Chem. Pharm. Bull.
33(4)1488—1495(1985)

Application of Principal Component Analysis to the Study of Quantitative Structure–Activity Relationships by Means of Multiple Regression Analysis[1,2)]

TANEKAZU KUBOTA,* JIRO HANAMURA, KENJI KANO, and BUNJI UNO

Gifu Pharmaceutical University, 6–1, Mitahora-higashi 5-chome, Gifu 502, Japan

(Received July 19, 1984)

Some important problems in the application of multiple regression analysis (MRA) to the study of quantitative structure–activity relationships (QSAR) are the effect of so-called collinearity among the explaining variables on MRA and the chance correlation. In order to reduce these effects on MRA we have here employed a combination of principal component analysis (PCA) and MRA. Firstly all the explaining variables (x_i) are normalized to the zero mean and one variance (x_i'), then converted to zero correlation coefficient by using the technique of PCA. Principal component scores pertinent to each principal component (Z_m) were next calculated, and MRA was carried out with a linear combination of Z_m's. Important Z_m's can easily be identified by applying the character of zero correlation coefficient among the variables. The above multiple regression equation is rewritten as a linear combination of x_i' or x_i by using the transformation matrix between Z_m and x_i'. This type of equation also seems to be useful for the purpose of predicting new drug structures. Actual calculation results are presented for some drug series. Finally, classification of the explaining variables was done by focusing on the factor loading values of the variables.

Keywords——principal component analysis; principal component regression; multiple regression analysis; quantitative structure–activity relationship; explaining variable; classification by factor loading; antileukemic activity; antibacterial activity; 2,5-bis(1-aziridinyl)-*p*-benzoquinone; quinoline derivative

The study of quantitative structure–activity relationships (QSAR) is important in relation to new drug research, and various kinds of approaches, such as physical, physicochemical, quantum-chemical, biological, and statistical treatments, have been applied for this purpose.[3,4)] One of the techniques frequently used in QSAR studies is multiple regression analysis (MRA), in which a linear combination of explaining variables (x) is employed to describe the biological activities (y: dependent variable) of drugs, as given by Eq. 1.[5)]

$$\hat{y}_\alpha = b_0 x_0 + \sum_{i=1}^{p} b_i x_{i\alpha} \tag{1}$$

where $\hat{y}_\alpha$ is the predicted value of the observed biological activity of the α-th drug y_α and estimated from the right-hand-side terms of Eq. 1, $b_0 x_0$ being the constant term (x_0 is a dummy variable always having the value 1). On the basis of general mechanisms of drug activity in biological systems, various kinds of physicochemical and quantum-chemical parameters have been proposed for x's. Among them, $\log P$ (P: partition coefficient) or π (Hansch–Fujita hydrophobic parameter) is the most popular as a descriptor of transport processes.[6,7)] The accuracy and significance of $\hat{y}$ and the partial regression coefficient b_i are usually assessed in terms of the multiple correlation coefficient (r), F-test, t-test, *etc.*[8)] However, if the variables for describing the drug transport processes and the electronic and steric terms pertinent to the drug–receptor interactions are quite large in number compared to

the number of test drugs, then the problem of so-called chance correlation, *i.e.*, physically meaningless correlation, may occur.[3,4,6,8-10] Further, collinearity among the explaining variables for MRA should be avoided.[3,4] It is thus desirable that the extra information quantities contained in the explaining variables employed for MRA should be removed and the correlation among the variables should be small or zero, *i.e.*, mutually independent variables.[3,4,6,8-12] In order to satisfy these conditions for overcoming as far as possible the above problems and also to make the selection and classification of the variables easier, one of the authors (T. K.) proposed in early reviews the usefulness of the technique[13] of the combination of principal component analysis (PCA) and MRA.[8,11,13] By this approach, important explaining variables seem to be easily selected. In this paper, actual examples of the PCA–MRA calculation and the advantages of this technique are reported in detail and applied to the QSAR of some drug series.

Calculation Methods

Calculation of Principal Component Analysis——Although the mathematical treatment of PCA and its application to MRA are well known in the field of multivariate analysis in mathematics,[13] the actual application of this technique to the investigation of QSAR has hardly been attempted as far as we know, except for ref. 12 (see the later discussion). Thus, the theoretical background for PCA–MRA is dealt with briefly here to aid an understanding of the actual calculation results (*vide infra*). Now, before PCA calculation the explaining variables $x_{i\alpha}$ are firstly normalized to zero mean and one variance by transforming the $x_{i\alpha}$ to $x'_{i\alpha}$ according to Eq. 2.

$$x'_{i\alpha}=\frac{(x_{i\alpha}-\bar{x}_i)}{\sqrt{V_{ii}}} \tag{2}$$

where $x_{i\alpha}$, $\bar{x}_i$, and V_{ii} mean the i-th explaining variable for the α-th drug, the mean value of x_i, and the variance of the variable x_i, respectively. Since $x_{i\alpha}$ having different dimensions are frequently employed as explaining variables in QSAR studies, the transformation to the dimensionless $x'_{i\alpha}$ is very convenient from the viewpoint of interpretation of calculation results. In addition, the variance and covariance matrix ($\boldsymbol{V}$) of $x'_{i\alpha}$ comes out equal to the correlation matrix ($\boldsymbol{R}$) of $x_{i\alpha}$. Resolving the determinant $|\boldsymbol{R}-\lambda\boldsymbol{E}|=0$ under the conditions of normalization and orthogonality, we can easily obtain the eigenvalue λ_m, the corresponding coefficient ψ_m, and the principal component Z_m, where m runs from 1 to p, p being the number of explaining variables. We can now write ψ_m as $\psi_m=\sum_{i=1}^{p} l_{mi}$, and the principal component Z_m and the score $Z_{m\alpha}$ pertaining to the α-th drug of Z_m are now given by $Z_m=\sum_{i=1}^{p} l_{mi}x'_i$ and $Z_{m\alpha}=\sum_{i=1}^{p} l_{mi}x'_{i\alpha}$, respectively. The $Z_{m\alpha}$ is also easily rewritten as Eq. 3[12] by using the factor loading r_{mi} that corresponds to the correlation coefficient between principal component Z_m and the explaining variable x'_i. PCA theory also tells us that

$$Z_{m\alpha}=\lambda_m^{-1/2}\sum_{i=1}^{p} r_{mi}x'_{i\alpha} \tag{3}$$

the relations, $v_i=\sum_{m=1}^{p} r_{mi}^2=1$, $\sum_{i=1}^{p} r_{mi}^2=\lambda_m$, $\sum_{m=1}^{p}\lambda_m=p$, and $\sum_{i=1}^{p} v_i=p$, should be satisfied. These mutual correlations due to the PCA theory are important for the present purpose for the following reasons: (i) the correlation coefficient between the scores of any two principal components Z_m and Z_k becomes zero; (ii) the variance of Z_m is equal to λ_m; (iii) $(\lambda_m/p)\cdot 100$ gives the percent contribution of Z_m to the total information quantity p; (iv) the value $\sum_{m=1}^{n} r_{mi}^2\cdot 100$ gives the total contribution of Z_1, Z_2, $\cdots Z_n$ to the information quantity of x'_i, *i.e.* 1, since the variance $V(x'_i)=1$. As is clear from the above discussion the factor loading r_{mi} plays an important role in PCA, and we can easily classify the explaining variables x'_i on the basis of the factor loading contributing to each of the principal components Z's (see later sections).

Application of Principal Component Analysis to Multiple Regression Analysis——Since the correlation coefficient between the scores $Z_{m\alpha}$ and $Z_{k\alpha}$ of any two principal components Z_m and Z_k is zero, the application of PCA to MRA is very useful for the actual analysis of QSAR. We can now write Eq. 4 instead of Eq. 1 by taking a linear combination of Z's,

$$\hat{y}_\alpha = a_0 Z_0 + \sum_{m=1}^{p} a_m Z_{m\alpha} \quad (4)$$

where $a_0 Z_0$ is the constant term (Z_0 is a dummy variables always having the value 1). Since the correlation between any two Z's is zero, the partial regression coefficient a_m in Eq. 4 turns out to be the same as that given by the simple regression Eq. 5. Further, this a_m in Eq. 5 does not change with the addition of the other terms $Z_{p\alpha}$ where p runs from 1 to p except m. Moreover, it is very convenient that the a_m in Eq. 5 is simply determined by Eq. 6.[13] These

$$\hat{y}_\alpha = a_0 Z_0 + a_m Z_{m\alpha} \quad (5)$$

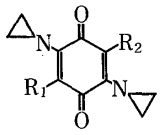
$$a_m = \frac{S_{my}}{S_{mm}} = \frac{\sum_{\alpha=1}^{n} (Z_{m\alpha} - \bar{Z}_m)(y_\alpha - \bar{y})}{\sum_{\alpha=1}^{n} (Z_{m\alpha} - \bar{Z}_m)^2} \quad (6)$$

circumstances make it very useful to adopt Eq. 4 instead of Eq. 1 from the viewpoint of MRA, and the important Z values contributing to Eq. 4 can be easily identified by focusing attention on the partial regression coefficients (a_m: $m=1$—p), the factor loading of the principal component, the correlation coefficients between Z_m and y and also between y and $\hat{y}$, and so on.[14] The significance level of Eq. 4 itself and of individual partial regression coefficients can be established by applying the so-called F-test and Students' t-test, respectively.[8] Usual examination based on the above $r_{y\hat{y}}$ and standard deviation is also very effective. In order to apply Eq. 4 to drug design, the form of Eq. 1 seems to be sometimes more convenient than that of Eq. 4, since the contribution of each explaining variable to the observed drug activity y_α is easily understood from each coefficient of Eq. 1 This kind of transformation from Eq. 4 to Eq. 1 is simply carried out as follows: the aforementioned relation $Z_{m\alpha} = \sum_{i=1}^{p} l_{mi} x'_{i\alpha}$ is introduced into the term $Z_{m\alpha}$ in Eq. 4, then the equation is rewritten as a function of x'_i (*vide infra*).[14d]

All the calculations written in this paper were performed on NEC PC-8001 and PC-9801 personal computers. The "basic" program pertinent to the principal component analysis was rewritten by us from that given in the literature[15] and combined with the multiple regression analysis programs used hitherto in our laboratory.

Results and Discussion

In this report on PCA–MRA calculation we will mainly deal with two examples, which have already been described by other authors in connection with QSAR studies by MRA. One example is the data on antileukemic activities of 2,5-bis(1-aziridinyl)-*p*-benzoquinones (I) reported by Yoshimoto,[16] where compounds with various substituents R_1 and R_2 were tested



I

for biological activities. Antileukemic activities adopted for the present computation were those in single injection against lymphoid leukemia L-1210 in BDF_1 mice. Here, two kinds of data, the minimum effective dose (MED) giving a 40% increase in life span (ILS) compared to the controls and the optimal dose (OD) giving maximum ILS, were subjected to PCA–MRA calculation. For the y corresponding to the predicted value $\hat{y}$ (see Eqs. 1, 4, 5), $\log(1/c)$ was employed, c (mol/kg) being the MED or OD. The parameter values pertinent to the explaining variables x were finally listed up as MR_1, $MR_{1,2} = MR_1 + MR_2$, π_2, $\pi_{1,2} = \pi_1 + \pi_2$, $F = F_1 + F_2$, and $R = R_1 + R_2$ for the total 35 and 37 compounds for MED and OD, respectively. We applied all these data to the PCA–MRA. The details are given for the results of OD-single injection, since the outcomes for the MED-single injection were almost the same as those for OD. Table I shows the calculation results of eigenvalue λ_m, principal component Z_m, and the factor loading r_{mi}, where CR and CCR are the information contribution ratio

TABLE I. Calculation Results of Eigenvalues (λ_m), Eigenvector (ψ_m) Corresponding to Principal Components (Z_m), and Factor Loading (r_{mi}) for the Antileukemic Activities (OD Data[a]) of 2,5-Bis(1-aziridinyl)-*p*-benzoquinones[16,b]

$m\rightarrow$ / $i\downarrow$		Z_1	Z_2	Z_3	Z_4	Z_5	Z_6
	λ_m	2.615	1.366	1.151	0.499	0.310	0.060
				Eigenvectors: ψ_m			
x_1	$MR_{1,2}$	0.450	0.180	−0.471	−0.119	−0.727	−0.028
x_2	$\pi_{1,2}$	0.532	−0.339	0.234	0.112	0.103	−0.724
x_3	π_2	0.450	−0.409	0.424	0.079	−0.135	0.652
x_4	MR_1	0.409	−0.070	−0.606	−0.101	0.637	0.212
x_5	F	−0.274	−0.530	−0.398	0.680	−0.154	0.002
x_6	R	0.264	0.632	0.133	0.704	0.116	0.066
				Factor loading: r_{mi}			
x_1	$MR_{1,2}$	0.728	0.210	−0.505	−0.084	−0.405	−0.007
x_2	$\pi_{1,2}$	0.860	−0.397	0.251	0.079	0.057	−0.177
x_3	π_2	0.728	−0.477	0.455	0.056	−0.075	0.159
x_4	MR_1	0.661	−0.082	−0.650	−0.071	0.355	0.052
x_5	F	−0.443	−0.620	−0.427	0.480	−0.086	0.001
x_6	R	0.427	0.739	0.142	0.497	0.065	0.016
	CR[c]	43.58	22.76	19.19	8.31	5.17	0.99
	CCR[c]	43.58	66.34	85.53	93.84	99.01	100.00

a) See the text for details. *b*) The notation x_i means the i-th physical constant from the top given in this table. This notation is used throughout this paper. *c*) $CR=(\lambda_m/p)\times 100$ and $CCR=(\sum_m \lambda_m/p)\times 100$ mean the contribution ratio and cumulative contribution ratio of the information quantity concerning the explaining variables, respectively. In the present case $p=6$.

TABLE II. Pearson Correlation Matrices among the Raw Data[16] of Six Explaining Variables and Those among Six Principal Components Plus log $(1/c)$[16,a]

		$MR_{1,2}$	$\pi_{1,2}$	π_2	MR_1	F	R
x_1	$MR_{1,2}$	1.0000					
x_2	$\pi_{1,2}$	0.3867	1.0000				
x_3	π_2	0.2243	0.9018	1.0000			
x_4	MR_1	0.6546	0.4431	0.2024	1.0000		
x_5	F	−0.2422	−0.2093	−0.1881	−0.0290	1.0000	
x_6	R	0.3262	0.1505	0.0489	0.1176	−0.4749	1.0000

	log $(1/c)$	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6
log $(1/c)$	1.0000						
Z_1	−0.6004	1.0000					
Z_2	0.3000	0.0000	1.0000				
Z_3	0.0575	0.0000	0.0000	1.0000			
Z_4	−0.6242	0.0000	0.0000	0.0000	1.0000		
Z_5	−0.0085	0.0000	0.0000	0.0000	0.0000	1.0000	
Z_6	0.0306	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000

a) See the title and footnote *b*) in Table I.

$(100\cdot\lambda_m/p(=6))$ and the cumulative contribution ratio $(100\cdot(\sum_m\lambda_m)/p)$, respectively. Table II lists the correlation matrices among the 6 variables before PCA and among the 6 principal components after PCA, plus log$(1/c)$. It is clear from Table II that the correlation between

TABLE III. Calculation Results of Multiple Regression Analysis Applied to the Principal Component Scores with Various Selections of Number of Principal Components[a]

$$\log(1/c)^{b)} = -0.220Z_1 + 0.152Z_2 + 0.032Z_3 - 0.524Z_4 - 0.009Z_5 + 0.074Z_6 + 4.724$$

PC[c]	$m^{d)}$ $(n-m-1)$	$t_{n-m-1,0.05} \cdot s\sqrt{c^{ii}}$ [e]						$s^{f)}$	$r^{f)}$	$F^{f)}$	$CCR^{g)}$
		Z_1	Z_2	Z_3	Z_4	Z_5	Z_6				
1—6	6 (30)	0.054	0.075	0.081	0.123	0.157	0.357	0.260	0.9189	27.129	100.00
1	1 (35)	0.101						0.487	0.6004	19.727	43.58
2	1 (35)		0.166					0.581	0.3000	3.460	22.76
4	1 (35)				0.225			0.476	0.6242	22.342	8.31
1, 2	2 (34)	0.095	0.131					0.458	0.6711	13.933	66.34
1, 4	2 (34)	0.063			0.146			0.309	0.8660	51.025	51.90
2, 4	2 (34)		0.127		0.211			0.446	0.6925	15.668	31.07
1, 2, 4	3 (33)	0.052	0.072		0.119			0.251	0.9166	57.779	74.66
1—4	4 (32)	0.052	0.072	0.079	0.120			0.252	0.9184	43.076	93.84
1—4, 6	5 (31)	0.053	0.073	0.080	0.121		0.351	0.256	0.9189	33.621	94.84

a) Principal component scores are calculated based upon the results given in Table I. *b*) Regression equation applied to the six principal components. Note that each coefficient is the same as that given by the simple regression equation (see the text). *c*) Principal components used for multiple regression analysis. *d*) "m" is the number of principal components, and "$n-m-1$" in parentheses is the degree of freedom of the regression equation, n being the number of samples. *e*) "$t_{n-m-1,0.05} \cdot s\sqrt{c^{ii}}$ means the 95% confidence region in the t-test, the value of which is required to be less than the value of the partial regression coefficient.[8] Here, $s\sqrt{c^{ii}}$ corresponds to the estimated standard error for the coefficient.[8] *f*) The values of s, r, and F are for the standard deviation, correlation coefficient, and F-value in F-test, respectively. *g*) See footnote *c*) in Table I.

two principal components is zero, so that the important principal components contributing to Eq. 4 are Z_1, Z_2, and Z_4, because the correlation coefficients (r_{yz}) of these three Z values against $\log(1/c)$ are considerably larger than for the other Z's. Recalling now that the above $|r_{yz}|$ is equal to the $r_{y\hat{y}}$ from Eq. 5, and also keeping in mind that the $r_{y\hat{y}}$ of Eq. 4 is written as the Pythagorean sum of the simple correlation coefficient r_{yz},[13,14] the $r_{y\hat{y}}$ for the total contribution from the above Z_1, Z_2, and Z_4 is straightforwardly calculated as $[(-0.6004)^2+(0.3000)^2+(-0.6242)^2]^{1/2}=0.9166$. These results of the regression analysis as well as those obtained from the other combination of the Z's are listed in Table III, in which the calculation results at 95% confidence in the t-test are also given for the partial regression coefficient. It is clear from Table III that the principal components Z_1, Z_2, and Z_4 make an important contribution to y ($\log(1/c)$) and the t-test is significant. However, the other Z values (*i.e.* Z_3, Z_5, Z_6) are rejected in the t-test and are not so important in the prediction of y. The $r_{y\hat{y}}$ value pertinent to the case where Z_1, Z_2, and Z_4 are used is 0.9166 and almost the same as the value of 0.9189 corresponding to the adoption of all the Z values. In addition, the evaluation based on the s and F values is rather better for the former than for the latter case. Thus, we may say that the six explaining variables used originally can be reduced to the three principal components without changing the predictive ability for biological activity. Thus, the information quantity (*vide ante*) is decreased from 100% for all the Z's to 74.65% for the above three Z's, the latter being easily divided into $R=16.26\%$ [$(0.427^2+0.739^2+0.497^2)\times 100=97.55/6$; see Table I], $\pi_{1,2}=15.06\%$, $F=13.52\%$, $\pi_2=12.68\%$, $MR_{1,2}=9.69\%$, and $MR_1=7.48\%$. The importance of the electronic and hydrophobic variables is well understood. From the viewpoint of drug design, however, the transformation from Eq. 4 to Eq. 1 is also very useful and can be carried out by the technique described in the foregoing section. The transformation equation for the present case (Eq. 7, see Table III) is now given by Eq. 8, which is exactly equivalent to Eq. 7 for predicting y. Here, x_i' is the

$$\log(1/c) = -0.2201Z_1 + 0.1522Z_2 - 0.5240Z_4 + 4.7238 \quad (7)$$

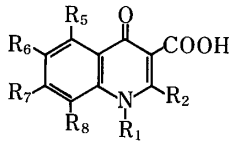
$$\log(1/c) = -0.0094x'_1 - 0.2273x'_2 - 0.2026x'_3 - 0.0477x'_4 - 0.3765x'_5 - 0.3307x'_6 + 4.7238 \quad (8)$$

i-th normalized variable according to Eq. 2; see Table I for the physical meaning pertaining to the x'_i. Again, the importance of the F', R', $\pi'_{1,2}$, and π'_2 terms is easily understood from the corresponding regression coefficient of Eq. 8. The conversion of Eq. 8 to the one with the original raw variables could be done by applying Eq. 2, and thus we get Eq. 9 from Eq. 8. Comparing Eq. 8 with Eq. 9 we can clearly see that the regression coefficients are

$$\log(1/c) = -0.0084x_1 - 0.1922x_2 - 0.2252x_3 - 0.0735x_4 - 3.1719x_5 - 1.4457x_6 + 4.6958 \quad (9)$$

quite different, particularly for the x_5 and x_6 descriptors, although the calculated values of $\log(1/c)$ are the same for both equations. This depends entirely on whether the explaining variables are normalized by Eq. 2 or not. For the purpose of drug design the expression of Eq. 8 is suitable, where the important variables are directly reflected in the regression coefficients. However, strictly speaking, Eq. 9 is a better type than Eq. 8 for calculating the $\log(1/c)$ of new compounds before experiments, because the values of $\bar{x}_i$ and $\sqrt{V_{ii}}$ in Eq. 2 would alter on adding the new compounds to the original data though the change is very small.[17]

Next, let us focus attention on the so-called "eigenvalue larger than one" criterion, which is frequently applied to PCA,[13] since the larger the eigenvalue, the larger is the information quantity from the principal component. Very recently Lukovits[12] has also reported the application of PCA to MRA independently from us, and he emphasized the usefulness of the above "eigenvalue larger than one" criterion. However, our calculations, including the results in this paper, have revealed many examples where the scores pertinent to the principal components with eigenvalues less than 1 also have a good correlation to the drug activity $\log(1/c)$. This seems to be quite reasonable because we have applied the PCA to the explaining variables alone, *i.e.* $\log(1/c)$ is excluded from PCA, so that there is a possibility of a good correlation of $\log(1/c)$ to some principal component scores that do not have large information quantities (*vide infra*). An example is the case of the application of our present method to the QSAR data reported by Koga.[18] His original treatment is as follows. In order to analyze the antibacterial activity $\log(1/c)$ (c: minimum inhibitory concentration) of a total of 71 samples with various substituents at the R_1, R_2, R_5, R_6, R_7, and R_8 positions of 1-substituted-1,4-dihydro-4-oxo-quinoline-3-carboxylic acid derivatives (II) he firstly selected total 50 explaining variables including hydrophobic, electronic, and steric parameters. After various preliminary



II

trials he applied the stepwise method for selecting the variables in MRA by focusing attention on the correlation coefficient to the $\log(1/c)$, the physical meaning of each variable, and the independency between explaining variables. As a result he finally derived 11 variables, use of which gave a good regression equation for $\log(1/c)$ with $n=71$, $s=0.274$, $r=0.964$. We have now carried out MRA after applying PCA to the above mentioned 11 explaining variables. The following results were obtained. When all 11 principal component values are taken into the calculation of MRA the resultant 11 a_m values in Eq. 4 are all significant in the t-test at the 95% confidence level, although for a simple regression equation (Eq. 5) the a_m values passing

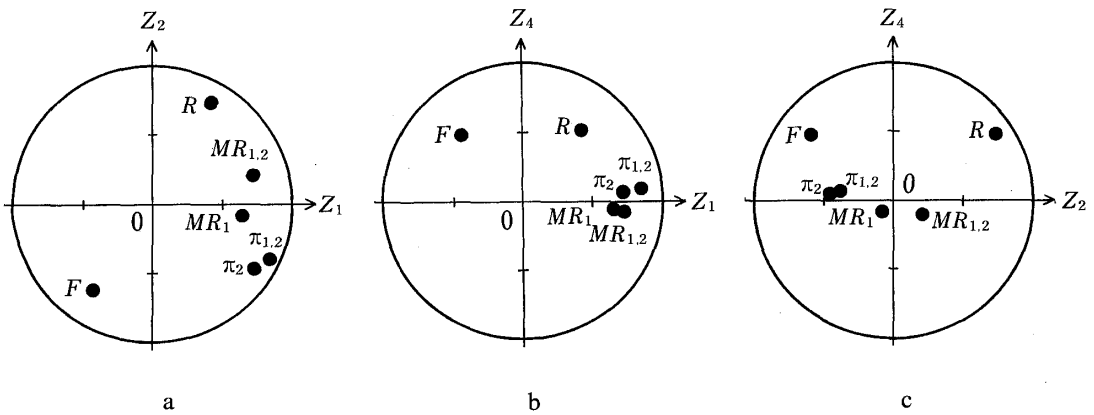
the t-test are those for Z_1, Z_4, Z_5, Z_7, and Z_{10}. Note here that the principal components having an eigenvalue larger than one are only Z_1, Z_2, Z_3, Z_4, and Z_5. These results may imply that there are no principal components which do not make an important contribution to $\log(1/c)$ from the viewpoint of the t-test, and all 11 variables selected by the above "stepwise method" are suitable and significant, so that we cannot delete any one of the 11 variables for the purpose of the multiple regression treatment. Therefore, we may say that the criterion of "eigenvalue larger than one" is not suitable.

Characterization of the Explaining Variables by Means of Factor Loading and Principal Component Scores

Since the square of factor loading, r_{mi}^2, corresponds to the contribution of Z_m to the information quantity of x_i', the correlation map of the factor loading values between any two different Z_m's may permit the classification of explaining variables into groups having similar characters. Examples are shown in Fig. 1a—c for the case of Eq. 7. Note that all the values pertinent to the explaining variables should fall into the circle of radius 1 because $\sum_{m=1}^{p} r_{mi}^2 = 1$. Therefore, the variables occupying positions near the circumference would make a large contribution to the principal components used as the two axes (Z_m, Z_n), since the square of the position vector $\vec{r}_i$ from the origin to a point (r_{mi}, r_{ni}) in the circle is given by $r_i^2 = r_{mi}^2 + r_{ni}^2$. As can be understood from Fig. 1a—c, the variables MR_1 and π_2 are in a closer position to $MR_{1,2}$ and $\pi_{1,2}$, respectively, so that these two pairs of variables seem to be quite similar in nature (see also Table II) and are mainly localized in the Z_1 component. On the other hand, the variables F and R occupy quite separate positions from each other, as Fig. 1a—c shows, indicating that the F and R values have quite different properties (see Table II). Also, we can see in Fig. 1 that the F and R contribute largely to the Z_1, Z_2, and Z_4, and in particular the component Z_4 consists mainly of the F and R variables alone with the same sign.

In turn, it is also worthwhile to plot the two principal component scores $Z_{m\alpha}$ and $Z_{n\alpha}$ in rectangular coordinates. The correlation coefficient between Z_m and Z_n is in principle zero, so that the scores of $Z_{m\alpha}$ and $Z_{n\alpha}$ ($\alpha = 1 \cdots n$) should be scattered at random. However, if there are extraordinarily separated points, or if the scores show some regular relationships in the figure, which would cancel out to make zero correlation coefficient, reevaluation of the explaining variables of the samples in question might be necessary.

In conclusion we would say that many explaining variables with physical meaning in QSAR studies can initially be selected unless special attention is paid to the correlation among the variables. After the variables are normalized by Eq. 2, PCA is applied to the descriptors



a b c

Fig. 1. Correlation Map of the Factor Loading Values between Two Principal Components

a, b, and c are respectively between the two axes of Z_1 and Z_2, Z_1 and Z_4, and Z_2 and Z_4 given in Table I.

x_i'. The scores ($Z_{m\alpha}$) of each principal component (Z_m), the factor loading (r_{mi}), and the transformation coefficient (ψ_m) are next calculated. The multiple regression analysis is now carried out using the Z_m's to predict the drug activity y_α, then the Z_m's making an important contribution to the MRA are identified. For the purpose of new drug design, it is convenient to rewrite the above regression equation in the form of a linear combination of x_i' or x_i by employing the transformation coefficient ψ_m.

References and Notes

1) Presented at the 10th Symposium on the Relationship between Structure and Activity, Kyoto, December 1983, and the 104th Annual Meeting of the Pharmaceutical Society of Japan, Sendai, March 1984, Abstract p. 704.
2) This paper forms Part V of "Studies on Structure–Activity Relationships;" Part IV: J. Hanamura, K. Kobayashi, K. Kano, and T. Kubota, *Chem. Pharm. Bull.*, **31**, 1357 (1983).
3) "Structure–Activity Relationship," Kagakuno Ryoiki, Zōkan, Vol. 122, ed. by the Research Group of Structure–Activity Relationship, Nankodo Press, Tokyo, 1979, (Review, in Japanese).
4) "Structure–Activity Relationship," Kagakuno Ryoiki, Zōkan, Vol. 136, ed. by the Research Group of Structure–Activity Relationship, Nankodo Press, Tokyo, 1982, (Review, in Japanese).
5) "Biological Correlations—The Hansch Approach," Adv. Chem. Ser., Vol. 114, Am. Chem. Soc., 1972.
6) W. P. Purnell, G. E. Bass, and J. M. Clayton, "Strategy of Drug Design—A Molecular Guide to Biological Activity," John Wiley & Sons, New York, 1973.
7) Various kinds of Hammett-type substituent constants as well as $\log P$ or π were first introduced in Eq. 1 by Hansch and Fujita, this treatment being called the Hansch approach.[5]
8) T. Kubota, see p. 43 of ref. 3. Also, see refs. 11, 13*a*, *b* and the introduction in ref. 13*c*, where the effect of the mutual correlation among explaining variables on the number of them used for MRA and also on the stability of the value of each partial regression coefficient is discussed in detail.
9) J. K. Topliss and R. J. Costello, *J. Med. Chem.*, **15**, 1066 (1972).
10) J. G. Topliss and R. P. Edwards, *J. Med. Chem.*, **22**, 1238 (1979).
11) T. Kubota, "Structure–Activity Relationship," Kagakuno Ryoiki, Zōkan, Vol. 122, 2nd ed., ed. by the Research Group of Structure–Activity Relationship, Nankodo Press, Tokyo, 1980, p. 405 (Review, in Japanese).
12) I. Lukovits, *J. Med. Chem.*, **26**, 1104 (1983).
13) *a*) T. W. Anderson, "Introduction to Multivariate Statistical Analysis," John Wiley & Sons, Inc., New York, 1970; *b*) T. Okuno, H. Kume, T. Haga, and T. Yoshizawa, "Multivariate Analysis," Nikkagiren Press, Tokyo, 1978 (in Japanese); *c*) T. Okuno, T. Haga, K. Yajima, C. Okuno, S. Hashimoto, and Y. Furukawa, "Multivariate Analysis—Continued," Nikkagiren Press, Tokyo, 1978 (in Japanese).
14) It should be noted[13] that (a) the a_0 in Eqs. 4 and 5 is the same as and equal to the mean value $\bar{y}$ of the observed biological activities of the total n drugs; (b) the multiple correlation coefficients $r_{y\hat{y}}$ for Eq. 4 are expressed by $r_{y\hat{y}}=\left[\sum_{m=1}^{p}(r_{m\cdot y\hat{y}})^2\right]^{1/2}$, where $r_{m\cdot y\hat{y}}$ is for the m-th simple correlation coefficient given by Eq. 5 and the summation is made over the number of Z_m in Eq. 4; (c) a dummy variable having, for example, a value of 1 or zero alone can be safely used as an explaining variable in Eq. 4, there being mathematically no problem in adopting this descriptor; and (d) however, it appears that each coefficient b_i of the type of Eq. 1, which is converted from Eq. 4 by using the transformation matrix (see the text), does not correspond to the unbiased estimate for the partial regression coefficient unless the above transformation is made by employing all the principal components Z_m ($m=1$—p).
15) T. Miyazawa, *ASCII*, **4**, 66 (1980). Note that a number of errors were found in the original version. Our program is available on request.
16) M. Yoshimoto, see p. 160 of ref. 3.
17) For the case of adopting all the principal components, the equations corresponding to Eqs. 7, 8, and 9 are respectively as follows with $n=37$, $s=0.2597$ and $r=0.9189$.

$$\log(1/c)=-0.2201Z_1+0.1522Z_2+0.0318Z_3-0.5240Z_4-0.0090Z_5+0.0743Z_6+4.7238 \quad (1)'$$

$$\log(1/c)=-0.0198x_1'-0.2746x_2'-0.1395x_3'-0.0569x_4'-0.3876x_5'-0.3226x_6'+4.7238 \quad (2)'$$

$$\log(1/c)=-0.0177x_1-0.2322x_2-0.1550x_3-0.0878x_4-3.2654x_5-1.4104x_6+4.7554 \quad (3)'$$

Here, Eqs. 2′ and 3′ are derived from Eq. 1′ by the technique described in the text. Of course, Eq. 3′ agrees completely with the one directly obtained from the MRA of the raw data given in the literature.[16]
18) H. Koga, see p. 177 of ref. 4.