Set of Ten Analogues Chosen from Cluster Analysis

var	mean	SD	sq mult correlat
π	0.59	1.26	0.47
S	0.31	0.75	0.38
P	0.16	1.52	0.38
MR	20.65	12.74	0.44

Series Chosen from Cluster Analysis. Hansch, Unger, and Forsythe[1] were the first workers to apply multidimensional mathematical techniques to the design of series of analogues. They used cluster analysis to group known substituents into groups of similar substituents. To design an optimal series one would select one analogue from each cluster. The first series to be discussed in this report is one chosen from their cluster analyses. Substituents were chosen from the reported clusters based on π, MR, $\mathcal{F}$, and $\mathcal{R}$. One substituent was chosen from each cluster: for multimembered clusters the middle member was chosen. The substituents are $CH{=}CH_2$, OCF_3, SCOMe, $NHCONH_2$, NHBu, OSO_2Ph, SO_2Ph, adamantyl, OH, and $CH{=}CHCOOH$.

The mean, standard deviation, and squared multiple correlation of the π, MR, S, and P values for this set are listed in Table VI. It can be seen that for all variables the standard deviation is larger than that in the whole data set and the squared multiple correlation is relatively small. The eigenvalues of the correlation matrix are 1.95, 1.22, 0.57, and 0.26. There are no pairs of substituents closer in standardized Euclidean distance than 1.0; each analogue is unique. Hence, the cluster analysis has allowed the selection of a set of substituents which span the parameter space in a relatively uncorrelated manner. Because only monosubstituted analogues are considered, it is possible that there is too little variation in S and P. Additionally, the assymetric values of the eigenvalues suggests that there is some remaining correlation between variables. In practice, one would usually use a sample size of 12–16 analogues to explore four properties. The added analogues should be selected to reduce the correlation between π and MR and that between S and P.

Analogues Designed by Multidimensional Nonlinear Mapping. Goodford et al. have reported on an interactive computer program which is used by the synthetic chemist to choose analogues for a series. The basis for selection of analogues is distance in multidimensional space and synthetic ease. Possible substituents for a phenyl ring are chosen from a data base within the program. The program was used by them to design a set of analogues of methoxychlor.[12] They also synthesized a



methoxychlor

second set of analogues because they were easy to make.

Table VII lists the mean, standard deviation, and squared multiple correlation for the variables on which their optimization was performed. These variables are π, summed $\mathcal{F}$ and $\mathcal{R}$ values weighted by position, MR of the ortho position (MR_o), and the sum of the MR of the meta and para positions (MR_{mp}). Note that the computer-designed set has a larger standard deviation for each of

Table VII. Means, Standard Deviations, and Squared Multiple Correlations of Physical Properties of Two Sets of Methoxychlor Analogues

var	mean	SD	sq mult correlat
A. computer designed series (n = 12)			
π	1.25	1.66	0.95
$\mathcal{F}$	0.66	0.57	0.68
$\mathcal{R}$	−0.41	0.30	0.54
MR_o	7.22	7.62	0.84
MR_{mp}	13.98	12.52	0.91
B. easy to synthesize analogues (n = 9)			
π	0.57	0.52	0.70
$\mathcal{F}$	0.34	0.31	0.26
$\mathcal{R}$	−0.23	0.21	0.74
MR_o	0.72	2.17	0.65
MR_{mp}	6.89	5.38	0.83

Table VIII. Rotated Factor Pattern of Methoxychlor Analogues

	factor			
var	1	2	3	4
A. computer designed series				
π	0.78	0.36	−0.42	−0.27
$\mathcal{F}$	0	0	0.97	0
$\mathcal{R}$	0	0	0	0.99
MR_o	0	0.94	0	0
MR_{mp}	0.95	0	0	0
B. easy to synthesize series				
π	0.95	0	−0.30	0
$\mathcal{F}$	0	0	0	0.98
$\mathcal{R}$	0	0.93	0	0
MR_o	−0.41	0	0.90	0
MR_{mp}	0.64	−0.64	−0.38	0

the variables but a much higher degree of colinearity than that of the easy to make set.

The eigenvalues of the correlation matrix of the computer-designed series are 2.21, 1.22, 1.09, 0.46, and 0.03; those of the easy to make set are 2.24, 1.71, 0.69, 0.26, and 0.09. In both cases, only four of the five eigenvalues are significant by any criterion.

The rotated factor patterns of the four factors of the two series are shown in Table VIII. In the computer-designed series, π is a linear function of all of the other variables; in the easy to make series, MR_{mp} is a function of π, $\mathcal{R}$, and MR_o.

In the computer-designed series there is one pair of close analogues; in the easy to make series, there are four such pairs.

It may be concluded that the multidimensional nonlinear mapping method, apparently because of the provision to delete hard to make analogues, may result in a series with too high a degree of multicolinearity. The authors themselves were aware of this problem.

Topliss Manual Method. Topliss has suggested that a useful strategy for the selection of substituents to optimize a lead is to first prepare the following specific set of five analogues: unsubstituted, 4-Cl, 3,4-Cl_2, 4-CH_3, 4-OCH_3. They are considered to usually be easy to synthesize. The substituents for the second small set of analogues are then selected on the basis of the apparent relationships between potency and physical properties which were revealed in the first set of analogues.[3] These relationships are considered to be suggestive only and not to form the basis of conclusions.

The various suggested sets (consisting of the original set plus, in turn, sets 2–8 of Table III in ref 3) were evaluated by factor and cluster analysis. The mean and standard deviation of the physical properties in each set are listed in Table IX. The π and σ are as given in the reference; MR values for the various positions have been used as steric parameters. It can be seen that in every case the standard deviation of at least one of the variables is less than ideal.

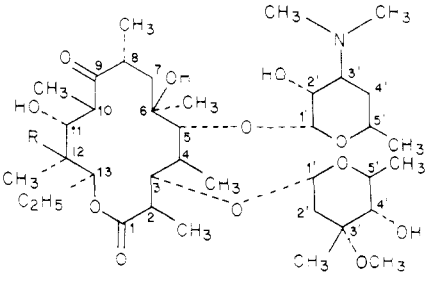
The squared multiple correlations of the variables are listed in Table X. The physical properties of several of the sets of analogues show a rather high multicolinearity.

The cluster analysis reveals that the initial set of five analogues contains two pairs with a distance of less than 1.0 standard deviation in standardized multidimensional space. This suggests that the initial set of five analogues does not contain as much variability in the four parameters as is theoretically possible. Since the follow-up set of analogues was designed to exploit relationships suggested by the original set, it is not surprising that every secondary set of analogues contains analogues which are close in multidimensional space to either one of the original five analogues or to another member of the same secondary set. The numbers of close analogues are listed in Table XI.

From this evaluation it may be concluded that, although the sets proposed by Topliss may be useful for their intended purpose, they are not ideal from the viewpoint of an uncorrelated spanning of substituent space in the minimum number of compounds.

Erythromycin Esters. This example was chosen to illustrate the contrast between a series with a high degree of colinearity and the same series augmented with additional analogues which eliminate this problem. Regression analysis and some aspects of the analysis of the design of this series have been discussed in a preliminary way previously.[6] The data are also listed in that source.

The first set consists of only those analogues in which one of the hydroxyl groups of erythromycin (structure II)



erythromycin

Table IX. Means and Standard Deviations of Physical Properties in Series Proposed by Topliss

set	π	σ	MR_4	MR_3	MR_2
1	0.50 ± 0.53	0.06 ± 0.32	5.32 ± 2.55	2.03 ± 2.24	
2	1.06 ± 0.81	0.27 ± 0.46	9.08 ± 7.85	2.21 ± 2.04	
3	1.08 ± 0.72	0.16 ± 0.35	13.19 ± 9.94	1.90 ± 1.95	
4	0.42 ± 0.81	0.30 ± 0.41	10.32 ± 7.97	1.83 ± 1.87	
5	0.78 ± 0.44	0.04 ± 0.29	7.29 ± 5.45	2.69 ± 2.31	
6	0.55 ± 0.47	0.16 ± 0.33	2.98 ± 2.76	5.67 ± 4.80	
7	0.43 ± 0.44	0.02 ± 0.27	3.41 ± 2.89	1.58 ± 1.67	2.84 ± 2.82
8	−0.41 ± 0.96	0.26 ± 0.34	8.83 ± 5.13	1.41 ± 1.39	

Table X. Squared Multiple Correlations of Properties in Table IX

set	π	σ	MR_4	MR_3	MR_2
1	0.82	0.84	0.40	0.70	
2	0.71	0.59	0.75	0.51	
3	0.84	0.66	0.90	0.42	
4	0.91	0.90	0.94	0.31	
5	0.73	0.46	0.67	0.66	
6	0.74	0.75	0.46	0.36	
7	0.75	0.78	0.59	0.59	0.57
8	0.74	0.32	0.56	0.51	

Table XI. Pairs of Close Analogues[a] in the Sets Proposed by Topliss

set	no. of analogues	pairs of close analogues
1	5	2
2	11	8
3	11	4
4	12	1
5	11	6
6	11	3
7	9	3
8	13	5

[a] Close is defined to be a distance of <1.00 in standardized Euclidean space. The two pairs of close analogues in set 1 are not included in the numbers listed for sets 2-8.

Table XII. Comparisons of the Mean, Standard Deviation, and Squared Multiple Correlation of the Physical Properties of Two Sets of Erythromycin Analogues

var	mean	SD	sq mult correlat
	A. series of alkyl esters		
log P	3.14	0.36	0.83
E_{s4}	−0.34	0.29	0.48
E_{s11}	−0.20	0.30	0.48
A	0.28	0.46	0.68
	B. total series		
log P	3.00	0.62	0.28
E_{s4}	−0.29	0.32	0.08
E_{s11}	−0.10	0.22	0.15
A	0.43	0.50	0.36

A or B is changed into an alkyl ester. It was one of the first series synthesized in the erythromycin-modification program at Abbott. Esterification was at either the 2′-hydroxyl, the 4′-hydroxyl, or the 11-hydroxyl; erythromycin A and B differ in the presence of the hydroxyl group at position 12. This set contains 28 compounds, which allowed the examination of seven variables (π_2, E_{s2}, π_4, E_{s4}, π_{11}, E_{s11}, A vs. B). If carefully chosen it could have also provided information on σ^* for each position of esterification. This would be ten variables for 28 analogues. The analysis here will consider those variables which were ultimately shown by regression analysis to be important determinants of relative potency. These important variables are log P, E_{s4}, E_{s11}, and the indicator variable A, which is equal to 1.0 if the compound is an erythromycin A analogue and 0.0 if it is an erythromycin B analogue.

The means and standard deviations of the physical properties in this data set are tabulated in the upper part of Table XII. The standard deviation of all physical properties is substantially lower than ideal.

The correlation matrix for this data set is shown in the upper part of Table XIII. There is no doubt that this series contains some correlation between variables. The question may be: How serious is this intercorrelation since the largest value corresponds to an R^2 of 0.61? The answer to this is found in Table XII; the squared multiple correlation of log P with the other three variables is 0.82. This high value clearly indicates a lack of independence of log P as a physical property. The eigenvalues of the correlation matrix also tell a similar story: they are 2.33, 0.996, 0.564, and 0.107. In practice, since log P is so strongly correlated with potency it was very difficult to decide by regression analysis if the steric terms and indicator variable are also important determinants of potency.

Table XIII. Correlation Matrices for Two Sets of Erythromycin Analogues

	log P	E_{s4}	E_{s11}	A
	A. series of alkyl esters			
log P	1.00			
E_{s4}	−0.48	1.00		
E_{s11}	−0.65	−0.16	1.00	
A	−0.75	−0.10	0.45	1.00
	B. total series			
log P	1.00			
E_{s4}	0.05	1.00		
E_{s11}	−0.28	0.04	1.00	
A	−0.50	0.22	0.37	1.00

Table XIV. Rotated Factor Pattern for Two Sets of Erythromycin Analogues

	factor			
var	1	2	3	4
	A. series of alkyl esters			
log P	−0.722	−0.472	−0.441	
E_{s4}	0	0	0.989	
E_{s11}	0.257	0.961	0	
A	0.969	0	0	
	B. total series			
log P	0	0.960	0	−0.244
E_{s4}	0	0	0.993	0
E_{s11}	0.978	0	0	0
A	0	−0.261	0	0.938

Table XV. Means, Standard Deviations, and Squared Multiple Correlation of the Physical Properties of a Set of Pargyline Analogues

var	mean	SD	sq mult correlat
MR_2	2.89	3.21	0.94
MR_3	2.41	2.39	0.97
MR_4	6.22	7.55	0.83
π_2	0.14	0.41	0.82
π_3	0.06	0.27	0.85
π_4	0.30	0.65	0.80
S_2	0.24	0.39	0.67
S_3	0.19	0.34	0.88
S_4	0.27	0.38	0.47
P_2	−0.20	0.41	0.94
P_3	−0.18	0.42	0.96
P_4	0.04	1.75	0.88

The rotated factor pattern is shown in the upper part of Table XIV. It shows very clearly the interrelationships between log P and E_{s11} and A. This pattern suggests that new analogues should be synthesized to reduce the correlation between log P and A and E_{s11}. Since the standard deviation of log P is rather low, it seems reasonable to include analogues of relatively lower log P in the enlarged data set.

The total set of analogues numbers 60. It includes derivatives of the ketone, the alkyl esters of the original

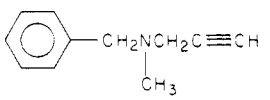
Table XVI. Correlation Matrix for Pargyline Data

var	MR_2	MR_3	MR_4	π_2	π_3	π_4	S_2	S_3	S_4	P_2	P_3	P_4
MR_2	1.00											
MR_3	−0.35	1.00										
MR_4	−0.29	−0.22	1.00									
π_2	0.34	−0.21	−0.22	1.00								
π_3	−0.14	0.50	−0.09	−0.08	1.00							
π_4	−0.22	−0.20	0.84	−0.08	0.04	1.00						
S_2	0.74	−0.38	−0.31	0.46	−0.14	−0.20	1.00					
S_3	−0.33	0.82	−0.20	−0.20	0.36	−0.19	−0.36	1.00				
S_4	−0.14	−0.00	0.45	−0.12	0.14	0.43	−0.10	0.00	1.00			
P_2	−0.86	0.29	0.23	0.12	0.11	0.23	−0.59	0.28	0.13	1.00		
P_3	0.27	−0.66	0.20	0.16	0.17	0.20	0.29	−0.40	0.10	−0.23	1.00	
P_4	−0.08	0.05	−0.23	0.00	−0.38	−0.10	−0.07	0.20	−0.30	0.09	−0.60	1.00

set, more varied types of esters, and several miscellaneous compounds. The structures of the compounds are listed in reference 6. Note in the bottom of Table XII that the standard deviation of log P is increased substantially. Additionally, the correlations between log P and the other physical properties are decreased (Table XIII). The eigenvalues of this correlation matrix are 1.80, 1.04, 0.73, and 0.42. The factor analysis for the total data set confirms that a slight relationship between log P and A remains, but the squared multiple correlations listed in Table XII show that this is not a large colinearity. Regression analysis of this data set was much less ambiguous than that of the alkyl esters alone.[6]

This example has shown the contrast between the factor analysis of a poorly designed and a more satisfactory data set.

Pargylines. This data was chosen as an example



pargyline

because it is a data set for which it has been very difficult to find a satisfactory regression equation.[13] The means, standard deviations, and squared multiple correlations are listed in Table XV. For only 2 of the 12 physical properties considered is the squared multiple correlation less than 0.80. This is in spite of the fact that only three pairwise correlations are >0.80 and none are >0.90 (Table XVI).

The last two examples suggest that one of the reasons that it may appear impossible to derive a regression equation for a data set may be that the physical properties are not independent.

Discussion

This work was initiated in response to comments of synthetic chemists that often it is impossible to synthesize any member of certain published[1] or internally generated clusters. This lack of attention to synthetic feasibility (proposed criterion 4) is one shortcoming of the cluster analysis approach to series design. A second shortcoming is the lack of attention to proposed criterion 2, sufficient variation in all properties. As noted above, this could result in too little variation in electronic properties.

In order to overcome the problem of synthetic difficulty, one may choose to do a customized cluster analysis on all the analogues which are synthetically feasible. Of course, this may involve a lot of work finding parameter values for compounds of which only a few will be synthesized. Additionally, it has been our experience that such a customized analysis often leads to series which contain too little variance and too high a degree of multicolinearity. Apparently, implicit in a synthetic chemists initial definition of easy to make is the assumption that the same reaction sequence will be used for all analogues; hence, only chemically similar analogues will be included. However, once attention is focused on the type of analogues required, alternate synthetic pathways may be perceived.

The multidimensional nonlinear mapping strategy solves the synthetic feasibility problem, but in so doing it may produce series with a high degree of multicolinearity. Although this approach considers the possibility of multiple substituents, explicit attention is not paid to the criterion of sufficient variability in key properties. In the examples published, the data base from which the substituents are chosen is restricted to substituents on aromatic rings. Finally, this program is not generally available.

The Topliss strategy was designed for other purposes, but if it is followed one will not have explored all of substituent space in an uncorrelated manner and some of the analogues will provide only redundant information. Hence, the full potential of the series will not have been explored. Additionally, it is even more rigid than the others with respect to the choice of analogues for synthesis.

The strategy proposed in this report results in a compromise series which meets all criteria satisfactorily. In particular, because of the early and frequent input of the synthetic chemist, synthetic feasibility is a key element in all decisions. A second strength of the approach is that attention is paid to the variability criterion with the result that multiply substituted analogues will be included when necessary. As was noted in the introduction, the proposed criteria can be applied to any molecular descriptors of interest. This is conveniently accomplished because no standard data base is used; rather, each series is analyzed by readily available BMDP programs. For the same reason, the strategy is not restricted to phenyl substituents only but can be as easily applied to any proposed series.

Appendix. Description of Factor Analysis

Factor or principal-component analysis is a mathematical method which is used to study the relationships between several properties which are associated with a series of observations.[14-16] In the subject of this report, factor analysis has been used to study the relationships (or lack of) between the physical properties of a set of substituents. The mathematical details are available elsewhere;[14] what follows is a discussion of the meaning of the various results of the calculations.

The first step in the calculation of the factors is the calculation of the correlation matrix. This matrix describes the degree of relationship between variables taken two at a time. For example, the correlation matrix for the factors in Table XIV is shown in Table XIII. Each entry is the correlation coefficient, r, between two variables. R^2 is the fraction of the variance in one property which may be explained by the variation in the other. For example, in

the series of alkyl esters 48% of the variation in E_{s4} may be explained by concurrent variation in log P; in the total series this value is reduced to 5%.

Although the correlation matrix is useful, it does not give a complete picture of the relationships between properties, because one property may be a linear function of more than one other property. Examination of the eigenvalues of the correlation matrix provides information on the number of truly independent variables. Every square matrix, including a correlation matrix, has an associated set of eigenvalues λ and eigenvectors. There are as many eigenvalues as there are variables in the data set. The sum of the eigenvalues is equal to the sum of the elements of the principal diagonal: for a correlation matrix this is equal to the number of variables. The eigenvalues are extracted in such a way that the first is the largest, etc. The proportion of the total variance in the data which is explained by a particular eigenvalue i is equal to the value of that eigenvalue divided by the sum of all the eigenvalues.

$$FV_i = \frac{\lambda_i}{\sum_n \lambda_i}$$

Thus, if one variable is a linear combination of two other variables, two rather than three significant eigenvalues would be associated with these properties. For example, in the analysis of the electronic properties σ_m, σ_p, $\mathcal{F}$, $\mathcal{R}$, S and P, only two of the six eigenvalues were nonzero. In chemical terms, there are only two electronic properties of substituents, field and resonance.

The determinant of a matrix is equal to the product of the eigenvalues. Hence, if the objective of a series design is to maximize the independence of the properties, then the determinant is also maximized.

Each eigenvalue has associated with it an eigenvector. This eigenvector consists of the coefficients of the contribution of each of the original variables to the eigenvalue.

Principal components are simply calculated from the eigenvectors as the product of the eigenvector and the square root of the corresponding eigenvalue. The principal components are also known as the unrotated factor pattern.

For further calculations one must choose how many factors are to be used. The choice depends on the values of the eigenvalues and the use to which the data will be put.[14] Since our objective is to look for unwanted relationships between physical properties, we decided that one measure of independence would be the number of eigenvalues necessary to explain 95% of the variance in the data. Factor analysis using this number of eigenvalues reveals if the properties are independent. If they are not, the relationship is highlighted by the rotated factor pattern.

Rotation of the factor pattern is performed to maximize the number of high and low loadings of variables on factors. Several methods are available; the varimax method was used in these examples.

The rotated factor pattern describes the relationships between the factors and the properties and, hence, between the properties also. A data set in which all properties are independent will have a factor pattern with as many factors as variables. Each variable will load on one factor only. An example is Table IV. On the other hand, highly related variables load similarly on all factors. For example, in Table I $\mathcal{R}$ and P and, to a somewhat lesser degree, $\mathcal{F}$ and S show a relationship.

The factor analyses described in this report were calculated with the BMDP4M program.[11] The initial factors were extracted by the principal-axis method and rotations were performed by the varimax method. Communalities were assumed to be 1.0.

References and Notes

(1) C. Hansch, S. H. Unger, and A. B. Forsythe, *J. Med. Chem.*, **16**, 1212 (1973).
(2) R. Wootton, R. Cranfield, G. C. Sheppey, and P. J. Goodford, *J. Med. Chem.*, **18**, 607 (1975).
(3) J. G. Topliss, *J. Med. Chem.*, **20**, 463 (1977).
(4) J. G. Topliss, and R. P. Edwards, personal communication.
(5) C. Hansch, *Drug Des.*, **1**, 315–332 (1971).
(6) Y. C. Martin, "Quantitative Drug Design. A Critical Introduction", Marcel Dekker, New York, 1978.
(7) C. G. Swain, and E. C. Lupton, *J. Am. Chem. Soc.*, **90**, 4328 (1968). For all calculations reported here, the values were calculated from the formula reported by C. Hansch, A. Leo, S. H. Unger, K. H. Kim, D. Nikaitani, and E. J. Lien, *J. Med. Chem.*, **16**, 1207 (1973).
(8) S. H. Unger, Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, Mass., 1971.
(9) R. W. Taft, in "Steric Effects in Organic Chemistry", M. S. Newman, Ed., Wiley, New York, 1956, pp 556–675.
(10) A. Verloop, W. Hoogenstraaten, and J. Tipker, *Drug Des.*, **7**, 165–207 (1976).
(11) W. J. Dixon, Ed., "BMDP Biomedical Computer Programs", University of California Press, Berkeley, Calif., 1975.
(12) P. J. Goodford, A. T. Hudson, G. C. Sheppey, R. Wootton, M. H. Black, G. J. Sutherland, and J. C. Wickham, *J. Med. Chem.*, **19**, 1239 (1976).
(13) Y. C. Martin, W. B. Martin, and J. D. Taylor, *J. Med. Chem.*, **18**, 883 (1975).
(14) J. E. Overall, and C. J. Klett, "Applied Multivariate Analysis", McGraw-Hill, New York, 1972, pp 24–136.
(15) E. R. Malinowski, in "Chemometrics: Theory and Application", B. R. Kowalski, Ed., American Chemical Society, Washington, D.C., 1977, pp 53–72.
(16) S. Wold, and M. Sjostrom, in ref 15, p 243–282.