Pergamon

0731-7085(95)01327-X

Journal of Pharmaceutical & Biomedical Analysis, Vol. 13, No. 4/5, pp. 369–376, 1995

0731-7085/95 $9.50 + 0.00

Use of multivariate characterization, design and analysis in assay optimization*

LARS STÅHLE,† ARSHAD MIAN and NATALIA BORG

Clinical Pharmacology, Karolinska Institute, Huddinge Hospital, S-191 86 Huddinge, Sweden

Abstract: A procedure is proposed for utilizing information from previous work on development of LC assays in order to facilitate the analysis of novel compounds related to those previously analysed. The procedure employs a multivariate method from the field of chemometrics, partial least squares analysis (PLS) to combine quantitative information on the chemical properties of a compound with a quantitative description of the column and the mobile phase and then to use this information to form a regression model for the retention time. A test of the procedure was made by using data on nucleoside analogues studied in our laboratory. Data obtained from chromatographic studies of seven compounds tested in a total of 28 combinations of columns and mobile phases (3–5 per compound) were used to calculate a PLS model. The model was then used to predict retention times of nine other substances and the results were compared with experimental data. The predictions were (115 ± 82%) (95% confidence interval) of the experimentally observed retention times. The results are encouraging and the method will be subject to further and extended investigations.

Keywords: *LC; experimental design; multivariate analysis.*

Introduction

LC is one of the most common tools in clinical pharmacology, pharmaceutical and biomedical laboratories doing analytical chemistry on biological samples to quantitate drug concentrations. Several methods have been proposed for the development of new LC assays and two methods seem to be the most prevalent. The most common is non-systematic experience (intuition combined with knowledge) and the second is expert systems. There are several examples also of experimental design based approaches with response surface modelling in the spirit of Box *et al.* [1]. In a separate paper [2] we have developed methods to optimize an LC assay for one or more compounds by using a multivariate method developed within the field of chemometrics [3], partial least squares analysis (PLS), which we have previously used extensively in pharmacology [4]. Here, we develop a method by which another, but related, problem can be approached.

In the development of new drugs it is by far the most common that a number of compounds within a chemical class of compounds, rather than structurally unrelated molecules, are studied. Several compounds within a series can be expected to reach the stage where a pharmacokinetic study is carried out necessitating the availability of an assay for the drug. Whenever a new compound in the class reaches the analytical laboratory it is naturally desirable that the experience gained from the previously investigated compounds (combined with the general knowledge of the analytical chemist) is utilized in the most efficient manner possible. In the present paper we use a similar approach as the one in the previous paper [2] to utilize knowledge on LC obtained through experience in a laboratory to identify the region from which an optimization can start. The approach will therefore have elements of an expert system since it provides knowledge-based suggestions by means of a computer program that uses a mathematical model to derive the suggestions. New features are that quantitative physicochemical characterizations of the drugs as well as the components of the LC system are used to the extent they are available. We illustrate the method by an example from our involvement in the development of nucleoside analogues.

*Presented at the Fifth International Symposium on Pharmaceutical and Biomedical Analysis, Stockholm, Sweden, September 1994.

†Author to whom correspondence should be addressed.

Mathematical Methods

Design

It will be assumed that considerations other than the design will govern the choice of compounds reaching the stage where assays need to be developed. While experimental design methods will have been used to optimize the LC method for each compound (that is, there will have been a systematic exploration of the influence of factors such as mobile phase and LC column) it cannot be assumed that factors other than the biological activity have been used to select a drug for further development. The physicochemical properties of the compounds in a series will therefore vary in a non systematic manner and collinearity may be a problem if multiple regression like data analysis methods are employed. The example used to illustrate the method is of this kind.

Multivariate analysis

The method chosen here for the analysis of the data is the partial least squares analysis (PLS). It has been described in considerable detail elsewhere [3, 4] and computer programs are commercially available. The algorithm is given below without further comments. Standard matrix algebra notation is used (see e.g. [5]) and all vectors are column vectors. The matrix containing the information about compounds, columns and mobile phases (described in detail below) is denoted **X** and the vector of regention times (or rather log K') is denoted **Y**. Weight vectors for **X** and **Y** are denoted **w** and **q** respectively, and the loading vectors for **X** are denoted **p**. Score vectors for **X** and **Y** are denoted **t** and **u**, respectively. The regression coefficients between **t** and **u** are denoted b. After normalization of **X** and **Y** to zero mean and unit variance (or other appropriate scaling) the PLS algorithm proceeds as follows

0. guess **u** as e.g. first column of **Y** (which here has only one column)
1. $\mathbf{w} = \mathbf{X}'\mathbf{u}/\mathbf{u}'\mathbf{u}$
2. $\|\mathbf{w}\| = 1$
3. $\mathbf{t} = \mathbf{X}\mathbf{w}/\mathbf{w}'\mathbf{w}$
4. $\mathbf{q} = \mathbf{Y}'\mathbf{t}/\mathbf{t}'\mathbf{t}$
5. $\|\mathbf{q}\| = 1$
6. $\mathbf{u} = \mathbf{Y}\mathbf{q}/\mathbf{q}'\mathbf{q}$
7. repeat 1–6 until convergence
8. $\mathbf{p} = \mathbf{X}'\mathbf{t}/\mathbf{t}'\mathbf{t}$
9. $b = \mathbf{u}'\mathbf{t}/\mathbf{t}'\mathbf{t}$.

We can predict **Y** from the model as

10. $\mathbf{Y}_{\text{predicted}} = b\mathbf{t}\mathbf{q}'$
11. $\mathbf{E} = \mathbf{X} - \mathbf{t}\mathbf{p}'$
12. repeat from 0 replacing **X** with **E** to get next dimension.

Maximally as many dimensions can be extracted as there are linearly independent columns in **X** or as the number of rows − 1, whichever is the smallest of the two. To determine the significant number of dimensions we use cross-validation as described in detail elsewhere [3, 4, 6]. The principle of cross-validation is to hold out part of the data, calculate the model for the remaining part and see if the model produces a smaller prediction error than a simpler model does. The process is repeated for the whole data set and the number of dimensions giving a significant reduction of the prediction error is chosen [6] with a significance level of 5%. To predict **Y**-values from new data these are scaled data in the same way as the model data and the **t**-score is calculated by means of the weight vector **w** and the b and **q**-values are used to predict **Y**-values which finally are rescaled to the original variables. To get a second dimension, residuals have to be calculated in the same way as in the algorithm by subtracting $\mathbf{e}_{\text{new}} = x_{\text{new}} - \mathbf{t}_{\text{new}}\mathbf{p}'$ and the process is repeated. A most important feature of the present methodology is that the number of significant PLS-dimensions usually is substantially smaller than the number of analytes, column and mobile phase characteristics.

Illustrative data

From the compounds belonging to the class of nucleoside analogues and their metabolites which we have studied in different pharmacokinetic experiments in our laboratory we have selected a subset to calculate a PLS-model. Nine of the remaining compounds were used to test the model. The training set consisted of the following substances for which the numbers given in parentheses corresponds to the numbers in Figs 1 and 2: alovudine-5′-glucuronide (19–21), AMT (2′3′-dideoxy-3′-aminothymidine) (26–28), FLG (2′3′-dideoxy-3′-fluoroguanosine) (1–5), H2G ((−)2-hydroxymethoxyhydroxybutyl-guanine) (22–25), FCdU (5-chloro-2′3′-dideoxy-3′-fluorouridine) (16–18), zalcitabine (6–10) and zidovudine (11–15). The test, or validation, set consisted

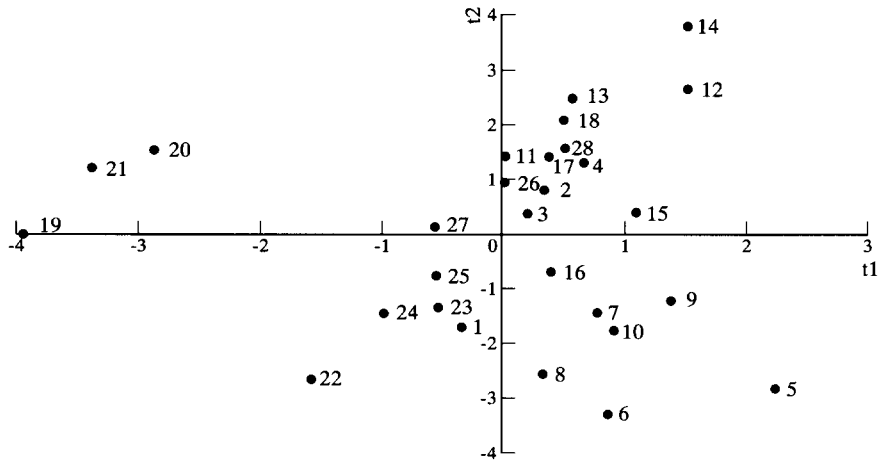


Figure 1
Plot of the t-scores of the first versus the second PLS component. The points are enumerated as given in the methods section.

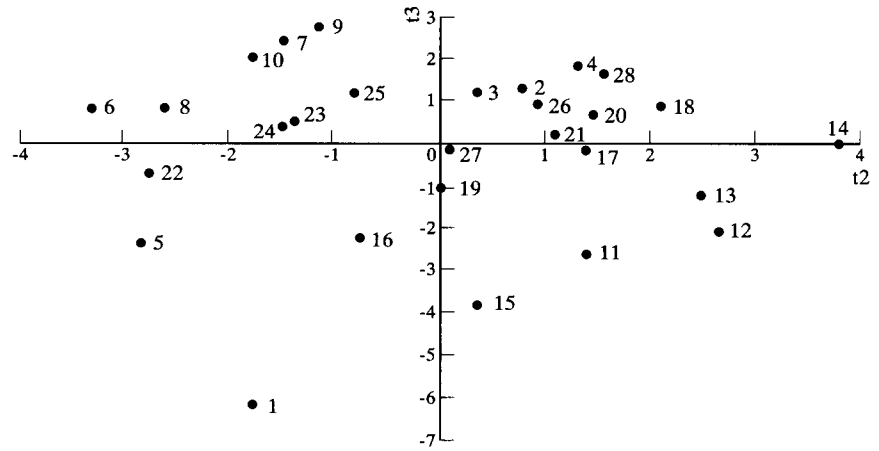


Figure 2
Plot of the t-scores of the second versus the third PLS component. The points are enumerated as given in the methods section.

of ACV (aciclovir), FLT (alovudine), BEA005 (2′3′-dideoxy-3′-methoxycytidine), dC (2′-deoxycytidine), dT (2′-deoxythymidine), FMMdU (5-methoxymethyl-2′3′-dideoxy-3′-fluorouridine), FEdU (5-ethynyl-2′3′-dideoxy-3′-fluorouridine), PCV (penciclovir) and AZT-G (zidovudine-5′-glucuronide). The compounds were physicochemically characterized by employing a substituent scale constructed for aromatic compounds by means of principal component analysis [7]. Three scales are used for each substituent. The 3′- and the 5-substituent of cytidine and uridine were quantitated in this manner. To distinguish between the bases cytosine, guanosine and uracil (which includes thymine) 0–1 indicator variables were used. The rational of this has been given by other authors [8]. Glucuronides and acyclic sugar were treated also by 0–1 indicator variables. Thus, the 14 variables (numbers 1–14 in Figs 3 and 4) used to describe the compounds were: glucuronide, molecular weight, partition coefficient octanol/water, pK_a, guanosine-base, cytidine-base, uridine-base, cyclic sugar, 3′-scale 1, 3′-scale 2, 3′-scale 3, 5-scale 1, 5-scale 2 and 5-scale 3. The 100 mm C18-columns were described by the 2 variables (numbers 15 and 16 in Figs 3 and 4): particle size (3 and 5 μm) and dead-volume time (0.68–2.15 min). The mobile phase was described by the 5 variables (numbers 17–21 in Figs 3 and 4): concentration of octane-sulph-

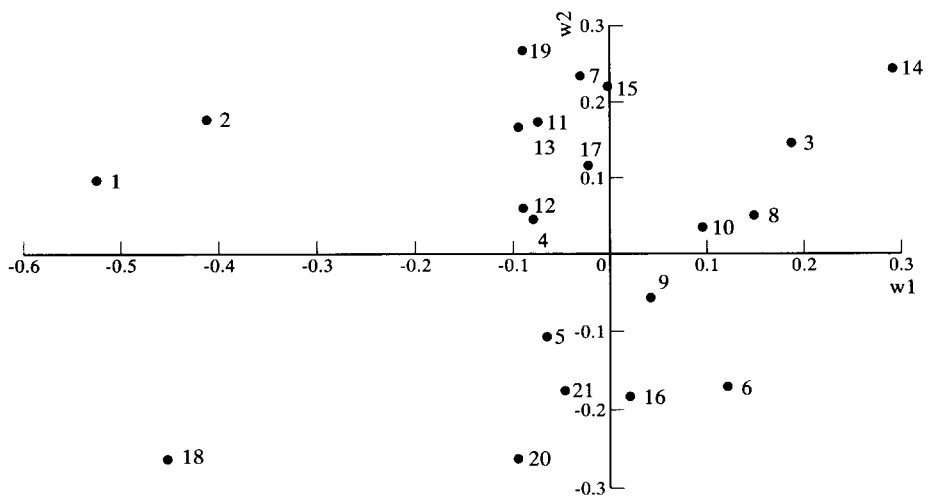


Figure 3
Plot of the w-scores of the first versus the second PLS component. The numbers in the figure correspond to the order of the variables in **X** as given in the methods section.

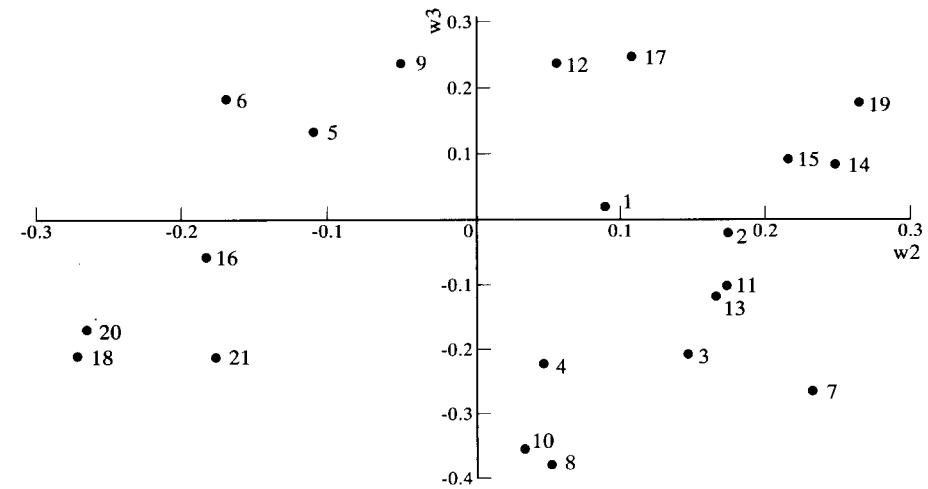


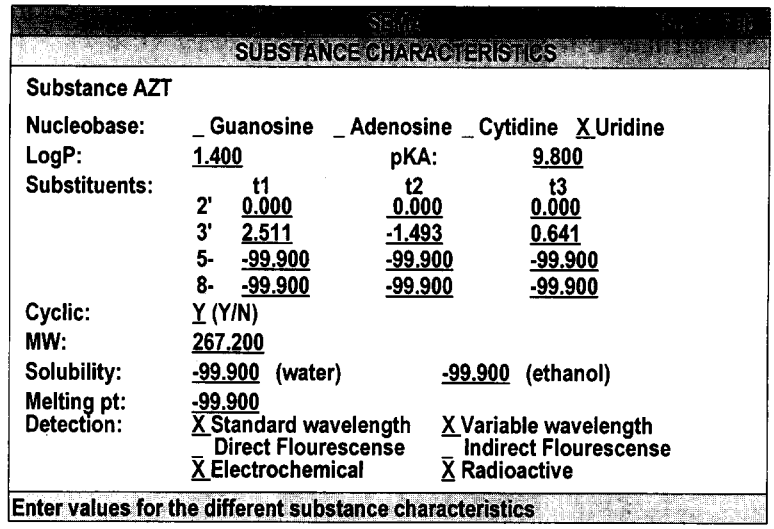
Figure 4
Plot of the w-scores of the second versus the third PLS component. The numbers in the figure correspond to the order of the variables in **X** as given in the methods section.

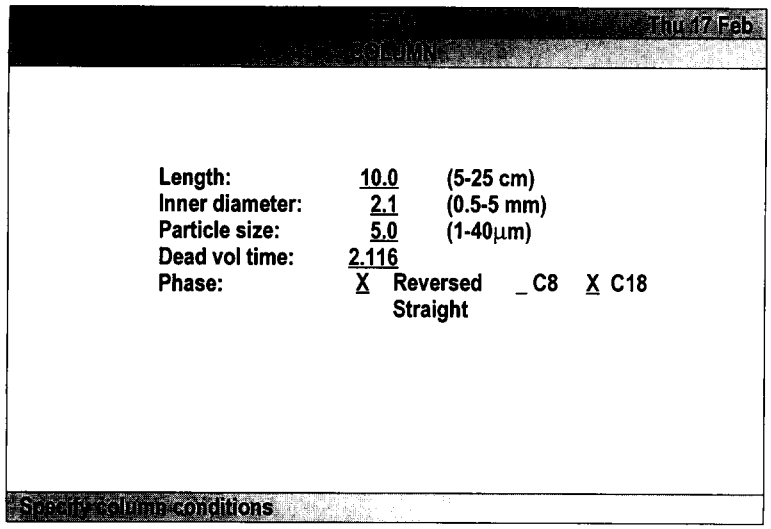
onic acid (0–5 mM), % organic solvent, solvent scale 1, solvent scale 2 and pH (2.5–6.0). The solvent scales were obtained from [9] and in this study we used methanol (4–20%) and 2-propanol (1.5%). A 0.05 M phosphate buffer was used throughout the experiments.

The data matrix **X** consisted of 21 columns, the descriptors enumerated in the preceding paragraph. Each row is a unique combination of a compound, a column and a mobile phase which result in a retention time (or rather log K') which make up **Y**. The results are presented as pair-wise plots of **t**-vectors and pair-wise plots of **w**-vectors. A plot of observed versus predicted retention times of the test, or validation, set is given together with a 95%-confidence interval for the predicted retention time relative to the observed retention time.

Computer program

Programs were written in Pascal (Turbo Pascal™) on IBM PC-compatible computers. The PLS algorithm has been validated both against commercially available programs and by internal checkups of orthogonality conditions, e.g. between **t** vectors and between **w** vectors. The final program allows an input of compounds, columns and mobile phases in





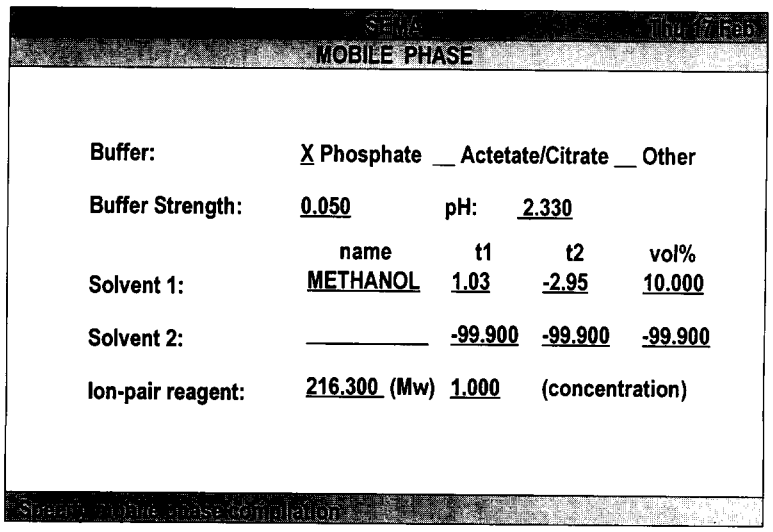


Figure 5
Sample screens from the computer program developed to handle the data obtained from nucleoside analogue chromatography.

three separate parts which can be uniquely combined later on and linked to a retention time. The program also provides suggestions to detection methods based on previous knowledge and it therefore has typical expert system characteristics. However, it does differ from conventional programs used as expert systems in the sense that it uses a mathematical model that provides a quantitative output and an interaction with the user in the prediction phase. A full description can be obtained upon request from one of the authors (AM). Selected screen outputs from the program are provided in Fig. 5.

Results

The score and weight plots calculated from the illustrative data are shown in Figs 1–4. From these plots the influence of different variables on the retention behaviour can be studied. We wish to point out that the influence of the variables must be interpreted together and that no assumption of independence is implicit in the model. However, from the score plot of the second versus the third dimension it can be seen that the "northwest–southeast" direction in the plot is due to differences among substances while the orthogonal direction is due to column and mobile phase.

In Fig. 6, the observed retention times of the validation set are plotted against the predicted retention times. The mean prediction of a retention time was 115% of the observed and the 95% confidence interval for the individual predictions was 33–197%. Thus, using the model presented here, there is a 95% chance that the predicted value lies between 33 and 197% of the retention time one would get if the same test had been carried out in the laboratory. The range of deviations observed in the present study was 75–165% and the distribution had a tendency to a skewness. The two compounds for which the prediction errors were largest are the 3′-hydroxylated nucleosides, dT and dC, and no such compound was used in the training set.

Discussion

The approach taken here can be used to facilitate the start of an assay development such that the region in which reasonable retention times are expected to be found can be identified. The example given in the results section shows that the method actually works in practice. Several factors will influence how well the method is going to work on a given set of compounds. Firstly, interpolation can always be expected to produce smaller prediction errors than extrapolation. Since the choice of the LC column and the mobile phase composition are under experimental control these factors do not constitute a difficult problem. In contrast, the compounds to be analysed have been chosen according to their pharmacological and possibly other properties but they have not been selected by considering their chromatographic properties. Therefore,

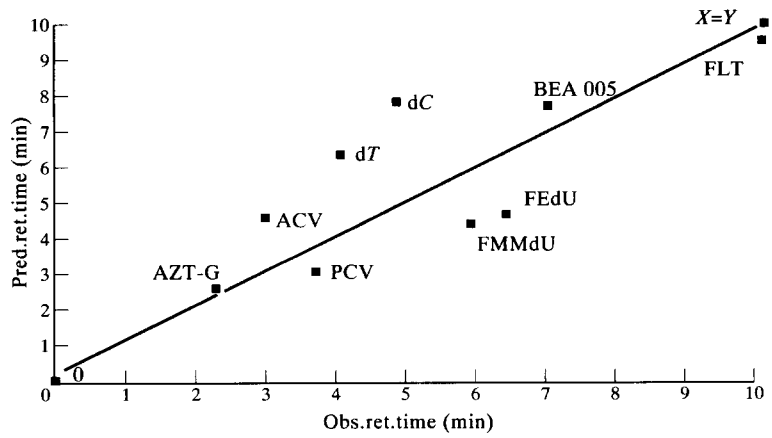


Figure 6
Plot of observed versus predicted retention times for nine compounds (abbreviations are given in the methods section) for which the mobile phase was chosen at random from those investigated. The model is based on seven other compounds (see the Methods section) with a total of 28 different chromatograms. Three to five chromatograms were selected for each compound taking experimental design considerations into account. The oblique line is $x = y$.

the compounds will appear to the analytical chemist as if they were, in principle, randomly selected and, therefore, interpolation will be an exception rather than the rule. Secondly, the fewer the compounds that have been investigated, the poorer the predictions are. At a very early phase it is probably better to guess than to use the design suggested here because the information available to the PLS prediction can be expected to be much smaller than the information available to the chemist. A possible improvement of our program is, therefore, to try to use information obtained from other series of compounds in order to use general information on retention behaviour in an LC system, e.g. the behaviour when pH varies around pK_a and the effect of lipophilicity of the compound. Thirdly, physicochemical characterizations of chromatographic columns, that can be used in the same way as the organic solvent descriptors, are not available and the reproducibility of some column materials is at present too poor to allow reasonably accurate predictions. An example that has been encountered in practice is that the position of analytes on cyanocolumns may switch when one column is replaced by another. Taking into consideration these difficulties we find it rather satisfactory that the predicted retention times found here result in a confidence region that is 33–197% of the true retention time. This provides us with the starting point we need to quickly develop an assay.

In the present paper some special features of PLS have been exploited in the coding of physicochemical properties that are of interest from a general point of view. In particular, the fact that the handling of missing values is part and parcel of the PLS algorithm has been used to handle different chemical skeletons, here represented by the different nucleoside bases. Mathematically, this is done by "assuming" that the missing values are exactly in accordance with the model i.e. the model is not influenced by the missing values. For substituents on molecules that are different, like the bases guanine and uracil, it is not possible to define the 5-substituent on both molecules such that they have the same physicochemical meaning. It is, however, possible to assign missing values to the structure lacking the substituent to be quantitated. In the present case, uridine and cytidine analogues have 5-substituents chemically corresponding to one another while guanosine analogues have no such substituent. We have therefore assigned missing values to the guanosine analogues in the three columns used to describe the 5-substituents. The approach worked in the present study but it is not yet known how extensively this method can be used and how disparate molecules that can be handled in the same analysis. Further research is needed to answer these questions.

Another possibility to improve upon the mathematical method used here is to include non-linear and interaction terms in the PLS-model. Such changes can easily be made and will also be included in future versions of the program. Another point worth mentioning is the common situation where some new approach has to be developed in order to improve upon the chromatography or to describe the compounds studied. Such changes can be incorporated into our method in a natural way provided they are carefully coded such that previously obtained data can be used in spite of the addition.

Finally, we must discuss the most desirable situation, i.e. how to choose a column and mobile phase in order to get a prespecified retention time. This can, of course, be achieved ultimately, but it requires that so much data has been collected in a systematic fashion that the whole property space is well spanned. This is, however, not realistic in a laboratory which spends most of its time and interest in analysing biological samples rather than developing new methods. In the more realistic situation the data will not span the property space and, therefore, it is not possible to get *one* unique combination of mobile phase and column to get a selected retention time for a specific compound. Instead, many combinations can give the same retention time and, at most, an equation for a subspace in the property space can be given which produces the desired retention time. We have not tried this approach yet, preferring the simulation-like situation presented in the present paper. A main reason for this is the simplicity of the approach, which we assume is attractive to many chemists and pharmacokineticists.

Acknowledgement — The present study was supported by grants from the Swedish Medical Research Council (grant 09069) and Medivir AB. The technical assistance of laboratory technician Eva Kristoffersson is gratefully acknowledged.

References

[1] G.E.P. Box, G.W. Hunter and J.S. Hunter, *Statistics for Experimenters*. Wiley & Sons, New York (1978).
[2] L. Ståhle, Å. Hallström, N. Borg and A. Carlsson, Manuscript (1994).
[3] S. Wold, A. Ruhe, H. Wold and W.J. Dunn, III, *SIAM J. Sci. Statist. Comput.* **5**, 735–743 (1984).
[4] L. Ståhle and S. Wold, *Progress in Medicinal Chemistry* Vol. 25, pp. 292–334, Elsevier, Amsterdam (1988).
[5] H. Anton, *Elementary Linear Algebra*, Wiley & Sons, New York (1987).
[6] L. Ståhle and S. Wold, *J. Chemometrics* **1**, 185–196 (1987).
[7] B. Skagerberg, D. Bonelli, S. Clementi, G. Cruciani and C. Ebert, *Quant. Struct. Act. Relat.* **8**, 32–38 (1989).
[8] J. Jonsson, Quantitative Sequence-Activity Modelling. Thesis, Univ. Umeå (1992).
[9] R. Carlsson, T. Lundstedt and C. Albano, *Acta Chem. Scand. B* **39**, 79–91 (1985).

[Received for review 22 September 1994; revised manuscript received 9 December 1994]