Cluster Significance Analysis Contrasted with Three Other Quantitative Structure-Activity Relationship Methods

James W. McFarland* and Daniel J. Gans

Central Research Division, Pfizer Inc., Groton, Connecticut 06340. *Received April 25, 1986*

Cluster significance analysis (CSA), a new statistical method to analyze structure–activity relationships in graphically displayed data, is contrasted with linear discriminant analysis, SIMCA, and the method of "relative odds". The data sets evaluated are as follows: antibacterial lasalocid derivatives, antimalarial naphthoquinones, and carcinogenic polycyclic aromatic hydrocarbons. CSA gives results comparable to these other methods, involves fewer assumptions, can be more reliable, and in general is easier to understand.

When seeking structure–activity relationships among congeneric compounds chemists will often plot the biological data against the physical properties. A common case is that in which the compounds fall into only two response classes: active and inactive, for example. It is then possible to use physical parameters as the coordinates of a graph on which the compounds are marked distinctively by class, such as triangles for actives and circles for inactives as in Figure 3. A noteworthy event occurs when the members of one class, e.g., the active compound group, tend to cluster in a relatively confined region of the graph.

Such clustering suggests that there is a connection between those particular physical properties and the biological activity of the compounds. Magee[1] has identified this method as a valid means to determine those parameters that are predictors of activity and has dubbed the procedure "parameter focusing". Knowledge of this type can be important in discovering the optimum members of a new class of biologically active compounds.

Having perceived an apparent cluster of actives, one also needs to determine whether it might have arisen by chance alone. We recently introduced a new statistical technique, cluster significance analysis (CSA), to help answer this question. CSA operates by assessing the tightness of the cluster of actives and determining the probability that a cluster as tight or tighter might have arisen purely by chance. Among its advantages is that it is relatively free from statistical assumptions.[2]

Our earlier paper presented the method in detail. Here, data from literature studies will be used to contrast CSA with three other analytical techniques.

Contrast with Linear Discriminant Analysis (LDA): Antibacterial Lasalocid Derivatives. In this first example we will apply CSA initially to a traditional approach to structure–activity analysis: observing features that seem to characterize the most active members of a series and concluding that those features are necessary for activity.

Westley et al.[3] prepared derivatives of the ionophore lasalocid and evaluated their antibacterial activities. To better understand the structure–activity relationships in this series, they also measured the $pK_{a'}$ values and partition coefficients (P's) of selected compounds. (However, in the ensuing calculations we will use log P.) The relevant data are presented in Table I. In commenting later on this group of compounds, Westley[4] considered separately the effects that $pK_{a'}$ and P had on antibacterial activity. He asserted "that any deviation in the acidity of the carboxyl group in lasalocid has a detrimental effect on the antibacterial activity of the antibiotic". This is a perfectly reasonable suggestion, but CSA now offers the possibility of testing it for statistical significance.

Table I. Apparent $pK_{a'}$, Logarithm of the Partition Coefficient (log P), and Relative in Vitro Antibacterial Activity of Lasalocid and Some of Its Derivatives[a]

compound	$pK_{a'}$	log P	in vitro act.[b]
lasalocid	4.4	2.83	1
bromolasalocid	3.9	3.25	1
chlorolasalocid	4.0	3.12	1
iodolasalocid	3.9	3.19	1
nitrolasalocid	2.4	2.46	0
aminolasalocid	6.0	0.28	0
N-(acetylamino)lasalocid	4.3	1.04	0
lasalocid acetate	4.15	0.95	0
bromolasalocid acetate	4.25	1.49	0
diazolasalocid	6.6[c]		0
benzylideneaminolasalocid		0.86	0
lasalocid methyl ether		2.54	0
lasalocid pentanoate		2.08	0
lasalocid octanoate		2.21	0
lasalocid decanoate		2.75	0
lasalocid bromobenzoate		3.48	0

[a] Adapted from data presented in the literature (see ref 3). Compounds for which there is neither $pK_{a'}$ nor log P data are omitted. [b] Activity = 1 where antibacterial potency is greater than 50% of that of lasalocid and 0 where it is less. There is a reasonably large gap between the potencies of the compounds so distinguished. [c] This $pK_{a'}$ value is taken from ref 4.

Figure 1 displays in one dimension the relationship between antibacterial activity and $pK_{a'}$. The more active members are clustered in the middle, but there are also some less active ones in this region as well. Therefore, the same degree of acidity as lasalocid does not guarantee good activity. Application of CSA to this set of data shows that such a clustering would have arisen under pure chance with a fairly high probability ($p = 0.143$). Therefore, a connection between $pK_{a'}$ and activity is not confirmed by the present data. To be sure it also has not been eliminated, but a proposal to prepare derivatives that are substantially more or less acidic than lasalocid should not be discouraged.

In assessing the influence of P on activity, Westley[4] concluded that there was at least some relationship: that those derivatives with P approximately the same or up to twice as large as that of lasalocid were likely to be the more active ones. The data for this relationship are shown in Figure 2. Here, the application of CSA supports his conjecture ($p = 0.002$).

We should also consider the possibility that there is a joint influence of $pK_{a'}$ and log P upon activity. In this case, however, there are measurements of log P and $pK_{a'}$ common to only nine of the compounds. The graphical representation of these data in two dimensions is given in

(1) Magee, P. S. In *IUPAC Pesticide Chemistry: Human Welfare and the Environment*; Miyamoto, J., Kearney, P. C., Eds., Pergamon: Oxford, 1983; p 251.
(2) McFarland, J. W.; Gans, D. J. *J. Med. Chem.* **1986**, *29*, 505.
(3) Westley, J. W.; Oliveto, E. P.; Berger, J.; Evans, R. H., Jr.; Glass, R.; Stempel, A.; Voldemar, T.; Williams, T. *J. Med. Chem.* **1973**, *16*, 397.
(4) Westley, J. W. In *Polyether Antibiotics*; Westley, J. W., Ed.; Marcel Dekker: New York, 1983; Vol. 2, p 65.

0022-2623/87/1830-0046$01.50/0

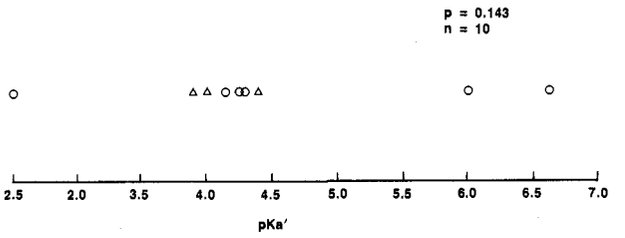

Figure 1. Lasalocid derivatives: antibacterial activity as a function of $pK_{a'}$; active compounds (Δ) and inactives (O). The triangle at $pK_{a'}$ = 3.9 represents two data points.

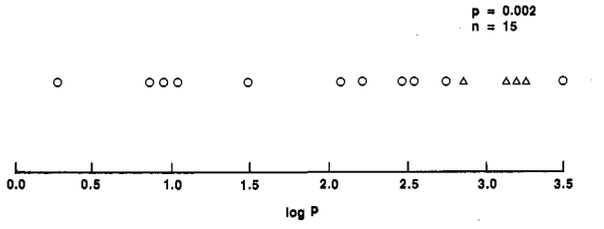

Figure 2. Lasalocid derivatives: antibacterial activity as a function of log P; active compounds (Δ) and inactives (O).

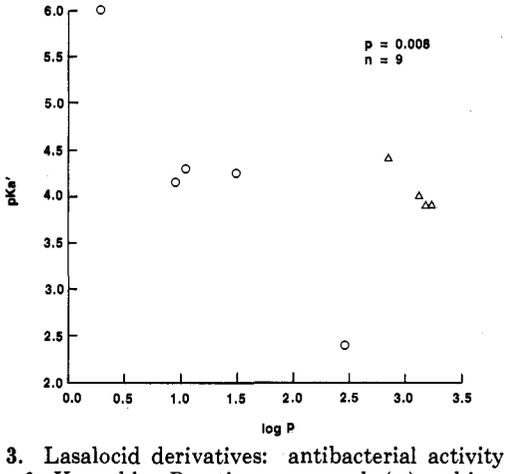

Figure 3. Lasalocid derivatives: antibacterial activity as a function of $pK_{a'}$ and log P; active compounds (Δ) and inactives (O).

Figure 3. The indicated probability (p = 0.008) would be considered significant normally, but in this case we note that this is a *higher* probability than that given by log P separately; i.e., two parameters operating together give a weaker result than one alone. In our previous work[2] we suggested that in situations of this kind, as a rough rule of thumb, where one parameter alone (but not the other) results in a lower probability than the two together, the parameter that results in the lower probability could be considered as contributory and the other possibly spurious. Only those parameters that together result in a probability lower than either one separately might be said to have a significant joint effect on activity. This guiding rule can extend to higher numbers of interacting parameters.

It is apparent from Figure 3 that linear discriminant analysis can be applied to separate the actives from the inactives. Significant two-dimensional separation of the two groups is indicated by the LDA-related Wilks' Λ test (p = 0.000 04). When all of the available $pK_{a'}$ data are analyzed in one dimension (see Figure 1), one fails to find significance (p = 0.483). On the other hand, when all of the available log P data are analyzed in one dimension, Wilks' test shows that this physical parameter is important (p = 0.026). The much greater strength of the two-dimension separation argues for $pK_{a'}$ adding to the discriminatory effect of log P alone, i.e., a joint effect.[5]

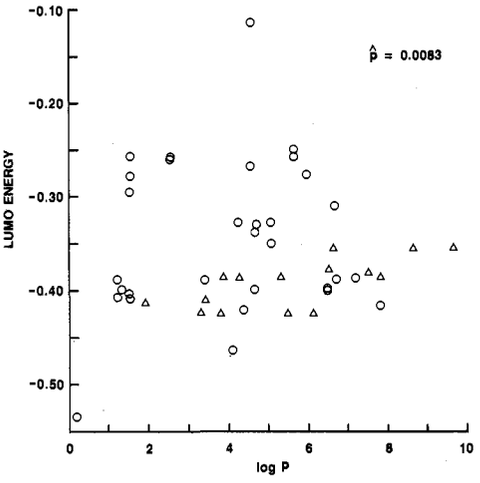

Figure 4. Quinones: antimalarial activity as a function of the energy of the lowest unoccupied molecular orbital (LUMO) and log P; active compounds (Δ) and inactives (O). Reprinted with permission from Dunn, W. J.; Wold, S. *J. Med. Chem.* **1980**, *23*, 595.

CSA and LDA agree that log P plays a role in determining antibacterial activity, but they diverge on whether jointly $pK_{a'}$ and log P do. Much of this difficulty may be due to the missing data in Table I. The fact that the most lipophilic derivative, lasalocid bromobenzoate, is not represented in Figure 3 enhances the importance of $pK_{a'}$ in the LDA approach. Further, the wide scatter of the inactive derivatives in contrast to the narrow range of the active group suggests that one of the required conditions for LDA may not be met: that the parameter vectors in the two classes follow multivariate distributions with the same covariance matrix.

We can come no further in our analysis. What is needed is more information. A good beginning would be to fill in the omitted data from Table I. Our purpose was to demonstrate that CSA is useful in appraising structure–activity relationships and compares favorably with LDA. In this example CSA has given credence to the importance of log P and reasonable direction for further work among these ionophores. Such work may resolve the disagreement about $pK_{a'}$, but better yet it may lead to the discovery of valuable therapeutic agents.

Contrast with SIMCA: Antimalarial Naphthoquinones. The SIMCA technique has advanced our abilities to deal with a particularly troublesome type of display that Wold and his co-workers[6-8] have termed the "asymmetric case". In this situation a well-delimited cluster of "active" compounds is embedded among a set of "inactives" that are scattered in various directions. Here it is difficult, if not impossible, to separate the two groups by a linear mathematical function such as would be em-

(5) Still within the framework of the assumptions involved in LDA, it is possible directly to judge, for the nine compounds with data on both parameters, the contribution to separating power made by $pK_{a'}$ when added to log P. This can be done by using an analysis-of-covariance approach (see W. J. Dixon et al., *BMDP Statistical Software 1985*; University of California Press: Berkeley, CA, p 520). This analysis shows a significant extra contribution (p = 0.0007) for $pK_{a'}$ beyond log P alone. CSA does not provide such a direct means for assessing the importance of one parameter when added to one or more others: hence, the use of the rule of thumb given above.

(6) Norden, B.; Eland, U.; Wold, S. *Acta Chem. Scand., Ser. B* **1979**, *32*, 1.
(7) Dunn, W. J., III; Wold, S. *J. Med. Chem.* **1978**, *21*, 1001.
(8) Dunn, W. J., III; Wold, S. *J. Med. Chem.* **1980**, *23*, 595.

ployed in LDA[9] or in certain other pattern recognition techniques.[10]

Dunn and Wold[8] have used SIMCA to characterize a structurally diverse set of 48 antimalarial naphthoquinones such that the activity class of any new naphthoquinone could be predicted by knowledge of its partition coefficient (log P) and the energy of its lowest unoccupied molecular orbital (LUMO). In this work, however, there appears to be inadequate consideration given to whether the pattern of observations could have arisen by chance alone.[11]

Dunn and Wold's data are displayed graphically in Figure 4. CSA shows that the clustering of the active compounds in the log P and LUMO space has an associated significance probability $p = 0.0083 \pm 0.0008$.[12] Hence, by the usual standards this would be called a "highly significant" relationship and fortuitous clustering would be considered unlikely.

However, further analysis reveals that only the LUMO energy makes a contribution. When visualized only along the axis of the LUMO energy, the active compounds are concentrated in the range −0.433 to −0.353. CSA shows that the probability of clustering as close as this is 0.0004 ± 0.0002 under chance alone, or some 20 times less likely than the aggregation of the actives in Figure 4. Visualizing in the dimension of log P, we find that the actives are scattered over a high proportion of the range. The significance probability associated with this weaker clustering is quite high: 0.422 ± 0.004.

We conclude, therefore, that the favorable association indicated in Figure 4 is borne almost entirely by the LUMO energy parameter and that log P adds nothing to the concentrating power of LUMO. In an earlier work Martin et al.[13] showed that among simple alkyl-substituted naphthoquinones there is a parabolic relationship between antimalarial potency and log P. Evidently, in the larger, structurally more complex set this effect is obscured by more dominant factors. Whether these factors include more than just the LUMO energy would have to be determined by measurements or estimates of additional physical properties and subsequent reanalysis.

SIMCA and CSA are not actually competing methods. SIMCA is a classification technique that seeks to define the region in parameter space wherein the active compounds lie. It does not appear, however, to furnish direct quantitative guidance in answering the prior question whether the active group could be an accidental association among all the compounds considered. It is not clear at present whether the technique cannot provide this or whether this capability simply has not been brought forward. CSA does address this issue, but leaves the region of activity undefined. With it one can comfortably predict the activity of a new congener when it is clearly inside or outside of the range of active compounds, but there is considerable uncertainty when a compound lies in the transition region. CSA is really a pattern verification rather than a classification technique. Thus, SIMCA and CSA complement one another and could be used in conjunction.

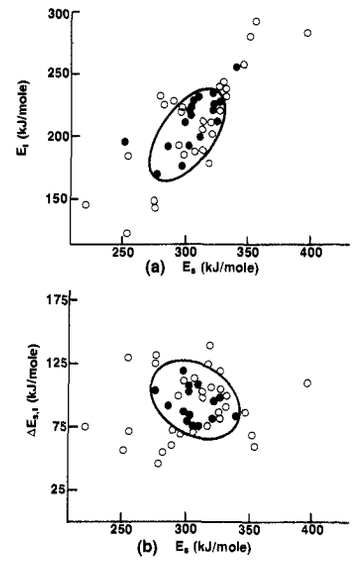

Figure 5. Polycyclic aromatic hydrocarbons: carcinogenicity as a function of the energies of electronically excited states E_s (singlet), E_t (triplet), and $\Delta E_{s,t}$ (difference between E_s and E_t); active compounds (●) and inactives (○). (a) Carcinogenicity as a function of E_s and E_t. (b) Carcinogenicity as a function of E_s and $\Delta E_{s,t}$. Reprinted with permission from Morgan, D. D.; Warshawsky, D.; Atkinson, T. *Photochem. Photobiol.* **1977**, *25*, 31. Copyright 1977 Pergamon Press.

Contrast with the Method of "Relative Odds": Carcinogenicity of Polycyclic Aromatic Hydrocarbons (PAHs). Morgan et al.[14] determined the energies of the lowest excited singlet (E_s) and triplet (E_t) states of 49 polycyclic aromatic hydrocarbons. Eighteen of these compounds are carcinogenic. These workers sought to justify statistically the hypothesis that one or both of these electronic excited states could be involved in the chemical events leading to carcinogenesis. Their plan was to correlate the energy values of the excited states with carcinogenic activity. This they did by dichotomizing the parameter space and comparing the proportions of carcinogens in the two regions.

In one dimension, either a single value was chosen to divide the parameter axis into regions above and below, or an interval on this axis was selected with the two regions constituting its interior and exterior. Correlation was measured by the "relative odds" of finding a carcinogen in one as against the other region, and its significance was assessed with Fisher's exact test. The authors interpreted their results to mean that E_s values are strongly correlated with carcinogenicity, but E_t and $\Delta E_{s,t}$ values are not. The "relative odds" for a compound falling in a "carcinogenic" region on a scale defined by E_s are 4.8 ($p = 0.015$) by the single value approach and 22.8 ($p = 0.000\,06$) by the interval approach.

They also plotted the compounds against combinations of these energy parameters two at a time. The carcinogenic compounds tended to cluster in the middle of the graphs of two of these combinations (see Figure 5). In two dimensions elliptical boundaries were used to dichotomize the parameter space. Again correlation was measured by the "relative odds", and Fisher's exact test was used to determine significance. In Figure 5a the coordinates are E_s and E_t while in Figure 5b they are E_s and $\Delta E_{s,t}$, the

(9) Martin, Y. C.; Holland, J. B.; Jarboe, C. H.; Plotnikoff, N. *J. Med. Chem.* **1974**, *17*, 409.
(10) Stuper, A. J.; Jurs, P. C. *J. Am. Chem. Soc.* **1975**, *97*, 182.
(11) SIMCA is actually a collection of related techniques following a hierarchy. The version used here, termed by its originators "level 2", is the one closest to CSA in that it deals with discrete classification.
(12) The CSA probabilities in this example and the next were estimated by the random sampling procedure described in ref 2. The values presented are the estimates $\hat{p}$ with 95% confidence limits.
(13) Martin, Y. C.; Bustard, T. M.; Lynn, K. R. *J. Med. Chem.* **1973**, *16*, 1089.
(14) Morgan, D. D.; Warshawsky, D.; Atkinson, T. *Photochem. Photobiol.* **1977**, *25*, 31.

Table II. Correlations among the Parameters Used To Analyze the Carcinogenicity Data of Certain Polycyclic Aromatic Hydrocarbons[a]

	E_s	E_t	$\Delta E_{s,t}$
E_s	1.000	0.782	0.097
E_t		1.000	−0.544
$\Delta E_{s,t}$			1.000

[a] Numeric values are the correlation coefficients (r) between the parameter pairs.

energy difference between the two excited states. The "relative odds" for "carcinogenic" compounds being inside the ellipse of Figure 5a are 9.71 (p = 0.0024), while those for Figure 5b are 20.6 (p = 0.0004).

Morgan and co-workers concluded that in some way the carcinogenicity of PAHs is related to some property or properties of the lowest excited singlet state. The major difficulty with this analysis is that in all cases the dichotomizing boundary was chosen from the data specifically to maximize the "relative odds", i.e., maximize the separation between the proportions of actives. Both the relative odds method and Fisher's test assume implicitly that a dichotomizing boundary is chosen a priori, or at least is not chosen in data-dependent fashion specifically to maximize the separation. Thus, the method Morgan et al. used will tend to overstate the relative odds and the significance emerging from Fisher's test.

CSA does not suffer from this problem because dichotomizing boundaries are not used. We analyze this same problem as follows.

To start, we observe in Figure 5a that the parameters making up the coordinates are strongly correlated (r = 0.78, see Table II) and are therefore largely a measure of the same thing. CSA on E_s and E_t jointly, or on E_t separately, may thus be expected to, and in fact does, largely reproduce the results on E_s alone. Either E_s or E_t might therefore be deleted from further analysis. $\Delta E_{s,t}$ correlates least with E_s, and thus we prefer to drop E_t. We will be concerned only with E_s and $\Delta E_{s,t}$ henceforth.

CSA shows that the differentiation between carcinogens and noncarcinogens in Figure 5b is highly significant (p = 0.010 ± 0.001). A single variable analysis of the data now shows that the p value associated with the clustering of actives along the dimension of E_s is 0.055 ± 0.002, a value short of significance at the 0.05 level but close enough to hold our interest. The corresponding probability for the arrangement of actives along the $\Delta E_{s,t}$ scale is 0.024 ± 0.001.

We conclude that there is reason to believe that among PAHs there is a significant relationship between carcinogenicity and the parameters E_s and $\Delta E_{s,t}$. Hence, the original conclusion of Morgan et al. is supported by CSA. While the inference derived is not different from that obtained by these earlier workers, CSA gives more trustworthy significance probabilities because it does not suffer from selection bias.

In summary, we have shown that CSA is useful in analyzing structure–activity data in which there are only two classes of biological responses. (In cases where there are more than two, the method can still be used if there is a natural and not outcome-driven way to combine the different classes into two.) The method has been contrasted with other approaches and has been shown to give similar results. In some cases it is able to make distinctions that the other methods cannot, or at least it can make assessments more reliably. The conceptual simplicity of CSA, its comparatively assumption-free nature, and its reliability bode well for its further application to drug design problems.

2,4-Diamino-6,7-dimethoxyquinazolines. 1. 2-[4-(1,4-Benzodioxan-2-ylcarbonyl)piperazin-1-yl] Derivatives as α_1-Adrenoceptor Antagonists and Antihypertensive Agents

Simon F. Campbell,* Michael J. Davey, J. David Hardstone, Brian N. Lewis (in part), and Michael J. Palmer

Departments of Discovery Chemistry and Discovery Biology, Pfizer Central Research, Sandwich, Kent, United Kingdom.
Received March 24, 1986

A series of 4-amino-2-[4-(1,4-benzodioxan-2-ylcarbonyl)piperazin-1-yl]-6,7-dimethoxyquinazoline derivatives was synthesized for evaluation as α-antagonists and antihypertensive agents. Most compounds displayed high (nM) binding affinity for α_1-adrenoceptors with no significant activity at α_2-sites. Selective antagonism of the α_1-mediated vasoconstrictor effects of norepinephrine is also characteristic of the series. Structure–activity relationships for α_1-adrenoceptor affinity are presented, and structural similarity between the 2,4-diamino-6,7-dimethoxyquinazoline nucleus and norepinephrine is established. An α_1-receptor model is presented in which charge-reinforced hydrogen bonding is important for binding of both antagonist and agonist molecules. Antihypertensive activity was evaluated after oral administration (5 mg/kg) to spontaneously hypertensive rats, and several compounds displayed similar efficacy to prazosin when assessed after 6 h. On the basis of α_1-adrenoceptor affinity/selectivity in vitro and duration of antihypertensive action in vivo, compound 1 (doxazosin) was selected for further evaluation and is currently progressing through phase III clinical trials.

Hypertension affects up to one-fifth of the adult population and is an important risk factor for various cardiovascular disorders.[1] Treatment of patients with marked blood pressure elevation has been routine clinical practice for some considerable time, and over the last few years, drug therapy for mild/moderate hypertension has also become more common.[2] Consequently, clinical interest in improved, novel antihypertensive agents has intensified.[3] Satisfactory blood pressure control invariably requires chronic therapy, and drug side effects must be minimal, particularly since most patients are asymptomatic. Furthermore, it is now realized that the major causes of

(1) For a review on coronary heart disease, see: *Am. J. Med.* **1984**, *76* (2A).

(2) For the Final Report of the Working Group on Risk and High Blood Pressure: *Hypertension* **1985**, *7*, 641.

(3) Graham, R. M.; Campbell, W. B. *Fed. Proc., Fed. Am. Soc. Exp. Biol.* **1981**, *40*, 2291.

0022-2623/87/1830-0049$01.50/0