Journal of Chromatography, 633 (1993) 43–56
Elsevier Science Publishers B.V., Amsterdam

CHROM. 24 743

Model building for the prediction of initial chromatographic conditions in pharmaceutical analysis using reversed-phase liquid chromatography

T. Hamoir, B. Bourguignon and D. L. Massart*

Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels (Belgium)

H. Hindriks

Analytical R&D Laboratories, Organon International BV, P.O. Box 20, 5340 BH Oss (Netherlands)

(First received February 5th, 1992; revised manuscript received October 8th, 1992)

ABSTRACT

The development of a first guess expert system in HPLC requires a rough estimate of retention properties. This paper investigates the extent to which the simplest possible descriptor, namely the total number of carbons in a molecule, can be used. For this purpose, experimental data acquired after investigation of this parameter on the retention behaviour of various acidic, basic and neutral compounds, for a mobile phase composed of methanol–phosphate buffer and a LiChrosorb CN column, were employed. The usefulness of the descriptor log P (calculated according to Rekker's fragment system) was also studied. Similar models were derived for both descriptors. Subsequently these models were used for the selection of initial chromatographic conditions. Both models were compared through a PRESS value. The regression equation including the descriptor log P was found to be more appropriate for the present purpose.

INTRODUCTION

Expert systems, an important application of artificial intelligence (AI), are becoming increasingly important in the field of analytical chemistry, particularly RPLC. Such computer systems contain the experts' chemical knowledge and can be applied by analytical chemists less experienced in the field. One of the possible functions of an RPLC expert system is the selection of initial chromatographic conditions. For this aim simple heuristics can be applied, as was performed in the expert system LABEL by considering the number of carbons in the solute(s). A molecule with a number of carbon atoms smaller than 10, for instance, was considered by LABEL to be non-hydrophobic [1]. The total number of carbons in a molecule is, of course, only a very rough indication of the polarity of a compound. Still, in the field studied (drugs with $M_r < 1500$) the approach used in LABEL [1] worked fairly well. A model for retention prediction would certainly be more convenient. Different approaches to retention prediction can be followed. A first point to be considered is whether to apply gradient elution or whether to use isocratic conditions in the final procedure. Certainly gradient elution is necessary if the retention range of the solutes is too wide. Moreover, the application of gradient elution presents some advantages, such as providing information on early- and late-eluting peaks for unknown samples, which under isocratic conditions can be lost owing to elution in the solvent peak or disappearance in the baseline, respectively. The gradient elution method is then best developed

* Corresponding author.

0021-9673/93/$06.00

as proposed by Dolan *et al.* [2], Chaminade [3] and Heinisch *et al.* [4].

Drylab and other such programs can also use the gradient results to predict optimum isocratic conditions. Often this works very well. However, sometimes one would like to develop the method using only isocratic elutions, as such an approach provides a better idea of peak shapes to be expected under those conditions. It is then necessary to have a good estimate of the solvent strength required to avoid unnecessary experimentation.

Several retention models have already been introduced for the prediction of retention in RPLC. However, these models are restricted to those cases where Martin's equation [5] applies, *i.e.*, to closely related compounds [6–14], and are therefore not applicable to a wide range of pharmaceutical compounds. The parameters should be simple and obtainable without additional experimentation. The most evident parameters to be included are the total number of carbons in the molecule, n_c, and the percentage of organic modifier in the mobile phase. A second descriptor will be considered, namely the log P value of the compound, which reflects the hydrophobic character of the molecule. Numerous publications have already described reasonable relationships between log $P_{o/w}$ of congeneric drugs and reversed-phase chromatographic retention data. The approach of Valko [15], however, was found to be applicable to structurally heterogeneous compounds. This approach was incorporated in a commercially available computer program to suggest HPLC eluent systems [16]. Recently, Kaliszan and Osmialowski [17] investigated the correlation between the chemical structure of structurally unrelated compounds and their retention on a polybutadiene-coated alumina column. The relationship between log k' extrapolated to pure water as the eluent (log k'_w) and log P, calculated by the fragmental method of Hansch and Leo [18], was found to be fairly satisfactory, considering the wide diversity of the structure. As the intention is to predict a mobile phase composition by computerized techniques, a computable descriptor is required. This is the case with log P, as it can be calculated by using the Hansch–Leo fragment system or the Rekker fragment system [19,20]. The latter method has been found to give a fairly good description of the hydrophobicity of more complex molecules [21]. Log P, calculated with Rekker's method, will therefore also be used as descriptor. Log P values can also be determined experimentally [22–24]. A database of experimental log P values, from which a first guess expert system would extract the value(s), seems an attractive approach. However, the insufficient availability of such values is an important problem with regard to our purposes and therefore not applicable in practice. The quality of the calculated log P values according to Rekker is not very good and better methods are available. However, we do not want lengthy and difficult calculations and the best possible log P values, as we only want to predict acceptable starting conditions that will be optimized in a later stage. Our purpose is therefore to investigate whether the relatively easy to obtain Rekker log P values are so much better than n_c that it is worth the trouble of performing log P calculations.

The aim of this study was to investigate whether it is feasible to derive a model for the retention prediction of a very diverse set of pharmaceutical compounds with the simple descriptors given, and subsequently to demonstrate the feasibility of the equation to determine initial chromatographic conditions. The quality of the prediction is expected to be superior using the descriptor log P. It should be stressed again that, in practice, chromatographers work in several steps. The initial guess is followed by an optimization of the separation and the aim is therefore not to obtain the best possible prediction, but a prediction good enough to guess with acceptable accuracy conditions for a first try.

THEORETICAL

The retention of the members of a homologous series has been described by Czok and Engelhardt [25] using an equation with four parameters:

$$\log k' = A_1 X_m + B_1 n_c + C_1 n_c X_m + D_1 \quad (1)$$

where X_m and n_c represent the percentage organic modifier in the mobile phase and the number of carbons in a homologous series, respectively. The quadratic dependence of log k' *versus* the percentage of organic modifier in a binary mobile phase, on the one hand, and the number of carbons in a homologous series, on the other, was shown by Schoenmakers [26] and Bogusz and Aderjan [27], respectively.

If quadratic terms are also included, the following equation is obtained:

$$\log k' = A_2X_m + B_2n_c + C_2n_cX_m + D_2X_m^2 + E_2n_c^2 + F_2 \quad (2)$$

By investigating whether the coefficients differ significantly from zero, the eventual equations can be obtained.

On the basis of previously published results [1], such equations can be derived for the descriptor n_c not only for homologous series, but also for a broader range of compounds and for the descriptor log P. It is the intention to investigate whether such equations permit the extent of retention of a specific compound as a function of the percentage of organic modifier to be predicted.

EXPERIMENTAL

High-performance liquid chromatography

The chromatographic system was a Varian 5000, equipped with a Rheodyne injector (sample loop 100 μl). The stationary phase was LiChrosorb cyanopropyl (particle size 5 μm; Merck) (250 × 4.0 mm I.D. column) and the mobile phase was mixtures of methanol and phosphate buffer (pH 3, ionic strength $u = 0.05$). The flow-rate was set at 1.0 ml/min. Detection was performed with a Perkin-Elmer LC 90 variable-wavelength UV detector at 254 μm and an attenuation of 0.05 a.u.f.s. Chromatograms were recorded with a Varian CDS 401 data system. All measurements were performed at 25°C in triplicate. The capacity factors, log k_i', were determined as defined by

$$\log k' = (t_r - t_0)/t_0 \quad (3)$$

The dead time of the system, t_0, was determined as the first distortion of the baseline after injection of methanol.

Standards and reagents

All drugs were of pharmaceutical purity. Stock solutions of 500 μl/ml in methanol were preserved at 4°C. Dilutions to the final injected concentrations were prepared freshly daily. Methanol, phosphoric acid, sodium phosphate and disodium phosphate, of analytical-reagent grade, were purchased from Merck. Purified water was obtained with a Milli-Q water purification system (Millipore). Buffers (ionic strength 0.05) were prepared using phosphoric acid, sodium phosphate and disodium phosphate. The pH of the buffers was adjusted to the final value by using an Orion Research 501 digital ionalyser and the electrodes were calibrated with standard buffer solutions (pH 3.00 and 7.00; Merck). Prior to use, the buffers were filtered through a membrane filter (0.2 μm).

Molecular descriptors

Retention data were correlated with the following molecular descriptors: n_c, which was obtained from the empirical formula; experimentally determined log $P_{0/w}$ values, taken from the literature; and calculated log P values, making use of Rekker's hydrophobic fragmental system.

The following equation is used for the calculation of log P values:

$$\log P = \sum_i a_n f_n \quad (4)$$

where f is the hydrophobic fragmental constant, the lipophilicity contribution of a constituent part of a structure to the total lipophilicity, and a is a numerical factor indicating the incidence of a given fragment in the structure.

Statistics

Statistical analysis of the retention data was performed with the Statistical Package for Social Sciences (SPSS) [28]. This program runs on an IBM PC or compatible computers. The orthogonal regression program was written in our laboratory. It is written in Basic and runs on an Apple II computer.

RESULTS AND DISCUSSION

First the usefulness of both descriptors was examined with literature data. As it is our intention to use the eventual model for the prediction of first guess conditions for structurally unrelated compounds, only data sets corresponding to this description were studied. Hafkenscheid and Tomlinson [29] investigated the relationship between experimental log P values of strongly basic compounds ($pK_a > 7.5$) and reversed-phase liquid chromatographic capacity factors (log k') of partially ionized solutes using an aqueous methanol mobile phase. Good correlations (multiple $R = 0.974$) were obtained after correction for ionization effects. This, however, required a

knowledge of the mobile phase pH and solute pK_a values (both under aqueous and mobile phase conditions). As the pK_a will often not be known to the chromatographer, we studied the same relationship without taking into account ionization effects. At pH 4.0 a good correlation was obtained (multiple $R = 0.944$). The equation was also found highly significant ($p < 0.00005$). At pH 7.0 similar results were obtained (multiple $R = 0.863$ and $p < 0.00005$). By using the descriptor n_c instead of log P, less satisfactory results were found, as expected. At pH 4.0 a correlation of 0.413 and a significance level of 0.0404 were obtained for the equation. At pH 7.0 similar results were derived (multiple $R = 0.547$ and $p = 0.0038$).

De Biasi and Lough [30] studied the suitability of retention data of non-ionized organic bases ($pK_a > 7$) for the estimation of the lipophilicity characteristics on a styrene–divinylbenzene stationary phase. A correlation coefficient of 0.906 was obtained for the relationship between log k' and experimentally determined log P values. The equation was found to be significant at $p = 0.0001$. Similar results were obtained when we used the calculated log P values (multiple $R = 0.843$ and $p = 0.0043$). When we replaced log P with n_c, a correlation coefficient of 0.632 and a significance level of 0.0276 were obtained for the equation. These results demonstrate that the descriptor log P is indeed more suitable than n_c for retention prediction purposes. The descriptor n_c can be used, but one must be aware that this is a very rough descriptor of a drug's polarity. As our purpose is not to obtain the best prediction, which would certainly require extensive structural information, but rather an acceptable prediction, the suitability of n_c for retention prediction purposes will be investigated further in this paper.

Most of the retention prediction studies were performed on reversed-phase C_{18} columns. Our laboratory has several years of experience with the LiChrosorb cyanopropyl (CN) column. It has been shown that most drug analyses can be carried out with this type of column [31]. This study was therefore performed with the LiChrosorb CN column. The presence of residual silanol sites on the surface of such a chemically bonded alkylsilica stationary phase plays an important role. The retention behaviour of a solute then becomes the result not only of a partition process between the stationary and the mobile phase (comparable to the log P value of a solute), but also of adsorption on the residual silanol sites. These silanophilic interactions therefore have to be eliminated either by the use of additives in the mobile phase, such as propylamine, or by the use of buffers [32,33]. Recently, polymeric columns, based for example on styrene–divinylbenzene polymers, were introduced. On such a type of column the silanophilic interactions are non-existent. However, polymeric columns suffer from a low plate number in comparison with traditional columns [34]. In this study, the LiChrosorb CN column was used in combination with a buffered mobile phase.

TABLE I

CHROMATOGRAPHIC DATA FOR THE ACIDIC COMPOUNDS ON A LICHROSORB CN COLUMN WITH THE MOBILE PHASE METHANOL–PHOSPHATE BUFFER: (A) 10:90, (B) 30:70 AND (C) 50:50

No.	Compound	Empirical formula	pK_a	Log P (exp.)	Log P (calc.)	Log k'		
						A	B	C
1	Salicylic acid	$C_7H_6O_3$	3.0	2.26	1.28	−0.069	−0.292	−0.398
2	Nipasol	$C_{10}H_{12}O_3$	8.4	3.04	2.70	0.399	−0.054	−0.309
3	Furosemide	$C_{12}H_{11}ClN_2O_5S$	3.9	−0.83	1.47	0.611	0.084	−0.280
4	Chlorthalidone	$C_{14}H_{11}ClN_2O_4S$	9.3	2.82	0.54	0.254	−0.169	−0.352
5	Flufenamic acid	$C_{14}H_{10}F_3NO_2$	3.9	5.62	3.91	1.360	0.553	−0.112
6	Bumetanide	$C_{17}H_{20}N_2O_5S$	N.A.[a]	N.A.	2.63	0.832	0.170	−0.245
7	Diethylstilbestrol	$C_{18}H_{20}O_2$	10.9	5.07	5.79	1.228	0.352	−0.239
8	Sulindac	$C_{20}H_{17}FO_3S$	4.7	3.01	3.81	1.145	0.371	−0.158

[a] N.A. = Not available.

TABLE II

CHROMATOGRAPHIC DATA FOR THE BASIC COMPOUNDS ON A LICHROSORB CN COLUMN WITH THE MOBILE PHASE METHANOL–PHOSPHATE BUFFER: (A) 10:90, (B) 30:70 AND (C) 50:50

No.	Compound	Empirical formula	pK_a	Log P (exp.)	Log P (calc.)	Log k'		
						A	B	C
9	Amphetamine	$C_9H_{13}N$	9.9	1.76	1.96	−0.332	−0.403	−0.391
10	Triamterene	$C_{12}H_{11}N_7$	6.2	0.98	–	0.113	−0.177	−0.297
11	Metoclopramide	$C_{14}H_{22}ClN_3O$	9.4	2.62	1.95	0.432	−0.037	−0.219
12	Diazepam	$C_{16}H_{13}ClN_2O$	3.4	2.80	2.95	0.755	0.277	−0.119
13	Triflupromazine	$C_{18}H_{19}F_3N_2S$	9.3	5.19	5.08	1.409	0.652	0.021
14	Mianserin	$C_{18}H_{20}N_2$	6.5	3.36	3.48	0.597	0.367	−0.062
15	Amitriptyline	$C_{20}H_{23}N$	9.4	5.04	5.55	1.136	0.490	0.000
16	Aprindine	$C_{22}H_{30}N_2$	10.0	4.76	5.92	0.346	0.086	−0.043
17	Mebeverine	$C_{25}H_{35}NO_5$	N.A.[a]	N.A.	6.27	1.175	0.402	−0.036

[a] N.A. = Not available.

and more specifically a phosphate buffer because, according to Wang and Lien [35], for acidic and neutral solutes this buffer appears to give partition coefficients closer to the values obtained with the *n*-octanol–water system than other buffers, such as acetate and hydrogencarbonate buffers.

The type of organic modifier also affects RPLC retention behaviour. The organic modifier methanol has been shown to interfere the least with the hydrophobic partition mechanism in RPLC among commonly used organic solvents, such as acetonitrile and tetrahydrofuran (THF) [36–39]. For this reason, an aqueous methanol mobile phase was used for this structure–retention study. The role of methanol may be much more active, owing to its interaction with surface silanols. Using methanol as solvent in a first guess study is compatible with optimization using the solvent strength, where acetonitrile and THF are introduced at a later stage. In fact, any solvent can be used initially using relationships from the literature that allow estimates of equivalent %B values for different solvents B to achieve the same retention time [26].

The investigation of the influence of the number

TABLE III

CHROMATOGRAPHIC DATA FOR THE NEUTRAL COMPOUNDS ON A LICHROSORB CN COLUMN WITH THE MOBILE PHASE METHANOL–PHOSPHATE BUFFER: (A) 10:90, (B) 30:70 AND (C) 50:50

No.	Compound	Empirical formula	Log P (exp.)	Log P (calc.)	Log k'		
					A	B	C
18	Phenacetin	$C_{10}H_{13}NO_2$	1.58	1.87	0.147	−0.177	−0.319
19	Pentoxifylline	$C_{13}H_{18}N_4O_3$	0.29	−1.67	0.179	−0.152	−0.277
20	Griseofulvin	$C_{17}H_{17}ClO_6$	2.18	1.16	0.882	0.225	−0.174
21	Testosterone	$C_{19}H_{28}O_2$	3.32	3.95	0.763	0.143	−0.210
22	Methyltestosterone	$C_{20}H_{30}O_2$	3.36	4.47	0.810	0.202	−0.180
23	Triamcinolone	$C_{21}H_{27}FO_6$	1.16	−3.08	0.120	−0.183	−0.346
24	Progesterone	$C_{21}H_{30}O_2$	3.87	4.43	N.A.[a]	0.372	−0.101
25	Betamethasone	$C_{22}H_{29}FO_5$	1.94	−0.93	0.400	−0.035	−0.289

[a] N.A. = Not available.

of carbons in the molecule, log P and the percentage of organic modifier (methanol) in the mobile phase on the retention was carried out with 25 pharmaceutical compounds of various polarities. Molecules with different numbers of carbons and also different functional groups or acidic–basic properties were selected. The chromatographic data are listed in Tables I, II and III for acidic, basic and neutral compounds, respectively.

Prior to the application of multiple regression techniques, a comparison was made between linear and orthogonal regression. The linear (multiple) regression techniques assume that the error exists only in the y direction (response). In fact, both descriptors are only indications of the hydrophobicity of a molecule and are therefore also subject to error. In such a situation orthogonal regression techniques [40], which are equivalent to determining the first principal component (PC1) of a matrix consisting of i rows (objects) and two columns (variables), are preferred. The results, including the slope and intercept, for the set of acidic compounds and a mobile phase composed of methanol–phosphate buffer (10:90), are listed in Table IV. The difference in the results obtained between the two regression techniques is clearly small. Similar results were found for the other chromatographic conditions (30 and 50% methanol) and also for the set of basic and neutral compounds. Further investigation was therefore restricted to the more usual multiple regression analysis.

Multiple regression analysis is a frequently used statistical method to study the relationships among variables. The aim is to describe a dependent variable by a set of independent variables, more specifically to investigate the following linear relationship:

$$Y_i = \sum_{j=1}^{k} \beta_j X_{ij} + e_i \qquad (5)$$

where Y_i represents the dependent variable, X_1 to X_k the different independent variables and e_i the residuals. The regression coefficients β_j are estimated by multiple linear regression. The most commonly used measure of the goodness of fit of the model is R^2 or its square root, called the multiple R. The adjusted R^2, however, reflects best the goodness of fit of the model in the population [28]. For the validation of linear models the F-statistic and its corresponding

TABLE IV

COMPARISON OF LINEAR *VERSUS* ORTHOGONAL REGRESSION FOR THE RELATIONSHIP LOG k' *VERSUS* THE DESCRIPTORS n_c AND LOG P (CALCULATED) FOR THE ACIDIC COMPOUNDS

The mobile phase was methanol–phosphate buffer (10:90).

Parameter	Linear regression		Orthogonal regression	
	n_c	Log P	n_c	Log P
Slope	0.0919	0.2389	0.0924	0.2431
Intercept	−0.5657	−0.0443	−0.5729	−0.0577

significance level p are available. These are described in most textbooks on regression analysis. Different procedures can be applied for the selection of appropriate variables, *i.e.*, stepwise regression, forward selection and backward elimination. Stepwise regression analysis is the most commonly used procedure. Less frequently used are the other two procedures, which need not result in the same equation. However, as stepwise regression investigates the significance of previously entered variables, which is not the case for the other two procedures, the stepwise technique is to be preferred.

Multiple regression was carried out by first considering the dependent variable log k' and the independent variables n_c, X_m, the interaction term and the quadratic terms.

The regression equation (eqn. 6) was derived for the acidic compounds by stepwise regression analysis (values of p-to-enter and p-to-remove 0.05 and 0.10, respectively). The regression coefficients are accompanied by the 95% confidence intervals according to the t-test. The number of data points (n), the standard deviation of the residuals (s), the multiple correlation coefficient (Mult. R), the adjusted R^2, the calculated F-value of the derived equation and its significance level (p) are also presented.

In eqn. 6, the retention is related to the number of carbons in the molecule on the one hand and the interaction between the percentage of organic modifier in the mobile phase and the number of carbons on the other. Both parameters were found to be significant at $p < 0.00005$. The quadratic terms and

$$\log k' = 0.1048\ (\pm 0.0271)n_c - 1.7694 \cdot 10^{-3}\ (\pm 0.4099 \cdot 10^{-3})n_c X_m - 0.5308\ (\pm 0.3524) \quad (6)$$

$n = 24$; $s = 0.2297$; Mult. $R = 0.9094$; Adj. $R^2 = 0.8105$; F(eqn.) = 50.19; $p < 0.00005$

$$\log k' = 0.0521\ (\pm 0.0250)n_c - 0.0198\ (\pm 0.0072)X_m - 0.0578\ (\pm 0.4926) \quad (7)$$

$n = 27$; $s = 0.2954$; Mult. $R = 0.8239$; Adj. $R^2 = 0.6521$; F(eqn.) = 25.36; $p < 0.00005$

$$\log k' = -0.0176\ (\pm 0.0061)X_m + 0.6226\ (\pm 0.2124) \quad (8)$$

$n = 23$; $s = 0.2269$; Mult. $R = 0.7953$; Adj. $R^2 = 0.6150$; F(eqn.) = 36.15; $p < 0.00005$

$$\log k' = 0.0249\ (\pm 0.0181)n_c - 0.0215\ (\pm 0.0050)X_m + 0.3978\ (\pm 0.3297) \quad (9)$$

$n = 47$; $s = 0.2738$; Mult. $R = 0.8080$; Adj. $R^2 = 0.6371$; F(eqn.) = 41.37; $p < 0.05$

$$\log k' = -9.9935 \cdot 10^{-3}\ (\pm 0.0108)X_m + 0.2995\ (\pm 0.1156)\log P - 5.2776 \cdot 10^{-3}\ (\pm 3.3843 \cdot 10^{-3})X_m \log P + 0.1029\ (\pm 0.3692) \quad (10)$$

$n = 24$; $s = 0.2073$; Mult. $R = 0.9305$; Adj. $R^2 = 0.8457$; F(eqn.) = 43.01; $p < 0.00005$

$$\log k' = -0.0149\ (\pm 0.0116)X_m + 0.0361\ (\pm 0.0187)\log P^2 - 3.0943 \cdot 10^{-3}\ (\pm 2.9687 \cdot 10^{-3})X_m \log P + 0.4519\ (\pm 0.3475) \quad (11)$$

$n = 21$; $s = 0.2622$; Mult. $R = 0.8932$; Adj. $R^2 = 0.7621$; F(eqn.) = 22.35; $p < 0.00005$

$$\log k' = -0.0198\ (\pm 0.0074)X_m + 0.1269\ (\pm 0.0652)\log P + 0.3518\ (\pm 0.3536) \quad (12)$$

$n = 27$; $s = 0.3042$; Mult. $R = 0.8121$; Adj. $R^2 = 0.6312$; F(eqn.) = 23.25; $p < 0.00005$

$$\log k' = 0.3352\ (\pm 0.0871)X_m - 5.1533 \cdot 10^{-3}\ (\pm 1.7315 \cdot 10^{-3})X_m \log P + 0.4067\ (\pm 0.2532) \quad (13)$$

$n = 24$; $s = 0.2414$; Mult. $R = 0.8729$; Adj. $R^2 = 0.7392$; F(eqn.) = 33.60; $p < 0.00005$

$$\log k' = -0.0182\ (\pm 4.76 \cdot 10^{-3})X_m + 0.0520\ (\pm 0.0283)\log P + 0.5811\ (\pm 0.1672) \quad (14)$$

$n = 23$; $s = 0.1763$; Mult. $R = 0.8880$; Adj. $R^2 = 0.7674$; F(eqn.) = 37.30; $p < 0.00005$

$$\log k' = -8.6645 \cdot 10^{-3}\ (\pm 8.1355)X_m + 0.2903\ (\pm 0.1237)\log P - 4.6905 \cdot 10^{-3}\ (\pm 3.4548 \cdot 10^{-3})X_m \log P + 0.0415\ (\pm 0.2826) \quad (15)$$

$n = 23$; $s = 0.1391$; Mult. $R = 0.9355$; Adj. $R^2 = 0.8554$; F(eqn.) = 44.37; $p < 0.00005$

also the term taking into account the pure effect of the percentage of organic modifier in the mobile phase were found to be insignificant. Initially the latter was introduced in the regression equation ($p < 0.00005$), but in a final step this variable was removed ($p = 0.8277$). The introduction of the interaction term in the equation was expected, because the change in log k' by changing the percentage of organic modifier depends on the number of carbons in the molecule (Fig. 1). This is true also for the basic and neutral compounds (Figs. 2 and 3).

About 80% of the model was explained by means of the selected parameters, which was considered satisfactory taking the diversity of the pharmaceutical compounds into account.

For the set of basic compounds the regression equation (eqn. 7) was obtained by stepwise regression (value of p-to-enter and p-to-remove 0.05 and 0.10, respectively). In comparison with eqn. 6, the term $n_c X_m$ was substituted by X_m. The latter term and the term n_c were found to be significant at $p < 0.00005$ and $p = 0.0002$, respectively.

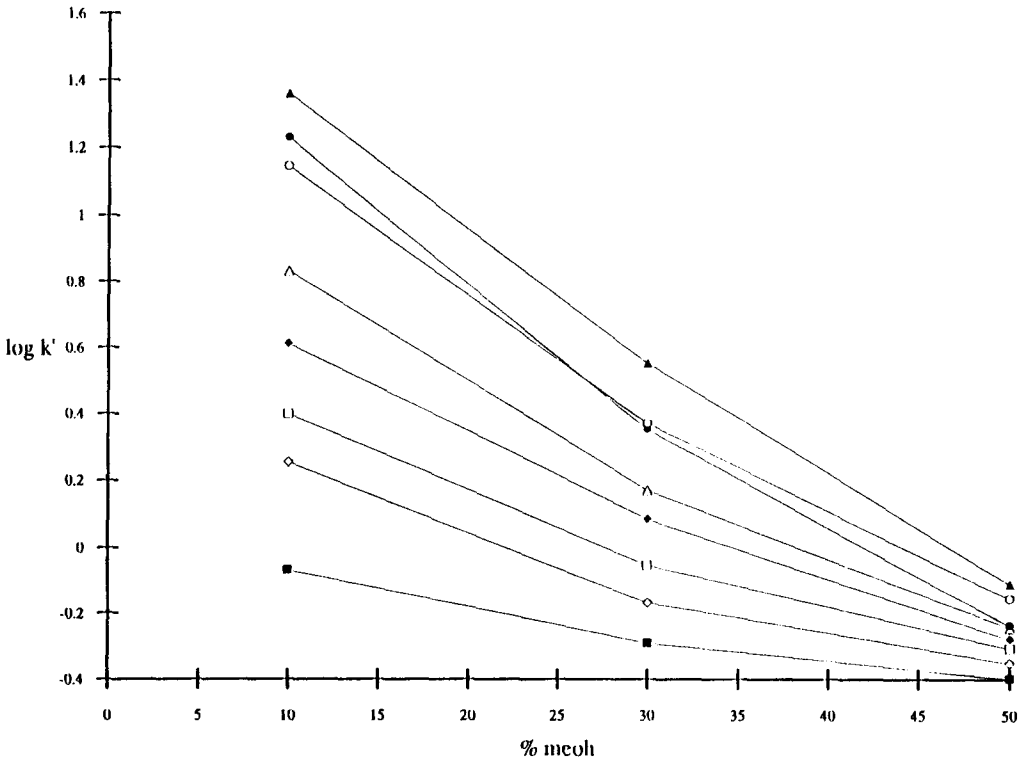

Fig. 1. Plot of log k' *versus* the volume percentage of methanol (meoh) with phosphate buffer (pH 3, $u = 0.05$) for acidic compounds. ■ = Salicylic acid; □ = nipasol; ◆ = furosemide; ◇ = chlorthalidone; ▲ = flufenamic acid; △ = bumetanide; ● = diethylstilbestrol; ○ = sulindac.

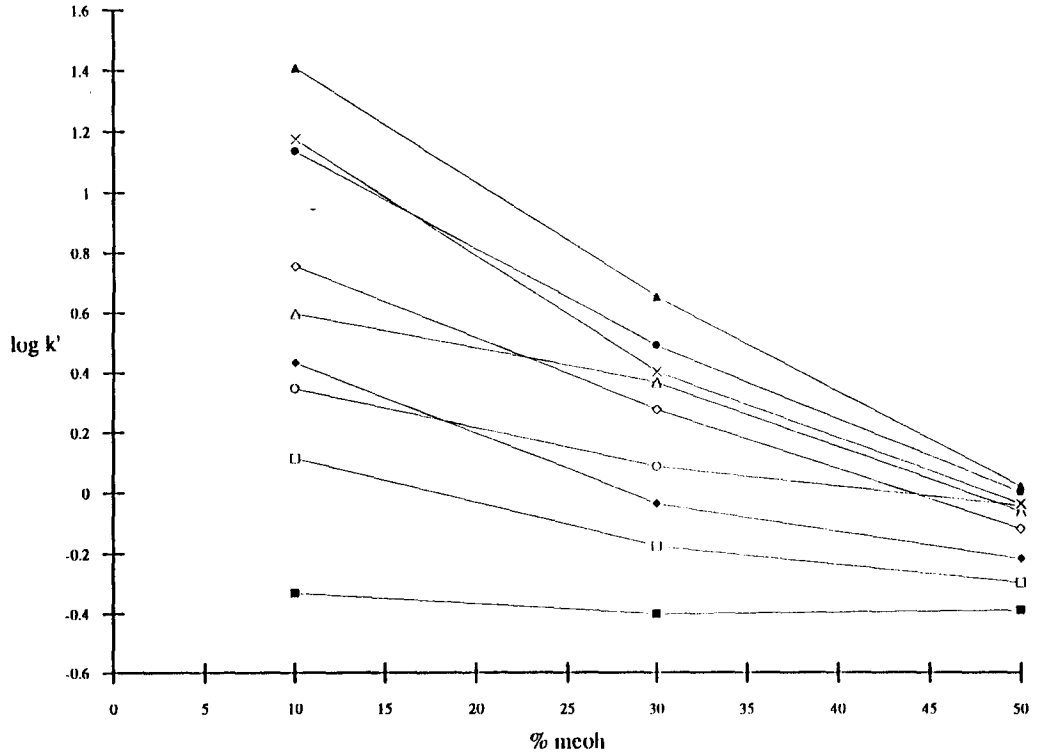

Fig. 2. Plot of log k' *versus* the volume percentage of methanol with phosphate buffer (pH 3, $u = 0.05$) for basic compounds. ■ = Amphetamine; □ = triamterene; ◆ = metoclopramide; ◇ = diazepam; ▲ = triflupromazine; △ = mianserin; ● = amitriptyline; ○ = aprindine; × = mebeverine.

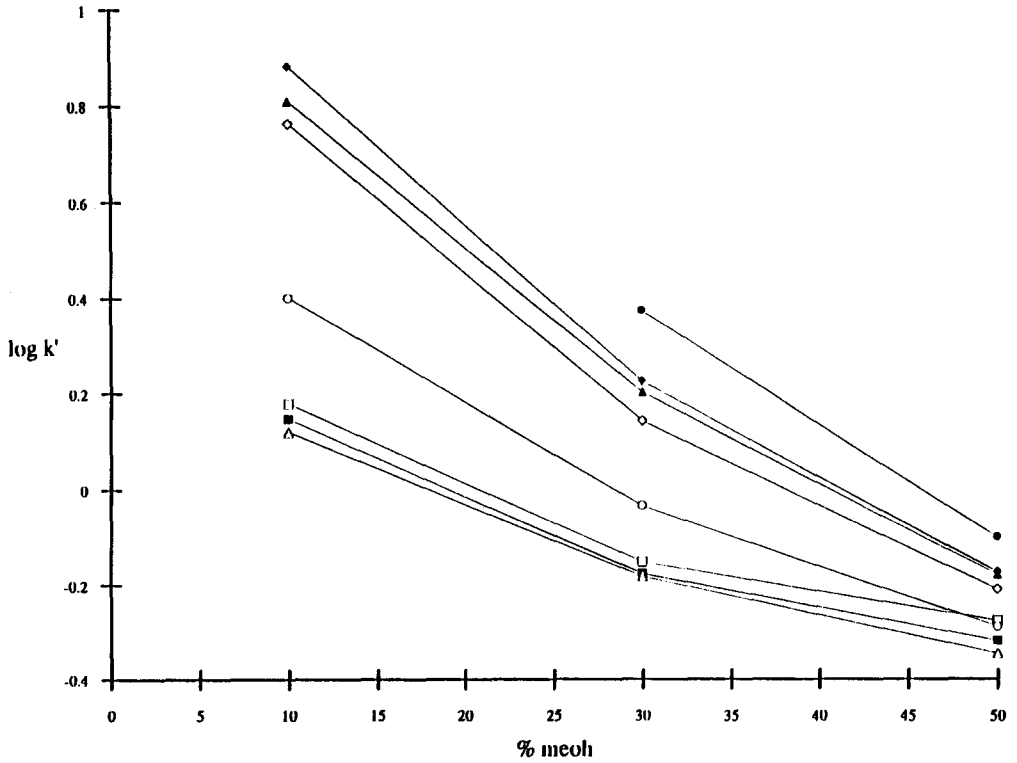

Fig. 3. Plot of log k' *versus* the volume percentage of methanol with phosphate buffer (pH 3, $u = 0.05$) for neutral compounds. ■ = Phenacetin; □ = pentoxifylline; ◆ = griseofulvin; ◇ = testosterone; ▲ = methyltestosterone; △ = triamcinolone; ● = progesterone; ○ = betamethasone.

For the neutral compounds eqn. 8 was derived by stepwise regression analysis (value of p-to-enter and p-to-remove 0.05 and 0.10, respectively). The only significant factor ($p < 0.00005$) is the volume percentage of organic modifier. The number of carbons in the molecule and the interaction term were found to be insignificant with p-values of 0.1145 and 0.8871, respectively, probably because the diversity in polarity for the set of neutral compounds studied was insufficient or that the dispersion of the results is too high.

The mobile phase consisted of a methanol–phosphate buffer (pH 3). At this pH most of the acidic molecules are present in non-ionic form, which is also the case for the neutral compounds. Therefore, the neutral and acidic compounds were considered together (eqn. 9). By stepwise regression the percentage of organic modifier ($p < 0.00005$) and the number of carbons in the molecule ($p = 0.0081$) were introduced into the regression equation. The interaction term was found to be insignificant ($p = 0.1676$).

In a second step, the usefulness of the descriptor log P, calculated according to Rekker's fragment system, was investigated. Regarding the application of Rekker's method, several problems were encountered. For the compound triamterene the contribution of the pteridine fragment was unavailable. As further division of this fragment would lead to a serious underestimation of the log P value, the experimental value was used. The application of Rekker's method to steroid compounds (testosterone, methyltestosterone, triamcinolone, progesterone and betamethasone) would certainly result in an overestimation of the log P value. For this reason, the log P values were calculated by using the experimental log P value of desoxycorticosterone ($\log P_{exp.} = 2.80$) as representative of the base structure of such compounds and also taking into account the Rekker fragment values for the differ-

ences in functional groups. As the experimental log P values were available for most compounds, these were also used for comparison with the calculated values (Tables I, II and III).

For the set of acidic compounds the regression equation (eqn. 10) was derived by stepwise regression analysis (value of p-to-enter and p-to-remove 0.05 and 0.10, respectively). In eqn. 10, the retention is related to X_m ($p = 0.0681$), log P ($p < 0.00005$) and the interaction term ($p = 0.0040$); the quadratic terms were found to be insignificant. The latter is in accordance with eqn. 1.

Similar results were obtained with the experimental log P values (eqn. 11). However, log P was not included in the regression equation ($p = 0.9437$), but rather its quadratic form ($p = 0.0008$). The terms X_m and the interaction term were found to be significant at $p = 0.0145$ and $p = 0.0420$, respectively.

For the set of basic compounds eqn. 12 was obtained by stepwise regression analysis (value of p-to-enter and p-to-remove 0.05 and 0.10, respectively). In eqn. 12, in comparison with the set of acidic compounds, the interaction term was found to be insignificant ($p = 0.0854$).

With the experimental log P values (eqn. 13), X_m ($p = 0.8022$) was replaced with the interaction term ($p < 0.00005$). The term log P was found to be significant at $p < 0.00005$.

For the neutral compounds, similar results to those for the basic compounds were derived (eqn. 14). The interaction term was found to be insignificant at $p = 0.0677$. In comparison with eqn. 8 the descriptor log P was introduced into the regression equation. This resulted in a change in R^2 of 0.1560, which was found to be significant at $p = 0.0010$. The term X_m was found to be significant at $p < 0.00005$.

Here also different results were obtained with the experimental log P values (eqn. 15). The interaction term was also found to be significant ($p = 0.0104$). The other terms, X_m, log P and the interaction term, were found to be significant at $p = 0.0380$, $p = 0.0001$ and $p = 0.0104$, respectively.

Comparison of the results obtained with the experimental log P values on the one hand and the log P values, calculated according to Rekker's fragment system, on the other demonstrates that the calculated descriptor log P values can be used instead of the experimental values.

Cross-validation of the model

Many regression analyses study only the estimation of coefficients and the quality of the fit. It is necessary, however, also to validate regression methods and to investigate how good the prediction is. This requires cross-validation. Cross-validation of a model is performed by leaving out each molecule in turn from the data set, by computing the regression equation without the molecule in question and then predicting the retention for the same molecule. This approach is called the leave-one-out method (LOOM). The predicted (estimated) value is then compared with the experimental (observed) value through a PRESS (prediction error sum of squares) value:

$$\text{PRESS} = \sum_i [y_{(i)} - \hat{y}_{(i)}]^2 \tag{16}$$

where $y_{(i)}$ and $\hat{y}_{(i)}$ represent the experimental value and the predicted value after the LOOM, respectively. PRESS is used to assess the predictive performance of the model. The model that produces the lowest PRESS value is preferred [41,42].

From the chromatographic point of view, the accuracy of the selection of initial chromatographic conditions with the model is most interesting. For this purpose eqns. 6 and 7 were rearranged as follows:

$$X_m = \frac{\log k' - An_c - C}{Bn_c} \tag{17}$$

$$X_m = \frac{\log k' - An_c - C}{B} \tag{18}$$

The following equations were derived for the model including the descriptor log P for acidic and basic (neutral) compounds, respectively:

$$X_m = \frac{\log k' - B\log P - D}{A + C\log P} \tag{19}$$

$$X_m = \frac{\log k' - B\log P - C}{A} \tag{20}$$

The cross-validation was hence performed by comparing the percentage of organic modifier required to obtain a given k', predicted by eqns. 17–20 and an experimental value. The latter was calculated by interpolation using the slopes and intercepts derived for the relationship of log k' *versus* X_m for

each compound. The k' values were selected to obtain a percentage of organic modifier situated in the investigated area, *i.e.*, between 10 and 50% methanol. The same k' values, together with the cross-validation regression coefficients and either the number of carbons in the molecule or the calculated log P value were used to predict the percentage of organic modifier from eqns. 17–20. These results are given in Tables V and VI. For amphetamine no interpolated value was obtained because of the very low retention of this molecule ($k' < 0.5$). For triflupromazine and flufenamic acid, which exhibit a very high retention ($k' > 20$) for a mobile phase containing a small amount of the organic modifier (10% methanol), the prediction was unsatisfactory in comparison with the other molecules. Similar results, but less pronounced, were obtained with the descriptor log P. For aprindine less satisfactory results were also found. These molecules, except amphetamine, were designated as outliers on a statistical basis. Such values certainly influence the quality of the prediction. However, outliers can also be considered as containing a lot of information and, bearing in mind the purpose of this study, these values were taken into account.

The same strategy was used to determine the quality of the prediction for the neutral compounds. For the determination of the percentage of organic modifier from eqns. 18 and 20, the cross-validation regression coefficients for the set of acidic and

TABLE V

ABSOLUTE DIFFERENCES BETWEEN THE INTERPOLATED VALUE (I) AND THE PERCENTAGE OF ORGANIC MODIFIER PREDICTED BY THE MODEL INCLUDING THE DESCRIPTOR n_c (II) OR CALCULATED LOG P (III) FOR ACIDIC COMPOUNDS

Compound	k'	Methanol (%)			|II − I|	|III − I|
		I	II	III		
Salicylic acid	0.5	36	47	48	11	12
Nipasol	0.5	48	46	50	2	2
	1.0	31	29	38	2	7
Furosemide	0.5	50	48	47	2	3
	1.0	36	34	29	2	7
	3.0	15	11	0	4	15
Chlorthalidone	0.5	44	51	44	7	0
	1.0	24	39	18	15	6
Flufenamic acid	1.0	46	36	40	10	6
	3.0	33	15	24	18	9
	5.0	27	6	16	21	11
Bumetanide	0.5	51	52	50	1	1
	1.0	39	42	37	3	2
	3.0	22	27	17	5	5
	5.0	13	19	7	6	6
Diethylstilbestrol	0.5	50	52	55	2	5
	1.0	42	43	49	1	7
	3.0	29	27	38	2	9
	5.0	23	20	33	3	10
	10.0	15	10	27	5	12
Sulindac	1.0	44	45	41	1	3
	3.0	29	32	25	3	4
	5.0	20	26	17	6	3
	10.0	13	17	7	4	6

TABLE VI

ABSOLUTE DIFFERENCES BETWEEN THE INTERPOLATED VALUE (I) AND THE PERCENTAGE OF ORGANIC MODIFIER PREDICTED BY THE MODEL INCLUDING THE DESCRIPTOR n_c (II) OR CALCULATED LOG P (III) FOR BASIC COMPOUNDS

Compound	k'	Methanol (%)			|II − I|	|III − I|
		I	II	III		
Amphetamine	0.5	N.A.[a]	42	49	N.A.	N.A.
Triamterene	0.5	46	45	39	1	7
	1.0	17	30	25	13	8
Metoclopramide	1.0	34	34	30	0	4
Diazepam	1.0	44	39	35	5	9
	3.0	22	14	11	8	11
Triflupromazine	1.0	50	43	49	7	1
	3.0	36	16	23	20	13
	5.0	30	4	10	26	20
Mianserin	1.0	49	44	39	5	10
	3.0	17	20	14	3	3
Amitriptyline	1.0	49	49	54	0	5
	3.0	32	24	28	8	4
	5.0	24	12	16	12	8
Aprindine	1.0	43	59	61	16	18
Mebeverine	1.0	47	71	61	24	14
	3.0	31	45	35	14	4
	5.0	24	33	23	9	1
	10.0	14	17	7	3	7

[a] N.A. = Not available.

TABLE VII

ABSOLUTE DIFFERENCES BETWEEN THE INTERPOLATED VALUE (I) AND THE PERCENTAGE OF ORGANIC MODIFIER PREDICTED BY THE MODEL INCLUDING THE DESCRIPTOR n_c (II) OR CALCULATED LOG P (III) FOR THE NEUTRAL MOLECULES

Compound	k'	Methanol (%)			|II − I|	|III − I|
		I	II	III		
Phenacetin	0.5	45	55	50	10	10
	1.0	20	32	39	12	19
Pentoxifylline	0.5	49	53	44	4	5
	1.0	23	35	28	12	5
Griseofulvin	1.0	43	38	34	5	9
	3.0	25	16	5	9	20
Testosterone	1.0	40	39	44	1	4
	3.0	20	20	16	0	4
	5.0	11	11	3	0	8
Methyltestosterone	1.0	41	40	45	1	4
	3.0	22	22	17	0	5
	5.0	13	13	4	0	9
Triamcinolone	1.0	18	42	24	24	6
Progesterone	1.0	46	39	44	7	2
	3.0	26	23	18	3	8
	5.0	16	15	5	1	11
Betamethasone	1.0	31	42	29	11	2

neutral compounds were used. These results are given in Table VII.

Through a PRESS value the predictive performance of the model was assessed and the most appropriate descriptor selected. From the results in Table VIII it can be concluded that the descriptor log P should be preferred to n_c. On the basis of these results and also the literature results the descriptor log P should be preferred.

TABLE VIII

PRESS VALUES FOR THE MODEL INCLUDING THE DESCRIPTOR n_c OR CALCULATED LOG P FOR THE SET OF ACIDIC, BASIC AND NEUTRAL COMPOUNDS

Group	PRESS	
	n_c	Log $P_{calc.}$
Acidic compounds ($n = 24$)	1468	1289
Basic compounds ($n = 18$)	2744	1701
Neutral compounds ($n = 17$)	1268	1439

TABLE IX

MEAN OF THE ABSOLUTE DIFFERENCES BETWEEN THE INTERPOLATED VALUE AND THE PERCENTAGE OF ORGANIC MODIFIER PREDICTED BY THE MODEL INCLUDING THE DESCRIPTOR LOG P (CALCULATED)

Group	Mean methanol content ± S.D. (%)
Acidic compounds ($n = 24$)	6 ± 4
Basic compounds ($n = 18$)	8 ± 5
Neutral compounds ($n = 17$)	8 ± 5

As the mobile phase consisted of methanol–phosphate buffer (pH 3), the ionization effects certainly play an important role. This is most certainly so for the basic compounds, which resulted in less satisfactory results for this type of compound in comparison with the acidic compounds. Other workers have also found that basic compounds cause problems [43]. Corrections for ionization effects are possible [29], but very impractical with regard to our purpose. Apart from ionization effects, stationary phase effects are likely to play an important role with basic compounds.

The results for the selection of initial chromatographic conditions are presented in Table IX as the mean (and standard deviation) of the absolute differences between the interpolated and the predicted value for the descriptor log P. The initial chromatographic conditions to obtain a certain k' will be under- or overestimated within a range of 6–8% methanol, which from the chromatographic point of view is satisfactory as a first guess, considering the diversity of molecular structures investigated. With the equation including the descriptor n_c, the range varied from 6 to 10% methanol but with a larger standard deviation of the mean prediction error.

CONCLUSIONS

The parameter log P, calculated according to the Rekker fragment system, has been found to be the

most suitable descriptor for the prediction of initial chromatographic conditions. The model, including the parameter log P, can cover a wide composition range (organic modifier contents between 10 and 50%) to predict a capacity factor situated between 1 and 10. However, in some instances, the prediction was less satisfactory. A source of error causing this bad prediction is the calculation of the log P values. Steric effects, for instance, cannot be taken into account. Hence, calculated log P values are certainly limited in their ability to predict retention as a function of the percentage of organic modifier in the mobile phase. However, it was necessary for this approach. Some suggestions are available in the literature for improving the prediction of log P values. However, our purpose was not to obtain the best prediction, but an acceptable one. Moreover, the regression results with the experimental log P values were very similar to these with the calculated log P values, indicating the possibility of applying the latter values for retention prediction and also the prediction of initial chromatographic conditions. Hence, log P values calculated according to Rekker seem sufficient for our purposes.

In the near future the retention prediction model will be incorporated into a first guess expert system. On the basis of the log P value of a compound the initial solvent composition can be predicted once the desired capacity factor has been defined. In some instances, however, Rekker's method is not applicable. The experimental log P can then be used, provided that this value is available. If not, the conditions can still be predicted on the basis of the total number of carbons in the molecule, *i.e.*, the retention model including the descriptor n_c will also be incorporated in the first guess expert system. The sample can, on the other hand, also contain different solutes of interest. In such a case, the average log P value of the compounds will be used to select the first guess.

ACKNOWLEDGEMENT

This work was financially supported by Fonds voor Geneeskundig en Wetenschappelijk Onderzoek (FGWO).

REFERENCES

1 M. De Smet, G. Musch, A. Peeters, L. Buydens and D. L. Massart, *J. Chromatogr.*, 485 (1989) 237.
2 J. W. Dolan, D. C. Lommen and L. R. Snyder, *J. Chromatogr.*, 485 (1989) 91.
3 P. Chaminade, *Ph. D. Thesis*, University of Paris-Sud, Paris, 1992.
4 S. Heinisch, J. L. Rocca and M. Kolosky, *Chromatographia*, 29 (1990) 483.
5 L. R. Snyder and J. J. Kirkland, *Introduction to Modern Liquid Chromatography*, Wiley-Interscience, New York, 1979.
6 P. Jandera, *Chromatographia*, 19 (1984) 101.
7 P. Jandera, *J. Chromatogr.*, 314 (1984) 13.
8 P. Jandera, *J. Chromatogr.*, 352 (1986) 91.
9 P. Jandera, *J. Chromatogr.*, 352 (1986) 111.
10 K. Jinno, M. Yamagami and M. Kuwajima, *Chromatographia*, 25 (11) (1988) 974.
11 K. Jinno, *A Computer-Assisted Chromatography System*, Hüthig, Heidelberg, 1990.
12 H. A. Cooper and R. J. Hurtubise, *J. Chromatogr.*, 360 (1986) 313.
13 S. F. Y. Li and H. K. Lee, *Chromatographia*, 25 (1988) 515.
14 R. M. Smith and C. M. Burr, *J. Chromatogr.*, 485 (1989) 325.
15 K. Valko, *J. Liq. Chromatogr.*, 7 (1984) 1405.
16 *EluEx 1.0*, CompuDrug Chemistry, Budapest, 1991.
17 R. Kaliszan and K. Osmialowski, *J. Chromatogr.*, 506 (1990) 3.
18 C. Hansch and A. Leo, *Substituent Constants for Correlation Analysis in Chemistry and Biology*, Wiley, New York, 1979.
19 R. F. Rekker, *The Hydrophobic Fragmental Constant —Its Derivation and Application*, Elsevier, Amsterdam, 1977.
20 R. F. Rekker and H. M. de Kort, *Eur. J. Med. Chem. Chim. Ther.*, 14 (1979) 479.
21 R. E. Koopmans and R. F. Rekker, *J. Chromatogr.*, 285 (1984) 267.
22 Y. C. Martin, *Quantitative Drug Design*, Marcel Dekker, New York, 1978.
23 T. Braumann, *J. Chromatogr.*, 373 (1986) 191.
24 R. S. Tsai, N. El Tayar, B. Testa and Y. Ito, *J. Chromatogr.*, 538 (1991) 119.
25 M. Czok and H. Engelhardt, *Chromatographia*, 27 (1989) 5.
26 P. J. Schoenmakers, *Optimization of Chromatographic Selectivity*, Elsevier, Amsterdam, 1st ed., 1986.
27 M. Bogusz and R. Aderjan, *J. Chromatogr.*, 435 (1988) 43.
28 *SPSS/PC for the IBM PC/XT, Release 1.0*; SPSS, Chicago, 1984, Ch. 17.
29 T. L. Hafkenscheid and E. Tomlinson, *J. Chromatogr.*, 292 (1984) 305.
30 V. De Biasi and W. J. Lough, *J. Chromatogr.*, 353 (1986) 279.
31 M. De Smet, G. Hoogewijs, M. Puttemans and D. L. Massart, *Anal. Chem.*, 56 (1984) 2662.
32 D. Chan Leach, M. A. Stadalius, J. S. Berus and L. R. Snyder, *LC · GC Int.*, 1, No. 5 (1988) 22.
33 J. W. Dolan, *LC · GC Int.*, 2, No. 7 (1989) 18.
34 R. Vervoort, H. Hindriks, M. Vrieling and F. Maris, *18th International Symposium on Chromatography, Amsterdam, 1990*, abstract Mo-P-092.
35 P. Wang and E. J. Lien, *J. Pharm. Sci.*, 69 (1980) 662.

36 N. Tanaka, G. Goodell and B. L. Karger, *J. Chromatogr.*, 158 (1978) 233.
37 T. Braumann, G. Werber and L. H. Grimme, *J. Chromatogr.*, 261 (1983) 329.
38 B. L. Karger, J. R. Gant, A. Hartkopf and P. H. Wiener, *J. Chromatogr.*, 128 (1976) 65.
39 A. Bechalany, A. Tsantili-Kakoulidou, N. El Tayar and B. Testa, *J. Chromatogr.*, 541 (1991) 221.
40 J. Mandel, *The Statistical Analysis of Experimental Data*, Wiley, New York, 1964.
41 F. Mosteller and J. W. Tukey, *Data Analysis and Regression*, Addison-Wesley, 1977.
42 G. H. Golub, M. Heath and G. Wakba, *Technometrics*, 21 (1979) 215.
43 J. A. Lewis, D. C. Lommen, W. D. Raddatz, J. W. Dolan and L. R. Snyder, *J. Chromatogr.*, 592 (1992) 183.