Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Chromatography A, 1055 (2004) 11–19
JOURNAL OF CHROMATOGRAPHY A
www.elsevier.com/locate/chroma

Multivariate adaptive regression splines (MARS) in chromatographic quantitative structure–retention relationship studies

R. Put[a], Q.S. Xu[a,b], D.L. Massart[a], Y. Vander Heyden[a,*]

[a] *ChemoAC, Department of Pharmaceutical and Biomedical Analysis, Pharmaceutical Institute, Vrije Universiteit Brussel (VUB), Laarbeeklaan 103, B-1090 Brussels, Belgium*
[b] *College of Mathematics and Econometrics, Hunan University, Changsha 410082, China*

Received 23 March 2004; received in revised form 1 July 2004; accepted 6 July 2004

Abstract

The multivariate adaptive regression splines (MARS) methodology was applied to build quantitative structure–retention relationships (QSRRs). The response (dependent variable) in the MARS models consisted of the logarithms of the extrapolated retention factors ($\log k_w$) of 83 structurally diverse drugs on a Unisphere PBD column, using isocratic elutions at pH 11.7. A set of 266 molecular descriptors was used as predictor (independent) variables in the MARS model building. The optimal MARS model uses 34 basis functions to describe the retention and has acceptable predictive properties for new objects. The molecular descriptors included in the model describe hydrophobicity, molecular size, complexity, shape and polarisability. Some additional MARS models were created using alternative strategies. These include models with $\log P$ as the single predictor and models obtained with only the three most important molecular descriptors. The use of classification and regression trees (CART) as feature selection technique for predictor variables used in the MARS model was also investigated. Further, it is also studied whether allowing quadratic terms instead of interaction terms might lead to better MARS models.

Keywords: Multivariate adaptive regression splines; Quantitative structure–retention relationships; Classification and regression trees; Molecular descriptors

1. Introduction

Reversed-phase high performance liquid chromatography (RPLC) is one of the most frequently used techniques to separate pharmaceutical mixtures. The enormous variety of stationary phases combined with a range of possible mobile phase compositions makes RPLC suited for almost any separation. However, the selection of an appropriate starting point, i.e. the initially selected chromatographic system, for further method development has become a crucial time-consuming step in RPLC method development, since in general a trial-and-error approach is followed [1].

The building of retention prediction models may open the possibility to predict the retention and consequently the separation of mixtures in a given chromatographic system. Several approaches for retention prediction in RPLC have been investigated, among which quantitative structure–retention relationships (QSRRs) are the most popular [2]. In QSRR, a model is built, in which the retention on a given chromatographic system is described as a function of solute (molecular) descriptors. The QSRR models described in the literature usually apply multiple linear regression (MLR) methods, often combined with genetic algorithms for feature selection [3-5]. Other frequently used approaches include artificial neural networks [5,6] and partial least squares (PLS) [7].

Multivariate adaptive regression splines (MARS) is a multivariate non-parametric regression procedure, which was proposed by Friedman [8]. The use of the MARS method in chemical studies was introduced by De Veaux et al. [9]. Later, MARS was successfully applied in a quantitative structure–activity relationship (QSAR) context by Nguyen-

* Corresponding author. Tel.: +32 2 477 47 34; fax: +32 2 477 47 35.
E-mail address: yvanvdh@vub.ac.be (Y. Vander Heyden).

0021-9673/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2004.07.112

Cong et al. [10] and Lahsen et al. [11]. Since in QSAR and QSRR, similar models are built to describe either biological activity or chromatographic retention, MARS might also be useful to construct QSRR models. The use of MARS to build QSRR models may show some advantages compared to the more traditionally used techniques like MLR and neural nets. Probably, the largest advantage is the fact that MARS is able to describe a given response (e.g. chromatographic retention) starting from a large number of predictors (e.g. molecular descriptors) from which the best are automatically selected. Compared to neural nets and PLS, the MARS models are easier to interpret, since the original variables can be directly found in the resulting model and even interactions between the variables are indicated. Thus, MARS is able to build flexible models without the disadvantages of the more 'black-box' methods, as PLS and neural networks are sometimes called.

The aim of this study was to evaluate the use of the MARS methodology in a QSRR context. MARS is applied to describe the chromatographic retention, on a given chromatographic system, as a function of a set of molecular descriptors. Additionally, the use of classification and regression trees (CART) [12] for feature selection of the predictors (i.e. molecular descriptors), prior to the application of MARS is studied. It was also researched whether the introduction of quadratic terms instead of interactions may lead to better models. A physicochemical explanation of the descriptors selected by MARS is also given and the resulting MARS models are compared with the CART model obtained in a previous study on the same dataset [13].

2. Theory

2.1. Multivariate adaptive regression splines

The MARS method is a local regression method that uses a series of local so-called basis functions to model complex (non-linear) relationships [8]. The space of the predictors is split into several (overlapping) regions in which so-called spline functions are fit. The global MARS model then consists of the weighted sum of the local models:

$$\hat{y} = a_0 + \sum_{m=1}^{M} a_m B_m(\boldsymbol{x}) \tag{1}$$

where $\hat{y}$ is the predicted response, a_0 the coefficient of the constant basis function, $B_m(\boldsymbol{x})$ the mth basis function, which may be a single spline function or a product (interaction) of two (or more) spline functions, a_m the coefficient of the mth basis function and M the number of basis functions included into the model.

In general, the MARS methodology consists of three steps: first a constructive phase, in which basis functions are introduced in several regions of the predictors and are combined in a weighted sum to define the global model (Eq. (1)). This model often contains too many basis functions, which leads to overfitting. Therefore, the constructive phase is followed by a pruning phase, in which some basis functions of the overfitting model are deleted. This leads to a sequence of consecutively smaller MARS models, from which the optimal one is selected in a third step.

In the first step, a model is created by stepwise adding basis functions. Each basis function covers a given domain of the response variable. The basis functions in MARS consist either of one single spline function or of the product of two (or more) spline functions for different predictors. The spline functions in MARS are piecewise polynomials, i.e. left-sided (Eq. (2)) and right-sided (Eq. (3)) truncated functions.

$$b_q^-(x-t) = [-(x-t)]_+^q = \begin{cases} (t-x)^q, & \text{if } x < t \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

$$b_q^+(x-t) = [+(x-t)]_+^q = \begin{cases} (x-t)^q, & \text{if } x > t \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where t is called the knot location; $b_q^-(x-t)$ and $b_q^+(x-t)$ are spline functions describing the regions to the left and the right of the given t; q indicates the power (>0) to which the spline is raised; the subscript "+" indicates a value of zero for negative values of the argument. An example of the graphical representation of a spline function can be found in Section 4.1. The first step consists of a "two-at-a-time" forward stepwise procedure which selects the best pairs of spline functions in order to improve the model. Each pair contains one left-sided and one right-sided truncated function defined by a given knot location. Thus, initially the best basis functions are added two by two in order to improve the description of the training data. The algorithm used evaluates all possible predictors, as well as all possible knot locations for each predictor. The search for the best predictor and knot location is performed in an iterative way. The predictor (knot location), which contributes most to the model, is selected first. Additionally, the algorithm checks at the end of each iteration whether the introduction of an interaction improves the model. While in successive iterations the basis functions are introduced in the model as single additive components, the interactions are expressed in the model as the product of two or more basis functions. The order of a MARS model indicates the maximum number of basis functions that interact (e.g. in second-order MARS the interaction order of the splines is not higher than two). The iterative building procedure continues until a user-defined maximum number of basis functions ($M_{\max}$) is included. The value of $M_{\max}$ should be considerably larger than the optimal model size M^* (typically the order of magnitude of $M_{\max}$ is twice the expected magnitude of M^*) [8]. This results in a model, which overfits the response.

The second step in the MARS methodology consists of a pruning step. A "one-at-a-time" backward elimination pro-

cedure is applied in which the basis functions with the lowest contribution to the model are excluded. Thus, in this step, the basis functions to be maintained in the final MARS model are selected from the set of all candidate basis functions, used in step 1. This pruning usually is based on the generalized cross-validation (GCV) criterion, but other approaches exist, such as *n*-fold cross validation [8]. The GCV parameter is an adjusted residual sum of squares, in which a penalty for the model complexity is incorporated. This criterion is used to avoid an excessive number of spline functions in the final model:

$$\mathrm{GCV}(M) = \frac{1}{n}\frac{\sum_{m=1}^{n}(y_i - \hat{y}_i)^2}{(1 - C(M)/n)^2} \quad (4)$$

where n is the number of objects studied, y_i the (experimental) response for object i, $\hat{y}_i$ the predicted response for object i and $C(M)$ a complexity penalty function, which is defined as:

$$C(M) = M + dM \quad (5)$$

with M the number of non-constant basis functions (i.e. all terms of Eq. (1) except a_0) in the MARS model and d a user-defined cost for each basis function optimisation. The higher the cost d gets, the more basis functions will be excluded. In practice, d is increased during the pruning step in order to obtain smaller models. Besides its use during the pruning step, the increase in the GCV value caused by removing a variable from the model is also used to evaluate the importance of the predictor variables in the final MARS model.

Eventually, the selection of the optimal model is performed in a third step. The selection is based on an evaluation of the predictive properties of the different models, which often are determined using cross-validation or a new independent test set.

Further details on MARS modelling are given in [8,14].

2.2. Molecular descriptors

A molecular descriptor can be defined as the result of a logical and/or mathematical procedure, which transforms chemical information encoded within a symbolic representation of a molecule into a useful number (theoretical descriptor), or as the result of some standardized experiment (experimental descriptor) [15]. The term "useful" means that the resulting number might contribute to an understanding of molecular properties and/or can be used in a model to predict properties of molecules. In the literature over 6000 descriptors are described, and the number still grows [15]. A further classification of theoretical molecular descriptors is based on the dimensionality of the molecular representation from which the descriptor is calculated [15]. One can distinguish zero-dimensional (0D), one-dimensional (1D), two-dimensional (2D) and three-dimensional (3D) molecular descriptors, which are derived from the chemical formula, a substructure list representation, a topological and geometrical representation, respectively.

More information about molecular descriptors can be found in [15].

3. Experimental

The chromatographic data used were obtained from Nasal et al. [16] and consisted of the logarithms of the extrapolated retention factors ($\log k_w$) for 83 basic drugs. The data concern the retention in buffer/methanol mixtures on a Unisphere PBD, a polybutadiene-coated alumina, column at pH 11.7 using isocratic elution. The composition of the methanol/aqueous buffer mobile phase ranged from 75:25 to 0:100 (v/v). To compare retentions, they were extrapolated to 0% organic modifier [16] and the $\log k_w$ values thus obtained are used.

In this study 266 0D, 1D and 2D theoretical descriptors were used. The $\log P$ values of the substances were obtained using the on-line interactive LogKow program of the Environmental Science Center of Syracuse Research Corporation, Syracuse, NY [17,18]. For all molecules, the geometrical structure was optimised using Hyperchem 6.03 Professional software (Hypercube, Gainesville, FL). Geometry optimisation was obtained by the molecular mechanics force field method (MM+) using the Polak-Ribière conjugate gradient algorithm with a RMS gradient of 0.05 kcal/(Å mol) as stop criterion. The Cartesian co-ordinates matrices of the atom positions in the molecule, resulting from this geometrical representation, were used to calculate the molecular descriptors by the Dragon 1.1 software [19]. The following groups of descriptors were calculated: 56 constitutional descriptors [15], 69 topological descriptors [20–24], 20 molecular walk counts [25], 21 Galvez topological charge indices [26], 96 2D autocorrelations [27–29] and 3 empirical descriptors [15]. A total of 266 descriptors were used.

The MARS models were built using an in-house algorithm, based on the original MARS method from Friedman [8], written in Matlab 5.3.1 environment (The Mathworks, Natick, MA). The extrapolated retention data were used as response to be modelled and the descriptors as predictors.

4. Results and discussion

4.1. The MARS model

A total of 266 molecular descriptors were used as predictor variables to build MARS models for the prediction of the extrapolated retention factors ($\log k_w$). Since second-order MARS was applied, the basis functions of the models consist of linear and second-order splines. M_{max}, the criterion to stop the maximum model building was set to 100. During the pruning step to obtain sequentially smaller models the generalised cross-validation criterion was alternated with 20-fold cross-validation. From the set of models with different complexity,

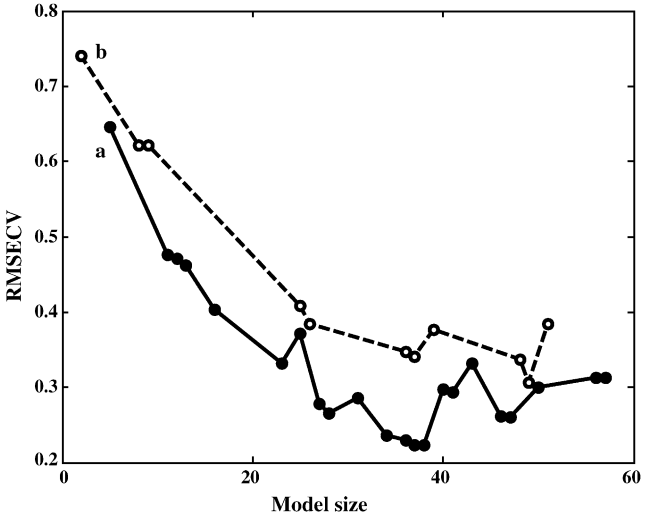

Fig. 1. RMSECV vs. model complexity: (a) for all predictors; (b) after feature selection by CART.

the optimal one was selected based on its predictive property estimated using leave-one-out cross-validation. Fig. 1a shows a plot of the root-mean-squared error of cross-validation (RMSECV) versus the complexity for the pruned models. The RMSECV shows a minimum (0.2231) for the model defined by 37 basis functions. However, a smaller model was selected as optimal, since it is less complex and shows a comparable RMSECV (0.2358). This optimal MARS model contains 34 basis functions (B_i), including the constant B_1, which equals 1 (Eq. (6)):

$$\log k_w = \sum_{i=1}^{34} a_i B_i \quad (6)$$

The basis functions represented by B_1, B_2, ..., B_{34} as well as their coefficients a_i are given in Table 1. Several interaction terms are integrated in the model. The model contains a constant B_1 (=1) and 10 basis functions that are single splines (B_2, B_3, B_4, B_5, B_{17}, B_{18}, B_{26}, B_{29}, B_{30} and B_{32}), defined by only one molecular descriptor. The other basis functions are second-order interactions of two molecular properties.

As an example of a basis function in the model, consider B_2:

$$(\log P + 4)_+ = \begin{cases} \log P + 4, & \text{if } \log P > -4 \\ 0, & \text{otherwise} \end{cases}$$

This means that, when $\log P > -4$, the second term of Eq. (6) is 0.8344 ($\log P$ + 4), otherwise it is 0 (Fig. 2). Since in this example the $\log P$ values exceed the knot value (=−4) for all molecules, the basis function is different from zero for all objects.

Table 1
List of basis functions B_i of the MARS model and their coefficients, a_i

B_i	Definition	a_i
B_1	1	−0.114
B_2	$(\log P + 4)_+$	0.834
B_3	$(\log P - 4)_+$	0.234
B_4	$(1 - \text{nR06})_+$	1.964
B_5	$(\text{nR06} - 1)_+$	0.837
B_6	$(\log P - 4)_+$ (GATS4p − 2.2340)$_+$	0.293
B_7	$(\text{nR06} - 1)_+$(0.1430 − MATS5m)$_+$	−1.210
B_8	$(\text{nR06} - 1)_+$(MATS5m − 0.1430)$_+$	−85.34
B_9	$(\text{nR06} - 1)_+$(0.3360 − ATS5v)$_+$	−119.7
B_{10}	$(1 - \text{nR06})_+$(0.2220 − MATS4e)$_+$	−2.557
B_{11}	$(1 - \text{nR06})_+$(MATS4e − 0.2220)$_+$	−48.36
B_{12}	$(\log P - 4)_+$(X1A − 0.4330)$_+$	4.016
B_{13}	$(\log P - 4)_+$(0.2890 − MATS8e)$_+$	0.568
B_{14}	$(\log P - 4)_+$(MATS8e − 0.2890)$_+$	3.803
B_{15}	$(\log P - 4)_+$(−0.0060 − MATS7p)$_+$	0.603
B_{16}	$(\log P - 4)_+$(MATS7p + 0.0060)$_+$	1.571
B_{17}	(2.6770 − GATS1p)$_+$	−1.558
B_{18}	(GATS1p − 2.6770)$_+$	−0.734
B_{19}	$(\text{nR06} - 1)_+$(ATS4m − 0.2540)$_+$	1.711
B_{20}	$(\log P - 4)_+$(0.0320 − JGI5)$_+$	8.260
B_{21}	(GATS1p − 2.6770)$_+$(1159.8 − TPCM)$_+$	−0.028
B_{22}	(GATS1p − 2.6770)$_+$(TPCM − 1159.8)$_+$	−0.002
B_{23}	$(\text{nR06} - 1)_+$(JGI3 − 0.5710)$_+$	−1.406
B_{24}	(2.6770 − GATS1p)$_+$(0.4520 − ATS2m)$_+$	16.05
B_{25}	(2.6770 − GATS1p)$_+$(ATS2m − 0.4520)$_+$	6.313
B_{26}	(0.4170 − ATS3m)$_+$	−5.840
B_{27}	$(\text{nR06} - 1)_+$(1.3460 − GATS7p)$_+$	−9.633
B_{28}	(ATS3m − 0.4170)$_+$(PJI2 − 0.8570)$_+$	−17.64
B_{29}	(0.7630 − X0A)$_+$	−6.969
B_{30}	(X0A − 0.7630)$_+$	−820.9
B_{31}	$(\log P - 4)_+$(ATS5v − 0.3170)$_+$	1.676
B_{32}	(MATS8p + 0.2580)$_+$	−0.189
B_{33}	$(\text{nR06} - 1)_+$(MATS5e − 0.0540)$_+$	−1.391
B_{34}	(0.4170 − ATS3m)$_+$(ATS3e − 1.300)$_+$	−443.9

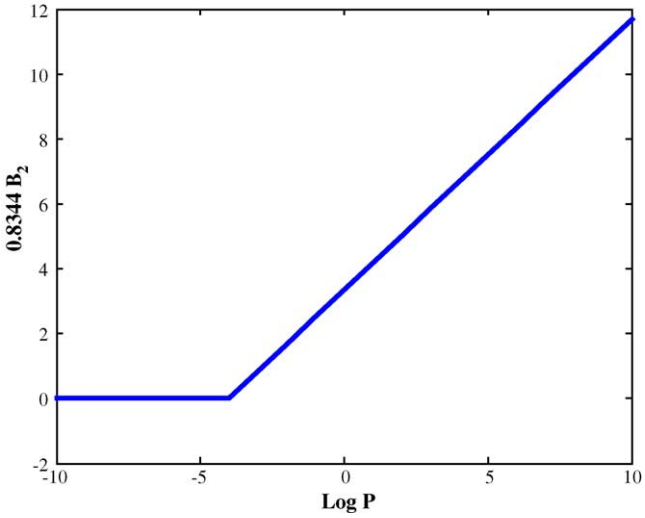

Fig. 2. Graphical representation of the second term of Eq. (6) (=0.8344 B_2) as a function of $\log P$.

As an example of an interaction between two molecular properties consider B_6:

$$(\log P - 4)_+(\mathrm{GATS4p} - 2.2340)_+ = \begin{cases} (\log P - 4)(\mathrm{GATS4p} - 2.2340), & \text{if } \log P > 4 \text{ and } \mathrm{GATS4p} > 2.2340 \\ 0, & \text{otherwise} \end{cases}$$

This means that, when $\log P > 4$ and GATS4p > 2.2340, the sixth term of Eq. (6) is 0.2933 $(\log P - 4)$ (GATS4p − 2.2340), otherwise it is 0.

Note that for a given molecule some terms of the model of Eq. (6) may equal 0, which means that those terms are not used to describe its retention.

4.2. Interpretation of the MARS model

Fig. 3 gives a graphical representation of the model structure, in which six main groups of basis functions can be distinguished. The groups indicated are defined based on the predictor variables (i.e. the molecular descriptors) used in the basis functions. Each group consists of one or two main basis functions and their interactions. The molecular descriptors used to differentiate between the six classes are the hydrophobicity parameter ($\log P$); the number of six-membered rings (nR06); the Geary autocorrelation coefficient of lag 1, weighted by the atomic polarisabilities (GATS1p): the Broto-Moreau autocorrelation coefficient of a topological structure of lag 3, weighted on the atomic masses (ATS3m); the average connectivity index chi-0 (X0A); and the Moran autocorrelation − lag 8/weighted by atomic polarisabilities (MATS8p) [15,19]. Since basis functions may equal 0 for given objects, one can define the size of a class based on the number of molecules it describes. Thus, a given class only contains those molecules for which one or more of its basis functions are different from zero. The class defined by ATS3m is the smallest and contains only 65 objects. The other classes contain either 77 (the class defined by MATS8p) or all, i.e. 83, objects (the classes defined by $\log P$, nR06, GATS1p and X0A).

Fig. 4 gives a plot of the importance of the molecular descriptors in the final MARS model, which is evaluated by the increase in the GCV value caused by removing the considered variables from the model. It can be observed that the hydrophobic properties ($\log P$) are the most important in the MARS model, followed by nR06. The contribution of MATS5m is much less important, i.e. about 20% of the importance of $\log P$. The remaining molecular descriptors used all have an importance less than 10% of $\log P$. MARS models created with only the most important molecular descriptors of Fig. 4 are evaluated in Section 4.4.

The hydrophobicity parameter ($\log P$) is selected as being the most important parameter as may be expected since hydrophobic interactions are the most important mechanisms to determine retention in reversed-phase liquid chromatography [1]. The fact that the MARS method selects $\log P$ out of more than 250 molecular descriptors, is an indication that the QSRR relationships obtained could be meaningful.

The number of six-membered rings (nR06) is a count descriptor that describes local chemical information in a way that can be related to both the volume of a molecule and its hydrophobicity. These both properties can be related to the molecule's chromatographic retention.

The descriptors GATS1p, ATS3m and MATS8p are autocorrelation descriptors. The GATS1p and the MATS8p both describe atomic polarisabilities, whereas the Broto-Moreau autocorrelation coefficient of a topological structure of lag 3, weighted on the atomic masses (ATS3m) describes the spatial distribution of the atomic masses in the molecule. However, the meaning of these descriptors from a practical point of view is not always that evident. These descriptors are more difficult to interpret directly in QSRR sense, but since they describe the atomic polarisabilities and the atomic masses, respectively, their selection could be understood. Polar interactions are important in RPLC and the atomic masses are related to the size and complexity of the molecules.

The selection of the average connectivity index chi-0 (X0A) can also be understood since it is a topological descriptor which is related to the molecular complexity, i.e. the molecular branching properties.

In the first class, eight interactions with $\log P$ are described, using six different molecular descriptors (GATS4p, X1A, MATS8e, MATS7p, JGI5 and ATS5v). Since GATS4p and MATS7p both describe atomic polarisabilities, only five different types of interactions with $\log P$ can be distinguished. The Geary autocorrelation descriptor GATS4p and the Moran autocorrelation descriptor MATS7p describe the spatial autocorrelation of the atomic polarisabilities. MATS8e describes the autocorrelation of the atomic electronegativities and the average Randic connectivity index (X1A) describes molecular branching and complexity. The Galvez topological charge index of order 5 (JGI5) is a descriptor proposed to evaluate the global charge transfer in a given molecule.

In the second class, the interactions of nR06 are with the autocorrelation descriptors (MATS4e, MATS5e, MATS5m, ATS5v, ATS4m and GATS7p) describing spatial autocorrelation of atomic electonegativities (MATS4e and MATS5e), atomic masses (MATS5m and ATS4m), atomic van der Waals volumes (ATS5v) and atomic polarisabilities (GATS7p), respectively. Only two molecular descriptors (i.e. ATS2m and TPCM) are selected to interact with GATS1p (class 3). The Broto-Moreau autocorrelation coefficient of a topological structure of lag 2, weighted by the atomic masses (ATS2m) describes the spatial distribution of the atomic mass. The total multiple path count (TPCM) is proposed to describe the complexity of a molecule. Also 2 interactions can be observed with ATS3m (i.e. PJI2 and ATS3e) (class 4). The 2D Petit-

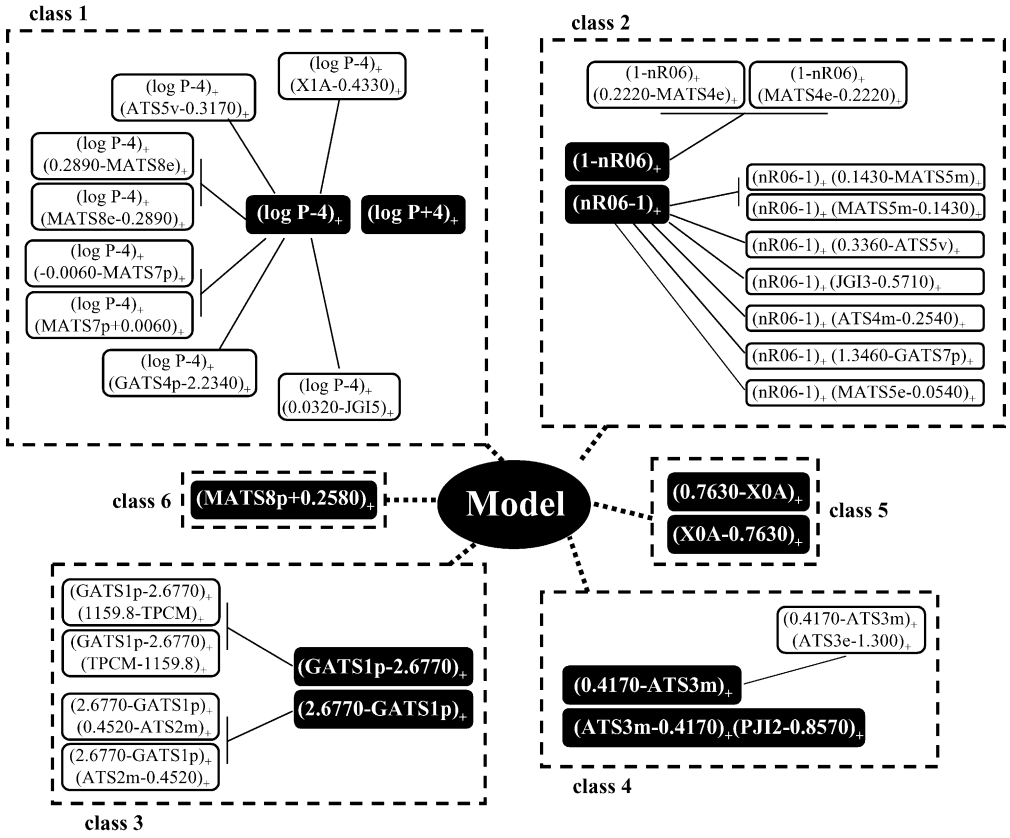

Fig. 3. Representation of the MARS model with the six classes of the basis functions.

jean shape index (PJI2), also called graph-theoretical shape coefficient, is proposed to describe the topological anisometry. This molecular shape descriptor describes the degree of deviation from a perfect cyclic topology. The Broto-Moreau autocorrelation coefficient of a topological structure of lag 3, weighted on the atomic electronegativities (ATS3e) describes the spatial distribution of the atomic electronegativities in the molecule.

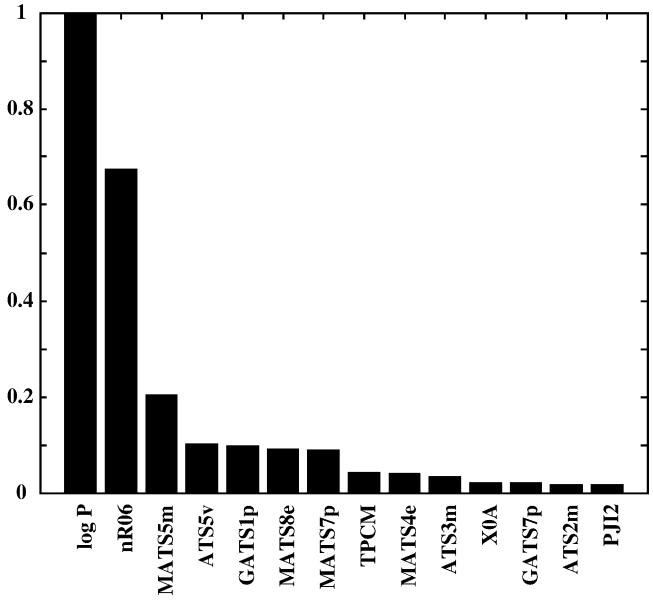

Fig. 4. Relative importance of the molecular descriptors selected in the MARS model.

Finally the two last classes, defined by the average connectivity index chi-0 (X0A) and by the Moran autocorrelation – lag 8, weighted by atomic polarisabilities (MATS8p) do not contain any interactions. MATS8p describes the spatial autocorrelation of the atomic polarisabilities and X0A describes the molecular branching of a molecule.

It can be concluded that the molecular properties selected to interact with the most important parameters of the model are not very easy to interpret, which often is intrinsic to their nature. However, they always are related to molecular properties that are important in RPLC such as molecular size, complexity and shape, polarisability and electronegativity.

4.3. Prediction properties of the MARS model compared to CART and MLR models

In order to evaluate the model, the retention of each molecule is predicted twice: once as part of the training set and once it is selected as a test object in leave-one-out cross-validation. The training set contains all 83 molecules and their retention is described using the model of Eq. (6). Using leave-one-out cross-validation each molecule is predicted

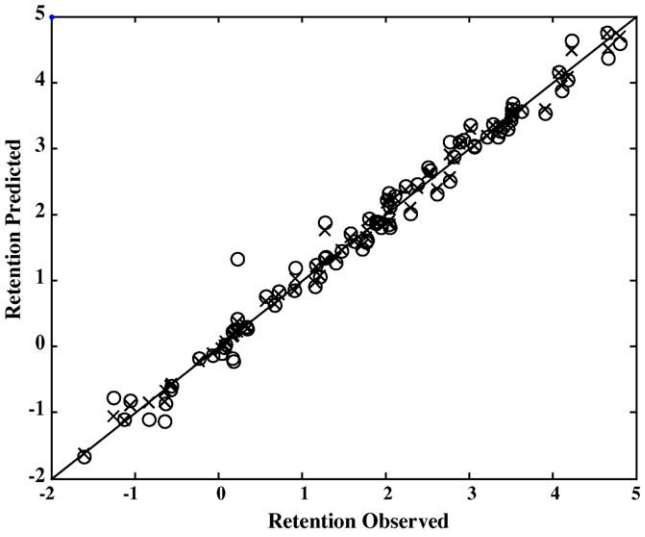

Fig. 5. Predicted versus observed retention for all objects, once using the training set (×) and once using leave-one-out cross validation (◯).

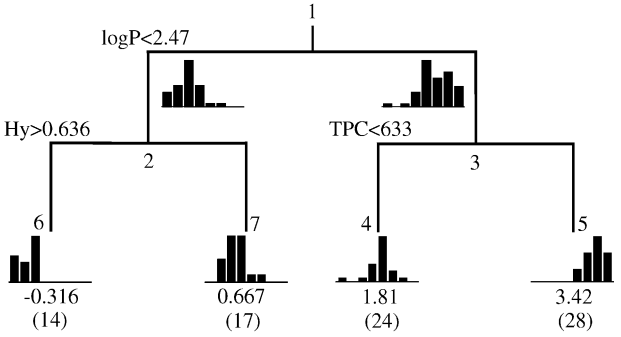

Fig. 6. The CART model obtained in [13] on the dataset used (Hy: hydrophilic factor; TPC: total path count).

once as a test object in a MARS model where the coefficients are determined based on the remaining 82 molecules.

The model of Eq. (6) describes the training set very well. A low value of the residual sum of squares (RSS = 0.1273) is obtained, the correlation coefficient is 0.9967 and R^2 equals 0.994. The predictive properties of the MARS model, evaluated using leave-one-out cross validation are also acceptable (RMSECV = 0.2358 and R^2 = 0.978). Fig. 5 shows a plot of the predicted versus the observed retention for both the training data and the test set. From a practical point of view these predictive values (for the test set) can be interpreted as a mean error equal to 13% of the real retention ($\log k'_w$).

In a previous study [13], classification and regression trees [12] were used to build regression trees for the retention prediction on the chromatographic system considered. The explanatory variables used were the 266 molecular descriptors also studied here. The final CART model was selected after cost-complexity pruning of the maximal tree using a 10-fold cross-validation [13]. Fig. 6 shows the selected optimal tree. The nodes of the tree are numbered according to the order of the tree growing. The splits, the average response values and the numbers of objects of the leaves are indicated. The histograms represent the distribution of the response for the objects within each leaf. Each bar covers a specific range of $\log k'_w$ values, with increasing retention towards the right part of the histograms. This allows to see clearly the partition in retention classes, i.e. low retention for nodes 6 and 7, intermediate for node 4 and long retention for node 5. Such retention classes are not obtained for the MARS model.

In the CART model, only three molecular descriptors were selected to describe the retention. The molecular descriptor selected first is the "hydrophobicity parameter ($\log P$)". The other descriptors are the "hydrophilic factor" (Hy) [30], representing hydrophilicity and the "total path count" (TPC) [31], related to the molecular size.

Besides these molecular descriptors, CART also provides lists of other important, so called, primary and surrogate variables [12,13]. The latter can be used when missing values occur in the splitting variables.

All molecular descriptors selected in the MARS model, with the exception of nR06, JGI5 and PJI2, were also selected by CART as either primary or surrogate variables. The $\log P$ is the main molecular descriptor both in MARS and CART. The autocorrelation descriptors, frequently selected by MARS were also found as surrogate and primary variables in the CART model.

The predictions obtained by the CART model were less good than with MARS since CART is only able to differentiate between classes with low, intermediate and high retention, whereas the MARS model is a real regression model. Moreover, the prediction power of the CART model could only be evaluated as a misclassification rate. However, when trying to link the selected theoretical descriptors with retention in a chromatographic system, it can be concluded that the MARS model is less easy to interpret than the CART trees.

In the literature, least squares univariate regression models that use $\log P$ as a predictor variable, to predict $\log k_w$ on polybutadiene-coated alumina columns, are described [32,33]. For the dataset studied here, the least squares univariate regression model obtained with only $\log P$ describes the training set less good (RSS = 0.8664 and R^2 = 0.6975) than the MARS model does. The predictive properties of this univariate regression model, evaluated using leave-one-out cross validation are also worse (RMSECV = 0.8900 and R^2 = 0.6879).

Moreover, stepwise regression [34] was used in order to include the descriptors with the highest correlation to $\log k_w$ in an MLR model. The MLR model obtained uses five molecular descriptors (log P, ATS3m, ATS7m, nOH: the number of hydroxyl functions, and SIC: structural information content, which is a structural complexity measure). Both the description of the training set (RSS = 0.5666 and R^2 = 0.8706) and

the leave-one-out cross validation predictive properties of the stepwise MLR model (RMSECV = 0.6145 and R^2 = 0.8519) have improved compared to the least squares univariate regression model with only log *P*, but the MARS model still performs better.

4.4. Additional MARS models

In accordance to the MLR models with only log *P*, we evaluated the performance of some simplified MARS models, using only log *P* or a small number of selected molecular descriptors, to describe retention. Additionally, it was researched whether it is useful to simplify the MARS models by banning interactions between predictor variables, but allowing quadratic terms of the basis functions. All models discussed below were created independently, which implies that always, first a maximum MARS model is created, which is then pruned back, and eventually the optimal model is selected. Analogously as in Section 4.3, the models were evaluated using the residual sum of squares for the training data and the leave-one-out cross validation to estimate the predictive power for new molecules.

The MARS model created using only log *P* uses only three terms ($\log k_w = 0.8705(\log P - 4)_+ + 0.3292(1.694 - \log P)_+ - 1.014$) but performs worse than the model of Eq. (6) (RSS = 0.8266 and R^2 = 0.725; RMSECV = 0.8538 and R^2 = 0.712). A model with six terms of log *P*, from which three are quadratic, describes the training data slightly better (RSS = 0.7847 and R^2 = 0.752), but its predictive performance is worse (RMSECV = 0.9476 and R^2 = 0.675).

MARS models with only the three most important molecular descriptors of Eq. (6) as predictor variables were also created. The other molecular descriptors, which all show an importance below 10%, relative to log *P* (Fig. 4), are not selected. The optimal model thus obtained uses eight terms and performs better than the models build with log *P* as the only predictor variable. RSS equals 0,5727 (R^2 = 0.868) and also the predictive power of the model is better (RMSECV = 0.6506 and R^2 = 0.847). Compared to the model build starting from all 266 molecular descriptors, both RSS and RMSECV are considerably higher, but the complexity of the model is much lower (model size = 8).

Since CART selects a number of molecular descriptors during the tree building for the prediction of retention classes, it could also be used for feature selection. The CART method is known to be very efficient to evaluate a large number of predictor variables for the description of a given response [12]. Since the screening of all possible predictors is much faster in CART than in MARS, it could be time saving and more efficient to perform feature selection by means of CART, prior to the MARS model building. The 32 molecular descriptors selected by CART as primary or surrogate splits in the model of Section 4.3, were used to build a MARS model. Independently from the previous MARS models, a new maximal MARS model was build. After pruning back the optimal model uses 48 basis functions to describe the retention. A total of 18 descriptors is selected in the optimal model, which includes 37 interaction terms. The MARS model obtained describes the training data better (RSS = 0.0693 and R^2 = 0.998) than before, but uses more basis functions. The correlation coefficient equals 0.9967. However, the predictive performance of the model (RMSECV = 0.3373 and R^2 = 0.954) is worse than that obtained without feature selection. In Fig. 7a, plot of the predicted retention versus the observed is shown. The retention of each molecule is again predicted twice: ones as a test molecule during leave-one-out cross-validation and once as a training set member. The prediction of dilevalol, when it is test sample, seems to be an outlier. After its elimination, the retention prediction of the other molecules becomes better (RSS = 0.0697 and R^2 = 0.998; RMSECV = 0.1638 and R^2 = 0.989) than from the model without feature selection. However, for the latter MARS model dilevalol was not detected as an outlier.

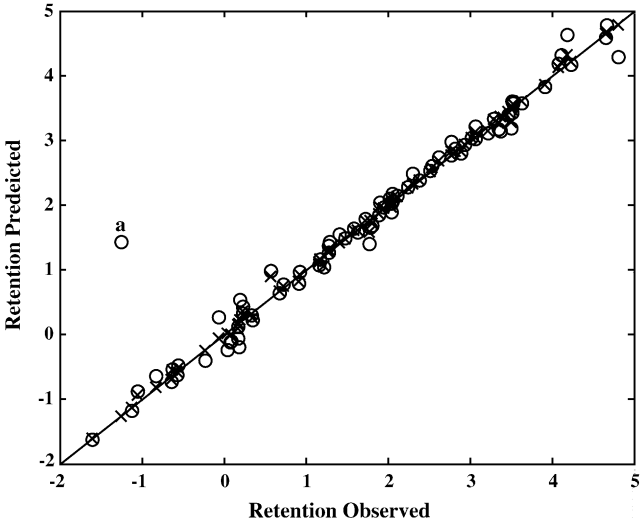

Fig. 7. Predicted versus observed retention for all objects using cross-validation for the MARS model obtained after feature selection by CART (a: dilevalol; ◯: test set; ×: training set).

A last possibility studied is the use of quadratic terms instead of interactions between different predictors. Starting from all 266 molecular descriptors, the optimal model uses 43 terms, from which 5 are quadratic. Compared to the model of Eq. (6), nine additional terms are included. Consequently, one could conclude that this model is not really simpler than the original one. However, the model describes the training data very good (RSS = 0.0745 and R^2 = 0.998) and the same goes for the prediction of the retention for new molecules (RMSECV = 0.1716 and R^2 = 0.988).

One may conclude that feature selection of predictors with CART prior to MARS may be useful to improve the interpretability of the MARS model, but it also may remove some valuable predictors, which also may decrease the predictive ability of the model. For the models built starting from all data, we may conclude that the use of quadratic terms instead of interactions may improve the model performance.

5. Conclusion

The performance of the MARS methodology to construct QSRRs was evaluated. The optimal MARS model obtained shows good predictive properties estimated with leave-one-out cross-validation and performs better than classical MLR. Moreover, the molecular descriptors selected in the MARS model are meaningful from a chromatographic point of view. The use of CART for variable selection may be useful since MARS is more computer intensive than CART. However, if computing time is not the main concern, variable selection is not advisable since valuable predictor variables may be lost. The introduction of quadratic terms of single molecular descriptors instead of interactions between molecular descriptors could also be useful to obtain better models and should be considered. Overall, it is concluded that MARS is a valuable technique to build QSRRs.

References

[1] K. Jinno, A Computer-Assisted Chromatography System, Hüthig, Heidelberg, 1990.
[2] R. Kaliszan, J. Chromatogr. B 715 (1998) 229.
[3] R. Kaliszan, J. Chromatogr. A 656 (1993) 417.
[4] R. Kaliszan, Quantitative Structure–Chromatographic Retention Relationships, Wiley–Interscience, New York, 1987.
[5] Y. Wang, X. Zhang, X. Yao, Y. Gao, M. Liu, Z. Hu, B. Fan, Anal. Chim. Acta 463 (2002) 89.
[6] Y.L. Loukas, J. Chromatogr. A 904 (2000) 119.
[7] L.I. Nord, D. Fransson, S.P. Jacobsson, Chemom. Intell. Lab. 44 (1998) 257.
[8] J.H. Friedman, Ann. Stat. 19 (1991) 1.
[9] R.D. De Veaux, D.C. Psichogios, L.H. Ungar, Comput. Chem. Eng. 17 (1993) 819.
[10] V. Nguyen-Cong, G. Van Dang, B.M. Rode, Eur. J. Med. Chem. 31 (1996) 797.
[11] J. Lahsen, H. Schmidhammer, B.M. Rode, Helv. Chim. Acta 84 (2001) 3299.
[12] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth, Monterey, CA, 1984.
[13] R. Put, C. Perrin, F. Questier, D. Coomans, D.L. Massart, Y. Vander Heyden, J. Chromatogr. A 988 (2003) 261.
[14] S. Sekulic, B.R. Kowalski, J. Chemom. 6 (1992) 199.
[15] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley–VCH, Weinheim, 2000.
[16] A. Nasal, A. Bucinski, L. Bober, R. Kaliszan, Int. J. Pharm. 159 (1997) 43.
[17] W.M. Meylan, P.H. Howard, J. Pharm. Sci. 84 (1995) 83.
[18] SRC, Interactive LogKow (KowWin) Demo, http://esc.syrres.com/interkow/kowdemo.htm.
[19] R. Todeschini, V. Consonni, Dragon Software version 1.1, http://www.disat.unimib.it/chm/Dragon.htm.
[20] L.B. Kier, L.H. Hall, Molecular Connectivity in Structure–Activity Analysis, Research Studies Press, Letchworth, 1986.
[21] D. Bonchev, Information Theoretic Indices for Characterization of Chemical Structures, Research Studies Press, Letchworth, 1983.
[22] E.V. Kostantinova, J. Chem. Inf. Comput. Sci. 36 (1997) 54.
[23] D. Bonchev, D.H. Rouvray (Eds.), Chemical Graph Theory—Introduction and Fundamentals, Gordon and Breach, New York, 1991.
[24] N. Trinajstic, Chemical Graph Theory, CRC Press, Boca Raton, FL, 1992.
[25] G. Rucker, C. Rucker, J. Chem. Inf. Comput. Sci. 33 (1993) 683.
[26] J. Galvez, R. Garcia, M.T. Salabert, R. Soler, J. Chem. Inf. Comput. Sci. 34 (1994) 520.
[27] P. Broto, G. Moreau, C. Vandycke, Eur. J. Med. Chem. 19 (1984) 66.
[28] P.A.P. Moran, Biometrika 37 (1950) 17.
[29] R.C. Geary, Incorp. Stat. 5 (1954) 115.
[30] R. Todeschini, P. Gramatica, Quant. Struct.-Act. Relat. 16 (1997) 120.
[31] M. Randic, G.M. Brissey, R.B. Spencer, C.L. Wilkins, Comput. Chem. 3 (1979) 5.
[32] A. Detroyer, Y. Vander Heyden, I. Cambré, D.L. Massart, J. Chromatogr A 986 (2003) 227.
[33] R. Kaliszan, K. Ośmiałowski, J. Chromatogr. 506 (1990) 3.
[34] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part A, Elsevier, Amsterdam, 1997.