Table VII. Comparison of Pharmacological Activity of 44 and Indomethacin

	44	indomethacin
LD_{50} in mice, mg/kg, po	>1000	24.3
LD_{50} in rats, mg/kg, po	1215	12.0
PG synthetase inhibn, IC_{50}, μM	2.5	1.0
carrageenan edema, ED_{50}, mg/kg, po	29.6	3.5
nystatin edema, ED_{50}, mg/kg, po[a]	26.5	>6.0
analgesic act., ED_{50}, mg/kg, po[b]	68.5	6.5
antipyretic act., MED, mg/kg, po[c,d]	12.4	1.6
gastroulcerogenic effect, ED_{50}, mg/kg, po[e,f]	170.0	4.6
adjuvant arthritis, MED, mg/kg, po[g]		
continuous dosing[h]	3.0	1.0
therapeutic dosing[i]	10.0	<1.0
prophylactic dosing[j]	10.0	2.0[k]

[a] E. Arrigoni-Martelli, P. Schiatti, and D. Selva, *Pharmacology*, 5, 215 (1971). [b] K. F. Swingle, T. J. Grant, and D. C. Kvam, *Proc. Soc. Exp. Biol. Med.*, 127, 536 (1971). [c] R. D. Sofia, W. Diamantis, R. Gordon, and M. Kletzkin, *Eur. J. Pharmacol.*, 26, 51 (1974). [d] MED = minimal effective dose; i.e., the lowest dose preventing any further increase of body temperature. [e] Groups of six Sprague-Dawley rats dosed orally once daily for 3 days. Examination for lesions of the internal surface of the stomachs 24 h after the last dose. [f] ED_{50} = dose causing lesions macroscopically appreciable in 50% of rats. [g] MED = minimal effective dose, i.e., the lowest dose causing significant ($p < 0.05$) inhibition of the swelling of the noninjected paw on day 28 post-adjuvant. [h] Daily dosing from the day of adjuvant injection to day 28. [i] Daily dosing from day 16 to day 28 postadjuvant. [j] Daily dosing from day 5 preadjuvant to day 5 postadjuvant. [k] Not active at 2.0.

adjuvant injection to day 28 postadjuvant, unless otherwise indicated.

Synthesis. Melting points were uncorrected and recorded with a Büchi 510 apparatus. Elemental analyses for C, H, N, S, halogen, and H_2O were performed by G. Cornali and W. Egger and were within ±0.4% of the calculated values, unless otherwise noted. IR and NMR spectra were obtained for all compounds and were consistent with assigned structures. A Perkin-Elmer 457 spectrophotometer was used to obtain IR spectra. NMR spectra were obtained with Varian A 60 A, JEOL JNM-PMX 60, and JEOL JNM-FX 100 spectrometers.

***N*-Cyclohexyl-*N'*-4-(2-methylquinolyl)carbodiimide.** A mixture of *N*-cyclohexyl-*N'*-4-(2-methylquinolyl)urea[11] (142.0 g, 0.50 mol), Ph_3P (150.0 g, 0.57 mol), CCl_4 (50.0 mL, 0.52 mol), and Et_3N (75.0 mL, 0.53 mol) in CH_2Cl_2 (1.0 L) was refluxed for 2 h. After evaporation of all solvent, the residue was triturated with four portions of petroleum ether (2.0 L). The combined extracts were evaporated in vacuo to yield the crude carbodiimide (110.0 g, 83.0%) as a yellow oil, which was not further characterized.

Method A. *N-tert*-Butyl-*N''*-4-(2-methylquinolyl)-*N'*-2-thiazolylguanidine (16, Table I). 4-Amino-2-methylquinoline (0.6 g, 5 mmol) was added to crude *N-tert*-butyl-*N'*-2-thiazolylcarbodiimide[7] (1.2 g, 6.6 mmol). The mixture was heated on a steam bath for 30 min and allowed to cool to room temperature. After 12 h, the mixture was triturated with Et_2O. The crystalline precipitate of **16** was collected.

Method B. *N*-Cyclohexyl-*N''*-4-(2-methylquinolyl)-*N'*-2-thiazolylguanidine (44, Table III). To crude *N*-cyclohexyl-*N'*-4-(2-methylquinolyl)carbodiimide (4.0 g, 15 mmol) in toluene (10 mL) was added 2-aminothiazole (1.0 g, 10 mmol), and the solution was refluxed for 1 h. After the solution was left standing at room temperature overnight, the precipitate (sometimes the addition of petroleum ether induced crystallization) was filtered off and washed with toluene and Et_2O.

Method C. *N-tert*-Butyl-*N'*-2-thiazolyl-*N''*-4-[2-(trifluoromethyl)quinolyl]guanidine (93, Table V). 4-Amino-2-(trifluoromethyl)quinoline (2.0 g, 9.1 mmol) was stirred in dry DMF (25 mL). NaH (50% in mineral oil dispersion; 0.5 g, 10 mmol) was added, followed by *N-tert*-butyl-*N'*-2-thiazolylcarbodiimide (1.9 g, 10.5 mmol). The mixture was stirred overnight at room temperature, and ice-water was added (100 mL). The precipitate was filtered and washed with H_2O and petroleum ether.

Acknowledgment. We are indebted to Professor Ole Buchardt (University of Copenhagen) for a generous supply of some aminoquinolines. The authors are grateful to the following for their excellent technical assistance: Mr. Jens Enemark, Mrs. Birthe Ewertsen, Mrs. Anne-Dorte Riis Christensen, Mrs. Bente Nielsen, and Mr. H. J. Poulsen. We are also grateful to Mr. F. Lund and Mr. H. J. Petersen for encouragement and helpful discussions and to Mrs. Lis Finderup Johansen for preparing the manuscript.

Adaptive Least-Squares Method Applied to Structure–Activity Correlation of Hypotensive *N*-Alkyl-*N''*-cyano-*N'*-pyridylguanidines

Ikuo Moriguchi,* Katsuichiro Komatsu, and Yasuo Matsushita

School of Pharmaceutical Sciences, Kitasato University, Shirokane 5-chome, Minato-ku, Tokyo, 108 Japan.
Received April 2, 1979

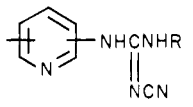
A method using an adaptive least-squares (ALS) technique has been developed for the discrimination of ordered categorical data. The method (ALS method) has the advantages of simultaneously considering any number of classes and of producing a single discriminant function which can place patterns in several classes. The ALS method was compared with linear discriminant analysis (LDA) in application to the problem of discriminating three-class hypotensive therapeutic indices of 76 *N*-alkyl-*N''*-cyano-*N'*-pyridylguanidines using nine descriptor variables. With the full data set and in the five leave-out runs, it was shown that the ALS method was superior and more stable in recognition and prediction. The structure–activity relationship is discussed on the basis of discriminant functions formulated.

The strength of drug action has been often recorded in a form of activity rating. To such ordered categorical data, linear discriminant analysis (LDA) has been applied for the structure–activity correlation.[1-4] For this purpose,

0022-2623/80/1823-0020$01.00/0

however, LDA is not considered to be suitable except in the two-group case because of the following reasons. (a) LDA is designed to deal with the problem of discrimination among independent categories and not designed for ordered categories. (b) The assumptions underlying LDA, i.e., multivariate normality and equality of the within-group covariance matrices, are not always fulfilled with structure–activity data of three or more groups.[5] (c) In LDA in the m-group case ($m \geq 3$), since m discriminant functions are derived, it is difficult to interpret in terms of a model of the structure–activity relationship. For the purpose of relating structure to activity rating, a new discrimination method was developed using an adaptive least-squares (ALS) technique, of which a preliminary report[6] had been published.

This article first describes the method (ALS method) and then reports a specific application of the ALS method to the problem of discriminating three-class hypotensive therapeutic indices of 76 *N*-alkyl-*N''*-cyano-*N'*-pyridylguanidines of general structure I reported by Petersen et



I

al.[7] The results are compared with those obtained by LDA and other related methods.

ALS Method. The ALS method makes decisions for ordered m-group ($m \geq 2$) discrimination by use of a single discriminant function as

$$L = w_0 + w_1x_1 + w_2x_2 + \ldots + w_px_p \quad (1a)$$

where x_k is the kth descriptor ($k = 1, 2, \ldots, p$) for the structure, w_k is the weight coefficient, and L is the discriminant score. For a set of n compounds, eq 1a can be rewritten as eq 1b. In the matrix $\mathbf{X}$, x_{ki} ($k = 1, 2, \ldots, p$ and $i = 1, 2, \ldots, n$) is the kth descriptor for the ith compound.

$$\mathbf{L} = \mathbf{XW} \quad (1b)$$

$$\mathbf{L} = \begin{bmatrix} L_1 \\ L_2 \\ \vdots \\ L_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{p1} \\ 1 & x_{12} & \cdots & x_{p2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{bmatrix}$$

Starting scores, a_j ($j = 1, 2, \ldots, m$), for the members of class j are assumed, and then cutting points, b_j ($j = 1, 2, \ldots, m - 1$), between classes are fixed in advance. The cutting points are usually not moved through the ALS calculation. In this study, a_j was assumed by eq 8 under Method, and b_j was taken as the midpoint between a_j and a_{j+1}.

The procedure begins with the setting of forcing factors $S_i^{(1)}$ ($i = 1, 2, \ldots, n$), which are taken to be

$$S_i^{(1)} = a_j \quad (2)$$

where a_j is the starting score for class j to which the ith substance of a set of n compounds was observed to belong. Generally, classes are numbered in ascending order of biological activity. By use of $S_i^{(1)}$ ($i = 1, 2, \ldots, n$) in place of $\mathbf{L}$ in eq 1b as $\mathbf{S}^{(1)} = \mathbf{XW}$, where $\mathbf{S}^{(1)} = (s_1^{(1)}, s_2^{(1)}, \ldots, s_n^{(1)})'$ (the prime denotes the transposition), the least-squares estimate of $\mathbf{W}$ is written to be $\mathbf{W}^{(1)}$ as

$$\mathbf{W}^{(1)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{S}^{(1)} \quad (3)$$

$\mathbf{W}^{(1)}$ is computed by use of an ordinary least-squares program and used for the initial weight vector.

Then, $L_i^{(1)}$ for each substance is calculated from eq 1a using $\mathbf{W}^{(1)}$ as

$$\mathbf{L}^{(1)} = \mathbf{XW}^{(1)} \quad (4)$$

All substances are classified on the basis of the values of $L_i^{(1)}$ and the cutting points as follows: If $L_i^{(1)} \leq b_1$, then assign the ith substance to class 1; if $b_1 < L_i^{(1)} \leq b_2$, then assign to class 2; ...; and if $L_i^{(1)} > b_{m-1}$, then assign to class m.

At iteration 2 and thereafter, the forcing factor $S_i^{(t+1)}$ ($t \geq 1$) is adapted as

$$\begin{aligned} S_i^{(t+1)} &= L_i^{(t)} \text{ (when correctly classified at iteration } t) \\ &= L_i^{(t)} \pm C_i^{(t)} \text{ (when misclassified)} \end{aligned} \quad (5)$$

$$i = 1, 2, \ldots, n$$

where $C_i^{(t)}$ is the correction term.

The correction term is taken to be a constant value or a function of the distance between $L_i^{(t)}$ and the nearer boundary of the observed class for the ith substance. In this study, $C_i^{(t)}$ was assumed as eq 10 under Method. The sign (±) for $C_i^{(t)}$ in eq 5 is chosen to correspond with that for $S_i^{(1)} - L_i^{(t)}$.

Then, the least-squares estimate of $w_k^{(t+1)}$ is computed from eq 6 and $L_i^{(t+1)}$ is calculated from eq 1 using $\mathbf{W}^{(t+1)}$ for classification.

$$\mathbf{W}^{(t+1)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{S}^{(t+1)} \quad (6)$$

The adaptation is repeated until all substances are correctly classified, or repeated given iteration times, and the best discriminant function is selected. Unlike the case of LDA, the discriminant function is nonunique in the ALS method.

The ALS method does not asssume any particular distribution of the data and can be considered to be a technique of pattern recognition.

Method

Activity Class. Although a large number of compounds possessing hypotensive potency have been reported, very few among them can be safely used for clinical therapy. The data of Petersen and his co-workers[7] contained 76 compounds to which six activity classes were alloted based on the minimum effective dose (MED) in relation to a rough estimate of the LD_{50} in rats. In this study, for simplicity of calculation, three-group analysis was done: class 1 (original classes 0 and 1), 32 compounds (poorly active); class 2 (original classes 2 and 3), 25 compounds (moderately active); class 3 (original classes 4 and 5), 19 compounds (very active).

Preliminary Selection of Descriptor Variables. For this purpose, the product of the weight coefficient (w_k) and the standard deviation for each descriptor was utilized as a measure of the contribution to the discriminant score (L). Descriptors whose value of the product exceeded 0.1 were preliminarily taken for this study.

ALS Calculation. (1) **Starting Score a_j.** There are several procedures for scoring ordered categories in the field of statistics.[8]

(1) Y. C. Martin, J. B. Holland, C. H. Jarboe, and N. Plotnikoff, *J. Med. Chem.*, **17**, 409 (1974).
(2) R. Franke and W. Meisske, *Acta Biol. Med. Ger.*, **35**, 73 (1976).
(3) E. M. Hodnett, G. Prakash, and J. Amirmoazzami, *J. Med. Chem.*, **21**, 11 (1978).
(4) G. Prakash and E. M. Hodnett, *J. Med. Chem.*, **21**, 369 (1978).
(5) I. Moriguchi in "Structure–Activity Relationships—Quantitative Approaches", T. Fujita, Ed., Nankodo, Tokyo, 1979, p 285.
(6) I. Moriguchi and K. Komatsu, *Chem. Pharm. Bull.*, **25**, 2800, 3440 (errata) (1977).
(7) H. J. Petersen, C. K. Nielsen, and E. Arrigoni-Martelli, *J. Med. Chem.*, **21**, 773 (1978).
(8) G. W. Snedecor and W. G. Cochran, "Statistical Methods", 6th ed, Iowa State University Press, Ames, Iowa, 1967, Chapter 9.

Table I. Structural Features and Hypotensive Activity of *N*-Alkyl-*N*''-cyano-*N*'-pyridylguanidines

			rating of therapeutic index[b]												
				recognized		predicted[f]									
compd[a]	pyridyl	R	obsd[c]	ALS[d]	LDA[e]	ALS	LDA	Nc	V_H[g]	*I*-1	*I*-2	*I*-3	*I*-4	*I*-5	*I*-6
1	3-	H	1	1	1	1	2	0	0	1	0	0	0	0	0
2	3-	CH_3	1	1	1	2	1	1	0	1	0	0	0	0	0
3	3-	*n*-C_3H_7	2	2	2	2	2	3	0	1	0	0	0	0	0
4	3-	*n*-C_4H_9	2	2	2	2	2	4	0	1	0	0	0	0	0
5	3-	*n*-C_5H_{11}	2	2	2	2	2	5	0	1	0	0	0	0	0
6	3-	*i*-C_4H_9	2	2	2	2	2	4	0	1	0	0	0	0	1
7	3-	*i*-C_5H_{11}	2	2	2	2	2	5	0	1	0	0	0	0	1
8	3-	neo-C_5H_{11}	3	3	3	3	3	5	0	1	0	0	1	0	0
9	3-	*n*-C_7H_{15}	1	1	1	1	1	7	0	1	0	0	0	0	0
10	3-	2-ethylhexyl	1	1	1	1	1	8	0	1	0	0	0	0	0
11	3-	*i*-C_3H_7	3	2	2	2	2	3	0	1	0	0	0	0	1
12	3-	*sec*-C_4H_9	2	2	2	2	2	4	0	1	0	0	0	0	0
13	3-	$CHEt_2$	1	2	2	2	2	5	0	1	0	0	0	0	0
14	3-	cyclopropyl	2	2	2	2	2	3	0	1	0	0	0	0	0
15	3-	cyclopentyl	2	2	2	2	2	5	0	1	0	0	0	0	0
16	3-	cyclohexyl	2	2	2	2	2	6	0	1	0	0	0	0	0
17	3-	$CH(Me)CMe_3$	3	3	3	3	3	6	0	1	0	0	1	0	0
18	3-	$CH(i\text{-}C_3H_7)_2$	3	2	2	2	2	7	0	1	0	0	0	0	1
19	3-	*t*-C_4H_9	3	3	3	3	3	4	0	1	0	0	1	0	0
20	3-	*t*-C_5H_{11}	3	3	3	3	3	5	0	1	0	0	1	0	0
21	3-	$C(Me_2)(CH_2)_2CH_3$	3	2	2	2	2	6	0	1	0	0	0	0	0
22	3-	$C(Me_2)CHMe_2$	3	3	3	3	3	6	0	1	0	0	1	0	1
23	3-	$C(Me)Et_2$	3	3	3	3	3	6	0	1	0	0	1	0	0
24	3-	$C(Me_2)CH_2CHMe_2$	3	2	2	2	2	7	0	1	0	0	0	0	1
25	3-	$C(Me_2)CMe_3$	1	1	1	1	1	7	0	1	0	0	0	0	0
26	3-	CEt_3	3	3	3	3	3	7	0	1	0	0	1	0	0
27	3-	1-Me-c-Bu	1	2	2	2	2	5	0	1	0	0	0	0	0
28	3-	$C(Me_2)CH_2CMe_3$	1	1	1	1	1	8	0	1	0	0	0	0	0
29	3-	1-adamantyl	1	1	1	1	1	10	0	1	0	0	0	0	0
30	3-	$CH_2C(Me)=CH_2$	2	2	2	2	2	4	0	1	0	0	0	0	0
31	3-	$(CH_2)_2OEt$	1	1	1	1	1	4	0.48	1	0	0	0	0	0
32	3-	$(CH_2)_2NEt_2$	1	1	1	1	1	6	0.59	1	0	0	0	0	0
33	3-	tetrahydrofuryl-2-methyl	1	1	1	1	2	5	0.48	1	0	0	0	1	0
34	3-	furyl-2-methyl	2	2	2	2	2	5	0.10	1	0	0	0	1	0
35	3-	benzyl	2	2	2	2	2	7	0	1	0	0	0	1	0
36	3-	β-phenethyl	1	1	1	1	1	8	0	1	0	0	0	0	0
37	3-	phenyl	2	2	2	2	2	6	0	1	0	0	0	0	0
38	3-	2,6-Me_2-Ph	1	1	1	1	1	8	0	1	0	0	0	0	0
39	3-	$RH = (CH_2)_5$	2	2	2	2	2	5	0	1	0	0	0	0	0
40	2-Cl-3-	*t*-C_4H_9	3	3	3	3	3	4	0	1	0	0	1	0	0
41	5-Br-3-	*t*-C_4H_9	3	3	3	3	3	4	0	1	0	0	1	0	0
42	6-MeO-3-	*t*-C_4H_9	1	1	1	1	1	4	0	1	0	1	1	0	0
43	2,6-$(MeO)_2$-3-	*t*-C_4H_9	1	1	1	1	1	4	0	1	0	1	1	0	0
44	2,6-Cl_2-3-	*t*-C_4H_9	1	1	1	1	1	4	0	1	0	1	1	0	0
45	2,4,6-Me_3-3-	*t*-C_4H_9	1	1	1	1	1	4	0	1	0	1	1	0	0
46	2,4,6-Me_3-3-	*t*-C_5H_{11}	1	1	1	1	1	5	0	1	0	1	1	0	0
47	4-	neo-C_5H_{11}	2	3	3	3	3	5	0	0	1	0	1	0	0
48	4-	$CHEt_2$	2	2	2	1	2	5	0	0	1	0	0	0	0
49	4-	$CH(Me)CH_2CHMe_2$	1	2	2	2	2	6	0	0	1	0	0	0	1
50	4-	$CH(Me)CMe_3$	3	3	3	3	3	6	0	0	1	0	1	0	0
51	4-	$CH(Me)(CH_2)_3CHMe_2$	1	1	1	2	1	8	0	0	1	0	0	0	1

52	4-	CH(Me)CH(Me)Et	1	1	2	1	2	6	0	0	1	0	0	0	0
53	4-	$CH(Et)CHMe_2$	2	2	2	1	2	6	0	0	1	0	0	0	1
54	4-	$CH(i\text{-}C_3H_7)_2$	1	2	2	2	2	7	0	0	1	0	0	0	1
55	4-	CH(Me)-c-Pr	2	2	2	2	2	5	0	0	1	0	0	1	0
56	4-	$CH(c\text{-}Pr)_2$	2	2	2	2	2	7	0	0	1	0	0	1	0
57	4-	cyclohexyl	1	1	2	1	2	6	0	0	1	0	0	0	0
58	4-	cyclooctyl	1	1	1	1	1	8	0	0	1	0	0	0	0
59	4-	$t\text{-}C_4H_9$	3	3	3	3	3	4	0	0	1	0	1	0	0
60	4-	$t\text{-}C_5H_{11}$	3	3	3	3	3	5	0	0	1	0	1	0	0
61	4-	$C(Me_2)CHMe_2$	2	3	3	3	3	6	0	0	1	0	1	0	1
62	4-	$C(Me_2)CMe_3$	2	1	1	1	2	7	0	0	1	0	0	0	0
63	4-	$C(Me_2)CH_2CHMe_2$	2	2	2	1	1	7	0	0	1	0	0	0	1
64	4-	$C(Me_2)CH_2CMe_3$	1	1	1	1	1	8	0	0	1	0	0	0	0
65	4-	$CMeEt_2$	3	3	3	3	3	6	0	0	1	0	1	0	0
66	4-	CEt_3	3	3	3	3	3	7	0	0	1	0	1	0	0
67	4-	1-adamantyl	3	1	1	1	1	10	0	0	1	0	0	0	0
68	4-	benzyl	2	2	2	2	2	7	0	0	1	0	0	1	0
69	4-	α-phenethyl	2	2	2	2	2	8	0	0	1	0	0	1	0
70	4-	β-phenethyl	2	1	1	1	1	8	0	0	1	0	0	0	0
71	4-	1-Ph-*i*-Pr	1	1	1	2	2	9	0	0	1	0	0	1	0
72	4-	phenyl	1	1	2	1	2	6	0	0	1	0	0	0	0
73	2-	cyclohexyl	1	1	1	1	1	6	0	0	0	0	0	0	0
74	2-	$t\text{-}C_4H_9$	1	1	1	1	3	4	0	0	0	0	1	0	0
75	5-Cl-2-	$t\text{-}C_4H_9$	1	1	1	1	3	4	0	0	0	0	1	0	0
76	4,6-Me$_2$-2-	$t\text{-}C_4H_9$	1	1	1	1	1	4	0	0	0	1	1	0	0

[a] Numbered as in Table I of ref 7. [b] Class 1, $MED/LD_{50} \geqslant 1/30$; class 2, $1/30 > MED/LD_{50} \geqslant 1/200$; class 3, $MED/LD_{50} < 1/200$. [c] From ref 7. [d] Using eq 16. [e] Using eq 18–20. [f] From the leave-out test (Table V). [g] From ref 16.

We used simple integer scores in the previous study.[6] However, since the data can be very skew, we have used a modified "ridit" in this study. As a standard numerical score for ordered categories, "ridit" was proposed.[9] It is defined in eq 7, where r_j is the ridit

$$r_j = \left(\sum_{i=1}^{j-1} n_i + n_j/2\right)/n \qquad (7)$$

for group j, and n_i and n_j are the size of groups i and j, respectively. For simplicity's sake, we modified the ridit as eq 8 for use as the

$$a_j = 4r_j - 2 = 2\left(2\sum_{i=1}^{j-1} n_i + n_j\right)/n - 2 \qquad (8)$$

starting score a_j. From eq 8, the mean value of a_j over n compounds becomes zero, and $a_1 = -1$ and $a_2 = +1$ for two groups of the same size. With the whole data set of 76 compounds in this study, $a_1 = -1.158$, $a_2 = 0.342$, $a_3 = 1.500$, and the cutting points $b_1 = -0.408$ and $b_2 = 0.921$. Choice of the ridit as the numerical score is based on the assumption that only the potency order of groups is reliable, i.e., quantitative differences in potency between the different groups and between the different compounds within a group are uncertain in the data to be analyzed.

(2) Correction Term $C_i^{(t)}$. In the previous study,[6] the function shown in eq 9 (which is rewritten in the form used in this report)

$$C_i^{(t)} = 0.1/(0.5 + \delta_i)^2 \qquad (9)$$

was empirically chosen for $C_i^{(t)}$, where $\delta_i = |L_i^{(t)} - b_k|$, and b_k is the cutting point [nearer to $L_i^{(t)}$] of the observed class for the ith substance. Although eq 9 is simple, it has been proved that eq 9 is effective only to the case where most of the misclassified substances are located near their boundaries. Therefore, in this study, eq 9 was modified as eq 10. The following values of the

$$C_i^{(t)} = 0.1/(\alpha + \delta_i)^2 + \beta(\alpha + \delta_i)^2 \qquad (10)$$

constants α and β in eq 10 were empirically selected for effective adaptation: $\alpha = 0.45$ and $\beta = 0.00, 0.01, 0.02, 0.03$, and 0.04. In each run, ALS iteration was performed 30 times.

(3) Criteria of the Best Discrimination. For the selection of the best discriminant function, the "ϵ" value corresponding to the mean square of errors and the Spearman rank correlation coefficient, R_S, were computed at each ALS iteration time. The ϵ value is empirically defined by

$$\epsilon = \sum_{i=1}^{n} e_i^2/(n - p - 1) \qquad (11)$$

where

$$e_i = 0 + |S_i^{(1)} - L_i| \text{ (if correctly classified)}$$
$$= 1 + |S_i^{(1)} - L_i| \text{ (if misclassified)} \qquad (12)$$

In eq 12, $S_i^{(1)}$ is the starting score for the ith substance; p in eq 11 is the number of descriptor variables used.

R_S is calculated as[10]

$$R_S = \frac{(n^3 - n)/6 - \sum T_x - \sum T_y - \sum d_i^2}{[[(n^3 - n)/12 - \sum T_x][(n^3 - n)/12 - \sum T_y]]^{1/2}} \qquad (13)$$

where

$$d_i = x_i - y_i \qquad (14)$$
$$i = 1, 2, ..., n$$

where x_i and y_i are the estimated and observed ranks for the ith subject, respectively, and

$$T = (t^3 - t)/12 \qquad (15)$$

where t is the number of tied observations, i.e., the group size.

(9) I. D. J. Bross, *Biometrics*, 14, 18 (1958).
(10) A. L. Delaunois, Ed., "Biostatistics in Pharmacology", Vol. 2, Pergamon Press, Oxford, 1973, p 943.

Table II. Squared Cross-Correlation Matrix of Descriptor Variables for Equation 16

	Nc	$(Nc)^2$	V_H	*I*-1	*I*-2	*I*-3	*I*-4	*I*-5	*I*-6
Nc	1.000								
$(Nc)^2$	0.935	1.000							
V_H	0.005	0.009	1.000						
I-1	0.101	0.087	0.030	1.000					
I-2	0.159	0.145	0.024	0.797	1.000				
I-3	0.059	0.064	0.004	0.019	0.045	1.000			
I-4	0.069	0.101	0.021	0.001	0.005	0.186	1.000		
I-5	0.034	0.031	0.029	0.026	0.042	0.010	0.054	1.000	
I-6	0.007	0.003	0.009	0.009	0.021	0.016	0.019	0.022	1.000

The summations, $\sum T_x$ and $\sum T_y$, are done over all estimated and observed groups, respectively.

In practice, the discrimination giving the maximum R_S value was selected as the best. When there was more than one discriminant function giving the maximum R_S, the function giving the lowest ϵ value among them was chosen.

LDA Calculation. The procedure generally used for several groups[11] was employed. Prior probabilities were taken to be proportional to group size.

Nonelementary Discriminant Function. The so-called nonelementary discriminant function was calculated using Shiba's program,[12] which gave the first weight vector mathematically equivalent to that[13] obtained by Läuter's method.[14]

***K*-Nearest-Neighbor Method (KNN).** In this study, the closest neighbor (k = 1) determined the identification of the unknown.[15] All descriptors used were normalized before the calculation of Euclidean distances between data points in the p-dimensional space.

Computation. The calculations were performed with a JEOL digital computer, Model JEC-7E, using double precision.

Results and Discussion

Discrimination with the Full Data Set. From a preliminary study of descriptor variables described under Method, the following nine descriptors were finally selected: the number of carbon atoms included in R (Nc) and its square [$(Nc)^2$], the hydrophilic effect[16] for R (V_H), and six indicator variables (*I*-1, *I*-2, ..., *I*-6). The indicator variables *I*-1 and *I*-2 are assigned a value of 1 for 3-pyridyl- and 4-pyridylguanidines, respectively, *I*-3 is for the presence of a 6-substituent on the pyridyl, *I*-4 is for R where the number of carbon atoms in the longest chain ≤3 and one tertiary carbon atom is included, *I*-5 is for R where at least one ring is attached to the α carbon, and *I*-6 is for R where $-CH(CH_3)_2$ is present in the terminal. The values of those descriptor variables, except $(Nc)^2$, are listed in Table I along with the structural and biological data.

Table II shows the squared cross-correlation matrix of the descriptor variables. The correlation of Nc with $(Nc)^2$ is highly significant (R^2 = 0.935). For fear of any trouble caused by the high colinearity between Nc and $(Nc)^2$, ΔNc (=Nc − 5) and $(\Delta Nc)^2$ were tentatively used instead. The value of R^2 between ΔNc and $(\Delta Nc)^2$ was 0.112. However, the discriminant functions obtained were completely equivalent and the resultant classifications were identical in both cases. Therefore, Nc and $(Nc)^2$ were used in this study for simplicity.

The best discriminant function with the full data set was eq 16, which was derived at iteration 10 with α = 0.45 and

$$L = 0.121Nc - 0.014_8(Nc)^2 - 1.713V_H + 1.799(I\text{-}1) + 1.716(I\text{-}2) - 1.993(I\text{-}3) + 1.436(I\text{-}4) + 0.303(I\text{-}5) + 0.128(I\text{-}6) - 2.344 \quad (16)$$

$$n = 76,\ n_{mis} = 13\ (1),\ \epsilon = 1.380,\ R_S = 0.832\ (p < 0.001)$$

β = 0.04 in the adaptation (eq 10) and where n_{mis} is the number of compounds misclassified and the figure in parentheses after this is the number misclassified into the next class but one. The calculated classes are listed in Table I.

The iterative development of eq 16 is shown in Table III. The R_S and ϵ values become lowest at iteration 10.

Using ΔNc instead of Nc in the ALS calculation, we obtained eq 17 which is mathematically equivalent with eq 16.

$$L = -0.027_2\Delta Nc - 0.014_8(\Delta Nc)^2 - 1.713V_H + 1.799(I\text{-}1) + 1.716(I\text{-}2) - 1.993(I\text{-}3) + 1.436(I\text{-}4) + 0.303(I\text{-}5) + 0.128(I\text{-}6) - 2.111 \quad (17)$$

$$n = 76,\ n_{mis} = 13\ (1),\ \epsilon = 1.380,\ R_S = 0.832\ (p < 0.001)$$

On the basis of eq 16 (or 17), we can draw the following conclusions concerning the in vivo rat activity (therapeutic index) for structure I. (1) As for substituent R, the optimum number of carbon atoms included is 4 or 5 (4.1). The $Nc/(Nc)^2$ parabolic relationship may represent both hydrophobic and steric effects. Further, activity is enhanced by compact substituents including a tertiary carbon atom, a ring structure attached to the α carbon, and a terminal $-CH(CH_3)_2$ group. Activity is decreased by hydrophilic substituents. (2) As for the pyridyl group, 3-pyridyl is the most favorable, followed by 4-pyridyl. Activity is decreased in the presence of 6-substituents.

Since the ALS method makes no assumption about the underlying statistics of the data, generally statistical tests are not made like other pattern-recognition techniques. However, as for significance of results of discrimination, the nonparametric test[10] for R_S can be applied. In eq 16 (and 17), R_S was highly significant at $p < 0.001$. As for descriptor variables, the F test for significance of adding one additional independent variable in stepwise multiple regression analysis[17] may be practically utilized by use of the rank coefficient (R_S) in place of the regression coefficient in the case of a large data set, though we have not sufficient ground for arguing this. The F statistics for every descriptor variable used in eq 16 was greater than F (p = 0.01).

(11) W. J. Dixon, Ed., "BMD-Biomedical Computer Programs", University of California Press, Berkeley, Calif., 1973, p 221.
(12) S. Shiba, "Correlation Analysis in Behavioral Science", Tokyo University Press, Tokyo, 1967, p 234.
(13) S. Dove, R. Franke, O. L. Mndshojan, W. A. Schkuljev, and L. W. Chashakjan, *J. Med. Chem.*, **22**, 90 (1979).
(14) H. Ahrens and J. Läuter, "Mehrdimensionale Varianzanalyse", Akademie-Verlag, Berlin, GDR, 1974.
(15) P. C. Jurs and T. L. Isenhour, "Chemical Applications of Pattern Recognition", Wiley, New York, 1975, p 78.
(16) I. Moriguchi, Y. Kanada, and K. Komatsu, *Chem. Pharm. Bull.*, **24**, 1799, (1976).
(17) Y. C. Martin, "Quantitative Drug Design", Marcel Dekker, New York, 1978, p 376.

Table III. Iterative Development[a] of Equation 16

iteration time	intercept	Nc	$(Nc)^2$	V_H	I-1	I-2	I-3	I-4	I-5	I-6	n_{mis}	ϵ	R_S
1	−2.279	0.276	−0.029	−2.413	1.587	1.291	−2.236	1.384	0.500	0.386	26 (1)	1.839	0.670
2	−2.323	0.233	−0.026	−2.300	1.701	1.392	−2.186	1.409	0.531	0.431	24 (1)	1.737	0.687
3	−2.494	0.219	−0.026	−2.173	1.904	1.672	−2.107	1.400	0.508	0.431	18 (1)	1.526	0.765
4	−2.575	0.229	−0.026	−2.099	1.882	1.657	−2.085	1.426	0.511	0.405	18 (1)	1.525	0.766
5	−2.496	0.198	−0.023	−2.048	1.869	1.686	−2.070	1.425	0.493	0.355	18 (1)	1.526	0.765
6	−2.451	0.175	−0.021	−1.965	1.862	1.690	−2.069	1.444	0.443	0.318	17 (1)	1.492	0.779
7	−2.372	0.132	−0.016	−1.892	1.849	1.692	−2.065	1.476	0.451	0.278	20 (1)	1.612	0.738
8	−2.488	0.167	−0.019	−1.801	1.835	1.712	−2.051	1.492	0.393	0.213	14 (1)	1.406	0.820
9	−2.426	0.144	−0.016	−1.791	1.821	1.706	−2.023	1.460	0.366	0.164	17 (1)	1.518	0.777
10[b]	−2.344	0.121	−0.015	−1.713	1.799	1.716	−1.993	1.436	0.303	0.128	13 (1)	1.380	0.832
11	−2.283	0.090	−0.011	−1.694	1.784	1.675	−1.968	1.423	0.298	0.085	18 (1)	1.569	0.768
12	−2.217	0.065	−0.009	−1.576	1.756	1.713	−1.932	1.410	0.216	0.093	14 (1)	1.432	0.818
...	...	...	...	...	...	...	...	...	...	...	...	...	...

[a] Using $\alpha = 0.45$ and $\beta = 0.04$ in eq 10. [b] Equation 16.

Table IV. Compounds Misclassified by NDA and KNN (k = 1)

rating		compound no.	
obsd	calcd	NDA	KNN
1	2	**9**, 10, 13, 25, 27-29, 36, **38**, **49**, 51, 54, 74, 75	**49**, **54**, 71, 73
1	3		
2	1	**48**, 55, 56, 62, 68-70	5, 15, 39, **48**, 53, 61-63, 70
2	3	47, 61	6, 47
3	1		67
3	2	11, 21, 67	11, 21, 22, **59**, 60

Using the same descriptor variables, LDA yielded eq 18–20 with the full data set. The constants for the overall poorly active compounds

$$L(1) = 13.069\mathrm{Nc} - 1.025(\mathrm{Nc})^2 - 7.761V_H + 23.258(I\text{-}1) + 20.040(I\text{-}2) - 2.043(I\text{-}3) + 4.683(I\text{-}4) + 0.089(I\text{-}5) - 0.627(I\text{-}6) - 28.132 \quad (18)$$

moderately active compounds

$$L(2) = 14.987\mathrm{Nc} - 1.252(\mathrm{Nc})^2 - 19.890V_H + 27.281(I\text{-}1) + 24.213(I\text{-}2) - 6.589(I\text{-}3) + 4.279(I\text{-}4) + 3.526(I\text{-}5) - 0.022(I\text{-}6) - 34.568 \quad (19)$$

very active compounds

$$L(3) = 13.708\mathrm{Nc} - 1.074(\mathrm{Nc})^2 - 16.046V_H + 30.987(I\text{-}1) + 25.808(I\text{-}2) - 13.574(I\text{-}3) + 13.678(I\text{-}4) + 1.273(I\text{-}5) + 1.470(I\text{-}6) - 40.182 \quad (20)$$

discrimination were $n = 76$, $n_{mis} = 16$ (1), and $R_S = 0.790$ ($p < 0.001$). Since the covariances of the three groups were not equal, statistical tests based on Mahalanobis' D^2 or Hotelling's T^2 could not be done. The classification results for individual compounds are listed in Table I. The results were somewhat inferior to those by ALS method using a single equation (eq 16). Moreover, although information given by eq 18–20 is considered to be qualitatively similar to that by eq 16, they are quantitatively different. For example, the optimum number of carbon atoms of R estimated from the $\mathrm{Nc}/(\mathrm{Nc})^2$ parabolic relationship is 6 or 7 (6.0–6.4) by eq 18–20, whereas it is 4 or 5 (4.1) by eq 16. The fact[7] is that the compounds possessing the highest therapeutic index are **19** (R = t-C_4H_9) and **20** (R = t-C_5H_{11}).

If the so-called nonelementary discriminant analysis[14] (NDA) is used instead of LDA, the number of significant discriminant functions can be reduced and, consequently, interpretability of the model obtained may be improved. Dove et al.[13] successfully classified three groups of toluenesulfonylureas and toluenesulfonylthioureas using the first discriminant function derived by NDA. Therefore, using the same descriptors as those used in eq 16–20, the first discriminant function was calculated (eq 21). How-

$$L = 0.038\mathrm{Nc} + 0.002_1(\mathrm{Nc})^2 - 1.287V_H + 1.657(I\text{-}1) + 1.163(I\text{-}2) - 2.563(I\text{-}3) + 2.246(I\text{-}4) + 0.077(I\text{-}5) + 0.480(I\text{-}6) \quad (21)$$

$$n = 76,\ n_{mis} = 26\ (0),\ R_S = 0.694\ (p < 0.001)$$

ever, although the rank coefficient R_S was significant at $p < 0.001$, the values of R_S and n_{mis} were considerably inferior to those from both ALS and LDA calculations. It might be possible that the descriptors employed were unsuitable to NDA in the type (discrete or continuous) and selection. The results of discrimination using eq 21 are shown in Table IV. The results may be improved by using both the first and second discriminant functions, but, in-

Table V. Results of the Leave-Out Tests of the ALS Method and LDA

leave-out run	excluding compd no.	no. of training compds	indices of recognition: ALS n_{mis}	ALS ϵ	ALS R_S	LDA n_{mis}	LDA R_S	no. of predict. compds	no. of prediction errors: ALS	LDA
1	1, 6, 11, . . ., 71, 76	60	10 (1)[a]	1.298	0.810	14 (3)[a]	0.649	16	5 (0)[a]	5 (0)[a]
2	2, 7, 12, . . ., 72	61	9 (0)	1.176	0.888	9 (2)	0.803	15	5 (1)	6 (1)
3	3, 8, 13, . . ., 73	61	9 (1)	1.393	0.858	16 (1)	0.738	15	5 (0)	4 (0)
4	4, 9, 14, . . ., 74	61	11 (1)	1.413	0.821	13 (1)	0.786	15	3 (0)	4 (1)
5	5, 10, 15, . . ., 75	61	13 (1)	1.659	0.795	13 (1)	0.780	15	1 (0)	2 (1)
total								76	19 (1)	21 (3)

[a] The figures in parentheses are the number of compounds misclassified into the next class but one.

stead, the model of the structure–activity relationship may become less clear than that of the ALS method.

As already mentioned, the ALS method is considered to be a pattern-recognition technique. It would be interesting, therefore, to compare the results of the ALS method with those of any of other pattern-recognition techniques, though their applications have not been reported to the study of ordered categorical data of three or more groups. Thus, KNN (k = 1) was applied to the classification of the 76 compounds using the same descriptors. The results of KNN are listed in Table IV. The indices obtained are R_S = 0.747 and n_{mis} = 21 (1). These indices appear to be somewhat inferior to those from the ALS method, even though KNN is based on the leave one out prediction. Further, the weakest point with KNN is that KNN does not form any discriminant function describing a structure–activity relationship.

Validation of Discrimination Results. Validation of results is one of the most important aspects of any discrimination study. With the set of 76 hypotensive guanidines, a comparison in validity was made between the ALS method and LDA. Validation was done by leaving out every fifth compound and then classifying these 15 or 16 compounds on the basis of a discriminant function (or functions) derived from the training set composed of the remaining 61 or 60 compounds. The leaving out was then moved until each compound had been left out once and only once. Thus, a series of five leave-out trials was performed. The results are presented in Table V, and the predicted classes for individual substances are listed in Table I.

Indices obtained for the overall prediction were n_{mis} = 19 (1) and R_S = 0.775 with the ALS method, whereas n_{mis} = 21 (3) and R_S = 0.671 with LDA. A higher prediction rate of the compounds left out using the ALS method indicates that the ALS method gives a more stable class structure than LDA. In the five leave-out runs, the higher stability of the ALS method was also shown in the recognition: R_S = 0.795–0.888 for the ALS method whereas R_S = 0.649–0.803 for LDA.

In conclusion, the ALS method is considered to serve as a powerful tool for structure–activity studies in both recognition and prediction of ordered multicategory discrimination.