Pergamon

Journal of Pharmaceutical & Biomedical Analysis, Vol. 12, No. 7, pp. 867–874, 1994

0731–7085/94 $7.00 + 0.00

0731–7085(94)E0021–R

Statistical evaluation of molecular descriptors and quantitative structure–property relationship studies of retinoids

MARJA SALO,*† SEPPO SARNA‡ and HEIKKI VUORELA†

† *Department of Pharmacy, P.O. Box 15 (Fabianinkatu 35), FIN-00014 University of Helsinki, Finland*
‡ *Department of Public Health Science, University of Helsinki, Finland*

Abstract: Various molecular descriptors, including connectivity indices, sums of the intrinsic state values, electrotopological state indices, topological equivalence indices, kappa indices, normalized Bonchev-Trinajstic and Shannon information indices, Wiener and Platt's F numbers and molecular weight, were calculated for 73 retinoids, whose structures and properties were taken from the literature. A novel methodology using statistical analyses (cluster, factor and stepwise regression) in selecting relevant molecular descriptors for quantitative structure–property relationship (QSPR) studies has been developed. The analyses were used in correlating molecular structure with affinity, pharmacokinetic properties and reversed-phase retention of retinoids.

Keywords: *Cluster analysis; factor analysis; molecular descriptors; quantitative structure–property relationship (QSPR); regression analysis; retinoids; topological indices.*

Introduction

Quantitative structure–activity (QSAR) and structure–property (QSPR) relationship studies concentrate on the interdependence of the chemical behaviour of a compound and its molecular structure. Quantitative structure–retention relationship (QSRR) studies conducted in reversed-phase liquid chromatography (RPLC) are used to determine the lipophilic nature of solutes, which mainly governs retention in RPLC. These studies have been extensively reviewed, e.g. by Braumann [1] and Kaliszan [2]. Molecular descriptors express the chemical structure of a molecule in a numerical form. The descriptors have proved useful in many QSAR and QSPR studies. An evaluation of the molecular descriptors serves two purposes: more reliable correlations can be achieved, and the physico-chemical significance of the indices can be understood more clearly.

The molecular structure of a molecule is most often described with topological indices, starting from its chemical, two-dimensional structure. The first topological index put forward was the Wiener number, W, which is based on the distance matrix of the molecule [3]. The most widely used topological indices are the connectivity indices, χ, introduced by Randic [4] and further developed by Kier and Hall [5]. Other topological indices include kappa indices, κ, which describe the molecular shape [6], the topological equivalence indices, T, describing similarities between atoms in the molecule [7], and electrotopological state indices, S_i, which give information about the topologic and electronic environment of a particular atom in a molecule [8]. A number of indices based on the information theory, e.g. the Shannon index [9] and Bonchev-Trinajstic indices [10], exist, also.

Because of its established correlation with biological and pharmacological parameters, the partition coefficient is the most extensively used structural parameter in medicinal chemistry. The partition coefficient describes the partition of a compound between a lipophilic and a hydrophilic phase; 1-octanol and water are the most commonly studied phases. The partition coefficient is usually expressed as the logarithm of the partition coefficient, logP or as Hansch's hydrophobic parameter π, which is derived from logP. LogP is a hydrophobic parameter, which mainly describes the ability of a molecule to take part in nonspecific interactions. LogP values are correlated with bulk properties, such as molecular weight and

* Author to whom correspondence should be addressed.

molar volume [11], and also with many molecular descriptors, such as the connectivity indices [11, 12]. The latter are more strongly dependent on the steric and electronic properties of the molecules, and therefore can describe the molecule's interactions with an enzyme or a receptor better than logP [13]. The QSRR and QSAR studies with topological indices relate molecular structure directly to chemical and pharmacological properties and make it possible to predict properties of hypothetical molecules [14].

A correlation analysis of topological indices used as independent variables is necessary [15]. Cluster analysis is a multivariate procedure for defining optimal grouping of data, based on the measurement of similarities of variables [16]. Cluster analysis has proved to being useful in classification of octadecyl stationary-phases [17] and antiviral compounds [18]. According to Massart, a hierarchical method is adequate for a rough classification of the data [19]. If the categories are vague and overlapping, factor analysis produces more information than the cluster analysis [20]. By factor analysis multidimensional data structure can be presented by two or three dimensions, called factors [21]. Loading values are the correlations between the variables and factors [22] and provide information on the physicochemical significance of the factors [23]. Factor analysis has been used in classification of molecular descriptors [11] and octadecyl stationary-phases [17], and in optimization of chromatographic separations [24].

It is easy to get excellent but insignificant correlations simply by adding enough variables into the equation. The selection of independent variables can be achieved by a stepwise regression procedure, by which the calculation of all possible variable combinations can be avoided, and also the risk of overfitting the data [25].

Retinoids are used in the treatment of severe skin diseases, such as psoriasis and cystic acne [26], and have a possible use in the treatment or prevention of cancer [27]. The structures and properties of 73 retinoids were for the most part taken from the literature: the pharmacokinetic properties of 19 retinoids, most of which were acetylenic [28], the binding affinity and teratogenicity of 26 retinoids [29] and the reversed-phase chromatographic retention of 24 [30] and 11 retinoids [31].

The aims of this work were to study the advantages of using molecular descriptors in describing the molecular structure of retinoids and to combine different statistical methods in evaluation of the descriptors and variable selection for QSPR studies.

Experimental

Experimental data were taken from the literature [28–31]. The data on the acetylenic retinoids included total body clearance (CL_T), mean residence time (MRT), distribution volume at steady state (V_{SS}), elimination half-life ($t_{1/2}$) and the logarithm of distribution coefficient (logDC), of which no experimental conditions were given [28]. The structure–affinity studies consisted of the concentration of retinoid required to displace 50% of bound, labelled all-*trans* retinoic acid (DC_{50}) and the median effective single oral, maternal dose administered for induction of terata in hamsters [29]. The chromatographic retention of retinoids was measured on an octadecyl stationary phase with aqueous acetonitrile and methanol as the mobile phases [30, 31].

The values for the molecular descriptors were calculated by MOLCONN-X v1.0 (L.H. Hall, Quincy, MA, USA) and their statistical analysis was conducted by SYSTAT v.5.1 (Systat Inc. Intelligent Software, USA) computer program. The programs were run on a Macintosh LC microcomputer. The following molecular descriptors were studied: molecular weight (MW), connectivity indices ($^{0-9}\chi$, $^{0-9}\chi^{v}$, $^{3,4}\chi_{c/pc}{}^{v}$) [5], graph complexity (GrComp), kappa indices ($^{1-3}\kappa$, $^{0-3}\kappa_{\alpha}$) [6], topological equivalence index (T) [7] and total topological equivalence indices (TTS, simple index and TTD, valence index), sum of the intrinsic state values I (sI) [8], sum of delta-I values (sdI) [8], electrotopological and total electrotopological state indices (S, TETS) [8], Shannon information index (Sh) [9], normalized Bonchev-Trinajstic information indices (nI(G), nIW(G)) [10], information content (InfC), Platt's F number (PF) [32, 33], Wiener number (W) and Wiener's P number [3]. The electrotopological and the topological state indices S and T, respectively, included in the study are the indices of the carbon connecting the ring and the polyene chain (S1, T1) and of the carboxyl carbon (S2, T2).

The advantages of using molecular descriptors in structure–property relationship studies were examined by correlation, factor, cluster

and stepwise regression analyses. For every variable included in the multiple regression equations, there existed at least five data points. The intercorrelation of the selected variables was studied. The structure–affinity relationship analyses were performed by two procedures. The descriptors related to the magnitude of the affinity/teratogenicity were studied by including only those structures, that had a numerical DC_{50}- or ED_{50}-value (referred to as the limited data set). The descriptors related to the existence of an effect was studied by providing hypothetical high values of 1000 μM and 10 000 μmol kg^{-1} for the non-competitive and inactive compounds, respectively, in order that those could be included in the calculations (referred to as the extended data set). The regression equations were calculated using the limited data set without the inactive or non-competitive compounds. For every independent variable included in the multiple regression equations, there were at least five data points. The intercorrelations of the selected variables was studied.

Results

A detailed description of the statistical results was regarded inappropriate because of the large amount of data obtained. The more general relationships between the descriptors will be discussed in the Discussion section of this paper.

Structure–pharmacokinetics relationships

The pharmacokinetic parameters had no correlation with the descriptors (n = 18) in the correlation analysis. In cluster analysis (n = 18) the descriptors and the pharmacokinetic parameters formed two groups. Total body clearance (CL_T) and distribution volume at steady state (V_{SS}) belonged to a different group than mean residence time (MRT) and half-life. Three factors were needed to explain 84.5% of the total variance of the data, of which the third factor explained 9.0%. The connectivity indices had large loading values for the 1st factor and the kappa indices for the 2nd. The pharmacokinetic parameters had high loading values only for the third factor (n = 18).

The regression analyses were performed with the logarithm of the distribution coefficient (logDC) and variables selected by stepwise regression, cluster and factor analyses for all compounds and also for ethyl esters and acids separately. The descriptors, which were most closely related (cluster) or had the same loading values (factor) were chosen, but in most cases produced very poor correlations. When variables for ethyl esters and acids were selected separately, better correlations were obtained. The highest value of MRT would have influenced the regression equation excessively (92 min, others 3–29 min), and was excluded from the calculations. The variables and correlation coefficients (r^2 for simple and R^2 for multiple regression) for all compounds, esters and acids are presented in Tables 1–3, respectively. The regression equations, with best correlations, are given below (standard error of regression coefficients given in parentheses, the intercorrelations of the variables presented by r_{ij}):

Acids and esters:

$$\log DC = 5.33\,(0.76) \times S2 + 4.26\,(1.91) \times {}^4\chi_{pc} + 0.59\,(0.15) \times {}^1\kappa - 4.13\,(3.24)$$
$$(n = 18,\ R^2 = 0.937,\ \text{F-ratio} = 69.04,\ P < 0.001,\ r_{ij} = -0.164 - 0.577).$$

Acids:

$$\text{half-life(min)} = -129.91(41.06) \times S1 + 98.92(33.37) \times {}^3\chi_c + 17.22(115.86)$$
$$(n = 10,\ R^2 = 0.858,\ \text{F-ratio} = 21.16,\ P < 0.005,\ r_{ij} = -0.601).$$

Esters:

$$\text{half-life(min)} = 43.93(5.22) \times {}^2\kappa - 375.82(49.72)$$
$$(n = 8,\ r^2 = 0.922,\ \text{F-ratio} = 70.95,\ P < 0.001).$$

$$CL_T\text{(ml/min)} = -4508.57\ (880.01) \times S2 - 1322.66\ (285.03)$$
$$(n = 8,\ R^2 = 0.814,\ \text{F-ratio} = 26.25,\ P < 0.005).$$

Calculated and observed values are plotted in Fig. 1. The regression equation of logDC gave a good fit with the experimental values, but the correlation between the pharmacokinetic parameters and logDC were almost insignificant, and not improved by treating esters and acids separately. Dividing the compounds according to their functionality to esters and acids improved the correlations

Table 1
Regression variables and correlation coefficients (r^2/R^2) from the regression analyses of the pharmacokinetic data of acetylenic retinoids

Dependent variable	Independent variables	Correlation coefficient (r^2/R^2)
1. logDC	S2, $^4\chi^v$, $^1\kappa$	0.937
half-life	S1, S2, $^1\kappa$	0.473
MRT	S1, S2, $^0\kappa_\alpha$	0.528
CL_T	S2, $^0\kappa_\alpha$, $^2\kappa_\alpha$	0.264
V_{ss}	TTD, $^6\chi^v$, $^4\chi_{pc}{}^v$	0.342
2. logDC	nI(G)	0.853
half-life	sdI	0.001
MRT	sdI	0.001
CL_T	S2	0.139
V_{ss}	S2	0.233
3. logDC	Sh, $^3\kappa$, $^2\kappa$	0.619
half-life	$^1\kappa_\alpha$, S1, $^6\chi^v$	0.200
MRT	$^1\kappa_\alpha$, sI, $^4\chi^v$	0.060
CL_T	$^1\kappa_\alpha$, sI, sdI	0.480
V_{ss}	$^1\kappa_\alpha$, sI, sdI	0.328
4. half-life	logDC	0.043
MRT	logDC	0.116
CL_T	logDC	0.071
V_{ss}	logDC	0.027

Variables were chosen by: 1. stepwise regression, 2. cluster analyses, 3. factor analyses, 4. correlation with logDC.

Table 2
Regression variables and correlation coefficients (r^2/R^2) from the regression analyses of the pharmacokinetic data of acetylenic retinoid esters

Dependent variable	Independent variables	Correlation coefficient (r^2/R^2)
1. logDC	$^0\chi^v$	0.511
half-life	$^2\kappa$	0.922
MRT	S1	0.025
CL_T	S2	0.814
V_{ss}	S2	0.485
2. logDC	S2	0.488
half-life	$^2\kappa$	0.922
MRT	$^2\kappa$	0.002
CL_T	S1	0.025
V_{ss}	S1	0.001
3. logDC	$^1\kappa$	0.364
half-life	MW	0.295
MRT	$^0\chi^v$	0.305
CL_T	S2	0.814
V_{ss}	sdI	0.041
4. half-life	logDC	0.056
MRT	logDC	0.142
CL_T	logDC	0.642
V_{ss}	logDC	0.736

Variables were chosen by: 1. stepwise regression, 2. cluster analyses, 3. factor analyses, 4. correlation with logDC.

between the pharmacokinetic parameters and molecular descriptors of esters. With acids, good correlations were obtained only for half-life.

Table 3
Regression variables and correlation coefficients (r^2/R^2) from the regression analyses of the pharmacokinetic data of acetylenic retinoid acids

Dependent variable	Independent variables	Correlation coefficient (r^2/R^2)
1. logDC	$^1\chi^v$, $^8\chi^v$	0.926
half-life	S1, $^3\chi_c$	0.858
MRT	S1, $^1\kappa$	0.739
CL_T	$^4\chi^v$, $^4\chi_c$	0.753
V_{ss}	$^1\kappa$	0.664
2. logDC	nIW(G)	0.870
half-life	W	0.538
MRT	W	0.619
CL_T	Sh	0.346
V_{ss}	Sh	0.518
3. logDC	nIW(G)	0.870
half-life	$^1\chi^v$, $^4\chi_{pc}{}^v$	0.597
MRT	$^2\kappa$, $^8\chi^v$	0.037
CL_T	S2, sdI	0.044
V_{ss}	$^0\chi^v$, $^3\chi^v$	0.631
4. half-life	logDC	0.117
MRT	logDC	0.304
CL_T	logDC	0.573
V_{ss}	logDC	0.005

Variables were chosen by: 1. stepwise regression, 2. cluster analyses, 3. factor analyses, 4. correlation with logDC.

Structure–affinity relationships

DC_{50} was defined as the concentration of retinoid required to displace 50% of bound, labelled all-*trans*-retinoic acid and ED_{50} as the median effective single oral, maternal dose for induction of terata in hamster. DC_{50} and ED_{50} were not related in the cluster analysis of the limited data set (n = 11 and 19, respectively), but they were closely related with the extended data set (n = 25 and 21, respectively). Three factors explained 82.6% of the total variance of the data in the extended data set and two factors 87.1% in the limited data set. The factor analyses of the different data sets produced similar results with regard to the loading values. DC_{50} and ED_{50} were mostly related to kappa and S_i indices, which had large loading values for the 2nd factor.

Variables for the regression analyses were selected by a stepwise regression procedure, cluster analyses and factor analyses, performed for both the limited and extended data sets. The variables were selected as for acetylenic retinoids. Variables used and correlation coefficient are presented in Table 4. The good correlation of DC_{50} with $^1\kappa$ and $^1\kappa_\alpha$ was due to the large leverage of the highest value on the regression equation (110 μM, other 0.3–

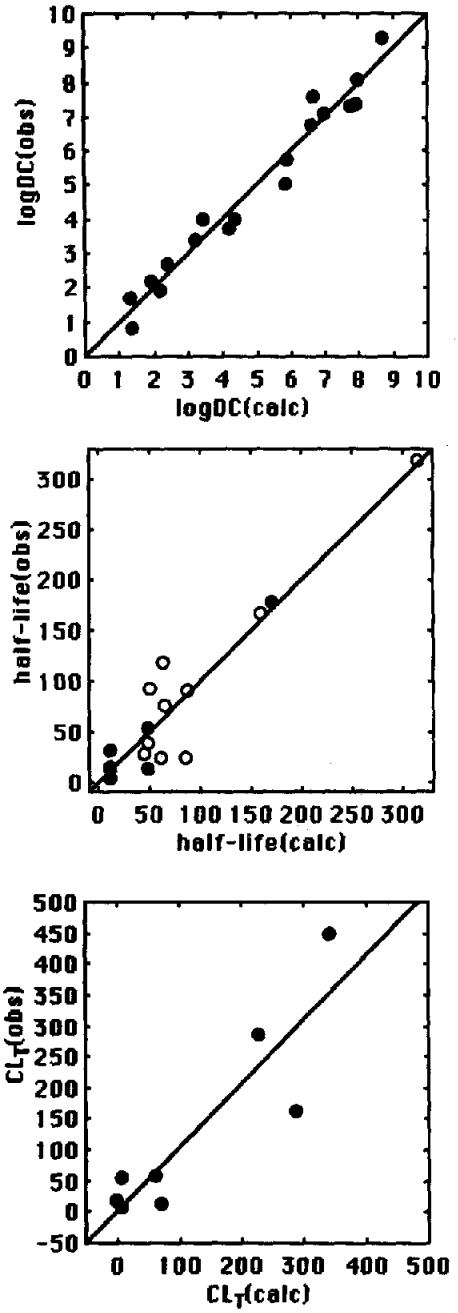

Figure 1
Observed vs calculated values of (a) distribution coefficient (logDC) for acetylenic retinoids; (b) half-life for acetylenic acids (○) and esters (●); (c) total body clearance (CL_T) for acetylenic esters.

19 μM). With 10 data points the correlation coefficient (R^2) was 0.471.

Structure–retention relationships of retinoids

Of the four mobile phases used in the study, sufficient retention data for statistical analyses were available for two, mobile phase B (water–acetonitrile 2:98) and D (10% aqueous acetic acid–methanol 1:99), the type of percentage

Table 4
Regression variables and correlation coefficients (r^2/R^2) from the regression analyses performed with limited data sets

Dependent variable	Independent variables	Correlation coefficient (r^2/R^2)
1.1 DC_{50}	$^1\kappa$, $^1\kappa_\alpha$	0.961
ED_{50}	T2	0.183
1.2 DC_{50}	$^0\chi^v$, $^1\kappa$	0.901
ED_{50}	$^4\chi_c^v$, $^1\kappa$	0.056
2.1 DC_{50}	$^1\kappa$, nIW(G)	0.761
ED_{50}	S1	0.078
2.2 DC_{50}	S2	0.342
ED_{50}	$^2\kappa$	0.007
3.1 DC_{50}	$^0\chi$, $^1\kappa_\alpha$	0.426
ED_{50}	$^2\kappa$, $^3\chi_c^v$	0.036
3.2 DC_{50}	S2, nI(G)	0.375
ED_{50}	S2, nI(G)	0.141

Variables were chosen by: 1. stepwise regression: 1, limited data set, 2. extended data set, 2. cluster analyses: 1, limited data set, 2. extended data set and 3. factor analyses: 1. limited data set, 2. extended data set.

was not given. The retention was expressed as retention volume (ml, V_B and V_D, respectively). The largest data point of V_D was excluded in order to avoid its too large leverage on the regression analysis. The numbers of observations were 11 and 13 in mobile phases B and D, respectively.

The molecular descriptors were relatively highly intercorrelated. The descriptors and the retention were clearly divided into two groups in the cluster analysis. The retention data of V_B were insufficient for factor analysis. Two factors explained 87.1% of the total variance of the descriptors and three, 91.9%. The retention was clearly related to kappa indices and the polarity of the end group (S2), which formed a separate group both in cluster and factor analyses, with large loadings for the 2nd factor.

The regression analyses were performed with descriptors chosen by cluster, factor (greatest loadings) and stepwise regression analyses. The descriptors chosen and the correlation coefficients are presented in Table 5. The following regression equation had the greatest correlation coefficient (standard error in parentheses):

$$V_D(\text{ml}) = 5.88\ (1.38) \times {}^0\chi^v - 0.09\ (0.05) \times \text{MW} - 47.49\ (9.45)$$
$$(n = 13,\ R^2 = 0.821,\ \text{F-ratio} = 22.95,\ P < 0.001,\ r_{ij} = 0.881).$$

Table 5
Regression variables and correlation coefficients (r^2/R^2) from the regression analyses

Dependent variable	Independent variables	Correlation coefficient (r^2/R^2)
1. V_B	$^0\chi^v$, $^4\chi^v_{pc}$	0.626
V_D	MW, $^0\chi^v$	0.821
2. V_B	$^1\kappa_\alpha$, S2	0.538
V_D	$^1\kappa_\alpha$, S2	0.395
3. a. V_B	$^1\kappa_\alpha$, S2	0.538
V_D	$^1\kappa_\alpha$, S2	0.395
b. V_D	$^9\chi^v$, $^4\chi^v_c$, $^2\kappa$	0.842

Variables were chosen by: 1. stepwise regression, 2. cluster analyses, 3. factor analyses: (a) descriptors with the greatest loading values; (b) descriptors with loading values closest to retention volume in mobile phase D, the dependent variables are retention volumes in mobile phase B (water–acetonitrile 2:98) (V_B) and D (10% acetic acid–methanol 1:99) (V_D).

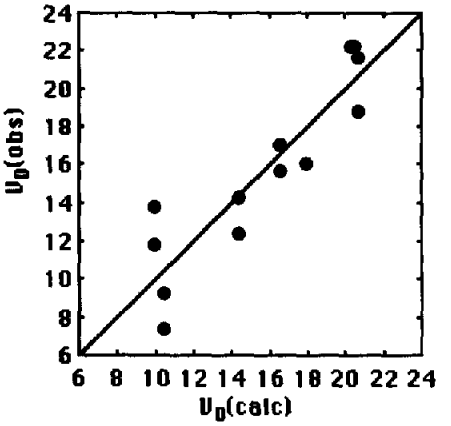

Figure 2
Observed vs calculated values of retention volume of mobile phase D (10% acetic acid–methanol 1:99, V_D).

Although the correlation coefficients were rather poor, the equations produced reasonably good estimates of the retention volumes (Fig. 2). The largest deviations between observed and calculated retention volumes were found in the compounds with an unsubstituted carboxyl group.

Structure–retention relationships of retinoates, retinoic acid and retinol

The number of observations was 11 for each mobile phase composition. Most of the descriptors correlated highly with each other. Similar results were obtained from the cluster analysis. Two and three factors explained 85.9 and 91.6% of the total variance, respectively. The $\chi_{c/pc}$, S_i and ^3K and the retention ($\ln k'$) were separated from the other indices by the analyses. Retinoates are closely congeneric, and this classification of descriptors expresses the relationship between the descriptors and hydrophobicity.

The variables for the regression equations were chosen as previously described (Variables and the correlation coefficients are presented in Table 6).

For ACN 92% simple regression with $^1\kappa_\alpha$ was preferred over the regression with $^1\kappa_\alpha$ and $^9\chi$: F-ratios 63.31 and 33.34, respectively). The following equations provided the best correlations with standard error given in parentheses:

$$\ln k'(\text{MeOH } 94\%) = 0.24(0.03) \times {}^1\kappa_\alpha - 4.04\ (0.60)$$
$$(n = 11,\ r^2 = 0.891,\ \text{F-ratio } 73.36,\ P < 0.001),$$

Table 6
Regression variables and correlation coefficients (r^2/R^2) from the regression analyses for retention of retinoates ($\ln k'$)

Dependent variable	Independent variables	Correlation coefficient (r^2/R^2)
1. $\ln k'$ (MeOH 94%)	$^1\kappa_\alpha$	0.891
$\ln k'$ (MeOH 86%)	$^1\kappa_\alpha$	0.899
$\ln k'$ (ACN 92%)	$^1\kappa_\alpha$	0.876
$\ln k'$ (ACN 82%)	$^1\kappa_\alpha$	0.909
2. $\ln k'$ (MeOH 94%)	$^3\kappa_\alpha$	0.717
$\ln k'$ (MeOH 86%)	$^3\kappa_\alpha$	0.770
$\ln k'$ (ACN 92%)	$^3\kappa_\alpha$	0.791
$\ln k'$ (ACN 82%)	$^3\kappa_\alpha$	0.792
3. $\ln k'$ (MeOH 94%)	$^8\chi$, sI	0.657
$\ln k'$ (MeOH 86%)	$^0\chi$, TETS	0.874
$\ln k'$ (ACN 92%)	$^8\chi^v$, $^1\kappa$	0.833
$\ln k'$ (ACN 82%)	$^0\chi^v$, TETS	0.934

Variables were chosen by: 1. stepwise regression, 2. cluster analyses, 3. factor analyses.

$$\ln k'(\text{MeOH } 86\%) = 0.32(0.04) \times {}^{1}\kappa_{\alpha} - 4.36\ (0.76)$$
$$(n = 11,\ r^2 = 0.899,\ \text{F-ratio } 79.87,\ P < 0.001),$$

$$\ln k'(\text{ACN } 92\%) = 0.21(0.03) \times {}^{1}\kappa_{\alpha} - 3.31\ (0.55)$$
$$(n = 11,\ r^2 = 0.876,\ \text{F-ratio } 63.31,\ P < 0.001),$$

$$\ln k'(\text{ACN } 82\%) = 0.49(0.06) \times {}^{0}\chi^{v} - 0.11\ (0.05) \times \text{TETS} - 4.06(0.62)$$
$$(n = 11,\ R^2 = 0.934,\ \text{F-ratio } 57.00,\ P < 0.001,\ r_{ij} = 0.774).$$

The equations predicted the retention of retinoates very well (Fig. 3), with slightly better correlation found in aqueous methanol.

Discussion

The study included a structurally diverse set of retinoids. The pharmacokinetic parameters were measured for acetylenic compounds and three reference compounds. Structure–affinity studies and a major part of the chromatographic work were conducted with a very diverse set of retinoids, however, the retinoates form a closely congeneric series.

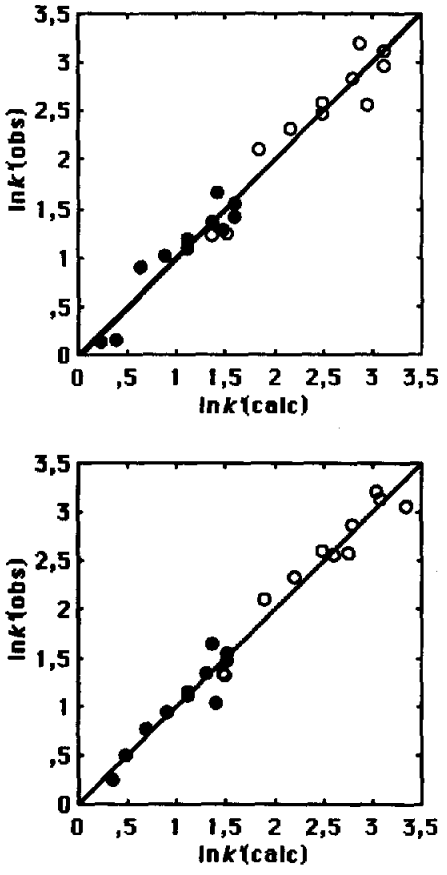

Figure 3
Observed vs calculated values of (a) retention (lnk') in aqueous methanol (● 94% MeOH, ○ 86% MeOH); (b) retention (lnk') in aqueous acetonitrile (● 92% ACN, ○ 82% ACN).

Most of the molecular descriptors included in this study are closely related to the shape and size of the molecule. To these belong the connectivity indices, kappa indices, molecular weight and information indices. The electrotopological indices are more related to the polarity of molecules or single atoms. In the factor analysis of retention data, most of the descriptors had a high loading value for the 1st factor, and only the kappa and electrotopological state indices (S_i) for the 2nd. The retention is dependent on the hydrophobic nature of the solutes, and the 1st is probably related to the molecular size and the 2nd to molecular shape and electronic properties. With pharmacokinetic and affinity data the 1st factor correlated with the connectivity and the total topology indices, but descriptors such as kappa indices, S_i indices, molecular weight and Wiener number with the 2nd. The factor pattern of structure-dependence is more complicated than in retention studies, and the physico-chemical significance is more difficult to explain.

The electronic properties of the molecules, especially the polarity of the carboxylic acid or ester group, were related to the pharmacokinetic parameters and logDC. The molecular shape expressed by kappa indices, proved to be more useful in describing the pharmacokinetic properties than logDC. The predictive power of descriptors was significantly enhanced by treating the acids and esters separately. The shape and size of the retinoid molecule significantly affected the affinity and teratogenecity of the compounds, but the descriptors failed in predicting the effects quantitatively. The descriptors are unable to describe the three-dimensional molecular structure essential in molecule–receptor interactions. The RPLC-retention is closely related to the hydrophobic properties of the solutes and could be predicted with relative accuracy. When the electronic properties of the compounds were taken into account, the retention could be predicted for a non-congeneric set of retinoids. With retinoates, which form a congeneric series, the size and shape of the structures were enough for an accurate prediction of retention.

Of the total 51 descriptors studied, 34 appeared in the results at least once, with kappa and electrotopological indices mentioned most often (42 and 27 times out of 105, respectively). S_i index is related to polarity and inductive effects caused by heteroatoms [8]. The kappa indices have been used in relating surface tension to molecular structure [34] and in the QSAR studies of toxicological data [35].

Factor and cluster analyses give valuable information about the physico-chemical background of the molecular descriptors and experimental data. The stepwise regression procedure was best suited for variable selection in QSPR and linear regression analysis. Even when other analyses expressed a very close relationship between certain descriptors and a specific property, in most cases those descriptors failed completely as quantitative parameters, and only in one instance produced a better correlation than variables chosen by the stepwise regression procedure.

Conclusions

Statistical methods in evaluation of molecular descriptors and topological indices have not previously been used in the extent presented in this study. Molecular descriptors are well suited for characterization of the molecular structure in RPLC-studies even with non-congeneric molecules, and also in pharmacokinetic studies with a congeneric set of molecules. The descriptors are unable to handle the complicated factors controlling the receptor-binding affinity and teratogenecity. The descriptors give helpful qualitative information about the relationships between the structure of a molecule and its biological and chromatographic properties.

Acknowledgements — The authors thank Dr Fritz Erni (Sandoz Pharma, Basle, Switzerland) for encouraging discussions concerning this paper.

References

[1] T. Braumann, *J. Chromatogr.* **373**, 191–225 (1986).
[2] R. Kaliszan, in *Quantitative Structure–Chromatographic Retention Relationships*, pp. 232–267. Wiley, New York (1987).
[3] H. Wiener, *J. Am. Chem. Soc.* **69**, 2636–2638 (1947).
[4] M. Randic, *J. Am. Chem. Soc.* **97**, 6609–6615 (1975).
[5] L.B. Kier and L.H. Hall, *Molecular Connectivity in Structure–Activity Analysis*, pp. 10–22. Research studies press, Letchworth (1986).
[6] L.B. Kier, *Acta Pharm. Jugosl.* **36**, 171–188 (1986).
[7] L.H. Hall and L.B. Kier, *Quant. Struct.-Act. Relat.* **9**, 115–131 (1990).
[8] L.B. Kier, L.H. Hall and J.W. Frazer, *J. Math. Chem.* **7**, 229–241 (1991).
[9] C. Shannon and W. Weaver, *Mathematical Theory of Communication*, Univ. Illinois Press, Urbana (1949).
[10] D. Bonchev and N. Trinajstec, *J. Chem. Phys.* **67**, 4517–4533 (1977).
[11] M. Tichy, *Int. J. Quant. Chem.* **16**, 509–515 (1979).
[12] W.J. Murray, L.B. Kier and L.H. Hall, *J. Pharm. Sci.* **64**, 1978–1981 (1975).
[13] S.C. Basak, D.K. Harriss and V.R. Magnuson, *J. Pharm. Sci.* **73**, 429–437 (1984).
[14] M. Randic, *J. Chem. Inf. Comput. Sci.* **31**, 311–320 (1991).
[15] R. Kaliszan, *Quantitative Structure–Chromatographic Retention Relationships*, p. 75. Wiley, New York (1987).
[16] D.L. Massart, B.G.M. Vandegiste, S.N. Deming, Y. Michotte and L. Kaufman, *Chemometrics: A Textbook*, p. 323. Elsevier, Amsterdam (1988).
[17] M.F. Delaney, A.N. Papas and M.J. Walters, *J. Chromatogr.* **410**, 31–41 (1987).
[18] M. Randic and P.J. Jurs, *Quant. Struct.-Act. Relat.* **8**, 39–48 (1989).
[19] D.L. Massart *et al.* ibid. p. 375.
[20] D.L. Massart *et al.* ibid. p. 325.
[21] D.L. Massart *et al.* ibid. p. 339.
[22] D.L. Massart *et al.* ibid. p. 352.
[23] C.H. Lochmuller, S.J. Breiner, C.E. Reese and M.N. Koel, *Anal. Chem.* **61**, 376–385 (1989).
[24] M. Righezza and J.R. Chrétien, *Chromatographia* **36**, 125–129 (1993).
[25] M.N. Hasan and P.C. Jurs, *Anal. Chem.* **55**, 263–269 (1983).
[26] J.E. Dowling and G. Wald, *Proc. Natl. Acad. Sci. USA* **46**, 587–608 (1960).
[27] M.B. Sporn, N.M. Dunlop, D.L. Newton and J.M. Smith, *Fed. Proc.* **38**, 1332–1338 (1976).
[28] S.S. Liu, R. Sandri and D.D.-S. Tang-Liu, *Drug Metab. Disp.* **18**, 1071–1077 (1990).
[29] C.C. Willhite, A. Jurek, R.P. Sharma and M.I. Dawson, *Toxicol. Appl. Pharmacol.* **112**, 144–153 (1992).
[30] M.I. Dawson and P.D. Hobbs, in *Methods Enzymol.*, Vol. 189, (Retinoids, Part A), (L. Packer, Ed.), pp. 21–25. Academic Press, San Diego (1990).
[31] M. Salo, H. Vuorela and J. Halmekoski, *Chromatographia* **36**, 147–151 (1993).
[32] J. Platt, *J. Chem. Phys.* **15**, 419–420 (1947).
[33] J. Platt, *J. Phys. Chem.* **56**, 328–336 (1952).
[34] D.T. Stanton and P.C. Jurs, *J. Chem. Inf. Comput. Sci.* 109–115 (1992).
[35] L.H. Hall and L.B. Kier, *Environ. Toxicol. Chem.* **8**, 19–24 (1989).

[*Received for review 20 October 1993; revised manuscript received 7 February 1994*]