(Model ZBI). Since the Daudi cells were grown in suspension culture, they were counted without trypsinization. The average number of cells/mL (in duplicate) for each drug was plotted against the concentration and response curves obtained. The concentration at 50% inhibition is the (ID_{50}) cell growth concentration. Results of this group of cells are given in Table I.

Growth Inhibition of Human Mammary Cells (in Vitro) (Table II). (a) **Human mammary SW-613 cells (adenocarcinomas)** were supplied by Dr. E. M. Jensen, of the E.G. & G/Mason Research Institute, Rockville, MD. The cloning efficiency of these cells in soft agar was 9%, and they were not responsive to estrogens. Their tumorgenicity in nude mice was excellent. For test purposes, 5×10^6 SW-613 cells were grown as a monolayer in 35-mm petri dishes in a medium containing RPMI 1640, 10% fetal calf serum, and neomycin. The analogues were tested as described previously.[15]

(b) **Human mammary MCF-7 cells (Scirrhous carcinoma of the breast)** were supplied by Dr. E. M. Jensen. These epithelial cells grow in monolayers and are sensitive to estrogens. For test purposes, 5×10^6 MCF-7 cells were grown as a monolayer in Eagle's minimal essential medium (MEM) with Hanks' balanced salt solution (BSS), 10% calf serum, 10 μg/mL insulin, 100 units/mL penicillin, and 100 μg/mL streptomycin and incubated in culture in the same manner as in the preceding tumor cells. The controls had a doubling time of 36 h. The workup and assay were done in the same manner as in the case of the SW-613 cells.[15]

In Vivo Tumor Cells Systems. Mouse Leukemia, L-1210 (DBA/2 Ha Mice). Female DBA/Ha mice, 6–8 weeks old (19–20 g), were obtained from the RPMI breeding colony. Each animal was injected ip with 10^6 cells of leukemia L-1210 and treated with the drug, once daily (or twice daily if fractionated or divided doses are used) for 5 consecutive days starting the day after inoculation. The drugs were in aqueous solution, or in the case of sparing solubility as N^6-benzyladenosine, in aqueous dispersion containing Tween 80. Survival time was noted for each animal. Tests were run in groups of five mice on three separate occasions. Results are given in Table III.

Tests with *carcinoma TA3 and Taper hepatoma* were carried out using A/ST female mice and HA/ICR Swiss female mice, respectively, in the same manner as in the preceding test, but no activity was obtained here which resulted in increased life spans (ILS).

Wistar–Furth Rats Bearing TW-98 Mammary Tumors. Two groups of ten rats each were injected (ip) with 24 mkd of N^6-allyladenosine five times per week for 3 weeks. Tumors were measured throughout. After termination of the experiment, the tumors were excised and weighed. Upon comparison with the controls, the latter were found to be growing at four times the rate as those of the treated animals.

Tumors Derived from Spontaneous Mammary Tumors of the DBA/2 HaDD Mouse (Table IV). The isolation and establishment of these tumor lines are discussed above. These tumors are not responsive to steroid drugs (estrogens). Tumor fragments, approximately 1 mm in diameter, were transplanted by standard trocar technique subcutaneously in the milk gland line into groups of five female mice. On day 3, the tumor was measured by calipers and the mice were weighed. Normally the drug was injected on 5 successive days. In the case of slow growing tumors, the first injection of the drug was given on day 6 and was continued for 5 successive days, although this regimen was varied as noted below. The tumors were measured twice a week, normally ending with the death of the mouse. The mortality was checked daily, 7 days a week. If the drug was given twice a day, it was done in the morning and late afternoon. The data for this study are given in Table IV. In some cases the tumor half-growth of the treated mice equals that of the control, and yet the treated animals have increased survival time. This is due to an increase in growth rate right after treatment is stopped, which slows down after half-growth rate is reached.

Acknowledgment. The authors are grateful for the help received from their associates in the course of this work. They thank Nancy Porter and Patricia Dix for assistance in tissue culture studies, Dr. Julius Horoszewicz and Dr. Susan Leong of the Department of Medical Viral Oncology for the work on the epithelial tissues, Dr. Margot Ip for the work with breast tumors in rats, and Alice Atwood for assistance with the tumor studies in mice. The work was supported in part by Grants CA-13038 from the National Cancer Institute and CA-19814 from the Breast Cancer Task Force, NCI.

(15) R. Bernacki, C. Porter, W. Korytnyk, and E. Mihich, *Adv. Enzyme Regul.* 16, 217 (1978).

On the Rational Selection of Test Series. 1. Principal Component Method Combined with Multidimensional Mapping

W. J. Streich, S. Dove, 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 March 14, 1980*

A method for the rational selection of optimal test series with high data variance and low collinearities is presented (PCMM method). The method combines the technique of multidimensional mapping originally introduced by Wootton and colleagues with the principal component method, and it is superior to other selection methods with respect to its collinearity decreasing power. Two examples of the application of PCMM are given, and the results are compared with corresponding results from other selection techniques.

Quantitative structure–activity relationships (QSAR) have developed into an important tool to rationalize drug design. The evaluation of meaningful QSAR and their effective application in the process of lead optimization requires that, in a first step, a test series with optimal properties is selected. This is not an easy task, especially for a batch situation (synthesis quicker than biological testing[1]) where this selection has to be performed prior to any experimental work. An optimal test series must provide a maximum of information with the smallest possible number of analogues, and this supposition can only be fulfilled if all physicochemical and structural properties governing biological potency are varied systematically and independently from each other over a sufficiently large range. That means that the analogues comprising the test series must be selected in such a way that the variances of hydrophobic, electronic, and steric molecule parameters are maximized and, at the same time, collinearities between these parameters are minimized. The analogues must, furthermore, span a sufficiently large part of the parameter space with sufficiently large and about equal Euclidean distances between them within this space. This is not only

0022-2623/80/1823-1452$01.00/0

important for the Hansch analysis but also for other QSAR methods using extrathermodynamic parameters as, for instance, discriminant analysis.

The importance of these requirements has been stressed by several workers,[1-10] and different methods for the planning of an optimal test series have been proposed[1-4,7,8,11] and more or less successfully applied.[12-15] We feel, however, that the available selection methods are either not efficient enough for practical work or do not suffice to fulfill all of the above-mentioned conditions of an optimal test series equally well. For this reason, we wish to present two new selection methods. The method presented in this paper is an extension of the Wootton approach allowing, in addition to the maximization of data variance, a systematic minimization of multiple collinearity (PCMM method). In the second part of this note, a novel approach particularly suitable for smaller sets of possible analogues will be presented.

The PCMM Method

Suppose that N ($j = 1, \cdots, N$) derivatives of a lead compound are synthetically accessible and that a test series of n analogues is to be selected from these N compounds in such a way that the above criteria are fulfilled. The N analogues can be presented as points in a space spanned by variables x_i describing hydrophobic, electronic, and steric molecular properties. Since the x_i may be on quite different scales, it is advantageous to normalize the variables to a mean of zero and a standard deviation of unity. If multiple collinearities are present in the data, a hyperplane can be fitted to the points corresponding to the analogues (objects); the fit will be better the higher the collinearities are. It is obvious that the collinearities are mainly due to those ojects which are close to this hyperplane. Collinearities could thus be eliminated by deleting such objects provided that the remaining objects cannot be fitted to a new hyperplane. Such a procedure is not acceptable, however, since the information contained in the objects close to the hyperplane would be completely lost, and it would be no longer possible to select a test series truly representative for the whole parameter space. In order to do that, the Euclidean distances of all points from the hyperplane are calculated (for all objects) and are then used for dividing the objects into two sets.[21] One set contains the objects close to (set 1) and the other those distant from the hyperplane (set 2). The sizes of the two sets are adjustable; according to our experiences it is a good choice to have about 30% of the objects in set 1 and 70% in set 2.

Table I. Clustering of the 35 Substituents Investigated in Reference 4 in the Parameter Space F, R, π, and MR Obtained from Cluster Analysis[a]

cluster no.	substituent no.[b]
1	3-8
2	1, 2
3	12-14, 28
4	15-17
5	10, 11, 32
6	19, 20
7	18, 21, 24-26
8	22, 23, 31
9	27, 29, 30
10	9, 33-35

[a] Data for the substituent constants taken from reference 20. [b] 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.

To each of the two sets the multidimensional mapping technique of Wootton is now applied separately in such a way that a higher percentage of compounds is selected from set 2. In doing that, the possibility must be considered that the best fitted hyperplane changes its position during the selection of compounds. In order to account for this possibility, an automatic stepwise iteration procedure is used. In each step a certain percentage of compounds is eliminated and, after fitting a new hyperplane to the remaining compounds, resulting in two new sets (1 and 2), multidimensional mapping is repeated. The process is continued until the number of objects left becomes equal to n, the desired number of analogues in the test series. We found it convenient to eliminate about 20% of the compounds remaining after the previous step from set 1 and about 10% from set 2 in each cycle. These percentages can, however, be varied; the value of D_{min}[4] necessary to produce the desired elimination rates is automatically adjusted by the computer. The preference for compounds from set 2 in each step (higher size, lower elimination rate) guarantees an effective minimization of multiple collinearities with still a sufficient information about set 1 included into the test series.

The whole procedure is a combination of two basic steps, principal component method and multidimensional mapping, and will, therefore, be called the PCMM method. In order to check the effectiveness of PCMM in comparison to other methods, it was applied to the same examples as treated by Wootton et al.[4] and by Hansch and co-workers[3] using the parameter space spanned by the variables π, R, F, and MR; for the Hansch example, $\mathcal{F}$ and $\mathcal{R}$ were used instead of F and R. Ten test series, comprising ten compounds in each case, were selected by different methods and compared with respect to data variance and collinearities. The following methods were used: stochastic selection on the basis of uniform distribution (method I); stochastic selection from the clusters presented in ref 3 and from clusters calculated by cluster analysis for the data used in ref 4, which are summarized in Table I (10 cluster level; method II); multidimensional mapping according to Wootton et al.[4] (method III); PCMM method (method IV). As operational criteria to judge the quality of the series selected we have used the "variance coefficient", V_s, and

(1) J. G. Topliss, *J. Med. Chem.*, **15**, 1006 (1972).
(2) P. N. Craig, *J. Med. Chem.*, **14**, 680 (1971).
(3) C. Hansch, S. H. Unger, and A. B. Forsythe, *J. Med. Chem.*, **16**, 1217 (1973).
(4) R. Wootton, R. Cranfield, G. C. Sheppey, and P. J. Goodford, *J. Med. Chem.*, **18**, 607 (1975).
(5) C. Hansch, *Pharmacochem. Libr.*, **2**, 47 and 287 (1977).
(6) W. P. Purcell, *Eur. J. Med. Chem.*, **10**, 335 (1975).
(7) Y. C. Martin, Poster on the VIth International Symposium on Medical Chemistry, Brighton, Great Britain, 1977.
(8) Y. C. Martin and H. N. Panas, *J. Med. Chem.*, **22**, 784 (1979).
(9) Y. C. Martin, "Drug Design Methods: A Critical Introduction", Marcel Dekker, New York, 1978.
(10) R. Franke, "Optimierungsmethoden in der Wirkstofforschung: Quantitative Struktur-Wirkungs Analyse", Akademie-Verlag, Berlin, 1980.
(11) T. J. Mitchell, *Technometrics*, **16**, 203 (1974).
(12) 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).
(13) W. J. Dunn, M. G. Greenberg, and S. S. Callejas, *J. Med. Chem.*, **19**, 1299 (1976).
(14) K. J. Shah and E. A. Coats, *J. Med. Chem.*, **20**, 1001 (1977).
(15) R. Cranfield, P. J. Goodford, F. E. Norrington, and W. H. G. Richards, *Br. J. Pharmacol.*, **52**, 87 (1974).

Table II. Test Series of Ten Analogues Selected from the 90 Compounds Treated by Hansch and Co-workers[a] Using Four Different Methods[b] Together with the Corresponding Values of V_s and D_s

method	test series no.	substituents selected[c]	D_s	V_s
I	1	8, 9, 11, 31, 33, 35, 38, 44, 57, 86	0.505	0.793
	2	7, 19, 24, 32, 34, 36, 39, 48, 57, 61	0.275	0.492
	3	5, 9, 15, 24, 36, 43, 47, 53, 69, 78	0.651	1.007
	4	1, 15, 23, 28, 40, 43, 55, 63, 85, 86	0.436	1.126
	5	8, 39, 41, 50, 54, 58, 63, 64, 68, 83	0.587	0.841
	6	1, 6, 18, 20, 27, 50, 57, 60, 72, 87	0.592	0.524
	7	16, 17, 49, 50, 56, 60, 64, 65, 71, 84	0.391	0.622
	8	3, 6, 40, 48, 56, 62, 70, 80, 81, 89	0.410	1.159
	9	20, 25, 36, 41, 47, 50, 55, 68, 76, 88	0.300	1.199
	10	19, 23, 26, 31, 41, 48, 72, 78, 79, 83	0.193	0.562
II	1	3, 10, 15, 43, 52, 61, 69, 75, 84, 90	0.500	1.775
	2	5, 10, 13, 19, 21, 46, 62, 70, 77, 90	0.466	1.501
	3	5, 10, 20, 35, 40, 55, 66, 75, 88, 90	0.424	1.962
	4	10, 19, 36, 44, 55, 57, 63, 70, 78, 89	0.448	1.713
	5	6, 10, 18, 21, 28, 44, 65, 72, 81, 89	0.233	1.569
	6	2, 10, 13, 20, 24, 46, 63, 66, 77, 89	0.156	1.534
	7	8, 10, 14, 21, 43, 48, 62, 71, 78, 89	0.486	1.584
	8	6, 10, 27, 40, 48, 61, 62, 74, 81, 89	0.294	1.412
	9	2, 10, 23, 31, 41, 53, 63, 86, 87, 89	0.505	1.691
	10	4, 10, 19, 24, 33, 44, 73, 75, 86, 90	0.192	1.934
III	1	6, 10, 31, 44, 55, 63, 67, 75, 77, 89	0.425	2.208
	2	10, 22, 33, 43, 56, 68, 72, 75, 79, 89	0.574	2.184
	3	1, 10, 33, 41, 51, 62, 68, 75, 86, 89	0.591	1.998
	4	4, 10, 11, 21, 45, 63, 68, 75, 77, 88	0.715	2.055
	5	10, 20, 43, 55, 62, 68, 75, 81, 87, 89	0.705	2.188
	6	10, 12, 38, 44, 57, 68, 71, 75, 81, 89	0.526	2.117
	7	4, 9, 10, 38, 44, 53, 68, 75, 84, 89	0.711	2.110
	8	7, 10, 31, 40, 42, 56. 68, 72, 75, 89	0.645	2.268
	9	10, 13, 41, 56, 62, 68, 75, 77, 80, 89	0.654	2.093
	10	8, 10, 33, 40, 42, 54, 70, 75, 76, 89	0.583	2.242
IV	1	1, 18, 30, 38, 56, 68, 75, 76, 87, 88	0.898	1.582
	2	7, 10, 37, 47, 50, 62, 68, 75, 87, 88	0.841	1.970
	3	1, 10, 18, 30, 62, 68, 75, 77, 87, 88	0.732	1.895
	4	1, 10, 37, 50, 56, 62, 68, 75, 87, 88	0.902	1.873
	5	4, 9, 31, 37, 47, 49, 56, 67, 75, 87	0.761	1.581
	6	9, 10, 37, 50, 56, 62, 68, 75, 86, 87	0.932	1.878
	7	8, 38, 47, 52, 56, 62, 68, 75, 77, 79	0.836	1.674
	8	37, 40, 47, 54, 65, 68, 70, 75, 77, 87	0.796	1.647
	9	4, 8, 40, 49, 56, 63, 68, 75, 77, 84	0.863	1.667
	10	12, 37, 40, 54, 62, 68, 75, 76, 87, 90	0.690	1.733

[a] See reference 3. [b] Method I = stochastic selection; method II = stochastic selection from the clusters presented in reference 3 (10 cluster level); method III = multidimensional mapping according to Wootton et al.;[4] method IV = PCMM method presented in this paper. [c] 1 = $B(OH)_2$, 2 = 3,4-(OCH_2O), 3 = EtCOOH, 4 = PMe_2, 5 = Me, 6 = C_2H_3, 7 = Et, 8 = MeOH, 9 = H, 10 = C_2H_2COOH, 11 = CN, 12 = NO_2, 13 = CHO, 14 = COOH, 15 = COMe, 16 = CH_2Cl, 17 = C_2H, 18 = Cl, 19 = N_3, 20 = SH, 21 = SMe, 22 = CHNOH, 23 = CH_2CN, 24 = C_2H_2CN, 25 = OCOMe, 26 = $C_2H_2NO_2$ (trans), 27 = SCOMe, 28 = CO_2Me, 29 = SCN, 30 = $CONH_2$, 31 = CONHMe, 32 = SO_2NH_2, 33 = SO_2Me, 34 = SOMe, 35 = NHCHO, 36 = NHCOMe, 37 = $NHCONH_2$, 38 = $NHCSNH_2$, 39 = $NHSO_2CH_3$, 40 = F, 41 = OMe, 42 = NH_2, 43 = $NHNH_2$, 44 = OH, 45 = NHMe, 46 = NHEt, 47 = NMe_2, 48 = Br, 49 = OCF_3, 50 = CF_3, 51 = NCS, 52 = I, 53 = SF_5, 54 = SCF_3, 55 = SO_2F, 56 = SO_2CF_3, 57 = CH_2Br, 58 = $NCCl_2$, 59 = SeMe, 60 = C_2H_2COMe, 61 = NHCOOEt, 62 = CHNPh, 63 = SO_2Ph, 64 = OSO_2Me, 65 = 5-Cl-tetrazolyl, 66 = C_2H_2COOEt, 67 = NHCOPh, 68 = NCHPh, 69 = C_2H_2COPh, 70 = $NHSO_2Ph$, 71 = OSO_2Ph, 72 = COPh, 73 = N_2Ph, 74 = OCOPh, 75 = $POPh_2$, 76 = 3,4-$(CH_2)_3$, 77 = 3,4-$(CH_2)_4$, 78 = *n*-Pr, 79 = *i*-Pr, 80 = 3,4-$(CH)_4$, 81 = NH-*n*-Bu, 82 = NHPh, 83 = 2-thienyl, 84 = Ph, 85 = CH_2Ph, 86 = *t*-Bu, 87 = OPh, 88 = $SiMe_3$, 89 = ferrocenyl, 90 = adamantyl.

the determinant of the correlation matrix, D_s. The variance coefficient is defined as

$$V_s = \frac{1}{m}\sum_{i=1}^{m}\frac{s_{is}^2}{s_{ip}^2} \quad (1)$$

where s_{is}^2 expresses the variance of the ith variable within the test series, and s_{ip}^2 is the variance of this variable within the starting population of the N compounds. The variance coefficient represents an average of the test series variance as compared with the total variance in the starting population. It is self evident that this quantity is an operational measure of the informational content of selected test series in comparison with the starting set; its value should always be ≥ 1. The determinant D_s of the correlation matrix measures collinearities.[17,18] D_s becomes equal to unity if collinearities are completely absent and converges to zero as collinearities increase. In comparison with other parameters characterizing collinearities, D_s has two

Table III. Mean Values and Standard Deviations of D_s and V_s for Different Selection Methods[a] for the Starting Population of 90 Compounds

	$\bar{D}_s$	$s_{\bar{D}_s}$	$\bar{V}_s$	$s_{\bar{V}_s}$
starting population	0.68		1.00	
method I	0.43	0.15	0.83	0.28
method II	0.37	0.14	1.67	0.18
method III	0.61	0.09	2.15	0.09
method IV	0.83	0.08	1.75	0.14

[a] See Table II.

(16) H. H. Harman, "Modern Factor Analysis", 2nd ed, University of Chicago Press, Chicago, 1967.
(17) D. E. Farrar and R. R. Glauber, *Rev. Econ. Stat.*, **49**, 92 (1967).
(18) Y. Haitovsky, *Rev. Econ. Stat.*, **50**, 486 (1968).

Table IV. Test Series of Ten Analogues Selected from the 35 Compounds Treated by Wootton and Co-workers[a] Using Four Different Methods[b] Together with the Corresponding Values of D_s and V_s

method	test series no.	substituents selected[c]	D_s	V_s
I	1	1, 9, 12, 16, 19, 21, 22, 27, 30, 32	0.547	1.186
	2	1, 5, 10, 12, 21, 23, 25, 26, 28, 33	0.575	0.598
	3	1, 5, 6, 11, 16-18, 22-24	0.132	1.273
	4	3, 4, 8, 19, 26, 28-31, 35	0.235	1.088
	5	3, 12, 14, 15, 18, 20, 28, 31, 34, 35	0.284	0.701
	6	1, 2, 5, 6, 11, 16, 20-23	0.202	1.137
	7	6-10, 14, 21, 26, 31, 33	0.159	1.046
	8	3, 8, 9, 11-13, 23, 24, 26, 33	0.418	0.651
	9	11, 13, 14, 22, 24-26, 29, 30, 33	0.382	0.746
	10	2, 8, 10-12, 16, 22, 23, 25, 35	0.194	0.990
II	1	1, 5, 15, 20, 23, 25, 28, 30, 32, 34	0.641	0.967
	2	2, 7, 11, 13, 17-19, 29, 31, 33	0.406	1.185
	3	2, 8, 10, 16, 20, 23, 24, 27, 28, 33	0.418	1.077
	4	1, 7, 11, 14, 16, 19, 25, 30, 31, 35	0.396	1.262
	5	2, 8, 11, 13, 15, 20, 22, 26, 30, 33	0.461	1.087
	6	1, 3, 9, 12, 15, 20, 24, 29, 31, 32	0.471	1.042
	7	2, 6, 10, 14, 15, 20, 25, 27, 31, 35	0.440	1.103
	8	2, 4, 10, 16, 18, 19, 22, 27, 28, 35	0.450	1.143
	9	1, 4, 9, 13, 16, 19, 23, 24, 27, 32	0.565	1.104
	10	1, 3, 11, 12, 15, 19, 26, 29, 31, 34	0.486	0.971
III	1	2, 4, 10, 12, 15, 20, 22, 25, 29, 33	0.516	1.066
	2	2, 7, 9, 15, 18-20, 22, 28, 32	0.287	1.152
	3	2, 7, 9, 11, 15, 18-20, 22, 28	0.205	1.081
	4	2, 7, 10, 14, 16, 20-22, 27, 33	0.509	1.216
	5	1, 7, 9, 14, 16, 18, 19, 20, 28, 32	0.336	1.170
	6	2, 7, 12, 16, 19, 20, 22, 25, 30, 32	0.479	1.356
	7	2, 6, 9, 10, 14, 16, 18, 20, 22, 28	0.255	1.172
	8	2, 7, 9, 10, 12, 16, 20, 23, 26, 29	0.360	1.168
	9	1, 4, 10, 13, 17, 20, 26, 27, 31, 34	0.547	1.202
	10	3, 8, 9, 14, 16, 19, 21, 27, 32, 35	0.585	1.166
IV	1	1, 3, 8, 9, 11, 19, 20, 25, 27, 32	0.608	1.062
	2	1, 8, 10, 17, 20, 22, 26, 27, 29, 34	0.581	1.437
	3	2, 6, 9, 11, 16, 17, 21, 26, 27, 33	0.464	1.117
	4	4, 8, 12, 17, 20, 22, 24, 29, 33, 35	0.462	1.177
	5	8, 10, 17-19, 22, 27, 29, 33, 35	0.471	1.302
	6	1, 3, 9, 16, 18, 20, 23, 26, 27, 33	0.502	1.014
	7	1, 3, 7, 10, 12, 18, 20, 27, 33, 35	0.724	0.911
	8	1, 8-10, 18, 20, 27, 29, 32, 34	0.752	1.129
	9	2, 7-9, 11, 15, 20, 22, 27, 35	0.431	1.166
	10	2, 8, 10, 13, 20-22, 25, 27, 34	0.568	1.033

[a] See ref 4. [b] Method I = stochastic selection, method II = stochastic selection from the clusters presented in Table I, method III = multidimensional mapping according to Wootton et al., method IV = PCMM method presented in this paper. [c] See footnote c in Table II.

advantages: it is on a well-defined scale ($0 \leq D_s \leq 1$) and invariant against linear transformations of the variables; this is not generally true, for instance, for the determinant of the covariance matrix, which is also a common measure for collinearities. It should be stressed that neither V_s nor D_s is used within the selection methods. They only serve as criteria for the quality of test series after the selection. Test series selected from the same starting set will be better the higher the values of V_s and, especially, D_s are.

Results and Discussion

Table II summarizes the test series selected from the 90 compounds treated by Hansch and co-workers (for the original data, see ref 3 and 19) together with the corresponding values of V_s and D_s; the mean values of V_s and D_s are presented in Table III. As expected, stochastically selected test series (method I) are far from being optimal as judged by the low values of $\bar{V}_s$ and $\bar{D}_s$. The Hansch and the Wootton techniques (methods II and III) both yield test series with high variances. The values of $\bar{D}_s$, however, are not very high for such a large data set, indicating that collinearities are not as low as one would like them to be.

Table V. Mean Values and Standard Deviations of D_s and V_s for Different Selection Methods[a] for the Starting Population of 35 Compounds

	$\bar{D}_s$	$s_{\bar{D}_s}$	$\bar{V}_s$	$s_{\bar{V}_s}$
starting population	0.57		1.00	
method I	0.31	0.16	0.94	0.25
method II	0.47	0.08	1.09	0.09
method III	0.40	0.14	1.18	0.08
method IV	0.56	0.11	1.14	0.15

[a] See Table IV.

This is especially true for the Hansch technique, since $\bar{D}_s$ is even lower here than for stochastically selected test series. The PCMM method presented in this paper is much more effective in minimizing collinearities, as follows from the very high value of $\bar{D}_s$. The variance coefficient for PCMM is somewhat smaller than for method III and of about the same magnitude as for method II. It can be considered completely acceptable, since it is distinctly higher than the value for stochastically selected test series (method I).

A similar picture emerges if the example treated by Wootton et al. is considered (for the data, see ref 4 and 20). The test series selected, together with the corre-

(19) C. Hansch, E. Leb, S. H. Unger, K. H. Kim, D. N. Nikaitani, and E. J. Lien, J. Med. Chem., 16, 1207 (1973).

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