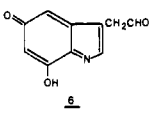
result was significant ($p < 0.001$) and, further, completely reminiscent of the norepinephrine depletion of ~80% of controls 10 days after injection reported by Massotti et al.[37] using a 21-μg dose of 5,7-DHT. However, the 20-μg dose of **5a** did not produce any substantial decline of 5-HT levels (mean ± SEM = 96 ± 1% of controls) in contrast to those observed with the parent 5,7-DHT.[37] This latter result is presumed to be the direct result of the lack of selective uptake of **5a** by serotonergic neutrons. Indeed, destruction of neurons by these neurotoxins has uniformly been shown to involve fairly selective uptake of the toxin by the targetted neurons. And, seemingly minor changes in the structure of a compound are well known to severely alter its uptake. Thus, we assume the addition of the 4-oxo group of **5a** compared to 5,7-DHT has effected such a change in its uptake properties with respect to 5-HT neurons. Nonetheless, **5a** certainly exhibits a number of biological properties that support its role as an active intermediate in the observed neurotoxicity of 5,7-DHT. First, it displays a general toxicity, leading to death of the intracranially injected animal, that is more potent than 5,7-DHT. Secondly, its long-term depletion of norepinephrine is completely comparable to that produced by a similar dose of 5,7-DHT.

Conclusions

The suggestion has been made that under certain circumstances a defect in the metabolism of 5-HT might lead to the formation of more reactive, more highly hydroxylated but unspecified derivatives[1,2,4,9] that in some fashion leads to mental disorders. Minor oxidation products of 5-HT in rats and rabbits have been speculated to be 4,5- or 5,6-DHT.[3,4,6,9] However, formation of di- or trihydroxy derivatives of tryptamine from 5-HT has never previously been demonstrated either in vivo or in vitro. The work reported here provides strong evidence that 5-HT is electrochemically oxidized to 5,7-DHT, which in turn is immediately further oxidized to **5a**. These reactions have been shown to occur in acidic solution. It has not been possible to detect **5a** as an electrooxidation product of 5-HT at physiological pH. However, the liquid chromatographic techniques employed in this investigation probably would not permit the detection of very small amounts of **5a** that might be formed in such pH regions particularly if it was coeluted with one of the many other colored oxidation products of 5-HT. Formation of **5a** as an oxidation product of 5-HT in vivo even in trace amounts might result in serious neurological consequences because of its powerful neurotoxic properties.

It has also been shown that the neurotoxin **5a** is formed by electrochemical oxidation of 5,7-DHT at pH 2 and 7 and by autoxidation of 5,7-DHT at pH 7. It is well known that 5,7-DHT has a profound lesioning effect on 5-HT containing neurons and, to a lesser extent, that it can also damage noradrenergic neurons.[16,23,28,29] However, the mechanism of neurotoxic action of 5,7-DHT is not well understood. Since the neurotoxicity of 5,7-DHT is prevented by inactivation of monoamine oxidase, it has been suggested[40] that in vivo in the presence of the latter enzyme autoxidation occurs, giving the quinone imine aldehyde **6**.



6

The aldehyde residue in **6** and, probably, the C(4) position provide two electrophilic sites that might be attacked by nucleophiles such as thiol residues on nerve ending proteins, leading to irreversible cross-linking of the proteins.[40]

The results reported here show that **5a**, formed by oxidation of 5-HT and 5,7-DHT, is a very powerful neurotoxin. This raises the possibility that at least part of the neurotoxicity of 5,7-DHT might be due to the in vivo formation of **5a**. In additoin, our results indicate for the first time that an oxidative metabolic route for 5-HT proceeding through 5,7-DHT to **5a** is *chemically* feasible. This, in turn, suggests that certain neurological disorders might be related to the 5-HT reaction pathway shown in Scheme I, which proceeds via the neurotoxin 5,7-DHT to the neurotoxin **5c**.

Acknowledgment. This work was supported by NIH Grants No. GM-32367-02 and NS-16887-03. Additional support was provided by the Research Council of the University of Oklahoma. One of us (D.L.) thanks the North Atlantic Treaty Organization for award of a fellowship.

(40) Rotman, A.; Daly, J. W.; Creveling, R. C. *Mol. Pharmacol.* **1976**, *12*, 887.

On the Significance of Clusters in the Graphical Display of Structure–Activity Data

James W. McFarland* and Daniel J. Gans

Central Research Division, Pfizer Inc., Groton, Connecticut 06340. *Received April 29, 1985*

A method is presented to evaluate the statistical significance of an apparently clustered group in the graphical display of structure–activity data. Two variations are described; each is implemented by means of a computer program. The first is applicable in situations with relatively small sets of compounds where a complete enumeration of all possible clusters can be accomplished reasonably on a high-speed electronic computer. The second is applicable in cases where such a calculation would be too time consuming. This latter variation uses random sampling of the set of all possible clusters. An application for each variation is given: for the smaller case a reevaluation of a study on aminotetralin and aminoindan monoamine oxidase inhibitors; for the larger case the discovery of some physical parameters that influence mutagenicity among some aminoacridine derivatives. It is proposed that this new technique be called cluster significance analysis (CSA).

Graphics are used in analyzing structure–activity data because the visual display of the information often affords insights that are not obvious otherwise. Notable examples in recent years include the work of Cramer and co-workers in their study of the antiallergic pyranenamines,[1] the study of antimicrobial activity in tuberlin analogues by Harrison et al.,[2] and the work of Morgan et al. on the carcinogenicity

(1) Cramer, R. D.; Snader, K. M.; Willis, C. R.; Chakrin, L. W.; Thomas, J.; Sutton, B. M. *J. Med. Chem.* **1979**, *22*, 714.

0022-2623/86/1829-0505$01.50/0

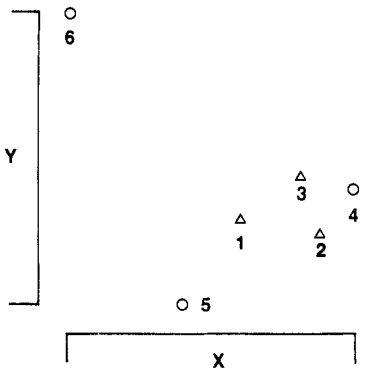


Figure 1. A two-dimensional plot of the six members of a hypothetical series of biologically active compounds: (△) the active members; (○) the inactive members. The x and y parameters are arbitrary. See text for a detailed explanation.

of polycyclic aromatic hydrocarbons in relationship to their electronic excited states.[3] The basic concept that we will expand upon here is derived from Magee's concept of "parameter focussing",[4] which concerns a set of congeneric compounds, the use of their sets of physical parameters as coordinates for two-dimensional plots, marking the compounds as active or inactive, and the chemist's discovery by this means that the active members are clustered in a relatively confined region of the graph. A reasonable assumption is that the more tightly clustered the active group, the more likely it is that the associated parameters are important determinants of activity. Putting it another way: for any relevant parameter there is an optimum region for active compounds.

This method has the advantage of being workable when only qualitative biological data (e.g., active or not active) are available and is likely to be more useful to medicinal chemists in the early stages of a developing drug series than is multiple regression analysis (MRA) where quantitative biological data are required.[5] Until now the disadvantages have been (i) the lack of a means to determine whether a "focussed" group of actives is merely a chance association and (ii) the limitation of needing to consider exactly two parameters at a time.

In this present work we will demonstrate two variations on a method for determining whether various clusters are chance occurrences, and that in principle this method can operate in one, two, three, and even more dimensions. Although quantitative *physical* data are important in this technique, quantitative *biological* data are in fact a hindrance. Thus, the method analyzes a series of compounds *qualitatively* by which of two response *classes* the members fall into but provides *quantitative* guidance as to whether the active class is fortuitously clustered. Certainly, this method is novel to quantitative structure–activity relationships (QSAR) and, as far as we are aware, to graphical analysis in general. We propose that it be designated as cluster significance analysis (CSA).

The Fundamental Concept

Figure 1 illustrates the basic idea in a simple imaginary case. There are six compounds: three actives (triangles) and three inactives (circles). The x and y axes represent arbitrary physical parameters. The actives appear to be clustered; the question is whether this apparent clustering is due to chance alone. What is needed first of all is a suitable definition of the tightness of a cluster. To this end the mean squared distance (MSD) among the three active compounds is calculated by taking the squared distance between each pair of points in the active group and then dividing the sum by the number of pairs:

$$\text{total squared distance} = (x_1 - x_2)^2 + (y_1 - y_2)^2 + (x_1 - x_3)^2 + (y_1 - y_3)^2 + (x_2 - x_3)^2 + (y_2 - y_3)^2 \quad (1)$$

$$\text{MSD} = (\text{total squared distance})/3 \quad (2)$$

Thus, the MSD is an index of tightness. Other definitions are conceivable—for instance the mean of the ordinary distances. The MSD, however, seems preferable to the latter because it puts greater weight on outliers.

Now, if x and y play no role in determining activity, the observed active cluster in Figure 1 is a chance aggregation, and all other possible clusters of the same size (three) are as likely to have arisen as the active one. In fact, there are 20 combinations in which these six items can be taken three at a time. The MSD for the active cluster having already been computed, the MSDs for the remaining 19 combinations are calculated in a same way and are compared to the MSD of the active group. The number of groups (including the active one itself) that have MSDs equal to or less than the MSD of the active group is designated as A. The probability (p) that a cluster at least as tight as the one observed would have arisen by chance alone then is given by

$$p = A/20 \quad (3)$$

This significance probability or p value thus indicates the significance of the relationship that x and y jointly have with activity. It has the same interpretation as, for example, the p value of an F test in multiple regression, because it gives the probability that a clustering at least as suggestive of relationship as the one actually obtained would have occurred by chance alone. As always, the lower the p value, the less tenable the chance explanation.

In Figure 1 the situation is so clear that the actual MSDs need not be computed. Compounds **1–3** comprise the active group. Only one group of three is more tightly clustered, compounds **2–4**. The groups composed of compounds **1**, **2**, and **4** and **1**, **3**, and **4** are close in size to the active group, but the members are somewhat farther apart. All other groups include compounds **5** and/or **6** and are therefore much more loosely clustered. Hence, the probability that clustering as tight as that observed occurs by chance is

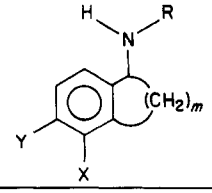
$$p = 2/20 = 0.10 \quad (4)$$

If we consider only p values at or below the 0.05 level as significant, we would judge this particular group of actives as possibly fortuitous, and the idea that the x and y parameters are indicators of activity is not confirmed.

While we have illustrated the concept with a two-dimensional example, other dimensions are easily treated. One has only to modify the definition of the squared distance appropriately. Thus, in the one-dimensional case, the terms containing y in eq 1 are dropped. For the three-dimensional situation, corresponding terms in z (a third parameter) are added to the equation. Higher dimensions are treated in the obvious way. Series containing greater numbers of compounds, both active and inactive, are of course more burdensome in terms of calculations,

(2) Harrison, I. T.; Kurz, W.; Massey, I. J.; Unger, S. H. *J. Med. Chem.* **1978**, *21*, 588.
(3) Morgan, D. D.; Warshawsky, D.; Atkinson, T. *Photochem. Photobiol.* **1977**, *25*, 31.
(4) Magee, P. S. In "IUPAC Pesticide Chemistry: Human Welfare and the Environment"; Miyamoto, J., Kearney, P. C., Eds.; Pergamon Press: Oxford, 1983; p 251.
(5) Hansch, C. In "Drug Design"; Ariens, E. J., Ed.; Academic Press: New York, 1971; Vol. 1, p 271.

Table I. Aminotetralins and Aminoindans as MAO Inhibitors[a]



compd no.	m	R	X	Y	Π	E_s^c	D[b]	RN[c]	act.[d]
7	2	CH_3	H	OCH_3	1.3	0.00	0	0.24	1
8	3	H	OCH_3	H	1.2	0.32	1	0.66	1
9	3	H	H	OCH_3	1.3	0.32	0	0.40	1
10	3	CH_2CH_3	H	OCH_3	2.2	−0.07	0	0.17	1
11	3	CH_3	H	OCH_3	1.7	0.00	0	0.58	1
12	3	CH_3	H	OH	1.0[e]	0.00	0	0.08	1
13	2	H	H	OCH_3	0.8	0.32	0	0.66	1
14	3	CH_3	OCH_3	H	1.7	0.00	1	0.46	0
15	3	$(CH_2)_2OCH_3$	H	OCH_3	1.7	−0.66	0	0.10	0
16	3	$(CH_2)_2CH_3$	H	OCH_3	2.7	−0.66	0	0.98	0
17	3	$(CH_2)_5CH_3$	H	OCH_3	4.2	−0.68	0	0.90	0
18	3	$CH_2C_6H_5$	OCH_3	H	3.5	−0.68	1	0.21	0
19	3	$(CH_2)_2OH$	H	OCH_3	1.0	−0.66	0	0.42	0
20	3	CH_3	OH	H	1.0[e]	0.00	1	0.63	0
21	3	$CH(CH_3)_2$	OCH_3	H	2.6	−1.08	1	0.76	0
22	3	$CH(CH_3)_2$	H	OCH_3	2.6	−1.08	0	0.21	0
23	2	$CH(CH_3)_2$	H	OCH_3	2.1	−1.08	0	0.54	0
24	2	H	OCH_3	H	0.8	0.32	1	0.65	0
25	3	$(CH_2)CH_3$	H	OCH_3	1.4	−0.66	0	0.08	0
26	3	$(CH_2)_6CH_3$	H	OCH_3	4.7	−0.68	0	0.96	0

[a] Modified from ref 7. [b] Dummy variable: $D = 0$ when X = H; $D = 1$ when X = OCH_3 or OH. [c] Computer-generated random number uniformly distributed between 0 and 1 (introduced in this work, not part of ref 7). [d] MAO activity in vivo: 0 indicates inactive compound; 1 indicates active compound. [e] Value is different from that of ref 7. In their study Martin and co-workers calculated Π by adding the lipophilic contributions of various structural elements using literature[10] values. In dealing with the X and Y substituents they assigned Π values of 1.7 to compounds 12 and 20 and also to compounds 11 and 14, yet the former pair differ from the latter in that they have hydroxy where the latter have methoxy substituents. For this reason we corrected these Π values to "1.0" to reflect a more realistic estimate of the hydrophobic effects involved. However, it should be noted that this change has little effect on the outcome of the CSA analysis: the results are essentially the same when the original values of 1.7 are employed. This is readily understood when it is observed that as the Π values for these two compounds change between 1.7 and 1.0 their markers in Figure 3 move away from those of like neighbors but at the same time move closer to others. Hence, there is overall no important change in the situation.

but the fundamental idea remains the same.[6] Helpful details on the mathematics involved are given in the Appendix.

The Procedure for Small Data Sets

In 1974 Martin and co-workers demonstrated that linear discriminant analysis (LDA) could be used successfully to establish which physical parameters are most influential in determining monoamine oxidase (MAO) activity among some aminotetralins and aminoindans.[7] The nature of the biological tests were such that there was no way to establish a scale of responses that could be graded quantitatively. In the final analysis there were only compounds that could be called either "active" or "inactive".[8]

Table I contains the relevant information for this analysis. Seven of the 20 compounds are classified as "active". Martin and co-workers found through LDA that the most important physical property related to activity

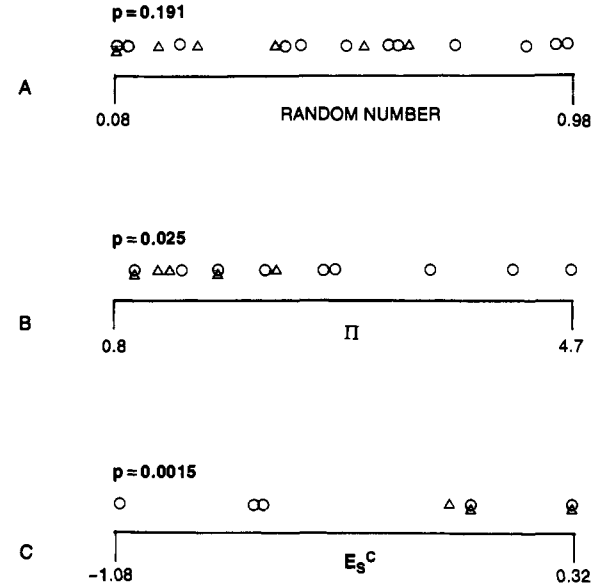


Figure 2. One-dimensional plots of active (△) and inactive (○) aminotetralin and aminoindan MAO inhibitors. Because some compounds share the same parameter values some of the symbols may represent more than one compound in this and the following figures. The scales are as follows: (A) the random number parameter RN, (B) the lipophilicity parameter Π, (C) the steric parameter E_s^c.

was the steric parameter E_s^c; secondarily, the dummy parameter D was also important (D is explained in a footnote to Table I). They found no relationship between the lipophilic parameter Π and biological activity.

(6) A potential problem in calculating the squared distances is that parameters are often given in diverse units. Even when the units are the same, in fact, one parameter may show a greater range than another and so will influence the entire evaluation more than the other. To avoid this, we standardize each parameter's values by subtracting the mean and dividing by the standard deviation before any distance computations are performed. The standardized parameters will then each have the same standard deviation (unity) and similar influence.

(7) Martin, Y. C.; Holland, J. B.; Jarboe, C. H.; Plotnikoff, N. *J. Med. Chem.* **1974**, *17*, 409.

(8) Of course, the method can be used on data with more than two values if there is a natural way to separate them into two classes.

Table II. Probabilities (p) for the Set of 20 MAO Inhibitors That Any Subset of Seven, If Selected by Chance Alone, Would Be at Least as Tightly Clustered as the Active Group in Various Parameter Spaces (Total Subsets: 77 520)

parameter(s)	A^a	p
D	21464	0.276 88
RN	14825	0.191 24
Π	1956	0.025 23
E_s^c	118	0.001 52
D, Π	1299	0.016 76
D, E_s^c	1175	0.015 16
RN, E_s^c	172	0.002 22
Π, E_s^c	71	0.000 92
RN, Π, E_s^c	151	0.001 95
D, Π, E_s^c	78	0.001 01

[a] Number of groups at least as tightly clustered as the active group.

As a first step in illustrating the method presented here, the same parameters considered by Martin and co-workers were evaluated graphically one at a time, the one-dimensional case. The results are shown in Figure 2. We also created a parameter from a computer-generated random number sequence to establish that nonrelevant information can be sifted out by the procedures described here (see Table I). It will become evident as we proceed that it is not always easy to identify such red herrings by inspection. Conversely, many groupings with severe outliers appear to be random, but even with these compounds included calculation often shows the clustering to be significant. A graph of the dummy parameter D is omitted because it would contain only two points and because each point would consist of both actives and inactives (triangles and circles).

To determine the significance probabilities relative to whether the active compounds are accidentally associated in Figure 2A–C and in terms of the dummy parameter, analyses were performed as in the imaginary example above, with a high-speed electronic computer used to make the calculations on the 77 520 combinations of the 20 compounds taken seven at a time.[9] The results are given in Table II. As can be seen, neither the dummy parameter D nor the random number parameter RN would be judged as significant determinants of MAO activity. However, the parameters Π and E_s^c both concentrate the active compounds sufficiently close to one another to give low probabilities of such tightness under chance association. On this basis, E_s^c gives the strongest evidence of association with activity, an observation that is in agreement with the literature. In contrast to the work of Martin and co-workers, this analysis selects Π but not D as another important determinant.

As an obvious next step, the compounds were plotted against E_s^c and Π, the two-dimensional case. The results are shown in Figure 3. Here the active compounds are concentrated in a small region in the upper left-hand portion of the figure, and by inspection one has no difficulty in accepting the idea that jointly E_s^c and Π are determinants of MAO activity. Calculation of the probabilities shows that of the reasonable combinations to consider, this combination of parameters results in the strongest evidence of association (lowest probability of tightness under chance). The addition of D or RN parameters in general results in marginally higher significance probabilities but does not obscure the basic underlying relationship. It appears, therefore, that irrelevant parameters may have to be identified by observing their effects in several combinations.

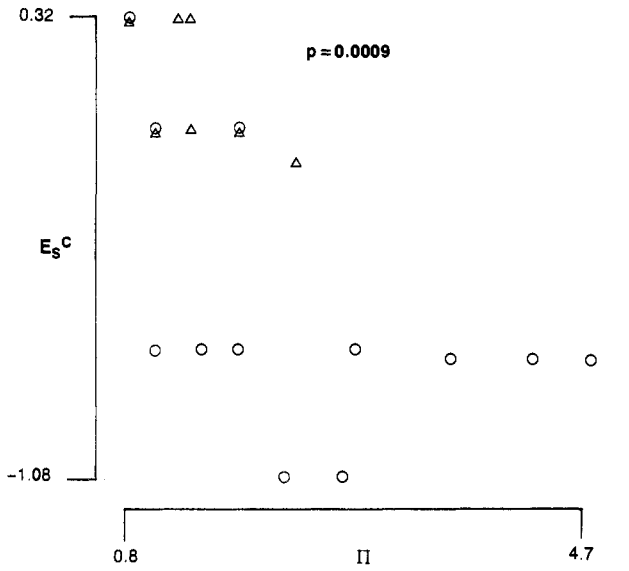


Figure 3. A two-dimensional plot of active (△) and inactive (○) aminotetralin and aminoindan MAO inhibitors: the lipophilic parameter Π vs. the steric parameter E_s^c.

A comparison of the LDA and CSA methods in analyzing these data on MAO inhibitors shows that both identify the most important factor in determining biological activity as being the steric parameter E_s^c. LDA also selects the dummy parameter D as a secondary factor of importance but does not identify Π as being significant. The reverse is true of CSA. These differences in no way invalidate the results of either procedure. It can be stated simply that LDA has not shown Π to be significant although it may be and that CSA has not shown D to be significant although it may be also. It is not unusual for distinguishable statistical procedures to give differing results, especially in somewhat marginal situations. Our own view is that, from both the graphical display of the data in Figure 3 and the statistical confirmation, Π should be accepted as a reasonable determinant of MAO activity.

The Procedure for Large Data Sets

The relationship between frameshift mutagenicity and the DNA-binding affinity of some aminoacridine derivatives was reported by Ferguson and Baguley in 1981.[11] They found that those compounds that caused a high maximal reversion frequency "clustered in a 'window' of DNA association constants between approximately 10^6 and 5×10^6". Interestingly, there were a significant number of compounds that resulted in low maximal reversion frequencies on either side of the "window". Ferguson and Baguley displayed this data graphically to good effect, allowing us to identify this group of derivatives as an excellent subject for the CSA method described here. They had no procedure to determine the significance of their results. The data with which we will be working is given in Table III; all compounds mentioned by Ferguson and Baguley are included.

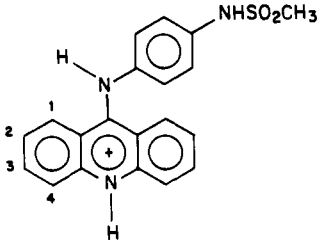
Of the 32 derivatives in the data set, 15 effected high maximal reversion frequencies. These data are presented as a one-dimensional plot against log K (logarithm of the

(9) Computer programs have been written for the two variations on the method described here. Annotated source code listings are available from us. These programs are written in ANSI FORTRAN 66 and operate on a DECsystem-10 computer.

(10) Hansch, C.; Leo, A.; Unger, S. H.; Kim, K. H.; Nikaitani, D.; Lien, E. J. *J. Med. Chem.* **1973**, *16*, 1207.

(11) Ferguson, L. R.; Baguley, B. C. *Mutat. Res.* **1981**, *82*, 31.

Table III. Aminoacridines: Mutagenicities, R_m's, pK_a's, and Group Dipole Moments (μ) for Various Derivatives[a]



compd no.	substit	max rev freq[b]	log K[c]	R_m	pK_a	μ[d]
27	3-SO_2CH_3	13.0	4.60	−0.39	5.65	2.365
28	3-aza	13.0	5.32	−0.35	5.32	1.100
29	3-CN	1.7	5.48	−0.15	5.75	1.965
30	2-$CONH_2$	1.7	5.54	−0.45	6.34	−1.885
31	2-Cl	1.9	5.62	0.16	6.42	−0.790
32	2-CH_3	15.0	5.70	0.19	7.23	−0.215
33	2-NO_2	0.5	5.84	−0.21	5.42	−1.965
34	1-Cl	1.0	5.90	0.01	5.86	−1.580
35	4-$CONH_2$	126.0	5.99	−0.47	6.12	3.770
36	1-OCH_3	185.0	5.99	−0.05		−1.250
37	3-$NHSO_2CH_3$	1.4	6.00	−0.26		
38	2-OCH_3	126.0	6.00	−0.04	6.87	−0.625
39	3-NO_2	224.0	6.04	−0.08	5.52	1.965
40	1-CH_3	3.3	6.04	0.19	6.43	−0.430
41	4-Cl	5.4	6.04	0.08	5.92	1.580
42	9-aminoacridine	219.0	6.08			
43	10-CH_3	212.0	6.18			
44	H	184.0	6.18	0.00	7.19	0.000
45	4-OCH_3	235.0	6.23	0.01	7.15	1.250
46	3-N_3	146.0	6.30	0.20	7.00	0.753
47	3-CH_3	147.0	6.52	0.24	7.49	0.215
48	4-CH_3	138.0	6.52	0.07	7.15	0.430
49	3-OCH_3	162.0	6.57	0.10	7.57	0.625
50	3-Cl	222.0	6.63	0.14	6.57	0.790
51	3-I	196.0	6.64	0.20	6.52	0.650
52	3-Br	263.0	6.67	0.16	6.56	0.780
53	4-NH_2	32.0	6.71	−0.19		1.560
54	3-$NHCOCH_3$	0.7	6.72	−0.12	7.34	1.860
55	3-NH_2	5.5	6.76	−0.18	9.80	0.780
56	2-NH_2	9.3	6.76	−0.32	7.15	−0.780
57	3-$NHCH_3$	7.5	6.87	−0.01	9.30	0.835
58	3-$NHCO_2CH_3$	2.6	7.00	−0.07	7.48	1.845

[a] From ref 11 and 12. [b] Maximum reversion frequencies are defined in ref 10; frequencies > 100 are considered to be "active", those < 100 are "inactive"; notice large gap between these two classes: lowest frequency for an "active" is 126; highest frequency for an "inactive" is 32. [c] Logarithm of the DNA-affinity association constant; from ref 11. [d] Group dipole moments are calculated from data in ref 15. It is assumed that the dipole of interest lies in the same direction as the 4-substituent; hence, the effective dipole moment will equal magnitude of the group dipole moment multiplied by the cosine of the angle made by the substituent of interest with the substituent at the 4-position: thus for 4-substituents the magnitude of the dipole is unchanged (cos 0° = 1); for 3-substituents, the magnitude is reduced by one half (cos 60° = 0.5); for 2-substituents, the magnitude is reduced by one-half and the sign is changed (cos 120° = −0.5); and for 1-substituents, the magnitude is unaltered but the sign is changed (cos 180° = −1).

DNA association constant) in Figure 4A. Because there are 565 722 720 combinations of 32 items taken 15 at a time, it is not feasible to compute the significance probability by exhaustive enumeration as above. Instead a modified method is employed in which the p value is estimated from a large random sample of the possible combinations.[9] The MSD of each member of this sample is compared to the MSD of the active group as before. The significance probability is then estimated on the basis of these comparisons in the same manner as earlier. This method is subject to sampling uncertainty so that the estimated p value is not exact. Statistical theory, however, provides for incorporating this uncertainty to obtain confidence bounds for the true p value. Thus, one can state within limits what the true p value (the value that would be obtained through exhaustive enumeration) would likely be. If the limits are felt to be too wide, they can be narrowed by increasing the size of the sample. It is usually not difficult, however, to obtain confidence bounds that are sharp enough to enable reasonably clear inferences to be drawn. Again, mathematical details are given in the Appendix.

Using this sampling method in the present example, we estimate the probability to be 0.000 03 ± 0.000 02 (95% confidence limits) that the 15 active compounds would be clustered along the log K scale as tightly as shown in Figure 4A by chance alone. Clearly, CSA has established that the relationship observed by Ferguson and Baguley is highly significant.

While this correlation is perfectly satisfactory for predicting the mutagenicity of other congeneric aminoacridines, we wanted to ask a new question. How is mutagenicity related to the more commonly measured or calculated physical constants used in QSAR? The answer to this would give us a better idea of the component forces associated with aminoacridine binding to DNA. Two sets of such constants have been reported by Ferguson and Denny: the R_m hydrophobicity constants for 30 of these compounds and the pK_as for 27 of them.[12] These workers made these determinations to perform MRA analyses on this group of compounds. This effort was only partially

(12) Ferguson, L. R.; Denny, W. A. *J. Med. Chem.* **1980**, *23*, 269.

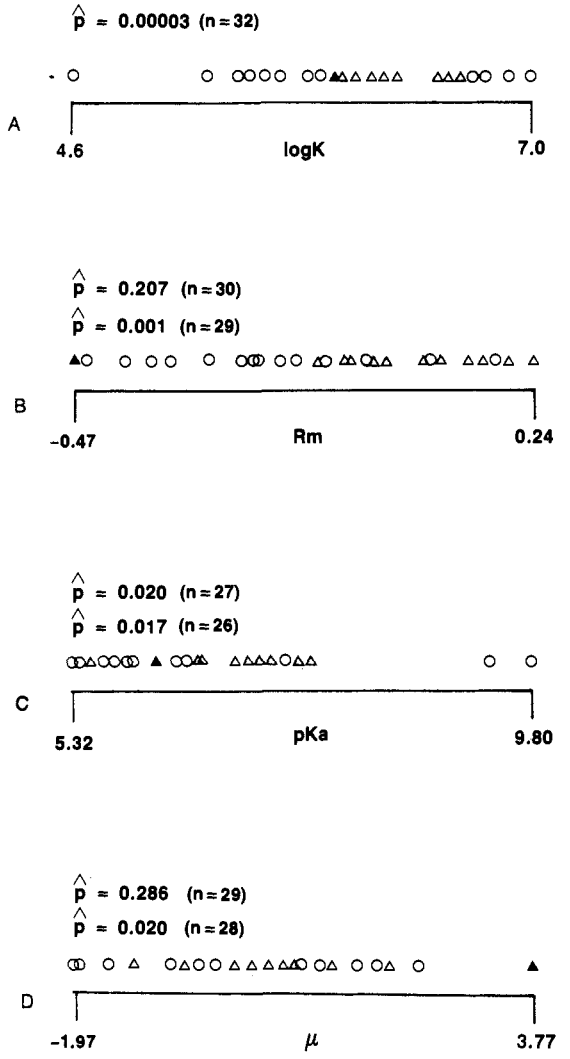


Figure 4. One-dimensional plots of active (△) and inactive (○) aminoacridine derivatives as frameshift mutagens. The scales are as follows: (A) the logarithm of the DNA association constant, (B) the lipophilicity parameter R_m, (C) the acidity parameter pK_a, (D) the group dipole moment μ. Compound **35** is represented by the solid triangle (▲) in this and the following figures.

successful: a moderate correlation (R = 0.79) between mutagenicity and R_m and σ_p was found among 19 members of the set, but this relationship could not be extended to the whole data base. From a theoretical point of view one might also consider that intermolecular forces associated with dipole moments could play a role, e.g., ion–dipole interactions.[13,14] Therefore, calculated group dipole moments (μ) have been included in Table III.[15] Hence, our goal now will be to find new relationships between mutagenicity and these readily available physical parameters.

Proceeding as before, we examine the one-dimensional plots of the compounds against these new parameters (see Figure 4B–D). It is not obvious from these graphs that the clusters of actives are sufficiently tight to indicate significance. However, estimation of the probabilities for these associations using the present random sampling method results in a much clearer picture (see Table IV). Initially, all of the available data were used: 30 compounds for the R_m series, 27 compounds for the pK_a series, and 29

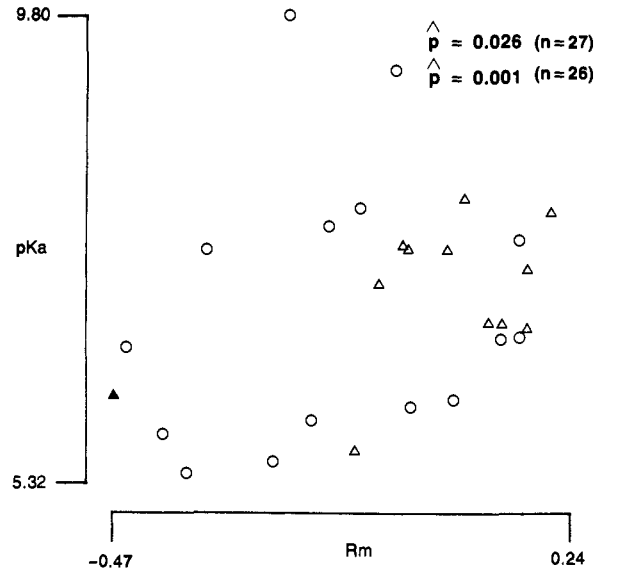


Figure 5. A two-dimensional plot of active (△) and inactive (○) aminoacridine derivatives as frameshift mutagens: the lipophilicity parameter R_m vs. the acidity parameter pK_a.

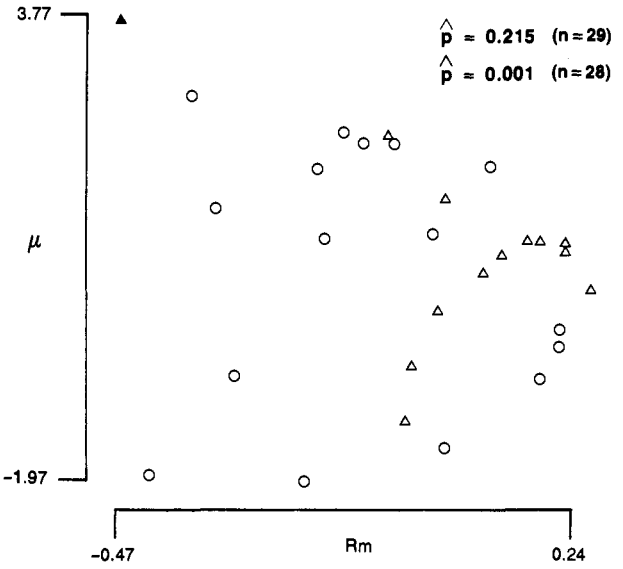


Figure 6. A two-dimensional plot of active (△) and inactive (○) aminoacridine derivatives as frameshift mutagens: the lipophilicity parameter R_m vs. the group dipole moment μ.

compounds for the group dipole moment (μ) series. Of these, only the pK_a series is significant ($p = 0.020 \pm 0.001$).

A closer look at the R_m and μ graphs reveals that in each case there is an active compound (filled triangle) that is quite remote from the main group of actives. In both instances the compound is the same: the 4-$CONH_2$ derivative (**35**). Because it fits very well into the relationship with log K, this compound cannot be considered an outlier by reason of unreliable data. However, it can be argued that there is a structural feature of the compound that allows it to associate with DNA more closely to the optimum than would be implied by the physical parameters under consideration here. A reasonable suggestion would be that the 4-$CONH_2$ group is favorably constituted and suitably located to interact with DNA by hydrogen bonding. The 2-$CONH_2$ derivative (**30**) is also favorably constituted but is not suitably located, while the 4-NH_2 compound (**53**) is suitably located but is not favorably constituted. No other groups appear to be reasonable candidates for hydrogen bonding in the way described. We can thus justifiably examine the relationships by omitting

(13) Tute, M. S. *J. Med. Chem.* **1970**, *13*, 48.
(14) McFarland, J. W. In "Progress in Drug Research"; Jucker, E., Ed.; Birkhaeuser Verlag: Basel, 1971; Vol. 15, p 123.
(15) McClellan, A. L. "Tables of Experimental Dipole Moments"; W. H. Freeman: San Francisco, 1963.

Table IV. Probabilities for Aminoacridine Derivatives That a Randomly Selected Group the Same Size as the Active Group Would Be at Least as Tightly Clustered as the Active Group in Various Parameter Spaces

parameter(s)	no. of compds		total combinations[c]	sample size[d]	no. as tight[e]	$\hat{p}$ (±95% CL)[f]
	N[a]	n[b]				
log K	32	15	565 722 720	200 000	6	0.000 03 ± 0.000 02
pK_a	27	12	17 383 860	50 000	986	0.019 72 ± 0.001 22
pK_a	26[g]	11	7 726 160	50 000	872	0.017 44 ± 0.001 15
R_m	30	13	119 759 850	50 000	10 339	0.206 78 ± 0.003 55
R_m	29[g]	12	51 895 935	50 000	67	0.001 34 ± 0.000 32
μ	29	13	67 863 915	50 000	14 299	0.285 98 ± 0.003 96
μ	28[g]	12	30 421 755	50 000	1006	0.020 12 ± 0.001 23
pK_a, R_m	27	12	17 383 860	50 000	1324	0.026 48 ± 0.001 41
pK_a, R_m	26[g]	11	7 726 160	50 000	63	0.001 26 ± 0.000 31
pK_a, μ	27	12	17 383 860	50 000	740	0.014 80 ± 0.001 06
pK_a, μ	26[g]	11	7 726 160	100 000	37	0.000 37 ± 0.000 12
R_m, μ	29	13	67 863 915	50 000	10 761	0.215 22 ± 0.003 60
R_m, μ	28[g]	12	30 421 755	50 000	50	0.001 00 ± 0.000 28
pK_a, R_m, μ	27	12	17 383 860	50 000	1367	0.027 34 ± 0.001 43
pK_a, R_m, μ	26[g]	11	7 726 160	200 000	28	0.000 14 ± 0.000 05

[a] Total number of compounds considered. [b] Number of actives in the total number considered. [c] Number of combinations of N things taken n at time. [d] Number of randomly sampled subsets evaluated to estimate the significance probability (p) that a clustering at least as tight as that observed for the active group would have arisen purely by chance. [e] Number of evaluated subsets found to be at least as tightly clustered as the active group. [f] Estimate with confidence limits at the 95% confidence level for the significance probability (p). [g] The 4-CONH$_2$ derivative **35** omitted from the data set.

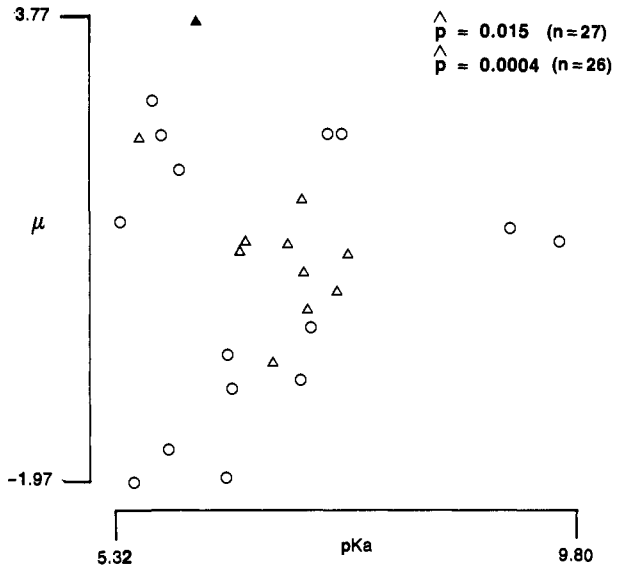


Figure 7. A two-dimensional plot of active (△) and inactive (○) aminoacridine derivatives as frameshift mutagens: the acidity parameter pK_a vs. the group dipole moment μ.

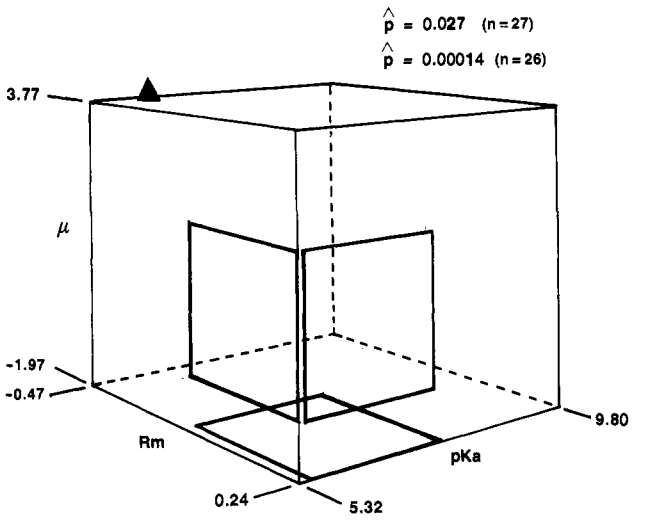


Figure 8. A three-dimensional plot of active and inactive aminoacridine derivatives as frameshift mutagens. To avoid confusion "windows" are placed on the surfaces of the cube to designate the ranges of the active compounds (with the exception of compound **35**, which is marked as the solid triangle ▲). The cube itself shows the parameter ranges for all compounds.

35 from our analyses.[16] The relationship with μ now becomes significant (p = 0.020 ± 0.001), that with R_m becomes highly significant (p = 0.0013 ± 0.0003), and that with pK_a is changed very little (p = 0.017 ± 0.001).

Figures 5–7 are two-dimensional plots of the various pairwise combinations of the parameters. The probability for the combination of R_m and pK_a is not much different from that given by pK_a alone when all possible compounds are considered, or from that given by R_m alone when **35** is omitted (see Figure 5). Nevertheless, two dimensions show the clustering of the active group better than one dimension does. A similar situation prevails for the combination of μ and R_m (Figure 6). However, there is a considerable lowering of both probabilities when μ and pK_a are combined (Figure 7). When all possible compounds are included, there is a modest decrease in the probability of tightness under chance association, but when **35** is omitted it drops to 0.0004 ± 0.0001, much better than for either parameter alone and the lowest value so far.

The rather good results that have been obtained with each of the possible two-dimensional plots lead naturally to the idea of combining all three into a grand three-dimensional summary. This is shown in Figure 8. Because actually plotting the points in this figure would confuse rather than clarify, we have chosen simply to put "windows" on the surfaces of the cubic figure to show the parameter ranges of the active compounds (with the exception of **35**). The cube itself is defined by the parameter ranges for all the compounds. Compound **35** is marked by the filled triangle; its location shows how truly remote this single outlier is in relation to the main body of actives. Again when including all possible compounds the probability of clustering of the actives under chance association in these three dimensions is similar to that given by pK_a alone. However, by omitting **35** from the analysis the probability becomes 0.000 14 ± 0.000 05, the lowest value of all. From this we conclude that it is not unreasonable to assume that **35** is exceptional because of its unique

(16) As always, however, one has to be cautious in discarding outliers. Almost any set of data (of a large enough size) can be made to yield statistically significant relationships if one allows oneself an unbridled license to do so.

Table V. A Comparison of Three Methods for Analyzing Structure–Activity Data

aspect	CSA	LDA	MRA
multiple dimensions	yes	yes	yes
significance testing	yes	yes	yes
qualitative interpretation	yes	yes	yes
sense of direction (predictability)	yes	yes	yes
quantitative correlation	no	no	yes
requires quantitative biological data	no	no	yes
role for "inactives"	yes	yes	no
can deal with "embedded" data	yes	no	yes
statistical assumptions: distributional	no	yes	yes
statistical assumptions: functional form	no	no[a]	yes
immediacy	yes	no	no

[a] It may appear at first sight that LDA involves an assumption of functional form in that the discriminant function is linear; this, however, is a consequence of the distributional assumptions made in LDA (stated in text) and not an *independent* assumption.

ability to hydrogen bond to DNA and that pK_a, R_m, and μ each play a significant role in determining mutagenicity in these aminoacridines.

Discussion

With these two examples we have shown that the CSA method is able to give insights into structure–activity data that have escaped notice previously. In the series of MAO inhibitors LDA was not able to detect the significance of lipophilicity upon activity,[7] while in the aminoacridine work only limited success was obtained using MRA.[12] It is also true that CSA did not confirm the previous finding of the importance of the dummy parameter D in the MAO series. In addition, this new method is somewhat difficult to apply to data in which the biological response varies either continuously or in evenly spaced discrete steps, selection of the "actives" and "inactives" being rather arbitrary. Hence, it appears that CSA is a new and useful tool complementary to those already in vogue. Table V is an attempt to summarize some of the strengths and weaknesses of CSA, LDA, and MRA. In the following comparisons it should be borne in mind that there is no one universally superior method; each has its advantages and disadvantages depending on the type of data being analyzed.

Each technique shares the ability to deal with multiple parameters and has associated with it a method to test the significance of the apparent relationship. Successful correlations from each technique can be interpreted qualitatively in terms of the physical parameters, and from such interpretations useful predictions can be made as to which new compounds will be active and which will not. This much the three analytical methods have in common.

MRA has an advantage over the other two in that it can predict the degree of activity. However, because quantitative biological data are frequently not available, LDA and CSA will often be successful where MRA cannot. In part, this may be owing to the fact LDA and CSA can make use—must make use—of the information contained in the inactive group of analogues. MRA is generally not suited to take advantage of these data because of the arbitrary nature of assigning a numeric value to a test result that only represents the highest dose evaluated (e.g., >200 mg/kg).

CSA distinguishes itself from LDA by virtue of its ability to treat "embedded" data. LDA works by defining a linear function that effectively separates items into two classes, e.g., "active" and "inactive". This function will separate by a point, a line, a plane, or a hyperplane, depending on the number of parameters involved, but it must divide the parameter space into two parts separating the actives from the inactives. If the members of one class of interest are "embedded" among the members of the other class such as we see in Figure 7, then no linear function can be found that will successfully accomplish this mission. Hence, CSA can be useful in such cases where LDA would be inappropriate. MRA treats embedded data by simply ignoring the other class, e.g., the "inactives".

Both LDA and MRA are tied to assumptions of statistical distribution. The former requires the parameter vectors in both classes to follow a multivariate normal distribution with the same covariance matrix. The latter requires one to assume that the errors of prediction in the biological response variable follow a univariate normal distribution with constant variance. Moreover, the validity of MRA depends on one having specified the correct functional form of the dependence of the response upon the predictors. CSA requires none of these. Thus, the CSA technique can be said to be both distribution-free and function-free. On the other hand, disadvantages of CSA are that, currently at least: (i) there is no way to assess exactly the contribution made by one parameter in the presence of others, because a p value is calculated for a whole set of parameters at once, and (ii) one can assess the significance of the relation between position in the graph and activity, but once that is established there is no equivalent of, say, a correlation coefficient to estimate the degree of precision in the relationship.

Finally, there is the issue of "immediacy". By this we mean the ability of those who have no special qualifications in statistical methods to interpret the results from these various analytical techniques. We submit that of these methods CSA is the most readily grasped intuitively. For favored simple cases both LDA and MRA may afford easily interpreted results. In more complicated cases both LDA and MRA require an advanced knowledge of statistics; however, CSA can be interpreted in such cases without introducing further complexity. Medicinal chemists are the principal generators and end users of structure–activity information, but they are for the most part not expert statisticians. For these reasons we believe that CSA will have a prominent future as a tool in drug design.

Appendix

We begin by presenting an efficient way of computing the MSDs for a large number of subsets (clusters). It is convenient to use a notation different from that of eq 1 above. Suppose there are k parameters numbered $1, \ldots, k$, and N compounds in all numbered $1, \ldots, N$. Let ξ_{ij} be the value of the ith parameter for the jth compound.

We suggest that the parameters be standardized ("autoscaled") to have standard deviation unity, to equalize the importance of each. Thus, we work with the values

$$x_{ij} = (\xi_{ij} - \bar{\xi}_i)/s_i \tag{5}$$

where $\bar{\xi}_i$ is the mean and s_i the standard deviation of the collection of values $\xi_{i1}, \ldots, \xi_{iN}$.

Let n be the size of the subset of active compounds and let S be any subset of size n of the set of integers $1, \ldots, N$. Thus, S represents an arbitrary subset of size n of the collection of all compounds in question. Indices in the sums below will be assumed implicitly to range as follows: i will range over $1, \ldots, k$; j will range over S; and j' will also range over S. (In the sole case of eq 7 an additional constraint is imposed.)

Let the squared distance between the jth and j'th compound be denoted by

$$d^2(j,j') = \sum_i (x_{ij} - x_{ij'})^2 \tag{6}$$

In S there are $n(n-1)/2$ distinct pairs of compounds; thus the MSD for S

$$M_S = \frac{2}{n(n-1)} \sum_{j<j'} \sum_{j'} d^2(j,j') \quad (7)$$

where the constraint $j < j'$ guarantees that in the double sum $j \neq j'$ and also that each pair is counted only once. However, eq 6 shows that $d^2(j,j') = 0$ if $j = j'$ and also that $d^2(j,j') = d^2(j',j)$ if $j \neq j'$. Thus eq 7 can be rewritten in the form

$$M_S = \frac{1}{n(n-1)} \sum_{j} \sum_{j'} d^2(j,j')$$

$$= \frac{1}{n(n-1)} \sum_{j} \sum_{j'} \sum_{i} (x_{ij} - x_{ij'})^2 \quad (8)$$

where now the lack of constraint on j and j' means that both indices range unrestrictedly over S. The last expression can be simplifed[17] to

$$M_S = \frac{2}{n-1} \sum_{i} \sum_{j} (x_{ij} - \bar{x}_i)^2, \quad (9)$$

where

$$\bar{x}_i = \frac{1}{n} \sum_{j} x_{ij}, \quad (10)$$

a considerable saving in computational effort.

The computational burden can be reduced further. By a standard identity[18]

$$\sum_{j} (x_{ij} - \bar{x}_i)^2 = \sum_{j} x_{ij}^2 - n\bar{x}_i^2 \quad (11)$$

for any value of i. Thus summing each term in eq 11 on i and multiplying by $2/(n-1)$ we arrive at

$$M_S = \frac{2}{n-1} \sum_{i} \sum_{j} x_{ij}^2 - \frac{2n}{n-1} \sum_{i} \bar{x}_i^2 \quad (12)$$

Let us put

$$Q_j = \frac{2}{n-1} \sum_{i} x_{ij}^2 \quad (13)$$

for each j. Because the sum on i is over all values $i = 1, \ldots, k$, the Q_j are fixed for a given problem. The N quantities Q_j, $j = 1, \ldots, N$, should be computed just once and stored.

Additionally, note that

$$\frac{2n}{n-1} \sum_{i} \bar{x}_i^2 = \frac{2n}{n-1} \sum_{i} \left(\frac{1}{n} \sum_{j} x_{ij} \right)^2$$

$$= \frac{2}{n(n-1)} \sum_{i} T_i^2 \quad (14)$$

where

$$T_i = \sum_{j} x_{ij} \quad (15)$$

for each i. Unlike the Q_j, the T_i have to be computed anew for each subset S, because the sum on j is over the values in S. The advantage of eq 14 is simply that it saves having to divide by n in computing each $\bar{x}_i$ each time a subset S is considered, a process that would involve k new divisions for each new S.

Using eq 13 and 14 in eq 12 there is obtained finally

$$M_S = \sum_{j} Q_j - \frac{2}{n(n-1)} \sum_{i} T_i^2 \quad (16)$$

To recapitulate, the T_i have to be computed anew for each S, while the Q_j are computed only once and then, given S, are retrieved for the values of j in S. Equation 16 represents a considerable saving in computation over eq 7, recalling that each d^2 term in the latter is itself a sum.[19] This saving allows one to use the exhaustive enumeration procedure, and thus obtain exact p values, for larger problems than otherwise would be the case.

For problems that are still too large for exhaustive enumeration, the random sampling method discussed in the main text can be employed. The technique just given for computing MSDs can be utilized without change; the major new point is the selection of random subsets rather than all possible ones.

In theory, greater statistical efficiency can be obtained in random sampling if one samples "without replacement", i.e., if the sampling process is such that it is impossible to sample the same subset twice. This would entail storing in the computer memory a complete list of all subsets sampled, however, which is not feasible in most settings. Even if it were possible, the savings in statistical efficiency would be more than outweighed by the greater complexity and hence longer running time of the subset selection algorithm.

Instead sampling "with replacement" is recommended. In this scheme at any point each of the $N!/[n!(N-n)!]$ possible subsets of size n is equally likely to be chosen, regardless of the past history of the selection process. Thus at any time each possible subset has probability $n!(N-n)!/N!$ of being chosen, whether or not it has been chosen previously.

A quick and simple way of effecting such selection has been described by Bebbington.[20] At each stage the compounds are considered in turn, proceeding in order from 1 to N. The probability of selecting compound 1 for membership in the subset is set at n/N. If compound 1 is in fact chosen, the selection probability for compound 2 is set at $(n-1)/(N-1)$; if not, it is set at $n/(N-1)$. As each compound is considered, the denominator of the fraction is reduced by 1. If the preceding compound was indeed selected, the numerator is also reduced by 1; if not, it remains unchanged. It can be shown that this algorithm places equal probability on the selection of each possible subset.[21] The actual random choices can be carried out

(17) Kendall, M. "Multivariate Analysis"; Charles Griffin and Co.: London, 1975; p 37.

(18) Dixon, W. J.; Massey, F. J., Jr. "Introduction to Statistical Analysis"; 3rd ed.; McGraw-Hill: New York, 1969; p 28.

(19) Efficiency in computational steps similar to that obtained with eq 16 can be had by calculating instead all $d^2(j,j')$ just once initially (using eq 6) and storing them. For each S one would then retrieve the appropriate d^2 terms and use eq 7 directly. An examination of the number of arithmetic and data-retrieval operations needed for each S suggests that, for each kind of operation, the number needed for this alternate algorithm is approximately $(n-1)/[2(k+1)]$ times the number needed for eq 16. By this yardstick sometimes the one or sometimes the other algorithm will be more efficient. The alternate algorithm, however, has the disadvantage that for very large N the two-dimensional array of stored d^2 values could be so large that its size itself would lengthen considerably the running time or for smaller computers make the alternate algorithm simply infeasible.

(20) Bebbington, A. C. *Appl. Statist.* **1975**, *24*, 136.

(21) Bebbington in his article refers to sampling "without replacement". This is not a contradiction; rather it corresponds to the fact that, in our terminology, a given *compound* cannot be selected more than once in a given *subset*. Accordingly, the sampling of compounds to form any one subset may be said to be performed without replacement. At the level of sampling successive subsets, however, sampling is still carried out with replacement.

with computer-generated random numbers.

Having obtained MSDs for a random sample of B subsets (not all of which need be distinct), we determine the number A among them (counting "repeats" as often as they occur) for which the MSD is at least as small as that of the subset of observed active compounds. As before the ratio

$$\hat{p} = A/B \quad (17)$$

is formed. Unlike the case of exhaustive enumeration, however, this ratio is not the true p value but simply an estimate of it (hence the caret).

The fraction of all subsets with an MSD at least as small as that of the "active" one is in fact the true but unknown p. Mathematically, the count A has a binomial distribution with probability parameter p and sample size parameter B. From this distribution confidence limits for p, incorporating the uncertainty due to sampling, can be obtained. Approximate confidence bounds, at the 95% level of confidence, are given by[22]

$$\hat{p} \pm 1.96[\hat{p}(1-\hat{p})/B]^{1/2} \quad (18)$$

Of course, the significance of the observed clustering must be evaluated in any given case by recalling that the true value of p, estimated by the bounds of eq 18, can be thought of as an ordinary significance probability.

Registry No. **7**, 52372-93-7; **8**, 52372-97-1; **9**, 52373-02-1; **10**, 52373-03-2; **11**, 52373-04-3; **12**, 52373-05-4; **13**, 52372-95-9; **14**, 52372-98-2; **15**, 52373-06-5; **16**, 52373-07-6; **17**, 52373-08-7; **18**, 52372-99-3; **19**, 52373-09-8; **20**, 52373-00-9; **21**, 52373-01-0; **22**, 52373-10-1; **23**, 52372-94-8; **24**, 52372-96-0; **26**, 52373-12-3; **27**, 61481-83-2; **28**, 72738-92-2; **29**, 53251-06-2; **30**, 72739-01-6; **31**, 61462-73-5; **32**, 53222-10-9; **33**, 58658-21-2; **34**, 72738-97-7; **35**, 72739-02-7; **36**, 61417-04-7; **37**, 72738-91-1; **38**, 53222-12-1; **39**, 59748-51-5; **40**, 61417-03-6; **41**, 61417-08-1; **42**, 90-45-9; **43**, 61417-13-8; **44**, 53478-38-9; **45**, 61417-05-8; **46**, 64894-90-2; **47**, 53478-39-0; **48**, 53221-79-7; **49**, 59748-95-7; **50**, 57164-73-5; **51**, 64894-94-6; **52**, 58682-45-4; **53**, 61417-07-0; **54**, 53222-14-3; **55**, 581-29-3; **56**, 58658-24-5; **57**, 66147-73-7; **58**, 72738-90-0; MAO, 9001-66-5.

(22) Reference 18, p 246.

Synthesis and Inhibition of Human Acrosin and Trypsin and Acute Toxicity of Aryl 4-Guanidinobenzoates

J. M. Kaminski,[†] L. Bauer,[‖] S. R. Mack,[†] R. A. Anderson, Jr.,[†,‡] D. P. Waller,[⊥] and L. J. D. Zaneveld*[§]

Departments of Obstetrics and Gynecology, Physiology, and Biochemistry, College of Medicine, Rush University, Chicago, Illinois 60612, and Departments of Medicinal Chemistry and Pharmacognosy, and Pharmacodynamics, College of Pharmacy, University of Illinois at Chicago, Chicago, Illinois 60680. *Received April 29, 1985*

The aryl 4-guanidinobenzoate, 4′-nitrophenyl 4-guanidinobenzoate (NPGB), is a potent inhibitor of sperm acrosin, an enzyme with an essential function in the fertilization process. NPGB prevents fertilization in a number of animal species and is a good lead compound for the development of contraceptive agents. In order to assess the efficacy of other aryl 4-guanidinobenzoates as acrosin inhibitors, 24 of these compounds were synthesized. Their inhibitory activity toward human acrosin was determined and compared with their activity toward human pancreatic trypsin in order to assess whether inhibitor sensitivity differed between these similar enzymes. Nine of the inhibitors were synthesized from phenols approved by the FDA for therapeutic use. The acute toxicity of these inhibitors in mice was determined and compared to that of nonoxynol-9, the most commonly used active ingredient in today's vaginal contraceptive preparations. All of the compounds proved to be potent inhibitors of human acrosin although 3 orders of magnitude difference were observed between the most and least effective inhibitors. Little specificity was present in regard to their inhibition of acrosin and trypsin. All the aryl 4-guanidinobenzoates synthesized from FDA-approved phenols were less toxic than nonoxynol-9, and it is concluded that these 4-guanidinobenzoates are of interest for further development and testing as nonhormonal contraceptive agents.

Acrosin, a serine proteinase with trypsin-like specificity and inhibitor sensitivity,[1] is associated with the sperm acrosome and has an essential function in the fertilization process. Spermatozoa appear to require acrosin for one or more of the following: (1) the sperm acrosome reaction, (2) sperm binding to the zona pellucida, the innermost of three layers surrounding the ovum during fertilization, and (3) lysis of a passage for the spermatozoon through the zona pellucida.[2] Thus, in the absence of acrosin, spermatozoa are unable to penetrate and fuse with the egg. Indeed, the addition of both naturally occurring and synthetic acrosin inhibitors to spermatozoa has been shown to prevent fertilization both in vitro and in vivo in the rabbit, rodent, and primate.[3]

Acrosin is specific to spermatozoa and makes an excellent target for the development of new, nonhormonal contraceptives (i.e., acrosin inhibitors). Several such inhibitors such as N^{α}-tosyl-L-lysine chloromethyl ketone (TLCK),[3] 4′-nitrophenyl 4-guanidinobenzoate (NPGB),[3] N-carbobenzoxy amino acid esters,[4,5] and sterol sulfates[6-8]

* Correspondence should be addressed to: L. J. D. Zaneveld, D.V.M., Ph.D., Professor and Director, Ob/Gyn Research, Rush–Presbyterian–St. Luke's Medican Center, Chicago, Illinois 60612.

[†] Department of Obstetrics and Gynecology, Rush University.

[‡] Department of Physiology, Rush University.

[§] Department of Biochemistry, Rush University.

[‖] Department of Medicinal Chemistry and Pharmacognosy, University of Illinois at Chicago.

[⊥] Department of Pharmacodynamics, University of Illinois at Chicago.

(1) Bhattacharyya, A. K.; Zaneveld, L. J. D. "Biochemistry of Mammalian Reproduction"; Wiley: New York, 1982; p 119.
(2) Rogers, B. J.; Bentwood, B. "Biochemistry of Mammalian Reproduction"; Wiley: New York, 1982; p 203.
(3) Zaneveld, L. J. D. "Human Semen and Fertility Regulation in Men"; C. V. Mosby: St. Louis, 1976; p 570.
(4) Hall, I. M.; Drew, J. H.; Sajadi, Z.; Loeffler, L. J. *J. Pharm. Sci.* **1979**, *68*, 696.
(5) Drew, J. H.; Loeffler, L. J.; Hall, I. H. *J. Pharm. Sci.* **1981**, *70*, 60.
(6) Burck, P. J.; Zimmerman, R. E. *J. Reprod. Fertil.* **1980**, *58*, 121.
(7) Burck, P. J.; Thakkar, A. L.; Zimmerman, R. E. *J. Reprod. Fertil.* **1982**, *66*, 109.

0022-2623/86/1829-0514$01.50/0