sponding values of D_s and V_s, are summarized in Table IV, and Table V contains the means and standard deviations of these quantities for the four methods applied. The selection of compounds from clusters (method II) yields somewhat better results in this than in the first example. The reason is simply that the number of objects per cluster is much smaller here so that the influence of chance is decreased.

Summarizing the results it can be concluded that the PCMM method effectively minimizes collinearities and still yields test series of sufficiently high data variance. Furthermore, these test series are truly respresentative for the parameter space considered so that all conditions for an optimal test series are fulfilled. With respect to the minimization of collinearities, this method is clearly superior to the methods introduced by Hansch and coworkers[3] and by Wootton et al.[4] For practical use, we recommend the selection of not only one but a whole set of possible test series, varying the starting points and the adjustable parameters of the procedure. This can easily be done, since the calculations are very rapid and do not need more than 20 s of computer time, on an average, for each run. As a result, a very clear picture on the optimality of test series obtainable from a given starting population emerges, and the particular properties of the population considered can be fully accounted for. In this way, the best possible test series can be found,[22] and sufficient freedom is left for a final decision by the synthetic chemist so that due regard can be paid to synthetic feasibility. Although only monosubstitutions have been considered in the present paper, the method can easily be extended to multiple substitutions in the same way as outlined in ref 4.

(20) F. E. Norrington, R. M. Hyde, S. G. Williams, and R. Wootton, *J. Med. Chem.*, 18, 604 (1975).

(21) The hyperplane can be found by spectral decomposition of the covariance matrix,[16] and the Euclidean distance $P(E,y)$ of a point, $y = (y_1, y_2, \cdots, y_m)$, from the hyperplane E is

$$P(E,y) = [\sum_{i=1}^{m}(y_i - \bar{x}_i)^2 - \sum_{k=1}^{m-1}(\sum_{i=1}^{m}(y_i - \bar{x}_i)\gamma_{ik})^2]^{1/2}$$

γ_{ik} is the ith element of the kth eigenvector of the covariance matrix C, calculated from the sample of remaining compounds after the previous iteration step, and $\bar{x}_i$ is the corresponding mean value of the ith variable. Consequently, in the first iteration step, C is the correlation matrix of the original sample, and $\bar{x}_i = 0$ ($i = 1, \cdots, m$).

(22) It must be pointed out that the parameter space considered in this paper is not the only possibility; different problems may require quite different variables (see part 2 of this series: S. Dove W. J. Streich, and R. Franke, *J. Med. Chem.*, following paper in this issue).

On the Rational Selection of Test Series. 2. Two-Dimensional Mapping of Intraclass Correlation Matrices

S. Dove, W. J. Streich, and R. Franke*

Academy of Sciences of the German Democratic Republic, Research Center for Molecular Biology and Medicine, Institute of Drug Research, 1136 Berlin, German Democratic Republic. *Received July 19, 1979*

A rational design of optimal test series can be performed by two-dimensional mapping of intraclass correlation matrices (TMIC method). The method results in a two-dimensional map from which substituents can be selected by simple inspection. Different test series can be obtained from the same map so that synthetic feasibility can easily be taken into account. The approach closely corresponds to the usual way of thinking of organic chemists, and the test series evaluated for an example show high data variance and low collinearities.

As already pointed out,[1] the selection of optimal test series with high information content is essential for a rational drug design. Such test series have to systematically explore a defined physical chemical parameter space important for biological activity in such a way that the variance of all parameters becomes sufficiently high and that collinearities between these parameters do not occur. Several series selection methods have been proposed in the literature (see ref 1). Some of these methods have the disadvantage that they appear to the synthetic chemist more or less as a black box; this is also true for the PCMM method outlined in the first part of this series.

Even if several runs are made with PCMM so that more than one test series and, thus, a fairly complete picture is obtained, the synthetic chemist may still feel unhappy with the thought that something even better may be hidden in this black box and that not enough room is left for this chemical intuition. We want to present, therefore, a completely different approach based on the spectral decomposition of intraclass correlation matrices (TMIC method) which allows one to present results in a very simple and instructive way in the form of a two-dimensional map. The test series are selected from the map by simple inspection in a way which is very close to the usual thinking of the organic chemist. The same map can be used to design different test series so that synthetic feasibility can easily be taken into account, and the resulting series show sufficient data variance and no collinearities of parameters.

Method

Correlations between the elements of groups of individuals with respect to a certain feature can be characterized by the intraclass correlation coefficient. If substituents or chemical compounds are treated as elements and the set of molecular parameters (x_i) characterizing their physical chemical properties is treated as individuals, the relatedness of two substituents, 1 and 2, can be expressed by the intraclass correlation coefficient, provided that the parameters are standardized to a mean of zero and a standard deviation of unity according to eq 1.

$$z_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i} \quad (1)$$

$\bar{x}_i$ and s_i are the mean and the standard deviation of the

(1) W. J. Streich, S. Dove, and R. Franke, *J. Med. Chem.*, preceding paper in this issue.

0022-2623/80/1823-1456$01.00/0

ith molecular parameter (variable; $i = 1, ..., m$), respectively, and x_{ij} is the value of this parameter for the jth compound or substituent (object; $j = 1, ..., N$). The intraclass correlation coefficient, $r_{I(1,2)}$, for two substituents, 1 and 2, with respect to m variables is defined as

$$r_{I(1,2)} = 2\sum_{i=1}^{m}(z_{i1} - \bar{z})(z_{i2} - \bar{z}) / \sum_{i=1}^{m}[(z_{i1} - \bar{z})^2 + (z_{i2} - \bar{z})^2] \quad (2)$$

with

$$\bar{z} = \frac{1}{2m}\sum_{i=1}^{m}\sum_{j=1}^{2} z_{ij} \quad (3)$$

and

$$-1 \leq r_I \leq 1 \quad (4)$$

The intraclass correlation coefficient is related to the variance between the two sets of data for the pair of compounds considered, s_B, and to the variance within the data pairs for both compounds, s_I, according to

$$r_{I(1,2)} = \frac{s_B^2 - s_I^2}{s_B^2 + s_I^2} \quad (5)$$

where

$$s_B^2 = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{z_{i1} + z_{i2}}{2} - \bar{z}\right)^2 \quad (6)$$

and

$$s_I^2 = \frac{1}{4m}\sum_{i=1}^{m}(z_{i1} - z_{i2})^2 \quad (7)$$

If two objects are identical with respect to all variables, s_B^2 becomes zero and $r_{I(1,2)}$ equal to +1. Clearly, the presence of such pairs is unfavorable for an optimal test series. This is, usually, also true if $r_{I(1,2)}$ becomes −1. If there is no relationship between the parameter values of two compounds, 1 and 2, s_I^2 will be equal to s_B^2, and $r_{I(1,2)}$ becomes zero.

Obviously, a good test series with low collinearities is characterized by low absolute values of the intraclass correlation coefficients for all possible pairs of compounds. As a rule, the data variance will also be high in such a series, since s_I^2 is directly related to the Euclidean distance d in the parameter space considered according to eq 8.

$$s_I^2 = d^2/4m \quad (8)$$

A good test series with high variance of all variables and low multiple collinearities can be obtained, therefore, if compounds or substituents are selected in such a way that the nondiagonal elements of the intraclass correlation matrix are close to zero. This could be done, in principle, by a trial and error procedure, but such a procedure would be much too laborous and inefficient for practical use. If the starting population of compounds from which the test series is to be selected is not too large (less than about 50 compounds), two-dimensional spectral mapping can advantageously be used here.

The intraclass correlation matrix is decomposed by the standard procedure of the principal component method,[9] and all objects to be considered are represented as points in the two-dimensional Cartesian space spanned by the first two components. Use is made here of the fact that these two components alone will adequately represent the largest part of information contained in the intraclass correlation matrix in many cases. This supposition can be checked by the corresponding eigenvalues λ_1 and λ_2 using the criterion of eq 9 (N = number of compounds in the starting population).

$$XI = (\lambda_1 + \lambda_2)/N \quad (9)$$

Equation 9 expresses the fraction of information, XI, accounted for by the first two components, which should be about 70% or greater in order to accept these components as representative for the whole data set. As a result of the whole procedure, a simple two-dimensional map is obtained in which compounds similar with respect to the molecular parameters considered (compounds with high positive values of $r_{I(1,2)}$ between them) are clustered together; the position of compounds with high negative values of the intraclass correlation coefficient is symmetrical with respect to the origin of the coordinate system. Geometrically, the intraclass correlation coefficient $r_{I(1,2)}$ approximately corresponds to the cosine of the angle between compounds 1 and 2. A good test series with high data variance and low collinearities will always be obtained if substituents distant from each other are selected in such a way that the whole space is systematically covered and, at the same time, the inclusion of points which can be reflected at the origin is avoided; this can simply be done by inspection of the map by eye. Since it is possible to obtain different test series from the same map, synthetic feasibility can always adequately be taken into account.[11]

(2) R. Wootton, R. Cranfield, G. C. Sheppey, and P. J. Goodford, *J. Med. Chem.*, **18**, 607 (1975).
(3) F. E. Norrington, R. M. Hyde, S. G. Williams, and R. Wootton, *J. Med. Chem.*, **18**, 604 (1975).
(4) C. Hansch, S. H. Unger, and A. B. Forsythe, *J. Med. Chem.*, **16**, 1217 (1973).
(5) A somewhat different approach can, in principle, be used here or in such cases where the criterion[9] is not fulfilled. Instead of considering only the first two components, all components up to a sum of eigenvalues greater than 95% of the sum of all positive and negative eigenvalues are included. The components are orthogonally rotated (Varimax rotation) in such a way that a data structure close to Thurstone's simple structure is obtained. This means that similar compounds which would appear in the same cluster in the two-dimensional map have high coefficients ("loadings") with respect to the same (rotated) components, whereas unconnected compounds have high loadings only in different components. A good test series can then be obtained by selecting one or two highly loaded compounds from each component.
(6) C. Hansch, *Pharmacochem. Libr.*, **2**, 47 (1977).
(7) E. J. Lien, in "Drug Design", Vol. V, E. J. Ariens, Ed., Academic Press, New York, 1975, p 81.
(8) P. J. Goodford, A. T. Hudson, G. C. Sheppey, R. Wootton, M. H. Blank, G. J. Sutherland, and J. C. Wickham, *J. Med. Chem.*, **19**, 1239 (1976).
(9) H. H. Harman, "Modern Factor Analysis", 2nd ed, University of Chicago Press, Chicago, 1967.
(10) Y. C. Martin and H. N. Panas, *J. Med. Chem.*, **22**, 784 (1979).
(11) A similar result can, in principle, be obtained from a spectral decomposition of a product-moment correlation matrix and a presentation of the objects in the space of the first two components. Since, however, the distances of objects in this space are closely related to Euclidean distances, this approach would simply result in a graphical version of the multidimensional mapping technique of Wootton et al.[2] This is the reason why we used the intraclass correlation coefficient as a basis for our analysis, which is more acceptable for comparing objects. It was our hope that this correlation coefficient would lead to test series with lower collinearities. Although we cannot offer a straightforward theoretical proof for this assumption, the results presented in Table II are, at least, not contradictory to such a view.

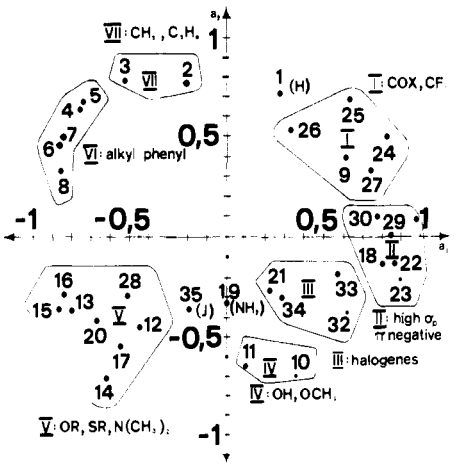

Figure 1. Two-dimensional map of substituents in the space spanned by the first two components obtained from spectral decomposition of the intraclass correlation matrix. The numbers refer to Table I.

This technique, which we will call the TMIC method, was applied to the same example as treated by Wootton et al.[2] Using the parameter space spanned by π, F, R, and MR, ten test series comprising ten compounds each were selected from the starting population of 35 compounds (for the data, see ref 3). The results are compared with other selection methods using the variance coefficient V_s and the determinant D_s of the correlation matrix as measures for data variances and collinearities, respectively (see ref 1).

Results and Discussion

The eigenvalues of the first two components are just high enough to fulfill the condition for XI with

$$\lambda_1 = 15.011 \qquad \lambda_2 = 7.710 \tag{10}$$

and

$$XI = 65\% \tag{11}$$

so that two-dimensional mapping can safely be applied here. The map resulting from the presentation of the 35 compounds in the space spanned by the first two components is presented in Figure 1. A distinct clustering of similar compounds becomes evident, which is chemically reasonable. The substituents in the clusters are as follows: cluster I, substituents with carbonyl functions and CF_3; cluster II, strongly electron-withdrawing substituents with negative π values and of similar size; cluster III, halogens (except iodine); cluster IV, hydroxy and methoxy groups; cluster V, alkoxy groups (except OCH_3) and SCH_3; cluster VI, phenyl and alkyl groups greater ethyl; cluster VII, methyl and ethyl groups. Hydrogen, iodine, and the amino group appear as single points. The ten test series selected from Figure 1, together with the corresponding values of V_s and D_s, are summarized in Table I; mean values and standard deviations of V_s and D_s for different selection methods are compared in Table II. Table II clearly shows that TMIC indeed yields test series of sufficient quality with high data variance and low collinearities. The value of $\bar{D}_s$ is of about the same magnitude as with PCMM and, thus, somewhat better than $\bar{D}_s$ from the Wootton approach or from cluster analysis and distinctly higher than for randomly selected test series. The variance coefficient obtained with TMIC is somewhat smaller than the corresponding values from multidimensional mapping or from PCMM. The differences are, however, only marginal. Test series selected by TMIC can, on an average, safely be considered as representative for the parameter space as covered by the starting population.

Table I. Test Series of Ten Substituents Selected by the TMIC Method (Figure 1) from the 35 Substituents Treated by Wootton et al., Together with the Corresponding Values of D_s and V_s

test series no.	substituents selected[a]	D_s	V_s
1	1, 2, 8, 10, 14, 20, 24, 30, 33, 35	0.740	1.119
2	1, 3, 8, 10, 12, 19, 24, 25, 29, 33	0.457	1.005
3	1, 2, 5, 11, 14, 18, 24, 28, 34, 35	0.566	0.649
4	1, 3, 8, 10, 15, 19, 24, 26, 30, 33	0.530	1.162
5	1, 3, 7, 11, 16, 19, 21, 22, 27, 32	0.581	1.243
6	1, 3, 8, 11, 17, 19, 27, 28, 29, 32	0.524	1.299
7	1, 3, 7, 10, 17, 19, 25, 28, 31, 33	0.332	1.186
8	1, 2, 5, 8, 11, 12, 18, 27, 34, 35	0.589	0.785
9	1, 2, 4, 8, 10, 20, 21, 25, 29, 32	0.408	1.106
10	1, 2, 6, 10, 15, 21, 26, 31, 32, 35	0.492	0.987

[a] 1 = H, 2 = Me, 3 = Et, 4 = *n*-Pr, 5 = *i*-Pr, 6 = *n*-Bu, 7 = *t*-Bu, 8 = Ph, 9 = CF_3, 10 = OH, 11 = OMe, 12 = OEt, 13 = *O*-*n*-Pr, 14 = *O*-*i*-Pr, 15 = *O*-*n*-Bu, 16 = *O*-*n*-Am, 17 = OPh, 18 = OAc, 19 = NH_2, 20 = NMe_2, 21 = NHAc, 22 = NO_2, 23 = CHO, 24 = Ac, 25 = COOMe, 26 = COOEt, 27 = $CONH_2$, 28 = SMe, 29 = SO_2Me, 30 = SO_2NH_2, 31 = CN, 32 = F, 33 = Cl, 34 = Br, 35 = I.

Table II. Mean Values and Standard Deviations of D_s and V_s for Different Selection Methods for the Starting Population of 35 Substituents

selection method	$\bar{D}_s$	$s_{\bar{D}_s}$	$\bar{V}_s$	$s_{\bar{V}_s}$
starting population	0.57		1.00	
stochastic selection[a]	0.31	0.14	0.94	0.17
cluster analysis[a]	0.47	0.08	1.09	0.09
Wootton method[a]	0.40	0.14	1.18	0.08
PCMM method[a]	0.56	0.11	1.14	0.15
TMIC method	0.52	0.11	1.05	0.20

[a] From ref 1.

In summary, TMIC can be recommended as a simple and comprehensive method for the selection of optimal test series if the number of compounds in the starting population is not too large (smaller than 50). It is as good as PCMM in such cases, and it has the advantage of a very simple and straightforward representation of the results in line with classical chemical thinking which should be attractive to synthetic chemists. Furthermore, the two-dimensional map gives a full picture of the data structure in the starting population which is not obtainable from multidimensional mapping or PCMM and only partly represented by the Hansch clusters (the within-cluster position of the objects is now known), and no a priori decisions as in the other methods (except for the parameter space to be considered) are necessary to establish that map. Different test series can be designed from the same map so that synthetic feasibility can readily be taken into account.

With large starting populations the map resulting from TMIC becomes too crowded so that a selection of test series by simple inspection is no longer possible.[5] Since the simplicity of TMIC is lost, PCMM is the method of choice in such cases.

Conclusion

A rational design of test series is of the utmost importance, even if the evaluation of QSAR is not attempted. This has also been stressed by Hansch in a recent paper:[6] "Such a systematic procedure is central to the task of reducing costs in research by maximizing the information obtained from each molecular probe in a set of congeners. If QSAR accomplishes nothing more than to get chemists to take a more thoughtful and logical attitude in deriva-

tizing a lead compound, it will have made an important contribution". Examples from the literature clearly indicate that hundreds of compounds have practically been investigated just for nothing and without any gain of information because of poor series design.

In order to demonstrate this point for the present example, ten test series (each comprising ten substituents) were constructed from such substituents which are close together in Figure 1 and, hence, very similar with respect to the parameter space considered. In this way a situation is simulated where the synthesis of analogues is mechanically performed always along the same route using similar precursors. As expected, the result was very poor with $\bar{D}_s = 0.150$ and $\bar{V}_s = 0.731$. The variance is so low and collinearities are so high that these test series are completely unacceptable. In order to avoid such results, a rational selection of test series is indispensable.

Although a good series design can indeed save hundreds of syntheses, it must be kept in mind that all series design methods have one critical point: a guess has to be made on the parameter space to be considered.[10] Fortunately, the space spanned by π, σ, and MR will be sufficient in many cases. It is probably also not too serious if parameters are included which are, in fact, not important. If, however, just one property essential for a particular biological activity is not adequately represented, the result of the series design may be poor with respect to the variance of the biological response data. For parabolic relationships between biological activity and certain variables (e.g., π or log P), quadratic terms of these variables must be included into the parameter space. That means that a test series optimal in a general sense does not exist, and for each particular case a new design problem may arise. Already existing QSAR can be very helpful here.

In addition to the computer work, it is, of course, necessary to adequately consider all other information available. If, for instance, drugs acting on the CNS are to be investigated, the ideal log P of about 2 for the penetration of the blood–brain barrier[7] is a good starting point to vary lipophilicity.

If nothing is known, it may be useful to design, in a first step, a small preliminary series using the parameters mentioned above, from which a tentative QSAR is then evaluated. This QSAR can aid in the design of the final test series in two ways: (1) with the information obtained from it the parameter space can be modified or completed, if necessary; (2) the synthesis of inactive compounds can be avoided. Such a strategy has been applied, for instance, by the Welcome group[8] in a study on methoxychlor analogues.

Additions and Corrections

1980, Volume 23

Schneur Rachlin,* E. Bramm, I. Ahnfelt-Rønne, and E. Arrigoni-Martelli: Basic Antiinflammatory Compounds. N,N′,N″-Trisubstituted Guanidines.

Page 14. In Scheme I, 2-Ath↓I should read 4-Amq↓IX in the reaction CII → **53–84**; R^1-QNH_2↓III should read R^2-QNH_2↓III in the reaction CI → **19–26**; and CIII should read CIII[b].

Françoise Heymans, Laurence Le Thêrizien, Jean-Jacques Godfroid,* and Pierre Bessin: Quantitative Structure–Activity Relationships for *N*-[(*N*′,*N*′-Disubstituted-amino)acetyl]arylamines for Local Anesthetic Activity and Acute Toxicity.

Page 187. In Table III, the anesthetic doses (AD), mM/L, should read: for compound **3**, 15.0; **4**, 35.4; **5**, 27.9; **6**, 55.8 (instead of, respectively, 34.5, 27.9, 55.8, and 20.9).

Josef Fried,* D. K. Mitra, M. Nagarajan, and M. M. Mehrotra: 10,10-Difluoro-13-dehydroprostacyclin: A Chemically and Metabolically Stabilized Potent Prostacyclin.

Page 235. In line 20 of column 2 and reference 15, tri-*sec*-butylaluminum hydride should read tri-*sec*-butylborohydride (K Selectride).

Barbara S. Rauckman and Barbara Roth*: 2,4-Diamino-5-benzylpyrimidines and Analogues as Antibacterial Agents. 3. C-Benzylation of Aminopyridines with Phenolic Mannich Bases. Synthesis of 1- and 3-Deaza Analogues of Trimethoprim.

Page 387. In Table II, the column heading which reads $I_{50} \times 10^6$ M should actually read $I_{50} \times 10^8$ M. Also, under this column heading all ~ signs should be replaced with @ (i.e., 11% @ 42000, etc.).

Masayoshi Murata, Prakash Bhuta, James Owens, and Jiři Žemlička*: Inhibition of Ribosomal Peptidyltransferase with 2′(3′)-*O*-Acetyl-2″(3″)-*O*-glycyl-1,2-di(adenosin-N^6-yl)ethane and -1,4-di(adenosin-N^6-yl)butane. Effect of Alkyl Chain Length.

Page 781. In line 3 of the abstract, the word "pyrazoline" should be changed to "imidazoline".

Page 784. In Figure 3, all concentrations $\times 10^{-3}$ M inside the graph should be corrected to $\times 10^{-4}$ M.

Yasunobu Sato,* Yasuo Shimoji, Hiroshi Fujita, Hiroshi Nishino, Hiroshi Mizuno, Shinsaku Kobayashi, and Seiji Kumakura: Studies on Cardiovascular Agents. 6. Synthesis and Coronary Vasodilating and Antihypertensive Activities of 1,2,4-Triazolo[1,5-*a*]pyrimidines Fused to Heterocyclic Systems.

Page 932. In Table VI, the R^5 group for compound **87** should be $(CH_2)_3$-c-$N(CH_2CH_2)_2N(CH_2)_3$-$C_8H_8N_5$, and $C_8H_8N_5$ should read as follows: 3-[4-[3-(7,8-dihydro-5-methyl-6*H*-pyrrolo[3,2-*e*][1,2,4]triazolo[1,5-*a*]pyrimidin-8-yl)propyl]-1-piperazinyl]propyl.

L. G. Abood: Annual Review of Neuroscience. Volume 3.

Page 1061. The title was incorrectly stated as "Annual Review of Neurochemistry". The correct title of the book should read "Annual Review of Neuroscience".