Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Chromatography A, 988 (2003) 261–276
JOURNAL OF CHROMATOGRAPHY A
www.elsevier.com/locate/chroma

Classification and regression tree analysis for molecular descriptor selection and retention prediction in chromatographic quantitative structure–retention relationship studies

R. Put[a], C. Perrin[a,b], F. Questier[a], D. Coomans[a,c], D.L. Massart[a], Y. Vander Heyden[a,*]

[a]*ChemoAC, Department of Pharmaceutical and Biomedical Analysis, Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels, Belgium*

[b]*Laboratoire de Chimie Analytique, Faculté de Pharmacie, Université Montpellier 1, 15 avenue Charles Flahault, BP 14 491, 34093 Montpellier Cedex 5, France*

[c]*Statistics and Intelligent Data Analysis Group, School of Mathematical and Physical Sciences, James Cook University, Townsville Q4814, Australia*

Received 11 November 2002; received in revised form 20 December 2002; accepted 20 December 2002

Abstract

The use of the classification and regression tree (CART) methodology was studied in a quantitative structure–retention relationship (QSRR) context on a data set consisting of the retentions of 83 structurally diverse drugs on a Unisphere PBD column, using isocratic elutions at pH 11.7. The response (dependent variable) in the tree models consisted of the predicted retention factor (log k_w) of the solutes, while a set of 266 molecular descriptors was used as explanatory variables in the tree building. Molecular descriptors related to the hydrophobicity (log *P* and Hy) and the size (TPC) of the molecules were selected out of these 266 descriptors in order to describe and predict retention. Besides the above mentioned, CART was also able to select hydrogen-bonding and molecular complexity descriptors. Since these variables are expected from QSRR knowledge, it demonstrates the potential of CART as a methodology to understand retention in chromatographic systems. The potential of CART to predict retention and thus occasionally to select an appropriate system for a given mixture was also evaluated. Reasonably good prediction, i.e. only 9% serious misclassification, was observed. Moreover, some of the misclassifications probably are inherent to the data set applied.

Keywords: Molecular descriptors; Retention prediction; Regression analysis; Structure–retention relationships

1. Introduction

High-performance liquid chromatography (HPLC) is the most widely used separation technique in pharmaceutical analysis. Its ability to analyse a wide polarity range of acidic, basic and neutral compounds, and its high separative capabilities combined with automation, make HPLC the most efficient technique for the analytical characterisation of the continuously growing number of samples, produced at the different stages of drug development [1].

Related to the application of combinatorial

*Corresponding author. Tel.: +32-2-477-4723; fax: +32-2-477-4735.

E-mail address: yvanvdh@fabi.vub.ac.be (Y. Vander Heyden).

0021-9673/03/$ – see front matter
doi:10.1016/S0021-9673(03)00004-9

chemistry and high-throughput techniques, rapid HPLC method development is clearly needed in the pharmaceutical industry. The selection of appropriate starting conditions for method development, among which the selection of the stationary phase, is then crucial to reduce the time dedicated to an analysis. A wide variety of chromatographic stationary phases, providing significantly different retention and selectivity, are commercially available and principally offer the opportunity to perform any separation. However, the retention mechanisms are still not exactly known [2,3] and many stationary phases present similar characteristics which makes the selection of a proper stationary phase difficult and problem dependent. The choice of the stationary phase is still often based on empirical knowledge of the analyst and/or on an experimental trial-and-error approach on a selected set of stationary phases. The selection is then time-consuming and cost demanding.

Consequently, the development of mathematical models to predict the retention of new molecules is of particular interest for the pharmaceutical industry. Several approaches have been investigated in HPLC, among which quantitative structure–retention relationships (QSRRs) are the most popular [2]. In QSRR analysis one models the retention (e.g. the retention factors, k) of solutes measured on a given stationary phase under specific conditions, as a function of structural descriptors of the solutes [4]. The models are usually constructed using multiple linear regression (MLR) methods [5,6]. However, this approach can only be used when the number of objects (i.e. molecules) is larger than the number of variables (i.e. molecular descriptors) and the variables are not highly correlated. Since hundreds of molecular descriptors have been developed [7], either feature (i.e. variable) selection methods [8] have to be applied prior to MLR or other modeling methods such as neural networks (NNs) [9,10], principal component regression (PCR) [11] or partial least squares (PLS) [12] have to be used. Since these latter methods use combinations of the original variables (latent variables), the understanding of the chromatographic retention mechanisms becomes almost impossible. Therefore, MLR with feature selection is usually the preferred approach [8]. Genetic algorithms have been used for feature selection in QSRR studies [13].

In this study, another approach, classification and regression tree (CART) analysis was investigated. CART is a statistical method that explains the variation of a response variable using a set of explanatory variables, so-called predictors [14]. The method is based on a recursive binary splitting of the data into mutually exclusive subgroups containing objects with similar properties. CART is extensively used for modeling and classification in several areas, such as medical diagnosis and prognosis [14–16], and ecology [17]. However, its use in analytical chemistry is very limited. A very interesting advantage of CART is the possibility to deal with large numbers of both categorical and numerical variables. Another advantage is that no assumption about the underlying distribution of the predictor variables is required (even categorical variables can be used). Eventually, CART provides a graphical representation, which makes the interpretation of the results easy. Therefore, we felt that CART could be a very interesting method to select and relate molecular descriptors with the chromatographic retention of the molecules.

The goal of this study was to explore the possibilities of CART to find relationships between chromatographic retention of solutes on a given chromatographic system and the selected molecular descriptors. Since, for a given molecule we are mainly interested in the prediction of a suitable chromatographic system, we focused on the ability of the methodology to distinguish between classes with respectively low, intermediate and high retention on the considered system, rather than on the exact retention prediction of the compounds. A physicochemical explanation of the selected descriptors is also given.

2. Theory

2.1. Classification and regression trees

In 1984, Breiman et al. [14] introduced a methodology for classification and modeling, called ‘‘classification and regression tree analysis’’. The goal of this statistical method is to explain the variation of a single dependent variable, the response variable, using a set of independent predictors, referred to as explanatory variables, via a binary

partitioning procedure. Both the response and the explanatory variables can be either categorical or numerical. A classification tree, equivalent to discriminant analysis [18], is grown when the response variable is categorical while a regression tree is obtained for a numerical response variable [14].

CART works by splitting the data into mutually exclusive subgroups, called nodes, within which the objects have similar values for the response variable. The process starts from the root or parent node, which contains all objects of the data set. CART uses a repeated binary splitting procedure, which means that the parent node is split in two nodes, called child nodes. The process is repeated by treating each child node as a parent node (Fig. 1). Each split is defined by a simple rule, usually based on a single explanatory variable. For numerical explanatory variables, a splitting value (cut point) is selected to form two groups, which contain objects with values smaller and larger, respectively, than the selected cut point. For categorical variables, a split is defined by relating one or more levels of the variable to a specific node. Trees are grown by selecting the splits in such a way that the so-called homogeneity and the impurity of the response variable within each node is maximized and minimized, respectively. To achieve this, CART looks at all possible splits for all variables included in the analysis. The resulting splits are compared and eventually, the best split is chosen by evaluation of the impurity of the formed nodes, according to statistical criteria. This procedure is repeated for each consecutive split made in the tree. The splitting procedure is continued until no further split can be performed, i.e. all child nodes are homogeneous, or contain one or a user-defined minimal number of observations. The tree thus obtained is called the maximal tree and the terminal nodes, the so-called leaves, represent the final groups formed by the tree. This maximal tree will usually contain too many leaves and will overfit the learning data set, which will cause poor predictive abilities for new samples [14]. Therefore, the selection of an optimal tree with a good compromise between model fit and predictive properties is required. Thus, in general, CART analysis consists of three steps: (1) the maximal-tree building, (2) the tree ''pruning'', which consists in the cutting-off of nodes to generate a sequence of simpler (i.e. smaller) trees, (3) the optimal-tree selection.

2.1.1. Maximal-tree building

The growing of the tree starts at the root node,

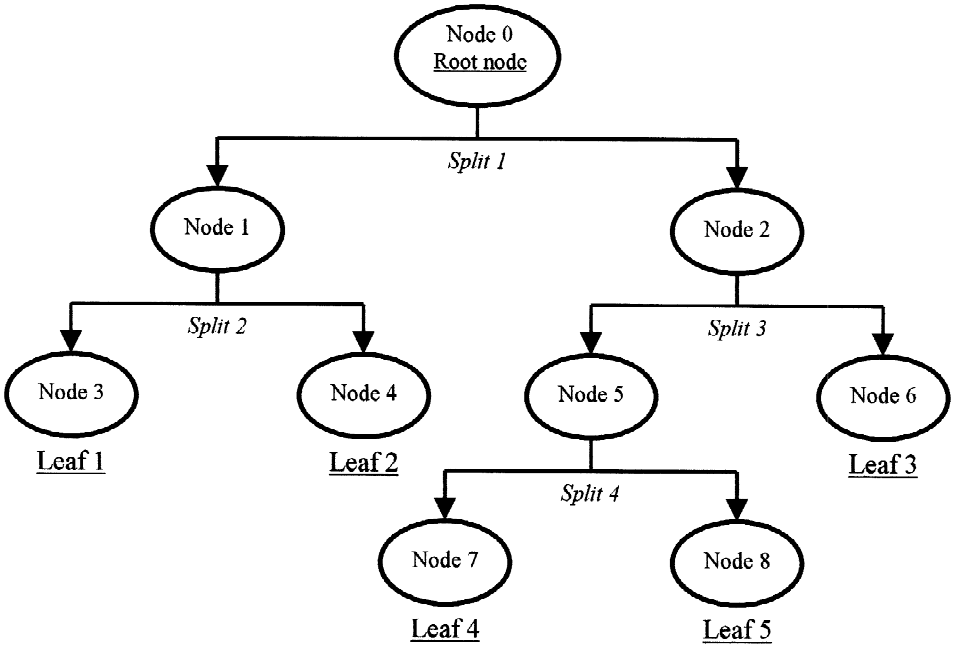

Fig. 1. Structure of a classification and regression tree.

containing all observations. CART is then looking for the best possible variable, so-called splitter, to divide the root node into two child nodes. To achieve this, the program looks at all possible variables, as well as at all possible values of the variable that can be used to split the data. The best splitter is defined as the variable (and associated splitting value) that will minimize the impurity, *i*, of the two child nodes. The goodness of a split is then defined as the impurity decrease between the parent node and its children:

$$\Delta i(s, t_P) = i_P(t_P) - p_L i(t_L) - p_R i(t_R) \quad (1)$$

where s is a candidate split, p_L and p_R are the fractions of observations of the parent node t_p that go into the child nodes t_L and t_R, respectively. The best splitter is the one that will maximize $\Delta i(s, t_p)$.

Different criteria to measure the impurity of a node have been proposed [14,17]. For regression trees, the total sum of squares of the response values about the mean of the node is the most popular measure of impurity [14]:

$$i(t) = \sum_{x_n \in t} (y_n - \bar{y}(t))^2 \quad (2)$$

where $i(t)$ is the impurity of node t; y_n, is the response value of observation x_n belonging to node t; $\bar{y}(t)$, the mean of all observations in node t. Absolute deviations about the node medians is another criterion which is used to build (robust) trees [14].

Once a split is made, a label or class is assigned to the child nodes. For regression trees, this is simply the mean within the node. For classification trees, the simplest rule is to assign the largest representation as the label (class) of a node. A label or class is assigned to every node of the tree since it is unknown which nodes finally will be kept in the optimal tree (see Sections 2.1.2 and 2.1.3).

2.1.2. Tree pruning

The resulting maximal trees are usually oversized and describe the training set perfectly. This is what in modeling is called overfitting [11,19]. Such trees often are difficult to interpret and their predictive ability for new observations is generally poor since they tend to fit also the noise in the data. The selection of a smaller tree, derived from the maximal one, is then necessary for predictive purposes. As with other modeling techniques, one is looking for the best compromise between model fit and prediction properties [19].

The selection of the optimal tree is done by a tree pruning procedure [14]. This procedure generates a sequence of smaller trees, which are obtained by removing successively branches of the maximal tree. The different subtrees are then compared to determine the optimal one.

Since several trees of the same size can be generated from the maximal tree, a procedure to determine the best one, is defined. Both accuracy, by some error measure, and complexity of the tree are considered. This is done by a cost-complexity measure, $R_\alpha(T)$, defined for each subtree, T, as:

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}| \quad (5)$$

with $R(T)$ the average within-node sum of squares, $|\tilde{T}|$ the tree complexity, which is equal to the total number of nodes of the subtree, and α the complexity parameter, which is a penalty for each additional terminal node [14]. During the pruning procedure the value of α will gradually be increased from 0 to 1. For each value of α, one can find a subtree, $T(\alpha)$, that minimizes $R_\alpha(T)$. The larger α becomes, the smaller $|\tilde{T}|$ should be to minimize $R_\alpha(T)$. Thus, by gradually increasing α, one generates a sequence of pruned subtrees starting from the largest tree.

2.1.3. Optimal tree selection

Eventually, the optimal tree is selected from the generated sequence of subtrees by evaluating the predictive error of the trees. The predictive error is often estimated using cross-validation, especially for small data sets [14]. In cross-validation, some samples are randomly drawn from the data set, to test the tree, which is built with the rest of the data [12]. In 10-fold cross-validation, the original data set is divided into 10 equal parts (test sets), each containing a similar distribution for the response variable. A tree is then built using 90% of the observations (learning set), while the remaining 10% (test set) are used to test the tree. This step is repeated 10 times using each time a different test set and the remaining observations as the learning set. The optimal tree is the one having the minimal cross-validation error (most accurate tree). In practice, the

optimal tree is chosen as the simplest tree with a predictive error estimate within one standard error of the minimum. In this way, the chosen tree is the simplest with an error estimate comparable to the one of the most accurate tree.

2.1.4. Variable ranking: selection of primary and surrogate splits

It is sometimes observed that a given variable x_2 does not occur in the final tree structure, while it prominently does when another tree, which is almost as accurate as the first one, is grown after removing a so-called masking variable x_1 from the data set. However, the variables x_1 and x_2 do not necessarily cause a similar split in the data set; they both cause a considerable decrease in impurity. Such variables are called primary variables and the splits they cause are the so-called primary splits. The importance of the explanatory variables to introduce a split in the tree is detected by the variable ranking method in CART. The most relevant properties to describe the response variable can then be identified, so that CART can be used for feature selection [14].

On the other hand, so-called surrogate splits are defined as splits causing a similar distribution of the objects in the groups obtained after splitting. The variables responsible for these similar distributions are called surrogate variables. When for an object the value of the splitting variable is missing, the value of a surrogate variable is then used to decide to which node the object is awarded.

2.2. Molecular descriptors

Molecular descriptors can be defined as the final result of a logical and 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) [7]. The term ''useful'' means that the resulting number can contribute to a better understanding of molecular properties and/or can be used in a model to predict properties of molecules.

In the literature over 6000 descriptors are defined, and the number still grows [7]. Several ways to classify molecular descriptors into groups exist. The simplest one is based on the nature of the descriptor, namely whether it is theoretical or experimental. A further classification of theoretical molecular descriptors is based on the dimensionality of the molecular representation [7]. A first class contains so-called zero-dimensional (0D) descriptors, which are derived from the chemical formula. The information considered here is, for instance, the number and type of atoms, the molecular mass, any function of atomic properties (e.g. sum of atomic van der Waals volumes). A substructure list representation of a molecule can be considered as a one-dimensional (1D) molecular representation and consists of a list of molecular fragments (e.g. functional groups, substituents, etc.). The derived molecular descriptors are called 1D-descriptors (e.g. count descriptors of functional groups, rings and bonds).

A molecular graph contains topological or two-dimensional (2D) information. It describes how the atoms are bonded in a molecule, both the type of bonding and the interaction of particular atoms. The derived molecular properties are called 2D descriptors (e.g. total path count; see Section 4.1). Another group of theoretical descriptors consists of three-dimensional (3D)-descriptors, which are calculated starting from a geometrical or 3D representation of a molecule. Finally the descriptors, which are derived from a stereo-electronic or lattice representation, are called four-dimensional (4D) descriptors.

In this study 0D, 1D, 2D molecular descriptors and four experimental descriptors (i.e. log *P*, the unsaturation index, the hydrophilic factor Hy (see Section 4.1) and the aromatic ratio) were used.

3. Experimental

The chromatographic data used were obtained from the paper by Nasal et al. [20] and consisted of the logarithms of the retention factors ($\log k_w$) for 83 basic drugs. They belonged to the following pharmacological classes: psychotropic drugs, drugs acting through α-adrenoreceptors (both agonists and antagonists), β-adrenolytics, antagonists of histamine H_1 receptors, histamine H_2 receptor antagonists and inactive phenothiazine derivatives. The data were obtained on Unisphere PBD, a polybutadiene-coated alumina column at pH 11.7 using isocratic elutions [20]. The dimensions of the column were 100×4.6

mm I.D., with a particle size of 8 μm. Since the solutes show a large diversity in molecular structure, it is not possible to measure the retentions for all molecules isocratically on the same chromatographic system. Therefore, the proportions (%, v/v) of methanol–aqueous buffer used range from 75:25 to 0:100 [20]. To compare the retentions measured, a hypothetical retention factor, log k_w, is then required. The log k values measured for individual solutes were regressed against the volume fraction of organic modifier in the eluent and the obtained line was extrapolated to a hypothetical capacity factor corresponding to 0% of organic modifier (100% buffer). This approach is, for instance, currently applied when one tries to predict log P values from chromatographic retention [21]. Therefore, it was also applied here in a more general QSRR context. More details of the chromatographic parameters can be found in Ref. [20]. A list of the molecules and their log k_w and log P values is shown in Table 1.

The log P values of the substances were calculated using the on-line interactive LOGKOW program of the Environmental Science Center of Syracuse Research, Syracuse, NY, USA [22,23].

For all molecules the geometrical structure was optimized using Hyperchem 6.03 Professional software (Hypercube, Gainesville, FL, USA). Geometry optimization was obtained by the Molecular Mechanics Force Field method (MM+) using the Polak-Ribière conjugate gradient algorithm with an RMS gradient of 0.05 kcal/(Å mol) as stopping criterion (1 cal = 4.184 J). The Cartesian coordinate matrices of the positions of the atoms in the molecule, which result from this geometrical representation, were used for the calculation of the molecular descriptors using the Dragon 1.1 software [24]. Out of the 853 molecular descriptors, which potentially can be calculated with this program, the 0D, 1D, 2D ones beside some experimental descriptors were selected. The following groups of descriptors, as defined in Dragon 1.1, were calculated: 56 constitutional descriptors [7], 69 topological descriptors [25–29], 20 molecular walk counts [30], 21 Galvez topological charge indices [31], 96 2D autocorrelations [32–34] and three empirical descriptors [7].

Regression trees were grown using the TreePlus add-on module [35] in the S-Plus 2000 environment (Mathsoft, Cambridge, MA, USA). The retention data were used as response variable and the selected descriptors as explanatory variables. Additional data plots were made using Matlab 5.3.1 (Mathworks, Natick, MA, USA).

4. Results and discussion

4.1. Building of the classification and regression trees

Trees were grown using the retention data (log k_w) of all 83 molecules on a Unisphere PBD column at pH 11.7. The chromatographic data investigated were chosen because the retention of a large diversity of chemical structures was measured. Since the response variable is continuous, the resulting trees are regression trees. The explanatory variables used belong to several classes of molecular descriptors as mentioned in Section 3. A total of 266 descriptors were used as explanatory variables.

The regression trees were grown using Eq. (2) as impurity measure. Ten-fold cross-validation was used to define the optimal tree. The latter was selected from the maximal tree, which was pruned back. The plot of the maximal regression tree is shown in Fig. 2. For this maximal tree the minimal number of objects per node, i.e. two in our study, was defined equal to log (n/2) with n the total number of objects [35]. For the abbreviations used for the different molecular descriptors, we refer to Ref. [24].

Fig. 3 shows a plot of the prediction error, calculated as the root mean squared error of cross validation (RMSECV), as a function of the size of the tree. A horizontal line indicates the selection limit, situated one standard error above the minimal RMSECV. Applying this selection limit suggests a tree size of four leaves as optimal. Fig. 4 shows both the tree with the minimal RMSECV (Fig. 4a) and the selected optimal tree (Fig. 4b). The nodes are numbered according to the order of the tree growing. The splitting rules, the average response value and the numbers of objects of the leaves are indicated similarly as in Fig. 2. Additionally, histograms are plotted that represent the distribution of the response for the objects within each node. Each bar covers a specific range of log k_w values, with increasing

Table 1
The extrapolated retention data log k_w, the log P values and the predicted retention classes (explanation: see Section 4.4) of the 83 drugs studied [20,23]

No.	Drug	Log k_w	Log P	Prediction class
1	Acebutolol	0.351	1.19	Very low or low
2	Acetopromazine	2.934	4.24	High or very high
3	2-Acetylphenothiazine	3.065	3.51	High or very high
4	Alprenolol[a]	1.720	2.81	Intermediate Intermediate or high
5	Antazoline	1.888	3.38	Intermediate or high
6	Astemizole	3.508	6.43	High or very high
7	Atenolol	−1.048	−0.03	Very low or low
8	Betaxolol	1.772	2.98	Intermediate or high
9	Bisoprolol	0.094	1.84	Very low or low
10	Brimonidine	0.178	−1.30	Very low or low
11	Bupranolol	2.055	3.07	Intermediate or high
12	Carbamazepine	0.926	2.25	Very low or low
13	Carteolol[a]	0.228	1.42	Very low Very low or low
14	Celiprolol[a]	0.232	1.93	Very low or low
15	Chloropyramine	2.767	3.37	Intermediate or high
16	Chlorpheniramine (+)	1.899	3.82	Intermediate
17	Chlorpheniramine (+/−)[a]	2.043	3.82	Intermediate or high
18	Chlorpromazine	4.076	5.20	High or very high
19	Chlorprothixene	4.235	5.14	High or very high
20	Cicloprolol	0.573	2.10	Very low or low
21	Cimetidine	0.724	0.57	Very low or low
22	Cinnarizine[a]	4.665	5.44	High or very high Intermediate or high
23	Cirazoline	1.583	3.22	Intermediate or high
24	Clomipramine	3.910	5.65	High or very high
25	Clonidine	1.283	1.89	Very low or low
26	Desipramine	2.888	4.80	High or very high
27	Detomidine[a]	1.627	3.29	Intermediate Intermediate or high
28	Dilevalol	−1.258	2.00	Very low or low
29	Dimethindene	2.240	4.98	Very high
30	Diphenhydramine	2.112	3.11	Intermediate or high
31	Doxazosin	2.823	2.09	Very low or low
32	Esmolol	0.916	2.00	Very low or low
33	Ethopropazine	4.181	5.47	High or very high
34	Famotidine	0.193	−0.65	Low
35	Fluphenazine	3.352	4.13	High
36	Imipramine	3.020	5.01	Intermediate or high
37	Indoramin	2.299	3.60	High or very high
38	Isothipendyl	2.535	3.94	High or very high
39	Ketotifen	1.950	3.64	High or very high
40	Lofexidine	1.410	3.58	High or very high
41	Medetomidine	2.516	4.50	Intermediate or high
42	Mepyramine	2.049	2.81	Intermediate or high
43	2-Methoxyphenothiazine	3.400	3.12	High or very high
44	Metiamide	0.044	0.52	Low

Table 1. Continued

No.	Drug	Log k_w	Log P	Prediction class
45	Metoprolol	−0.553	1.69	Low
46	Moxonidine	−1.125	0.24	Very low or low
47	Nadolol	−0.637	1.17	Low or intermediate
48	Naphazoline	1.476	3.52	High
49	Nifenalol	0.075	0.99	Low or intermediate
50	Nizatidine	−0.569	−0.67	Very low or low
51	Oxprenolol	1.218	1.83	Low or intermediate
52	Oxymetazoline	1.274	4.87	Intermediate or high
53	Perphenazine[a]	3.070	3.82	High or very high
54	Pheniramine	1.275	3.17	High or very high
55	Phenothiazine[a]	3.375	3.82	High or very high
				Intermediate or high
56	Phentolamine	−0.834	3.36	Intermediate or high
57	Pindolol	0.331	1.48	Very low or low
58	Pizotifen	3.465	5.51	Intermediate or high
59	Practolol	−0.627	0.53	Very low or low
60	Prazosin[a]	1.172	1.28	Very low or low
				Low or intermediate
61	Prochlorperazine[a]	3.523	4.79	Very high
				High or very high
62	Promazine	3.294	4.56	High or very high
63	Promethazine[a]	3.216	4.487	Very high
				High or very high
64	Propiomazine	3.497	4.66	High or very high
65	Propranolol[a]	2.038	2.60	Intermediate
				Intermediate or high
66	Ranitidine	1.779	0.29	Very low or low
67	Roxatidine acetate[a]	1.154	2.21	Low
				Very low or low
68	Sotalol	−1.602	0.37	Very low
69	Terazosin	0.167	1.47	Low
70	Tetryzoline	0.680	3.69	High or very high
71	Thioridazine	4.655	6.45	High or very high
72	Thiothixene-*cis*	2.770	3.14	High or very high
73	Tiamenidine[a]	−0.231	0.79	Low
74	Timolol	0.171	1.75	Very low or low
75	Tolazoline	−0.063	2.34	Intermediate
				Very low or low
76	Trifluoperazine	3.632	5.11	Very high
77	2-Trifluoromethylphenothiazine	4.804	4.79	Very low or low
78	Triflupromazine	4.117	5.52	Very high
79	Trimeprazine[a]	3.508	4.98	High or very high
80	Tripelennamine	1.807	2.73	Intermediate or high
81	Triprolidine[a]	2.618	3.70	Intermediate or high
82	Tymazoline	2.012	3.88	Intermediate or high
83	Xylometazoline	2.385	5.35	Intermediate or high

[a] The molecule was selected twice for the test set.

retention towards the right part of the plots. This allows to see clearly the partition in retention classes (i.e. low retention for nodes 6 and 7, medium for node 4 and long retention for nodes 5, 8 and 9).

For the optimal subtree with four terminal nodes, three molecular descriptors were selected to describe the retention data. The molecular descriptor, which is selected first is the ‘‘hydrophobicity parameter (log

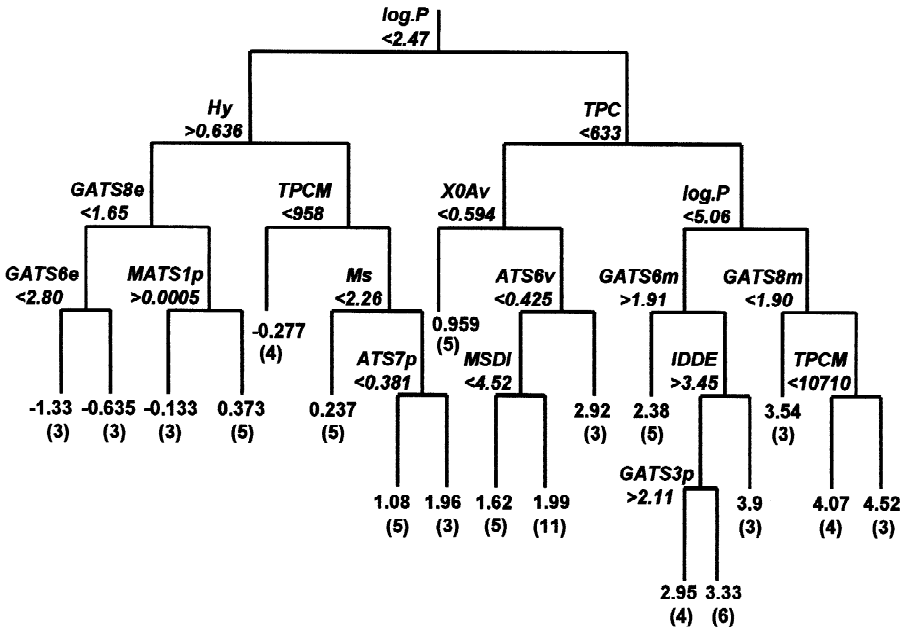

Fig. 2. Maximal regression tree, grown for the log k_w values of 83 drugs on a Unisphere PBD column at pH 11.7 using 266 molecular descriptors as explanatory variables. For each leaf the mean log k_w value is given, as well as the number of objects (molecules), between brackets. For each split the criterion that defines the left part is indicated.

P)''. For the tree with the minimal RMSECV this descriptor is even used twice: it defines both the first and the last split. The other selected molecular descriptors are the ''hydrophilic factor'' (Hy) [36] and the ''total path count'' (TPC) [37].

The use of log *P* to describe retention data can be expected. In the literature, log *P* indeed is often used in quantitative structure–retention relationships for RPLC, because retention is based on a partition mechanism, in which hydrophobic interactions are the most important [1]. The selection of log *P* out of more than 250 molecular descriptors indicates the ability of CART to relate chromatographic retention with molecular descriptors and its use for feature selection in QSRR.

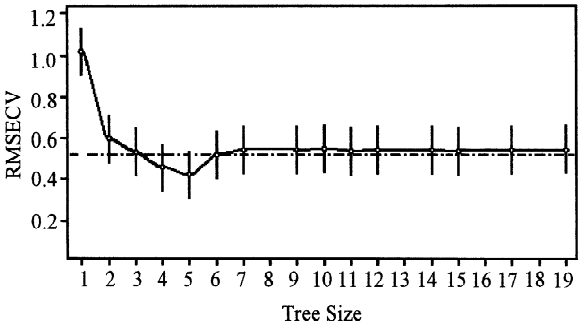

Fig. 3. RMSECV versus tree size. The tree size is defined as the number of leaves in a given tree. The dotted line represents the selection limit.

The hydrophilic factor (Hy) is directly, but in a negative way, correlated to log *P* and thus its selection also is not surprising. Hy is defined by Todeschini et al. [36] as an empirical index related to the hydrophilicity of compounds. It is based on count descriptors and can be calculated as:

$$\mathrm{Hy} = \frac{(1+N_{\mathrm{Hy}})\log_2(1+N_{\mathrm{Hy}}) + N_{\mathrm{C}}\cdot[(1+A)\log_2(1/A)] + \sqrt{N_{\mathrm{Hy}}/A^2}}{\log_2(1+A)} \quad (9)$$

where N_{Hy} represents the number of hydrophilic groups (–OH, –SH, –NH), N_{C} the number of carbon atoms and A the total number of atoms, hydrogens excluded.

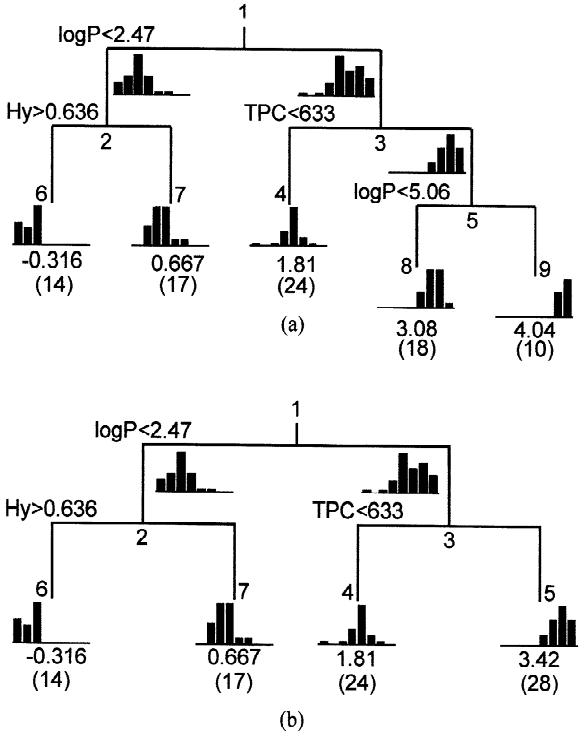

Fig. 4. Pruned regression trees, (a) with minimal RMSECV and (b) optimal tree. Data used: see text. For each leaf the mean log k_w value for its elements and the number of molecules is represented. The distribution of log k_w for each node is illustrated in a histogram. The criterion defining each split is also printed.

The selection of the total path count (TPC) on the other hand was not a priori foreseen. A molecular path count ${}^{m}P$ is defined as the total number of paths of length m in the graph [38]. The TPC is a descriptor obtained from the H-depleted molecular graph of a molecule and is calculated by summing all molecular path counts ${}^{m}P$ with $m = 0, 1 \ldots, L$ and L the length of the longest path in the graph [37]:

$$\mathrm{TPC} = \sum_{m=0}^{L} {}^{m}P \tag{10}$$

In general, the TPC is considered as a quantitative measure of molecular complexity [7]. Because of the fact that the TPC and the volume of the molecule are correlated, the TPC is related to the size of the molecule. This interpretation explains the selection of the TPC in the tree, since it is known that hydrophobicity and molecular size are two main discriminant properties for retention in RPLC [6].

4.2. Primary and surrogate splits

Log *P*, TPC and Hy define both the tree with the minimal RMSECV (Fig. 4a) and the optimal tree (Fig. 4b). However, these are not the only descriptors selected by CART. For each node splitting, CART provides a list of descriptors giving the most important improvement in node impurity. The corresponding splits are called primary splits and the one improving impurity most, is used to cause the effective split. The primary splits for the nodes of the tree with the minimal RMSECV (Fig. 4a) are listed in Table 2. As mentioned above, hydrophobic/hydrophilic properties (log *P*/Hy) are the most important for the first split (node 1). H-bonding properties (nHD) [39], molecular shape (PW5) [40] and molecular complexity (PCR, TPC) [7,37] also are variables causing a considerable decrease in the impurity. Notice that the molecular descriptors used to define all splits in the tree (i.e. log *P*, Hy and TPC) are already selected as primary splits of the first node. For the second node a more general index of spatial autocorrelation regarding the atomic masses (GATS5m) [34] is selected as primary variable, besides the hydrophilic properties (Hy). Molecular complexity is represented by TPCM, PCR and by the average valence connectivity indices X0Av and X1Av [41,42]. The third node has besides TPC, nR06 containing steric properties information [14], TPCM, PCD and the molecular walk counts (MWC05 and MWC06) [43], related to molecular size and molecular branching, as primary variables. Finally, the last split (node 5) selects analogous descriptors as before (log *P*, ATS8m, GATS2m and ATS5m) and additionally the topological charge indices GGI9 and JGI9 [31,44], which were proposed to evaluate the global charge transfer in the molecule.

From the above, it can be observed that primary variables do not necessarily describe the same properties. This is not surprising since their selection is only based on the improvement of the impurity criterion and they do not necessarily lead to a comparable distribution in the child nodes.

After removing log *P* from the data set, Hy is selected for the first split, as might be expected from the list of primary splits of the first node. In this new tree (i.e. without log *P*), steric (nR10) and H-bond-

Table 2
Molecular descriptors selected by CART. The node numbers refer to Fig. 4a. The surrogate splits are those for the most important primary variable

Node 1		
Primary splits	Importance	Definition descriptor
Log $P<2.469\rightarrow$left	0.5691	Hydrophobicity parameter
Hy$<0.2745\rightarrow$right	0.5335	Hydrophilic factor
nHD$<1.5\rightarrow$right	0.5074	Number of donor atoms for H-bonds
PW5$<0.097\rightarrow$left	0.4701	Path/Walk 5–Randic shape
PCR$<4.73\rightarrow$left	0.4684	Ratio of multiple path counts to path counts
TPC$<824.5\rightarrow$left	0.4626	Total path count
Surrogate splits	Agree	Definition descriptor
ATS 1e$<1.023\rightarrow$right	0.9157	Broto-Moreau autocorrelation of a topological structure (ATS)–lag 1/weighted by atomic Sanderson electronegativities (SEN)
Hy$<0.333\rightarrow$right	0.8916	Hydrophilic factor
ATS3e$<1.01\rightarrow$right	0.8795	ATS–lag 3/weighted by atomic SEN
ATS6e$<1.007\rightarrow$right	0.8795	ATS–lag 6/weighted by atomic SEN
nHA$<4.5\rightarrow$right	0.8675	Number of acceptor atoms for H-bonds
Node 2		
Primary splits	Importance	Definition descriptor
Hy$<0.636\rightarrow$right	0.2806	Hydrophilic factor
GATS5m$<1.762\rightarrow$left	0.2726	Geary autocorrelation (GATS)–lag 5/weighted by atomic masses
TPCM$<16860\rightarrow$left	0.2522	Total multiple path count
PCR$<8.762\rightarrow$left	0.2522	Ratio of multiple path counts to path counts
X0Av$<0.5675\rightarrow$right	0.2522	Average valence connectivity index chi-0
X1Av$<0.297\rightarrow$right	0.2522	Average valence connectivity index chi-1
Surrogate splits	Agree	Definition descriptor
nHD$<2.5\rightarrow$right	0.9355	Number of donor atoms for H-bonds
IVDE$<1.88\rightarrow$right	0.8065	Mean information content vertex degree equality
CIC$<1.037\rightarrow$left	0.7742	Complementary information content (neighborhood symmetry)
MATS1e$<-0.0565\rightarrow$left	0.7742	Moran autocorrelation–lag 1/weighted by atomic SEN
GATS7m$<1.875\rightarrow$left	0.7742	GATS–lag 7/weighted by atomic masses
Node 3		
Primary splits	Importance	Definition descriptor
TPC$<633\rightarrow$left	0.5360	Total path count
nR06$<2.5\rightarrow$left	0.5043	Number of 6-membered rings
TPCM$<3324\rightarrow$left	0.4990	Total multiple path count
PCD$<31.35\rightarrow$left	0.4990	Difference of multiple path counts to path counts
MWC05$<11.6\rightarrow$left	0.4683	Molecular walk count of order 5
MWC06$<4.75\rightarrow$left	0.4683	Molecular walk count of order 6
Surrogate splits	Agree	Definition descriptor
TPCM$<4472\rightarrow$left	0.9615	Total multiple path count
PCD$<43.76\rightarrow$left	0.9615	Difference of multiple path counts to path counts
MWC05$<11.45\rightarrow$left	0.9423	Molecular walk count of order 5
MWC06$<4.25\rightarrow$left	0.9423	Molecular walk count of order 6
MWC07$<1.45\rightarrow$left	0.9423	Molecular walk count of order 7

Table 2. Continued

Node 5		
Primary splits	Importance	Definition descriptor
Log $P<5.059\rightarrow$left	0.4166	Hydrophobicity parameter
ATS8m$<0.212\rightarrow$left	0.2831	ATS–lag 8/weighted by atomic masses
GGI9$<0.0455\rightarrow$left	0.2575	Topological charge index of order 9
JGI9$<0.0025\rightarrow$left	0.2575	Mean topological charge index of order 9
GATS2m$<1.492\rightarrow$right	0.2532	GATS–lag 2/weighted by atomic masses
ATS5m$<0.432\rightarrow$left	0.2532	ATS–lag 5/weighted by atomic masses
Surrogate splits	Agree	Definition descriptor
GATS5e$<1.167\rightarrow$right	0.8214	GATS–lag 5/weighted by atomic SEN
X0AV$<0.655\rightarrow$left	0.7857	Average valence connectivity index chi-0
SIC$<0.7545\rightarrow$left	0.7857	Structural information content (neighborhood symmetry)
BIC$<0.6835\rightarrow$left	0.7857	Bond information content (neighborhood symmetry)
PW3$<0.3205\rightarrow$right	0.7857	Path/Walk 3–Randic shape

ing properties (nHA) are selected besides the hydrophilic properties (Hy). Thus analogue properties are selected compared to the original tree. The descriptors selected always are related to the hydrophobic/hydrophilic properties of the molecule, its H-bonding properties and to its molecular/steric complexity.

Besides the primary splits, CART also provides surrogate splits for the most important primary variable in a node. This is another benefit of CART, because sometimes it is very likely that missing data occur when dealing with molecular descriptors (e.g. experimental descriptors). To appoint, for instance, molecules with missing log *P* values in the first split to the child nodes, the autocorrelation descriptor ATS1e is used as surrogate variable. The descriptors hydrophilicity (Hy), the autocorrelation descriptors (ATS3e and ATS6e) and H-bonding acceptor properties (nHA) also give a classification that is about 90% similar to the one obtained with log *P*.

The surrogate splits for node 2 descriptor Hy, consist of H-bonding donor properties (nHD), symmetry characteristics (IVDE and CIC) [45] and autocorrelation descriptors (MATS1e and GATS7m) [33,34]. The surrogate splits for TPC (node 3) are defined by molecular complexity (TPCM and PCD) [7,37] and molecular walk counts (MWC05, MWC06 and MWC07) [43]. In node 5 the surrogate splitters are GATS5e, X0Av, SIC, BIC, PW3, respectively.

Since surrogate variables cause a similar distribution of the objects in the groups obtained after splitting, one could expect that surrogate variables usually will represent similar properties. This can, for instance, clearly be seen for node 3, where several molecular complexity descriptors are selected as surrogates for TPC. A second benefit of the surrogate variables, besides indicating objects with missing values to a node, is the interpretation of properties described by a descriptor, since some molecular descriptors are easier to interpret than others.

Because CART provides lists of these primary and surrogate splits it is very efficient to evaluate all possible variables, which can be related to a certain property (response variable).

4.3. Evaluation of the splits in the tree with minimal RMSECV

Log k_w values from the parent nodes of Fig. 4a were plotted versus log *P* (twice), Hy and TPC, i.e. the variables causing the split into child nodes, to have a closer look at the introduced splits during the tree building. The relationships between the selected variables and log k_w are shown in Fig. 5. The limit values defining the splits are indicated by a vertical line. Only the molecules relevant for a specific node are plotted. In Fig. 5a, for instance, all 83 molecules are plotted, whereas only 31 molecules are represented in Fig. 5b.

The descriptor, selected by CART to define the first split (log *P*), is highly correlated with log k_w

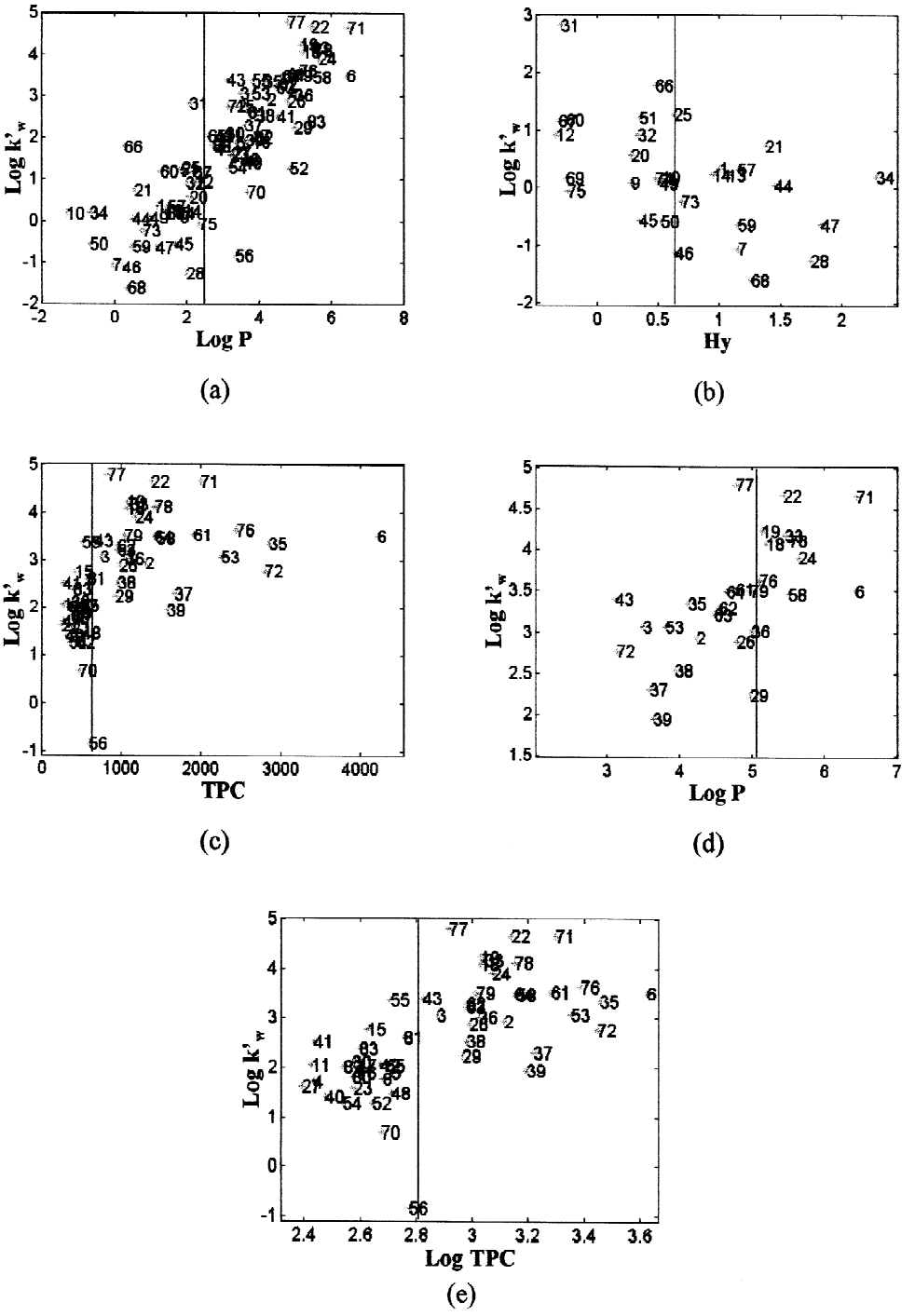

Fig. 5. Log k_w versus the explanatory variables causing the splits in Fig. 4a, (a) log *P*, (b) Hy, (c) TPC, (d) log *P*, (e) log TPC. The vertical line represents the limit value to divide into two child nodes.

($r = 0.84$) (Fig. 5a). The first split divides the data into two groups, which contain molecules with log P values below and above 2.5, respectively. This corresponds with log k_w values roughly below and above 1.5. The last split (Fig. 5d) is also defined by log P: values under 5.06 now form the first group with log k_w values below 3.5, while the other group contains molecules with log $P > 5.06$ and log $k_w >$ 3.5. As mentioned before, Hy and log P are correlated in a negative way as can be seen from Fig. 5a and b. The relation between Hy and log k_w seems to be rather linear ($r = 0.65$). The two groups defined by the Hy split show an overlap in log k_w values. For the low Hy values the log k_w range from -0.4 to 3, while for the high Hy values they are between -2 and 1.

Finally, the retention in Fig. 5c is non-linearly related to the TPC data. This is something that should not occur in a classical regression model. The log k_w values range from -1 to about 3 for lower TPC values, whereas they have values between 2 and 5 for the high ones. To obtain a more linear relation, the logarithm of TPC (log TPC) was plotted against log k_w. As can be seen from Fig. 5e the relationship between log TPC and log k_w becomes indeed more linear.

4.4. Prediction

To evaluate the predictive power of CART, 10-fold cross-validation was performed. Therefore, initially the molecules were ranked in ascending order of retention. Then the data were split into uniformly distributed test/calibration sets (10/73 objects). Trees were grown from the calibration sets, while the corresponding test samples were predicted. Since our main interest is the prediction of retention classes, rather than the exact log k_w values, the prediction of the test samples was evaluated in terms of misclassification rate instead of RMSECV. The criterion used to distinguish between well-classified and misclassified test samples was defined based on the observed distributions of the training substances within the four leaves of the minimal trees. The possibility exists that high retention values in a distribution overlap with low retention values of a distribution in a neighboring leaf. Therefore, a test sample is considered misclassified as CART predicts it in a leaf, covering a log k_w range that does not contain the experimental log k_w value of the considered test sample.

The minimal tree built from the training sets always contained four leaves. Arbitrarily we divided the retentions of all 83 molecules equally into five classes, which were called the very low, low, intermediate, high and very high retention classes. Then for a given training/calibration tree a leaf received a label, which was equivalent to one or two of the above classes depending on its content. For instance, if the members of a leaf mainly had retention parameters belonging to the class very low retention, the leaf was labeled ‘‘very low retention’’. If another leaf contained mainly substances from the classes intermediate and high retention, we gave it the label ‘‘intermediate or high retention’’. Thus, depending on the calibration tree considered the labeling of the leaves could be somewhat different. The above approach allowed us, as shown in Table 1, to indicate to which class a given substance was predicted to belong, for those situations it was a member of the test set during a cross-validation step.

The selected descriptors in the calibration trees were analogue to those defining the tree grown on all data. Four test sets showed a misclassification rate of one out of 10 test samples, twice two molecules were misclassified and for the remaining four test sets three test samples were misclassified. Relatively high misclassification rates may thus be obtained. Overall, for the 10 test sets a misclassification of 20% is observed. After further examination of the misclassifications, it was concluded that just nine molecules are more seriously misclassified while the remaining 11 molecules are situated just outside the domain of the correct nodes. Thus 80 out of 100 molecules are classified correctly, 11 are classified just outside the correct node and only nine are more seriously misclassified. For instance, the prediction of 2-trifluoromethylphenothiazine turned out to be very bad. It was predicted to have a (very) low retention, whereas the experimental log k_w value was the highest in the data set (log $k_w = 4.804$). This large error may be due to the fact that this high retention value, when occurring in a test set, always is situated outside the domain of the given training set and for its prediction one is extrapolating. Thus we may conclude that extrapolations in CART have to be

avoided as in other modeling methods. Therefore it is important that the training set covers all possible retention values.

Another possible explanation for some misclassifications may be found in the nature of the data. As mentioned in Section 3, the retention data were obtained using different mobile phase compositions, with proportions (%, v/v) of methanol–aqueous buffer ranging from 75:25 to 0:100 [20]. The measured log k values were regressed against the volume fraction of organic modifier in the eluent and the obtained line was extrapolated to a hypothetical capacity factor corresponding to 0% of organic modifier (100% buffer). Since it is known that the relationship between log k and the volume fraction of organic modifier may be non-linear, the extrapolation may introduce considerable errors in the used retention data [46]. Therefore retention values measured with only one mobile phase or obtained from interpolated values [47] might lead to better predictions (less serious misclassifications).

5. Conclusion

A published chromatographic data set was studied to demonstrate the CART methodology and its potential in QSRR. The chromatographic data investigated were chosen because they show a large diversity of chemical structures determined on a given reversed-phase chromatographic system. The selection of hydrophobic (log P) and molecular size descriptors to describe and predict retention validates the methodology, i.e. these descriptors are selected as the most relevant variables out of 266 descriptors. Moreover, after removing the descriptors used in the original tree from the data set, CART was able to select analogue descriptors describing both hydrophobic and H-bonding properties, and molecular complexity. All these properties are known to be important for retention in RPLC.

The predictive properties of the methodology also seem to be promising. However, to achieve global models that describe, explain or predict chromatographic behavior in a given system even better, additional descriptors and a still more diverse set of substances with more diverse retention might be needed. In a first instance, descriptors describing the charges or the electronic properties in a molecule might be considered, since in most RPLC systems one does not work in conditions were the charge or the dissociation of the drug molecule can be ignored, as was the case for the conditions considered here (test substances were bases measured at pH 11.7).

In summary, we feel that we have demonstrated the potential of the CART methodology as a tool to understand or to select chromatographic methods.

Acknowledgements

Investigation financed with a grant from the Research Council of the Vrije Universiteit Brussel (OZR-VUB), Belgium. Y.V.H. is a post-doctoral fellow of the Fund for Scientific Research (FWO), Vlaanderen, Belgium.

References

[1] P. Jandera, Separation methods in drug synthesis and purification, in: K. Valko (Ed.), Handbook of Analytical Separations, Vol. 1, Elsevier, Amsterdam, 2000, p. 1, Chapter 1.
[2] R. Kaliszan, J. Chromatogr. B 715 (1998) 229.
[3] L.A. Lopez, S.C. Rutan, J. Chromatogr. A 965 (2002) 301.
[4] R. Kaliszan, Crit. Rev. Anal. Chem. 16 (1980) 323.
[5] R. Kaliszan, J. Chromatogr. A 656 (1993) 417.
[6] R. Kaliszan, Quantitative Structure–Chromatographic Retention Relationships, Wiley–Interscience, New York, 1987.
[7] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley–VCH, Weinheim, 2000.
[8] S. Agatonovic-Kustrin, R. Beresford, J. Pharm. Biomed. Anal. 22 (2000) 717.
[9] J.M. Sutter, T.A. Peterson, P.C. Jurs, Anal. Chim. Acta 342 (1997) 113.
[10] J.M. Zurada, Introduction to Artificial Neural Systems, West, St. Paul, MN, 1992.
[11] 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.
[12] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part B, Elsevier, Amsterdam, 1998.
[13] F. Ros, M. Pintore, J.R. Chrétien, Chemometr. Intell. Lab. 63 (2002) 15.
[14] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth, Monterey, 1984.
[15] N. Lavrač, Artif. Intell. Med. 16 (1999) 3.
[16] R.J. Marshall, J. Clin. Epidemiol. 54 (2001) 603.

[17] G. De'Ath, K.E. Fabricius, Ecology 81 (2000) 3178.
[18] D. Steinberg, P. Colla, CART: Tree-Structured Non-Parametric Data Analysis, Salford Systems, San Diego, CA, 1995.
[19] Y. Vander Heyden, S.T. Popovici, P.J. Schoenmakers, J. Chromatogr. A 957 (2002) 127.
[20] A. Nasal, A. Bucinski, L. Bober, R. Kaliszan, Int. J. Pharm. 159 (1997) 43.
[21] M. Harnisch, H.J. Möckel, G. Schulze, J. Chromatogr. 282 (1983) 315.
[22] W.M. Meylan, P.H. Howard, J. Pharm. Sci. 84 (1995) 83.
[23] SRC, interactive LogKow (KowWin) demo, http://esc.syrres.com/interkow/kowdemo.htm.
[24] R. Todeschini, V. Consonni, Dragon software version 1.1, http://www.disat.unimib.it/chm/Dragon.htm.
[25] L.B. Kier, L.H. Hall, Molecular Connectivity in Structure Activity Analysis, Research Studies Press, Letchworth, 1986.
[26] D. Bonchev, Information Theoretic Indices for Characterization of Chemical Structures, Research Studies Press, Letchworth, 1983.
[27] E.V. Kostantinova, J. Chem. Inf. Comp. Sci. 36 (1997) 54.
[28] D. Bonchev, D.H. Rouvray (Eds.), Chemical Graph Theory–Introduction and Fundamentals, Gordon and Breach, New York, 1991.
[29] N. Trinajstic, Chemical Graph Theory, CRC Press, Boca Raton, FL, 1992.
[30] G. Rücker, C. Rücker, J. Chem. Inf. Comp. Sci. 33 (1993) 683.
[31] J. Gálvez, R. Garcia, M.T. Salabert, R. Soler, J. Chem. Inf. Comp. Sci. 34 (1994) 520.
[32] P. Broto, G. Moreau, C. Vandycke, Eur. J. Med. Chem. 19 (1984) 66.
[33] P.A.P. Moran, Biometrika 37 (1950) 17.
[34] R.C. Geary, Incorp. Statist. 5 (1954) 115.
[35] G. De'Ath, Ph.D. Thesis, James Cook University, Townsville, Australia, 1999.
[36] R. Todeschini, P. Gramatica, Quant. Struct.–Act. Relatsh. 16 (1997) 120.
[37] M. Randic, G.M. Brissey, R.B. Spencer, C.L. Wilkins, Comput. Chem. 3 (1979) 5.
[38] M. Randic, MATCH 7 (1979) 5.
[39] S. Winiwarter, N.M. Bonham, F. Ax, A. Hallberg, H. Lennernäs, A. Karlén, J. Med. Chem. 41 (1998) 4939.
[40] M. Randic, Acta Chim. Slov. 45 (1998) 239.
[41] L.B. Kier, L.H. Hall, J. Pharm. Sci. 70 (1981) 583.
[42] L.B. Kier, L.H. Hall, J. Pharm. Sci. 72 (1983) 1170.
[43] D.M. Cvetkovic, I. Gutman, Croat. Chem. Acta 49 (1977) 115.
[44] J. Gálvez, R. Garcìa-Domenech, V. De Julián-Ortiz, R. Soler, J. Chem. Inf. Comput. Sci. 35 (1995) 272.
[45] G.J. Klir, T.A. Folger, Fuzzy Sets, Uncertainty and Information, Prentice-Hall, Englewood Cliffs, NJ, 1988.
[46] T. Hamoir, Y. Verlinden, D.L. Massart, J. Chromatogr. Sci. 32 (1994) 14.
[47] A. Detroyer, Y. Vander Heyden, S. Carda-Broch, M.C. Garcia-Alvarez-Coque, D.L. Massart, J. Chromatogr. A 912 (2001) 211.