ELSEVIER

Journal of Pharmaceutical and Biomedical Analysis
15 (1997) 431–438

JOURNAL OF
PHARMACEUTICAL
AND BIOMEDICAL
ANALYSIS

Inclusion of the standard deviation of data in principal component analysis. A graphical approximation[1]

Sylvia Wallerstein, Tibor Cserháti*, Esther Forgács, Veronika Kiss

Central Research Institute for Chemistry, Hungarian Academy of Sciences, P.O. Box 17, 1525 Budapest, Hungary

Received for review 26 January 1996

Abstract

The adsorption capacity and specific adsorption surface area of 13 anti-hypoxia drugs were determined in three chromatographic systems using methanol–carbon tetrachloride, chloroform–carbon tetrachloride and acetonitrile–carbon tetrachloride mixtures as eluents. The retention behaviours of the anti-hypoxia drugs were compared using principal component analysis (PCA). A graphical approximation was used for the inclusion of the standard deviations of both the variables and observations in PCA and the results were visualized by two-dimensional nonlinear mapping and cluster analysis. The results indicated that the graphical approximation can be successfully used for the inclusion of the standard deviation of data in PCA calculations. Nonlinear mapping and cluster analysis resulted in similar, but not identical, classification of drugs and chromatographic systems, indicating that each multivariate method can be successfully used for the comparison of solutes and chromatographic systems.

Keywords: Cluster analysis; Graphical approximation; Principal component analysis

1. Introduction

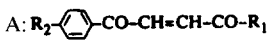
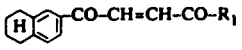
The development of new calculation methods to interpret large retention data matrices has been one of the major advances in chromatography during the last decade. The most notable features of this rapidly evolving field are: various automated chromatographic instruments, high-speed computers, and a variety of mathematical–statistical methodologies such as stepwise regression analysis [1,2], factor analysis [3,4], principal component analysis [5,6], canonical correlation analysis [7,8], spectral mapping techniques [9,10], etc. The evalution of data sets containing a considerable amount of information (i.e. retention times of a great number of compounds determined under several chromatographic conditions) is practically impossible using the traditional linear regression model. The modern multivariate mathematical–statistical methods make possible the

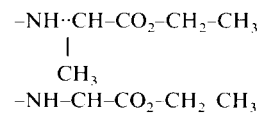
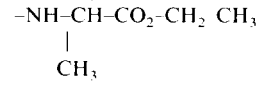
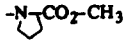
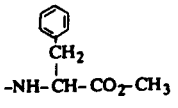
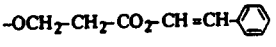
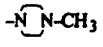
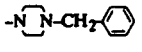
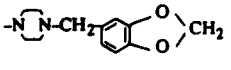
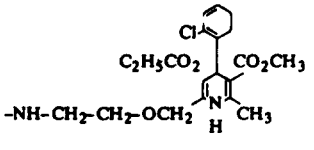
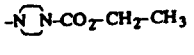
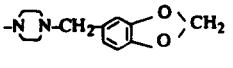
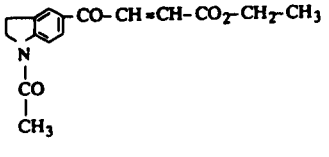
* Corresponding author.

[1] Presented at the Fifth International Symposium on Drug Analysis, September 1995, Leuven, Belgium.

PII S0731-7085(96)01862-6

Table 1
Chemical structures of anti-hypoxia drugs. General structures:

A:  ; B: 

Compound number	General structure	R_1	R_2
1	A		H
2	A		H
3	A		OCH_3
4[a]	A		H
5	A		H
6	A		H
7	A		H
8	A		H
9	A		H
10	B	–OH–	
11	B		
12	B		
13			

[a] Saturated bond in the butenoic side-chain.

Table 2
Parameters of the linear relationship between the methanol concentration in the methanol–carbon tetrachloride eluent mixture (C vol. %) and the R_M values of anti-hypoxia drugs. Numbers refer to the anti-hypoxia drugs in Table 1. ($R_M = a + b \cdot C$)

Parameter	Compound No.						
	1	2	3	4	5	6	7
n	12	12	12	12	12	10	12
R_{M0}	1.00	1.31	0.97	1.08	2.28	1.69	1.11
$-b$ ($\times 10$)	3.98	4.39	4.05	4.12	2.84	3.10	5.04
s_b ($\times 10^2$)	5.49	6.28	6.06	6.15	7.44	3.84	8.87
r	0.9164	0.9111	0.9041	0.9043	0.7703	0.9273	0.8737
Parameter	Compound No.						
	8	9	10	11	12	13	
n	11	10	9	12	12	12	
R_{M0}	0.98	1.74	2.06	1.15	1.19	1.17	
$-b$ ($\times 10$)	4.48	4.60	2.11	5.12	5.66	3.94	
s_b ($\times 10^2$)	6.24	5.80	1.79	9.21	10.61	5.81	
r	0.8513	0.9097	0.8092	0.8692	0.8600	0.9062	

simultaneous assessment of a practically unlimited number of variables (generally chromatographic parameters) that greatly facilitate the solution of both theoretical and practical problems. Multivariate methods have mainly been used in chromatography to identify basic factors which have a significant impact on solute–solvent interactions and to cluster solutes, supports and solvents into groups to expose similar retention characteristics. As each multivariate mathematical–statistical method highlights only one aspect of the chromatographic problem under investigation, the use of more than one method is the rule rather than the exception. However, the successful use of some multivariate mathematical–statistical methods (factor analysis, principal component analysis, cluster analysis) is hampered by the fact that the standard deviation of the raw data is generally not included in the calculations. Therefore, no significance test is available for evaluation of the similarities and dissimilarities of the results of such analyses. Thus, only one attempt has been made to include the standard deviation of data in principal component analysis [11].

The objectives of this work were: (i) to use a graphical approximation for the inclusion of the standard deviation of raw data in principal component analysis and cluster analysis using the retention paremeters of some anti-hypoxia drugs as data matrices; and (ii) to evaluate the performance of this approximation.

2. Experimental

The chemical structures of the anti-hypoxia drugs investigated are listed in Table 1. The compounds are the derivatives of 4-aryl-4-oxo-(2E)-butenoic acid. Compound 10 represents an example of a non-derivatized form which is the reaction product of terahydronaphthalene and maleic anhydride in a Friedel–Crafts acylation. The compounds were synthesized at the research laboratory of the Chemical Works of Gedeon Richter Ltd (Budapest, Hungary) [12,13]. Each compound has pharmacological activity in the normobaric hypoxia test. Polygram UV_{254} plates (Macherey-Nagel, Dürren, Germany) were used without pretreatment. The drugs were separately dissolved in acetonitrile to give a concentration of 5 mg ml^{-1} and 2 μl of solution was spotted onto the plates. Development was performed in sandwich chambers (22 × 22 × 3 cm^3) at room temperature, and the running distance was $\approx$ 15 cm. The chambers were not presaturated.

Table 3
Parameters of the linear relationship between the chloroform concentration in the chloroform–carbon tetrachloride eluent mixture (C vol. %) and the R_M values of anti-hypoxia drugs. Numbers refer to the anti-hypoxia drugs in Table 1. ($R_M = a + b \cdot C$)

Parameters	Compound No.						
	1	2	3	4	5	6	7
n	10	9	10	10	9	6	9
R_{M0}	1.86	2.23	1.92	2.03	0.91	2.49	2.43
$-b$ ($\times 10^2$)	2.04	2.17	2.08	2.22	1.49	0.87	1.87
s_b ($\times 10^3$)	2.91	2.93	3.14	2.72	2.67	1.52	2.20
r	0.9277	0.9417	0.9195	0.9450	0.9028	0.8600	0.9544
Parameter	Compound No.						
	8	9	10	11	12	13	
n	7	6	5	9	9	9	
R_{M0}	2.63	2.99	2.63	2.18	2.44	1.95	
$-b$ ($\times 10^2$)	2.03	2.08	1.13	2.13	1.91	2.22	
s_b ($\times 10^3$)	3.62	3.85	1.01	2.20	1.26	2.47	
r	0.9286	0.9377	0.9881	0.9645	0.9852	0.9594	

The eluents were methanol–carbon tetrachloride (methanol concentration 0–5 vol.%), chloroform–carbon tetrachloride (chloroform concentration 10–100 vol.%), and acetonitrile–carbon tetrachloride (acetonitrile concentration 2–100 vol.%) mixtures. After development the plates were dried at room temperature and the spots detected under UV light at 254 nm. Each determination was run in quadruplicate. The R_M values were calculated as $\log(1/R_f - 1)$ and extrapolated to zero concentration of the components with higher solvent strength:

$$R_M = R_{M0} + b \cdot C \quad (1)$$

where R_M is the actual R_M value of a compound determined at C vol.% concentration of the stronger component in the eluent, R_{M0} is the R_M value of the compound extrapolated to zero concentration of the stronger component in the eluent, b is the change in the R_M value caused by a 1% change in the eluent composition, and C is the concentration of the stronger component (vol.%). When the relative standard deviation of the parallel determinations was higher than 6% the data was omitted from the calculations. The intercept (R_{M0}) and slope (b) values of Eq. (1) were considered to be the best estimates of the adsorptive capacity and the specific hydrophilic surface area of the compounds respectively [14]. Calculations were carried out separately for each compound in both methanol–carbon tetrachloride and chloroform–carbon tetrachloride eluent mixtures. As the drugs showed irregular retention behaviour in acetonitrile–carbon tetrachloride mixtures, in this case the calculations were modified. Their retention decreased at the lower concentration range of acetonitrile, reached a minimum, and then increased again at the higher concentration range. This irregular retention behaviour was described by a quadratic function:

$$R_M = R_{M0} + b_1 \cdot C + b_2 \cdot C^2 \quad (2)$$

where the symbols are the same as in Eq. (1).

To find the similarities and dissimilarities between the TLC systems and the retention behaviour of drugs, principal component analysis (PCA) followed by two-dimensional non-linear mapping and cluster analysis were used. To include the standard deviation of parameters of Eqs. (1) and (2) PCA as been carried out in two different manners:

(i) The 13 drugs were the variables, the main value of the slope and the intercept values of Eqs. (1) and (2), the main value minus twice the standard deviation and the main value plus twice the standard deviation of the above parameters were the observations (21 observations in total).

Table 4
Parameters of the linear relationship between the acetonitrile concentration in the acetonitrile–carbon tetrachloride eluent mixture (C vol. %) and the R_M value of the anti-hypoxia drugs. Numbers refer to the anti-hypoxia drugs in Table 1. ($R_M = a + b_1 \cdot C + b_2 \cdot C^2$)

Parameters	Compound No.						
	1	2	3	4	5	6	7
n	14	15	15	15	13	12	15
R_{M0}	0.83	1.23	1.05	1.14	0.32	2.23	1.44
$-b_1$ ($\times 10^2$)	6.05	7.46	6.39	6.87	4.24	4.02	6.38
s_{b1} ($\times 10^3$)	1.22	1.10	1.03	1.00	0.75	0.99	1.03
$-b_2$ ($\times 10^4$)	4.83	5.92	5.09	5.40	3.34	3.24	5.26
s_{b2} ($\times 10^4$)	1.22	1.09	1.01	0.99	0.75	0.92	1.02
b'_1 (%)	55.58	55.51	55.39	55.74	55.90	53.54	54.54
b'_2 (%)	44.42	44.49	44.61	44.26	44.10	46.46	45.46
r^2	0.8357	0.8903	0.8705	0.8966	0.8832	0.7529	0.8512
$F_{calc.}$	25.42	44.62	36.99	47.67	34.04	12.19	31.47

Parameters	Compound No.					
	8	9	10	11	12	13
n	14	14	10	13	14	15
R_{M0}	1.49	1.65	2.14	1.21	1.42	1.21
$-b_1$ ($\times 10^2$)	5.84	8.36	2.16	5.68	6.50	7.11
s_{b1} ($\times 10^3$)	0.89	1.00	0.31	0.89	1.08	1.05
$-b_2$ ($\times 10^4$)	4.64	6.54	1.87	4.33	5.40	5.72
s_{b2} ($\times 10^4$)	0.86	0.97	2.83	0.87	1.05	1.04
b'_1 (%)	54.98	55.42	51.38	56.28	53.92	55.14
b'_2 (%)	45.02	44.58	48.62	43.72	46.08	44.86
r^2	0.8877	0.9335	0.8949	0.9166	0.8463	0.8841
$F_{calc.}$	39.52	70.19	25.56	49.47	27.53	41.94

(ii) The seven parameters of Eqs. (1) and (2) were the variables, the main value of the slope and the intercept values of Eqs. (1) and (2), the main value minus twice the standard deviation and the main value plus twice the standard deviation of the above parameters of each drug were the observations (39 observations in total).

The limit of the variance explained was set to 99% in both cases. The nonlinear mapping techniques [15] and cluster analysis were used for the visualization of the multidimensional matrices of PC loadings and variables. The iteration of the two-dimensional nonlinear maps was carried to the point when the difference between the last two iterations was lower than 10^{-8}. A circle can be formed from the mean value and the mean value $\pm$two standard deviations on the two-dimensional nonlinear map, the centre of the circle being the mean and the radius of the circle being represented by the mean $\pm$two standard deviations. It is assumed that the retention parameters (PCA I) or drugs (PCA II) are significantly different (significance level 95%) when the circles do not overlap. It was further assumed that the mean value and the mean value $\pm$two standard deviations of a parameter are close to each other (form a triad) on the cluster dendogram when they significantly differ from the others. The inclusion of both the nonlinear mapping technique and cluster analysis in the evaluation was motivated by the consideration that cluster analysis and the nonlinear mapping technique are theoretically similar, as they calculate and visualize the relative distances between the members of a multi-dimensional data matrix.

3. Results and discussion

The parameters of Eq. (1) determined in methanol–carbon tetrachloride and chloroform–carbon tetrachloride eluent mixtures are compiled in Tables 2 and 3. Eq. (1) fits the experimental data well, the significance level being over 95% in each instance. The considerable differences between the slope and intercept values of the anti-hypoxia drugs indicate that the solutes can be successfully separated in these TLC systems. The parameters of Eq. (2) describing the retention behaviour of drugs in acetonitrile–carbon tetrachloride eluents are listed in Table 4. The relationship between the acetonitrile concentration and the R_M value is not linear. This irregular retention behaviour is not clearly understood. The retention decreases with increasing acetonitrile concentration in the lower concentration range, reaches a minimum, and then increases

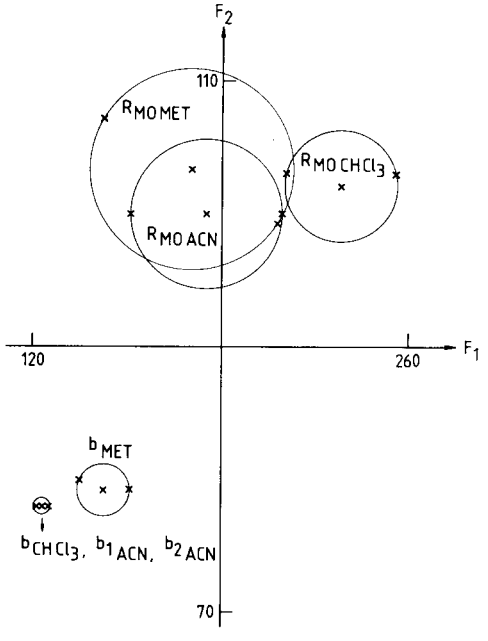


Fig. 1. Similarities and dissimilarities between the retention parameters of anti-hypoxia drugs. Two-dimensional nonlinear map of principal component variables calculated from PCA I. No. of iterations: 303; maximum error: 1.60×10^{-5}. R_{M0MET}, R_{M0CHCl_3}, b_{MET} and b_{CHCl_3} are the parameters of Eq. (1) using methanol–carbon tetrachloride and chloroform–carbon tetrachloride eluent mixtures. R_{M0ACN}, b_{1ACN} and b_{2ACN} are the parameters of Eq. (2) using acetonitrile–carbon tetrachloride eluent mixtures.

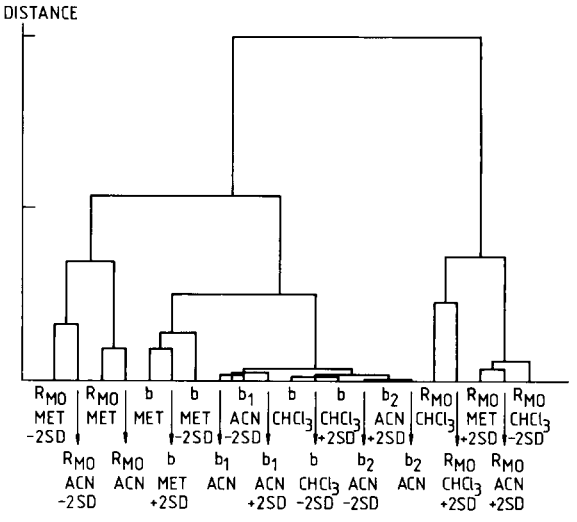


Fig. 2. Similarities and dissimilarities between the retention parameters of anti-hypoxia drugs. Cluster dendogram calculated from the principal component variables of PCA I. For symbols see Fig. 1. SD = Standard deviation.

with further increase in the organic modifier ratio. Similar anomalous retention behaviour has been observed in the reversed-phase chromatography of peptides [16] and was tentatively explained by the dual retention mechanism: the free silanol groups are more easily accessible for the solute molecules at the higher organic phase concentration. However, this explanation is hardly valid in this case as the layer was uncovered silica and the eluents were typical adsorption phase eluents. It is assumed that at lower concentrations acetonitrile acts as the stronger eluent component, linearly decreasing the R_M value of solutes. However, at higher concentrations it modifies the dissociation of the polar substructures in the drug molecules, in this manner influencing their capacity to interact with the acidic surface of the silica support.

The first principal component explained the overwhelming majority of variance in PCA I (eigenvalue 12.31; variance explained 94.72%). Each drug has a high loading in the first PC (0.83–0.99) indicating that the retention behaviour of drugs is similar in the three adsorption chromatographic systems. The mean of the retention parameters (R_{M0} and b of Eqs. (1) and (2)) and the circles representing the limit of the 95% significance level are shown on the two-dimensional nonlinear map of PC variables

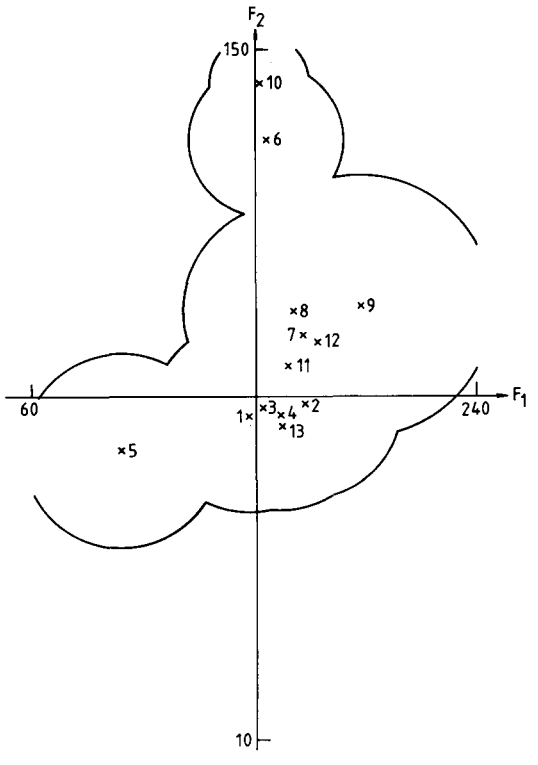


Fig. 3. Similarities and dissimilarities between the anti-hypoxia drugs. Two-dimensional nonlinear map of principal component variables calculated from PCA II. Numbers refer to antihypoxia drugs in Table 1. No. of iterations: 83; maximum error: 8.99×10^{-3}.

(Fig. 1). The circles of the R_{M0} and b values are well separated, indicating the significant differences between these two parameter sets. This finding further suggests that both parameter sets can be included as independent variables in future structure–retention behaviour calculations. The circles of R_{M0} values overlap, indicating that no significance differences can be observed between these extrapolated values. The specific hydrophilic surface areas of drugs in methanol–carbon tetrachloride (b_{MET}) are significantly different from those measured in other eluent systems. This discrepancy can be explained by the supposition that methanol binds to the adsorption centers on the surface of the silica support, in this manner modifying its capacity to interact with the solute molecules. The results of cluster analysis entirely support the previous conclusions (Fig. 2). The R_{M0} and b values are well separated; however, the mean R_{M0} values $R_{M0} \pm$ 2SD values are overlapping, again indicating their similarities. The corresponding triads of b values are well separated from each other, proving the different information contents of these retention parameters. Although cluster analysis and non-linear mapping give similar results, the application of two-dimensional non-linear mapping instead of cluster analysis is strongly advocated because of its higher dimensionality. It is assumed that the two-dimensional non-linear map may contain more information than the one-dimensional structure of clusters.

Three principal components explained 92.84% of the total variance in PCA II (first PC = 55.31%; second PC = 25.02%; third PC = 12.51%), indicating that the seven original variables can be substituted by three background (abstract) variables with only 7% loss of information. Unfortunately, PCA does not prove the existence of such background variables as concrete physicochemical entities, it only indicates their mathematical possibility. The mean values of drugs and the contour of the circles representing the 95% significance level are shown on the two-dimensional nonlinear map of PC variables (Fig. 3). In contrast to the two-dimensional nonlinear map of retention parameters the circles overlap, forming one continuous domain, suggesting that the retention behaviour of drugs is similar in each TLC system. This finding further suggests that each substructure has a similar impact on the retention behaviour of the drugs in adsorptive TLC and that the retention observed is the result of the interplay of the various forces occurring between the solutes and the stationary phase.

It is concluded that this graphical approximation represents a useful method for the inclusion of the standard deviation of both variables and observations in any data matrices evaluated by PCA folllowed by nonlinear mapping or cluster analysis.

Acknowledgements

This work was supported by the Polish–Hungarian Cooperation Project "Development of New Analytical and Evaluation Methods for the Assessment of Biological Effects of Xenobiotics and Pharmaceuticals".

References

[1] H. Mager, Moderne Regressionsanalyse, Salle, Sauerländer, Frankfurt am Main, Germany 1982, pp. 135–157.
[2] E. Forgács, J. Pharm. Biomed. Anal., 13 525–532.
[3] E.R. Malinowski and D.C. Howery, Factor Analysis in Chemistry, Wiley, New York, 1980.
[4] J.R. Chretien, M. Righezza, A. Hassam and B.Y. Meklati, J. Chromatogr., 609 (1992) 261–267.
[5] K.V. Mardia, J.T. Kent and J.M. Bibby, Multivariate Analysis, Academic Press, London, 1979, pp. 213–254.
[6] P. Karsnas and T. Lindblom, J. Chromatogr., 599 (1992) 131.
[7] L. Orloci, C.R. Rao and W.M. Stitiler, Multivariate Methods in Ecological Work, International Cooperative Publishing House, Fairland, MD, 1979.
[8] E. Forgács, T. Cserháti and B. Bordás, Anal. Chim. Acta, 279 (1993) 115.
[9] P.J. Lewi, Arzneim.-Forsch., 26 (1976) 1295.
[10] T. Hamoir, F.C. Sanchez, B. Bouguignon and D.L. Massart, J. Chromatogr. Sci., 32 (1994) 488–498.
[11] E. Forgács, T. Cserháti and K. Valkó, J. Chromatogr., 592 (1992) 75–83.
[12] L. Dobay, J. Fisher, E. Ezer, J. Matuz, L. Szporny and K. Sághy, Hung. Pat. 198012, (1980).
[13] J. Fisher, L. Dobay, Gy. Fekete, E. Ezer, J. Matuz and L. Szporny, Hung. Pat. 198294, (1982).
[14] T. Cserháti and K. Magyar, J. Biochem. Biophys. Methods, 24 (1992) 249–264.
[15] J.W. Sammon, Jr., IEEE Trans. Comput, (1969) 401–407.
[16] A. Nahum and C. Horváth, J. Chromatogr., 203 (1981) 53-63.