Strategy in Drug Design. Cluster Analysis as an Aid in the Selection of Substituents

Corwin Hansch,* Stefan H. Unger,

Department of Chemistry, Pomona College, Claremont, California 91711

and Alan B. Forsythe

Department of Biomathematics, University of California, Los Angeles, California 90024. *Received June 18, 1973*

The large number of possible substituents that might be selected for an initial set of derivatives presents a formidable problem in decision making. By factoring such a set into more or less homogeneous subgroups with respect to various physicochemical parameters of importance, one can then focus upon such special considerations as H-bonding effects, metabolic behavior, or ease of synthesis. If the clusters are formed by an objective procedure such as minimum Euclidian distance between the points in a parameter space, selecting one derivative from each cluster will tend to give a maximum range in parameter type and help to establish a viable structure-activity relationship more rapidly. Thus, by the use of hierarchical clustering, 90 substituents have been successively clustered into 5, 10, 20, and 60 clusters with respect to various combinations of the lipophilic π (and π^2) constant, electronic Swain and Lupton-type $\mathcal{F}$ and $\mathcal{R}$ constants, and the approximate steric MR (molar refractivity) and MW (molecular weight) constants. Clusters at the 60 level approach bioisosteric combinations, while clusters at the low cluster level (5-20) reflect increasing loss of information. Noncollinearity and variance of the substituents selected can be tested by a separate procedure.

Once a new lead molecule has been uncovered, be it from the folk literature, expedition up the Amazon, random screening, intuition, or whatever, the modern practice of drug research is to start a program of drug modification to find the most active yet least toxic derivatives. This has become an increasingly expensive and sophisticated undertaking; the cost of making and testing derivatives now averages \$2000-4000 per molecule. This makes the choice of derivatives much more crucial; one wishes to obtain the maximum amount of information from each probe. The information content of a derivative is its biological response (BR) in various systems and its relation to other derivatives. Thus, the choice of substituents directly determines the amount of information available to the investigator.

It has been pointed out[1-3] that the number of derivatives which can be made from a given set N of substituents is N^m, where m is the number of nonsymmetrical positions on the parent molecule. Thus, in the present study of 90 substituents, we have the possibility (on paper) of making 90, 8100, or 729,000 derivatives for one, two, or three positions on, for example, a benzene ring. What constitutes a representative selection of these possibilities? How can we obtain the maximum information from each of the derivatives actually synthesized and tested? What is the "best" selection of an initial set of derivatives (say five or ten), other factors being equal?

With the advent of large computers and extensive development of substituent constants[4,5] regression analysis has become an important tool in the elucidation of quantitative structure-activity relationships (QSAR).[3,6] QSAR attempt to explain the variance in BR $[\Sigma(y - \bar{y})^2/N - 1]$ for a set of derivatives in terms of the variation in extrathermodynamic substituent parameters (lipophilic, electronic, steric, etc.).

The importance of substituent constants in QSAR has led to considerable discussion of the interdependence of these[4,7-10] as well as to the proposal of criteria for selecting the best equations.[11-13] Recent work with electronic parameters has led to the forced orthogonalization of these in order to truly dissect field from resonance effects.[11-14,†]

Bioisosterism[15] has been redefined as the property of different molecules to induce equivalent BR's, for whatever reason.[16] Clearly, there are two types of bioisosterism: isometric and nonisometric. Isometric (equal measure) bioisosterism arises when all physicochemical parameters of importance have the same value for two or more substituents, while nonisometric bioisosterism can arise by (fortuitous?) cancellation of substituent effects. This is made clear by writing the QSAR in the following form

$$\begin{bmatrix} \log \mathrm{BR}_1 \\ \log \mathrm{BR}_2 \\ \cdots \\ \log \mathrm{BR}_N \end{bmatrix} = a\begin{bmatrix} \pi_1^2 \\ \pi_2^2 \\ \cdots \\ \pi_N^2 \end{bmatrix} + b\begin{bmatrix} \pi_1 \\ \pi_2 \\ \cdots \\ \pi_N \end{bmatrix} + c\begin{bmatrix} \mathcal{F}_1 \\ \mathcal{F}_2 \\ \cdots \\ \mathcal{F}_N \end{bmatrix} + d\begin{bmatrix} \mathcal{R}_1 \\ \mathcal{R}_2 \\ \cdots \\ \mathcal{R}_N \end{bmatrix} + e\begin{bmatrix} E_{s1} \\ E_{s2} \\ \cdots \\ E_{sN} \end{bmatrix} + f\begin{bmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix} + \epsilon \quad (1)$$

where π is the hydrophobic constant, $\mathcal{F}$ and $\mathcal{R}$ are Swain and Lupton type field and resonance parameters,[4,17] E_s is Taft's steric constant, and ϵ is the error. Equation 1 shows that the information (variance) contained in the BR on the left is partitioned into the six terms on the right, the last being the constant term. Each column in eq 1 may be considered a vector (ordered list of numbers) and together the six vectors are said to "span" substituent space. If the substituent constants for any two substituents are identical, then the log BR's will be the same and they will be isometrically bioisosteric. If the substituent constants are not all equal and $a\pi_i^2 + b\pi_i + c\mathcal{F}_i + d\mathcal{R}_i + eE_{si} + f = a\pi_j^2 + b\pi_j + c\mathcal{F}_j + d\mathcal{R}_j + eE_{sj} + f$, then i and j are nonisometrically bioisosteric.

Equation 1 is only an approximation since parameters have not been included for H bonding, special dipole interactions, side reactions, etc. Nonetheless, if each term is position dependent and we substitute at three positions, then, assuming one needs five data points to validate each term,[12] about 75 derivatives would be needed (15 × 5) and 32,767 equations ($2^{15} - 1$) would have to be examined in order not to miss significant correlations.[11] Fortunately, nature does not appear to be quite so complex (or our testing procedures are not that sensitive) and considerable reductions can occur through insensitivity to certain effects at certain positions, the ability to sum contributions, etc. Even for the simple case of a single effect at a single position, we still have a very large choice of potential sub-

[†] C. G. Swain, S. H. Unger, P. Strong, and N. R. Rosenquist, manuscript in preparation.

stituents from which to choose and the list grows rapidly as new constants are measured.

Topliss[18] has recently described a decision tree as an aid in the rational choice of substituents, leading rapidly to the most active congener. We would like to describe a more general approach toward the selection of the "best" initial set of substituents that will tend to provide maximum information to the researcher.

It should be clear that the initial testing of isometrically bioisosteric substituents should be avoided since they may all provide similar information. Therefore, an initial set of derivatives should, within the limits of synthetic availability, cost and other factors such as metabolic sensitivity be as different from each other as possible. (That is, the substituents should have qualitatively and quantitatively different physicochemical parameters, and these should be noncollinear.[13]) If we could restrict ourselves to two parameters, then a simple plot of the parameters *vs.* each other would suggest a good set of substituents: those that lie in different quadrants of the plot.[9] In the general case, however, where the QSAR is *not* known beforehand, there is little or no information concerning which parameters might be important. We should then consider, as a minimum, the five parameters indicated in eq 1. We cannot as yet parameterize all pertinent physicochemical properties which might play a role; however, lipophilic (both parabolic and linear forms), field and resonance electronic (or some linear combination such as σ_p), and steric parameters have generally been sufficient for the bulk of known QSAR. The problem then becomes one of plotting these five parameters and selecting substituents that are as distant from each other as possible, while also being noncollinear, a difficult task for three-dimensional man!

Fortunately, there is an objective numerical procedure for doing this and also for grouping the substituents into any number of objectively different subgroups. This procedure is hierarchical clustering,[19-21] a type of pattern recognition recently described for chemical problems by Kowalski and Bender.[22]

Suppose we have K parameters each with N substituents. Then the distance between points i and j is given by the Euclidian distance between them (eq 2). It is not necessary that the parameters x' be on the same scale, but the clustering will, in most cases, be more representative if the data are first standardized (eq 3). This equation measures the deviation of each point from the group average $\bar{x}'_i$ in units of the standard deviation s_i. Even though most substituent parameters represent relative free energy terms (log K_{rel} or log k_{rel}), they are not objectively scaled; *e.g.*, Hammett σ are corrected by dividing by a ρ determined for a limited, nonstandard set of substituents, sometimes even from a different solvent system, substrate, or reaction. Therefore, we have standardized all data, allowing inclusion of molar refractivities and molecular weights along with π and σ type parameters (see Method).

$$d_{ij} = [\sum_{k=1}^{K}(x'_{ik} - \bar{x}'_{jk})^2]^{1/2} \quad i, j = 1, 2, \ldots, N \quad (2)$$

$$x_{ik} = (x'_{ik} - \bar{x}'_i)/s_i \quad (3)$$

In the hierarchical clustering procedure, all interpoint distances in K space are found and the closest two points are clustered into a pseudopoint (see Figure 1 and Method). New distances are determined and the next closest point, including pseudopoints, is clustered. The procedure continues stepwise until one large cluster is formed. At any level of clustering the points within a cluster are objectively (by eq 2) the most similar to each other and different from all others (within group variance is minimized and between group variance maximized).[19,20] Hierarchical clustering is illustrated in Figure 1.

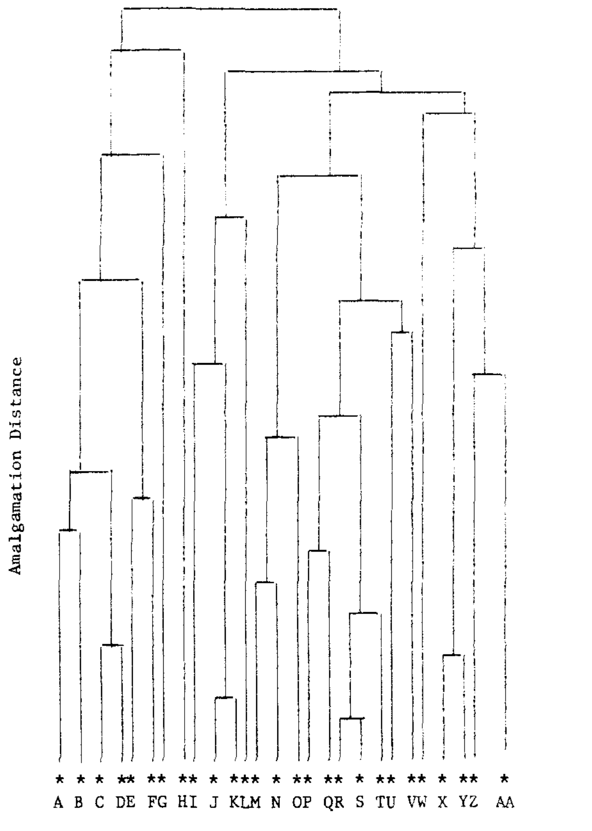

Figure 1. Symbolic representation of hierarchical clustering. Across the base of the diagram are 27 substituents, (A →AA) in 27 clusters. The two most similar substituents, R and S, first form a cluster of 2, followed by J and K and then X and Y. The RS cluster merges with T after C and D are clustered. Cluster RST is next merged with the PQ cluster and so on. The more unique substituents are clustered later. For example, H stands alone until the penultimate cluster.

It is quite important to recognize that the clusters are forced at each level and depend entirely upon the particular parameters and substituents selected. Clustering is a method of successively partitioning the total information (variance) provided. It is not a smoothing procedure (like least-squares fitting) but rather a way of objectively grouping the least dissimilar objects.

Therefore, if the QSAR is not known, we should select substituents on the basis of equally weighted parameters for lipophilic, electronic, and steric effects while including a parameter known not to be involved in the QSAR will serve no useful purpose and perhaps give misleading results. There is no single set of parameters that will be applicable for every situation.

Some pattern recognition and clustering procedures freeze the clusters at a certain level of information loss[21,22] and present two- or three-dimensional representations of the K space. For the purposes of selecting substituents, we feel the complete hierarchical procedure has the advantage of giving objective clusters for any number of substituents desired. Selecting one substituent from each cluster will help ensure the widest range in type. After a subset has been chosen, one should test for collinearity and variance (see Discussion).

This method is clearly only an aid in the rational choice of substituents; it does not represent an immutable law of nature and it would not be a disaster to "violate" the clustering since the activity of the compound and its parameter values will still contribute information. With large clusters from which to choose, other considerations must play their respective roles. It should also be clear that there are many diverse uses for clustering. For example, how does one choose the next set of substituents? Assuming that K parameters span the substituent space, selecting $K + 1$ substituents from $K + 1$ clusters will tend to give maximum variance. (One parameter defines a line, requiring two well-separated points in order to estimate the slope; two parameters define a plane, requiring three well-separated points, etc.) Therefore, selecting points near the original $K + 1$ will help to better determine the regression coefficients if the model (*e.g.*, eq 1) is correct; replicated points (or, approximately, bioisosteres) enable us to form an independent estimate of error variance. Selecting points from other clusters will help detect departures from the model such as nonlinearities (if, *e.g.*, π^2 were important but not included in the fitted equation).

A cautionary note is in order. It is not possible to avoid the QSAR; that is, one must use only relevant parameters in the clustering in order to obtain clusters that are relevant to the problem at hand.[‡] Therefore, the QSAR should be checked[11,13] at each stage. An initial set of five derivatives will not be sufficient[11,12] to determine more than perhaps which parameters are *least* important, but this will be valuable information in the selection of the second set. With about 10-20 derivatives one often has sufficient information to begin to rely more on regression analysis and clustering becomes less useful.

Method

We have limited ourselves to 90 uncharged aromatic-type substituents for which π, σ_m, σ_p, $\mathcal{F}$, $\mathcal{R}$, MR, and MW are available.[4,§] In addition, IO_2 was omitted because of its unique properties which, true also of charged substituents, tend to distort the analysis.

Three hierarchical clustering computer programs were investigated: the UCLA Biomed BMDP2M and BMDP1M and the Xerox Data Systems CLUSANL program based upon the algorithm of Ward and Hook.[19] The BMD programs ran in a matter of seconds on the UCLA IBM 360/90, while the CLUSANL took on the order of 6 min on the Pomona IBM 360/40, including printing time. The BMD programs provide a tree-like graphical printout (Figure 1) and amalgamation distance, while CLUSANL provides printed name clusters and an estimate of the information loss at each step (*e.g.*, total within group variance about the group means).[19,20] The programs differ in the treatment of pseudopoints, and, hence, in their clustering, especially at lower numbers of clusters. Intuitively reasonable results were obtained with the CLUSANL program which treats the pseudopoints as collections of individuals; BMDP2M lumps the members of a pseudopoint together and uses the centroid value. BMDP1M contains several options for distance and amalgamation criteria.

Lipophilic effects were parameterized by π in both parabolic and linear forms because both are found to be important in biological systems (π^2 is frequently important in *in vivo* systems). $\mathcal{F}$ and $\mathcal{R}$ were used because they provide the best available[4,11,14] factored electronic effects. Since we assume that the QSAR is not known beforehand, the use of these two factors instead of a blended value seems appropriate. Since σ_p contains about equal amounts of these two factors,[4,11,14,17] we have included some clustering using σ_p in supplementary tables.[§]

As discussed previously,[4] MR (molar refraction) and MW (molecular weight) are only related in a rough way (see Discussion) and therefore measure somewhat different aspects of steric "bulk." These are used in place of E_s values which are not available for the majority of substituents under study. Since steric effects usually are not as important as lipophilic and electronic effects, we have just used MR for all but the most general clustering where MW is included.

Additional computer programming involving the investigation of measures of noncollinearity and variance (see Discussion) was done at Pomona using the APL language on the APL*PLUS time-sharing system, courtesy of generous time grants from Scientific Time Sharing Systems, Inc.

Discussion

Results of the hierarchical clustering using CLUSANL with standardization are given in Table I. The 90 substituents have been successively clustered into 5, 10, 20, and 60 clusters with respect to four different sets of parameters. The numbers in Table I are arbitrarily assigned to the clusters at each level and apply only to a given column.

The variables (π^2, π, $\mathcal{F}$, $\mathcal{R}$, MR, MW) used to form the clusters of Table I are those that can at present be readily expressed in numerical terms. They by no means contain all of the information pertinent to the role of substituents in QSAR. However, cluster analysis does help to free the medicinal chemist's mind from concern about the proper choice of substituents with respect to general hydrophobic, electronic, and steric properties so that he is better able to separate other important factors. Once substituents have been clustered on the basis of these general parameters, one can then draw on his experience with bioorganic reaction mechanisms, drug metabolism, and the difficulties of organic synthesis to make more subjective judgments on which derivatives should first be prepared.

Set 1 of Table I represents the broadest set of parameters where lipophilic, electronic, and steric effects are weighted more or less equally in that two terms are used for each property. At the 60 level only nonsingle-membered clusters are shown. This higher cluster level was selected in order to indicate a more "natural" level of clustering; the members of these small groups would be expected to lie close to each other in K space whereas forced clustering at the 5-20 level builds groups in which the members are not necessarily very close to each other. The selection of the 60 level was made from an examination of plots of ln (cumulate information loss)[19] *vs.* number of clusters. This sigmoidal plot rises slowly between 90 and 60 clusters, is essentially linear from 60 to 20, and rises rapidly to the final value at 1 cluster.[=] Therefore, the 60-cluster level is a conservative estimate of the "natural" structure.[22] Figure 1 is a symbolic representation of the

[‡] For example, we have tested this on the result of Hussain and Lien[23] who determined LD_{50} for some cyclic ureas and thioureas in mice. When numbers of ring CH_2 groups, R_m, and MR of the substituent groups were "thrown in" the clustering, the most active congeners did not fall in the same higher numbered clusters; when restricted to the $(\log P)^2$, $\log P$ and dpm found most important in the QSAR, all the most active congeners fell in the same cluster. Of course, with a diverse set of derivatives, one might have little choice but to use a miscellaneous group of parameters such as MW, measured interatomic distances, nmr shifts, etc. The results will be no better than a QSAR based on these parameters. The advantage of clustering is in selecting substituents rationally before they are tested, when the QSAR is not known.

[§] Supplementary tables containing clustering based upon many subsets of the parameters under study are available from the Pomona College Medicinal Chemistry Project, Claremont, Calif. 91711. These will be updated from time to time as new substituents become available. The original analysis was completed in May of 1973.

[=] This smooth transition indicates that the substituents are fairly evenly distributed throughout the "volume" in K space.

Table I. Cluster Analysis for 90 Substituents by Various Parameter Combinations[a,d]

		Set 1 clusters				Set 2 clusters				Set 3 clusters				Set 4 clusters			
No.	Substituent	5	10	20	60[b]	5	10	20	60[b]	5	10	20	60[b]	5	10	20	60[b]
1	$B(OH)_2$	1[c]	1	1	1	1	1	1		1	1	1		1	1	1	
2	3,4-(OCH_2O)	1	1	1	1	1	1	1		1	1	1		1	1	1	
3	$CH_2CH_2CO_2H$	1	1	1		1	1	1	11	1	1	1		1	1	1	9
4	PMe_2	1	1	1		1	1	1		1	1	1		1	1	20	
5	CH_3	1	1	2	2	1	4	5	7	1	1	4	7	1	4	9	6
6	$CH{=}CH_2$	1	1	2	2	1	4	5	9	1	1	4	4	1	4	9	7
7	CH_2CH_3	1	1	2	2	1	4	5	7	1	1	4	7	1	4	9	6
8	CH_2OH	1	1	2		1	1	1		1	1	1		1	1	1	
9	H	1	1	2		1	1	1	11	1	1	4		1	1	1	9
10	$CH{=}CHCO_2H$	1	2	3		1	9	18		1	10	18		3	10	18	
11	CN	1	3	4	3	2	3	4	3	2	3	3		3	3	4	
12	NO_2	1	3	4	3	2	3	13	16	2	3	15	18	3	8	13	
13	CHO	1	3	4	4	2	3	4	4	2	3	3	3	3	3	5	3
14	CO_2H	1	3	4	4	2	3	4	4	2	3	3	3	3	3	5	3
15	COMe	1	3	4	4	2	3	4	4	2	3	3	3	3	3	5	3
16	CH_2Cl	1	3	5	5	2	5	6	5	1	1	4	5	1	4	6	4
17	$C{\equiv}CH$	1	3	5	5	2	5	6	5	1	1	4	5	1	4	6	4
18	Cl	1	3	5	6	2	2	2	1	2	2	2	1	2	2	2	1
19	N_3	1	3	5	6	2	5	6	8	2	2	2	19	1	4	6	15
20	SH	1	3	5	6	2	5	6	8	2	2	2	19	1	4	6	15
21	SMe	1	3	5	6	2	5	6	19	1	4	7	21	1	4	6	19
22	$CH{=}NOH$	1	3	5	7	2	5	8	6	1	4	6	6	1	5	8	5
23	CH_2CN	1	3	5	7	2	5	8	6	1	4	6	6	1	5	8	5
24	$CH{=}CHCN$	1	3	5	7	2	5	8	6	1	4	7	8	1	5	8	5
25	OCOMe	1	3	5	7	2	3	4	21	2	3	5	10	3	3	4	21
26	$CH{=}CHNO_2$ (trans)	1	3	5	8	2	5	6	8	1	4	7	8	1	4	6	
27	SCOMe	1	3	5	8	2	3	4	10	2	3	3	9	3	3	5	8
28	CO_2Me	1	3	5	9	2	3	4	10	2	3	3	9	3	3	5	8
29	SCN	1	3	5	9	2	3	4	10	2	3	3	9	3	3	5	8
30	$CONH_2$	1	4	6	10	3	6	7		2	3	5		3	3	7	
31	CONHMe	1	4	6	10	3	6	7		2	3	5	10	3	3	7	
32	SO_2NH_2	1	4	6	11	3	6	17		2	3	5	23	3	3	17	
33	SO_2Me	1	4	6	11	3	6	17		2	3	5	23	3	3	17	
34	SOMe	1	4	6	11	3	6	17		2	3	5		3	3	17	
35	NHCHO	1	4	7	12	3	6	16	17	1	4	6	6	1	5	16	16
36	NHCOMe	1	4	7	12	3	6	16	17	1	4	6	22	1	5	16	16
37	$NHCONH_2$	1	4	7		3	6	16		1	4	6		1	5	16	
38	$NHCSNH_2$	1	4	7	13	3	6	7		1	4	6		3	3	7	
39	$NHSO_2CH_3$	1	4	7	13	3	6	16	17	1	4	6	22	1	5	16	16
40	F	2	5	8		2	5	8		5	9	14		1	5	8	
41	OMe	2	5	8		5	8	15		5	9	14		5	9	15	
42	NH_2	2	5	8		5	8	14		5	9	16	20	5	9	14	
43	$NHNH_2$	2	5	8	14	5	8	15		5	9	16	20	5	9	15	
44	OH	2	5	8	14	5	8	15		5	9	14		5	9	15	
45	NHMe	2	5	9		5	8	14		5	9	16		5	9	14	
46	NHEt	2	5	9		5	8	14		5	9	16		5	9	14	
47	NMe_2	2	5	9		5	8	15		5	9	16		5	9	15	
48	Br	3	6	10	15	2	2	2	1	2	2	2	1	2	2	2	1
49	OCF_3	3	6	10	15	2	2	2	20	2	2	2	2	2	2	2	20
50	CF_3	3	6	10		2	2	3	2	2	2	2	2	2	2	3	2
51	$N{=}C{=}S$	3	6	10		2	2	2		2	2	2	17	2	2	2	
52	I	3	6	10		2	2	2	1	2	2	2	17	2	2	2	1
53	SF_5	3	6	10		2	2	3		2	2	2		2	2	3	
54	SCF_3	3	6	10		2	2	3		2	2	2		2	2	3	18
55	SO_2F	3	6	11		2	3	13	16	2	3	15	18	3	8	13	22
56	SO_2CF_3	3	6	11		2	3	13		2	3	15		3	8	13	22
57	CH_2Br	3	7	12	16	1	4	5		1	1	4	4	1	4	6	
58	$N{=}CCl_2$	3	7	12	16	2	5	6	8	1	4	7	21	1	4	6	15
59	SeMe	3	7	12	16	1	4	5	9	1	4	7	21	1	4	9	7
60	$CH{=}CHCOMe$	3	7	12	17	2	5	8		1	4	7	13	1	5	8	
61	$NHCO_2Et$	3	7	12	17	2	5	6	18	1	4	7	13	1	4	9	17
62	$CH{=}NPh$	3	7	13		2	3	4	4	4	7	12		3	3	5	3
63	SO_2Ph	3	7	13		2	3	13		4	7	12		3	8	13	
64	OSO_2Me	3	7	13		2	3	4	21	2	3	5	10	3	3	4	21
65	5-Cl-tetrazolyl	3	7	13		2	3	4	3	4	7	12		3	3	4	
66	$CH{=}CHCO_2Et$	3	8	14		2	5	6	14	4	6	10		1	4	6	13
67	NHCOPh	3	8	14		2	5	6	18	4	6	10		1	4	9	17
68	$N{=}CHPh$	3	8	14		5	8	15		5	9	19		5	9	15	
69	$CH{=}CHCOPh$	3	8	14	18	2	5	6	14	4	6	10	16	1	4	6	13
70	$NHSO_2Ph$	3	8	14	18	2	5	6	19	4	6	10	16	1	4	6	19
71	OSO_2Ph	3	8	14		2	2	2	20	4	6	11		2	2	2	20
72	COPh	3	8	15		2	2	3	2	4	6	11	14	2	2	3	2
73	$N{=}NPh$	3	8	15	19	2	2	3		4	6	11	14	2	2	3	18
74	OCOPh	3	8	15	19	1	4	10		4	6	10		4	7	11	
75	PO_2Ph	3	8	16		2	2	3	2	4	7	20		2	2	3	2

Table I (*Continued*)

No.	Substituent	Set 1 clusters				Set 2 clusters				Set 3 clusters				Set 4 clusters			
		5	10	20	60[b]	5	10	20	60[b]	5	10	20	60[b]	5	10	20	60[b]
76	3,4-$(CH_2)_3$	4	9	17		1	4	9		3	5	8		4	6	10	
77	3,4-$(CH_2)_4$	4	9	17		1	4	9		3	5	8		4	6	10	
78	Pr	4	9	17	20	1	4	10	12	3	5	9	11	4	7	11	10
79	*i*-Pr	4	9	17	20	1	4	10	12	3	5	9	11	4	7	11	10
80	3,4-$(CH)_4$	4	9	17	20	1	4	10		3	5	9	11	4	7	11	11
81	NHBu	4	9	17		1	4	9		3	5	17		4	6	10	
82	NHPh	4	9	17		1	4	10		3	5	17		1	4	9	
83	2-Thienyl	4	9	18		1	4	10		3	5	9	12	4	7	11	11
84	Ph	4	9	18	21	4	7	11		3	5	9	12	4	7	11	
85	CH_2Ph	4	9	18	21	4	7	11	13	3	5	9	15	4	7	11	12
86	*t*-Bu	4	9	18		4	7	11	13	3	5	9	11	4	7	11	12
87	OPh	4	9	18		4	7	19		4	6	10		2	2	19	
88	$SiMe_3$	4	9	18		4	7	12	15	3	5	9	15	4	7	12	14
89	Ferrocenyl	5	10	19		4	7	12	15	3	8	13		4	7	12	14
90	Adamantyl	5	10	20		4	10	20		3	8	13		4	7	12	

[a] CLUSANL with standardization. [b] Single member groups not indicated. [c] Numbering not significant, except to indicate clusters within a given column. [d] Parameters: for set 1, π^2, π, $\mathcal{F}$, $\mathcal{R}$, MR, MW; for set 2, π^2, π, $\mathcal{F}$, $\mathcal{R}$; for set 3, π, $\mathcal{F}$, $\mathcal{R}$, MR; for set 4, π, $\mathcal{F}$, $\mathcal{R}$.

Table II. Cross Correlations for 90 Substituents

	π^2	π	σ_m	σ_p	$\mathcal{F}$	$\mathcal{R}$	MR	MW
π^2	1.00	0.521	−0.239	−0.138	−0.274	0.028	0.390	0.289
π		1.000	−0.278	−0.153	−0.323	0.049	0.442	0.364
σ_m			1.000	0.884	0.959	0.468	−0.119	0.232
σ_p				1.000	0.716	0.827	−0.041	0.263
$\mathcal{F}$					1.000	0.200	−0.153	0.187
$\mathcal{R}$						1.000	0.067	0.221
MR							1.000	0.829
MW								1.000

entire hierarchical clustering procedure. This can be visualized by considering functions 84 and 85, Ph and $PhCH_2$, which are combined at the 60 level. At the 20 level they are forced to merge with *t*-Bu, OPh, and SiMe; at the 10 level, 3,4-$(CH_2)_3$, 3,4-$(CH_2)_4$, Pr, *i*-Pr, 3,4-$(CH)_4$, NHBu, NHPh, and 2-thienyl are forced into the cluster. This cluster remains intact at the 5 level. The sequence is: cluster "20" (60 level) → cluster "18" (20 level) → cluster "9" (10 level) → cluster "4" (5 level).

The strategy of selecting substituents will vary with different problems. For example, if the decision is made to make five derivatives using the first set of clusters of Table I, should one function be selected from each of the five clusters? Probably not, in this case, since ferrocenyl and adamantyl constitute a single cluster. These are difficult functions to introduce onto an aromatic ring and, although they would tend to provide maximum variance in properties of the six parameters determining these clusters, there is little in past SAR experience to incline one to overlook the large synthetic problems. It would be better to select two functions from different subclusters (10 level) of a large cluster (5 level).

In set 1, the cluster designated "1" contains 39 functions. At this low cluster level of 5 some strange bedfellows turn up. No one experienced with biochemical oxidations would consider Cl and SH or SMe equivalent; nor would one consider CH_2CH_2COOH and CH_2CH_3 equivalent, knowing that the acid would be completely ionized under physiological conditions. The cluster also contains strong hydrogen bonders such as CH_2OH, CH=NOH, etc., and functions with little or no propensity for hydrogen bonding.

There is also a large amount of information in the form of "chemical reactivity" which at present cannot be expressed by a single set or even a few sets of numerical constants. The various types of esters and amides possess quite different rates of hydrolysis, CH_2Cl undergoes a different type of nucleophilic substitution, and various types of C–H bonds are attacked at different rates by the mixed-function oxidases, etc. On all of these matters the medicinal chemist must exercise judgment.

As one goes to higher levels of factorization, the functions within a cluster seem more "reasonable." Clustering is not as forced and now at the 10 level, nine substituents are in the "1" cluster. Even at this level H is found along with CH_2CH_3 and $CH_2CH_2CO_2H$. At the 20 level, the cluster of nine substituents from the 10 level is broken into two clusters of four and five members which seem even more "natural." Finally, at the 60 level, clusters result in which the functions begin to appear "bioisosteric." Even at this level rather "strange" clusters result. In cluster "1," $B(OH)_2$ and 3,4-(OCH_2O) are found together. Looking at sets 2–4, it is seen that these two functions occur as individual clusters at the 60 level. This illustrates the supreme importance of the parameter set in determining the type of clustering which comes about at any given level.

Considering cluster "2" of set 1, one can understand why the three substituents, CH_3, $CH{=}CH_2$, and CH_2CH_3, are together. However, $CH{=}CH_2$ is chemically more different than CH_3 and CH_2CH_3. This comes out in sets 2–4 where CH_3 and CH_2CH_3 are in the same cluster and $CH{=}CH_2$ is placed in another cluster. Other substituents form more stable clusters: CHO, CO_2H, and COMe are in the same clusters in sets 1–4. Of course, CO_2H drops from this group when physiological solutions are considered and CHO and COMe separate when biochemical oxidations are considered as a variable. Again, looking at the 60 level, CH_2Cl and $C{\equiv}CH$ cluster in sets 1–4. No bioorganic chemist would entertain the hope that these two functions would form derivatives having equivalent biological activity. While the difference in chemical reactivity is immediately apparent from the chemical formulas, the similarities in general hydrophobic, electronic, and steric factors are not readily perceived. Cluster analysis gives a new perspective into the nature of substituents not obvious from classical organic symbolism.

Set 2 shows the effects of ignoring steric parameters. (Note that the clusters are no longer contiguous in order to conform to the order obtained for the case of set 1.) For example, CH_3, $CH{=}CH_2$, and CH_2CH_3 are now clustered with Pr and *i*-Pr; F is now clustered with $NHSO_2Ph$, etc.

Set 3 contains results which are probably more easily appreciated since only four linear parameters of the more familiar kind are involved.

Studying the clusters at the 20 and 60 levels, one will recognize many instances of functions which are often found to be bioisosteric; *e.g.*, Br, CF_3; C_6H_5, C_5H_5S; NO_2, CN; etc. It is of interest that for CH_2X, Br is "equivalent" to a double bond but Cl to a triple bond. For COX, equivalency is found when X = H, OH, or CH_3. When attached to O or NH, SO_2CH_3 and $COCH_3$ are "equivalent" as are NH_2 and CH_3 when attached to SO_2. However, the results of Table I should not be taken to imply that functions in the same cluster, even at the 60 level, will usually be bioisosteric.

A correlation matrix for the eight parameters (π^2, π, σ_m, σ_p, $\mathcal{F}$, $\mathcal{R}$, MR, and MW) for 90 substituents is given in Table II. The arcosines of these values give the angles of separation between the vectors[11] and, therefore, Table II is useful in understanding the taxonomy of substituent space.[4,14] The results agree essentially with those given in the previous paper for larger numbers of substituents. The significant observations are that $\mathcal{F}$ and $\mathcal{R}$ are more orthogonal than σ_m and σ_p and that the only other important collinearity is between MR and MW. Note that π is not highly correlated with MR. Therefore, the use of π^2, π, MR, and MW in the same set is reasonable as these parameters are sensibly independent for these 90 substituents.

As stated above, the initial group of compounds is used to estimate the QSAR, which is then helpful in estimating the BR for other compounds. A good initial choice of substituents, other factors being equal, would enable us to determine which parameters are of importance or, at least, which are *not* important. The choice of substituents can be poor in two respects. (a) Collinearity in parameter values will not enable an accurate choice of parameters to be made (collinearity problem).[13] (b) A narrow range in value for a given parameter will not enable it to be estimated with accuracy (variance problem).

If we let X be the matrix of parameter values selected (an $N \times 4$ matrix for the case of N substituents selected from set 3, for example), expressed relative to the substituent (column) means, then $(X'X)$ is the variance-covariance matrix (times a constant) for this set. Off-diagonal elements reflect the mutual covariance of the points (collinearity), while diagonal elements represent the variance of the points. Thus, the determinant, $\det(X'X)$, is an overall measure of both the variance and collinearity of the data selected. Anderson[24] gives a discussion of this entity in geometric terms as the volume parameter space spanned by the selected set of substituents. Thus, one can evaluate alternative sets of substituents, using the set for which $\det(X'X)$ is largest.**

Acknowledgment. This work was supported in part by Grants CA 11110 and RR-3 from the National Institutes of Health. We wish to thank Pomona College and UCLA Health Sciences Computing Centers and Scientific Time Sharing, Inc., for generous computer time allotments.

References

(1) C. Hansch, *Ann. N. Y. Acad. Sci.*, **186**, 235 (1971).
(2) C. Hansch, *Cancer Chemother. Rep.*, **56**, 433 (1972).
(3) C. Hansch, *Advan. Chem. Ser.*, **No. 114**, 20 (1972).
(4) C. Hansch, A. Leo, S. H. Unger, K. H. Kim, D. Nikaitani, and E. J. Lien, *J. Med. Chem.*, **16**, 1207 (1973).
(5) N. B. Chapman and J. Shorter, "Advances in Linear Free Energy Relationships," Plenum Press, New York, N. Y., 1972.
(6) C. Hansch in "Drug Design," Vol. I, E. J. Ariëns, Ed., Academic Press, New York, N. Y., 1971, p 271.
(7) P. N. Craig, *Advan. Chem. Ser.*, **No. 114**, 115 (1972).
(8) A. Cammarata and S. J. Yau, *J. Med. Chem.*, **13**, 93 (1970).
(9) P. N. Craig, *ibid.*, **14**, 680 (1971).
(10) A. Leo, C. Hansch, and C. Church, *ibid.*, **12**, 766 (1969).
(11) S. H. Unger and C. Hansch, *ibid.*, **16**, 745 (1973).
(12) J. G. Topliss and R. J. Costello, *ibid.*, **15**, 1066 (1972).
(13) D. E. Farrar and R. R. Glauber, *Rev. Econ. Stat.*, **49**, 92 (1967).
(14) S. H. Unger, Ph.D. Thesis, Massachusetts Institute of Technology, Sept 1970.
(15) A. Burger in "Medicinal Chemistry," Vol. I, 3rd ed, Wiley-Interscience, New York, N. Y., 1970, p 72.
(16) C. Hansch, *Intra-Sci. Chem. Rep.*, in press.
(17) C. G. Swain and E. C. Lupton, Jr., *J. Amer. Chem. Soc.*, **90**, 4328 (1968).
(18) (a) J. G. Topliss, *J. Med. Chem.*, **15**, 1006 (1972); (b) Y. C. Martin, W. J. Dunn III, *ibid.*, **16**, 578 (1973).
(19) J. H. Ward and M. E. Hook, *Educ. Psych. Measurement*, **23**, 69 (1963).
(20) D. J. Veldman, "Fortran Programming for the Behavioral Sciences," Holt, Rinehart and Winston, New York, N. Y., 1967, Chapter 12.
(21) R. C. Tryon and D. E. Bailey, "Cluster Analysis," McGraw-Hill, New York, N. Y., 1970.
(22) B. R. Kowalski and C. F. Bender, *J. Amer. Chem. Soc.*, **94**, 5632 (1972); **95**, 686 (1973).
(23) M. H. Hussain and E. J. Lien, *J. Med. Chem.*, **14**, 138 (1971).
(24) T. W. Anderson, "An Introduction to Multivariate Statistical Analysis," Wiley, New York, N. Y., 1958, p 166.

** For the special case of an initial set of five substituents (selected from the 90 under consideration), Mr. Laszlo Engleman of the UCLA Health Sciences Computing Facility has developed a method for finding the set maximizing $\det(X)$ when four, three, two, one, or zero specific substituents must be included. In this case the parameter values are centered about those for one substituent by subtracting its values from all the others. Then the "sweep out" method of determinant evaluation is used to select substituents and to evaluate the determinant. For set 3, when no substituents were forced in, $CH{=}CHCOOH$, adamantyl, NH_2, $POPh_2$, and SO_2F were selected. These represent the extreme outliers of data set 3. Note that each comes from a different cluster at the 5 level, an independent confirmation of the hierarchical clustering procedure.