Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Pharmaceutical and Biomedical Analysis 39 (2005) 91–103
JOURNAL OF PHARMACEUTICAL AND BIOMEDICAL ANALYSIS
www.elsevier.com/locate/jpba

Classification of drugs in absorption classes using the classification and regression trees (CART) methodology

E. Deconinck [a], T. Hancock [b], D. Coomans [b], D.L. Massart [a], Y. Vander Heyden [a,*]

[a] *Department of Pharmaceutical and Biomedical Analysis, Pharmaceutical Institute, Vrije Universiteit Brussel (VUB), Laarbeeklaan 103, B-1090 Brussels, Belgium*
[b] *Statistics & Intelligent Data Analysis Group, James Cook University, Townsville 4814, Australia*

Received 24 December 2004; accepted 22 March 2005
Available online 8 June 2005

Abstract

Classification and regression trees (CART) were evaluated for their potential use in a quantitative structure-activity relationship (QSAR) context. Models were build using the published absorption values for 141 drug-like molecules as response variable and over 1400 molecular descriptors as potential explanatory variables. Both the role of two- and three-dimensional descriptors and their relative importance were evaluated.

For the used dataset, CART models showed high descriptive and predictive abilities. The predictive abilities were evaluated based on both cross-validation and an external test set. Application of the variable ranking method to the models showed high importances for the *n*-octanol/water partition coefficient (log *P*) and polar surface area (PSA). This shows that CART is capable of selecting the most important descriptors, as known from the literature, for the absorption process in the intestinal tract.

Keywords: QSAR; Drug absorption; In silico prediction; CART; Variable ranking

1. Introduction

A major problem in drug discovery is that molecules positively screened for their interaction with target molecules fail to become drugs because of non-proper absorption, distribution, metabolisation, elimination and toxicity (ADMET)-properties. Therefore, different methods are being developed to screen these molecules for their ADMET-properties in an early stage of the drug development. This work considers the problem of screening for absorption properties in the gastro-intestinal tract. Different in vitro techniques, like Caco-2 membrane permeability, Parallel Artificial Membrane Permeability Assay (PAMPA) and animal tissue based methods, have been discussed by several authors [1,2] but these techniques are not very valuable in the high-throughput screening of diverse candidate drug molecules. Next to these techniques based on artificial membranes, isolated gut-segments or cultivated intestinal cells, HPLC-methods were developed, using either classical reversed-phase conditions [3], special stationary phases or special mobile phases [1,2,4,5]. A method with a special stationary phase is immobilized artificial membrane (IAM)-chromatography [1,2], which uses a stationary phase containing fosfatidylcholine groups, to mimic the lipophilic environment of the celmembranes [4]. This method considers that a molecule with a relatively high retention on the IAM-column should have a good permeability through celmembranes. It should be noted that molecules with very high retention times will dissolve in the lipophylic environment of the celmembrane and show a low permeability. It has been reported that this technique can give a classification from low to high absorption in a dataset of related molecules, but not in more diverse datasets [1]. This is because the IAM-technique

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

0731-7085/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpba.2005.03.008

only considers passive diffusion and not carrier-mediated or active transport through the celmembrane. Detroyer et al. [5] reported the use of a special mobile phase, in example a micellar liquid chromatography technique, which gives a good correlation between the retention factors and the *n*-octanol/water partition coefficient, $\log P$. Since $\log P$ is considered to be an estimate related to the partitioning over bio-membranes [6–8], the MLC technique can possibly be used to screen the absorption of molecules [9].

In drug development the application of in silico techniques is of special interest because they can be used in the very first stages. Different mathematical models were developed in the past years. These models are called "quantitative structure-activity relationships" (QSAR) and relate molecular activities, like absorption, to the structural properties of the molecule, described by molecular descriptors. An example of such a relationship is the frequently used Lipinski rule of five [6], which considers that poor absorption is more likely when a molecule answers positive to two or more of the following rules: more than five H-bond donors, more than 10 H-bond acceptors, the molecular weight is higher than 500 and the calculated $\log P$ is higher than 5. Another example is the linear free energy relationship (LFER) of Abraham et al. [10]. This relationship relates activities, like solubility, partitioning between hydrophylic and lipophylic phases, blood–brain distribution, cell permeability, and human intestinal absorption to five molecular descriptors. The relationship can be written as:

$$\mathrm{SP} = c + eE + sS + aA + bB + vV$$

where SP is the response variable or molecular activity, E the excess molar refraction, S the solute polarity/polarisability, A the solute H-bond acidity, B the solute H-bond basicity and V is the Mc Gowan characteristic molar volume, while c, e, a, b and v are regression coefficients [10]. In the literature many QSAR-models can be found using advanced regression techniques like multivariate linear regression (MLR), principal component regression (PCR), partial least squares (PLS) and neural networks (NN) [11–15].

This paper wants to introduce a new approach in QSAR. Where the models described above try to exactly predict the activities (here absorption) of molecules, we aim at predicting absorption classes, ranging from low to high absorption. A relatively new technique in QSAR, classification and regression trees (CART) [16], will be evaluated on its ability to classify molecules in absorption classes and on its predictive power. All previously mentioned models require variable selection before modeling can be started. In CART, the variable selection is part of the methodology. This means that modelling can be started with an extended set of descriptors. The CART-methodology was already used successfully in quantitative structure-retention relationships (QSRR) by Put et al. [17]. The latter relationships relate the retention of molecules on chromatographic systems to molecular descriptors [18].

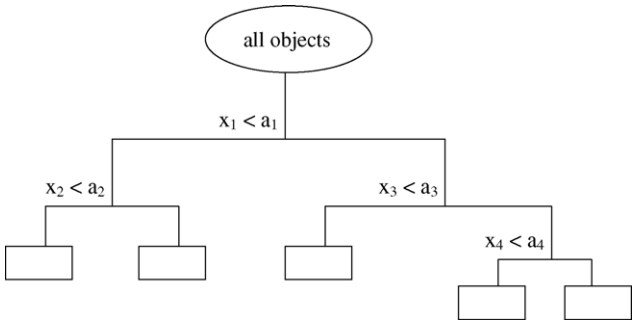

Fig. 1. General structure of a CART-model, x_i is the selected split variable and a_i is the selected split value.

2. Theory

2.1. Classification and regression trees

CART is a non-parametric statistical method, which uses a decision tree to solve classification and regression problems using both categorical and continuous variables. The method was developed by Breiman et al. [16] in order to build a decision tree which describes one response variable (univariate CART), e.g. absorption as a function of different explanatory variables, e.g. molecular descriptors (Fig. 1). When the dependent variable is categorical, CART produces a classification tree, when it is continuous it will lead to a regression tree.

A CART analysis generally consists of three steps. In a first step an overgrown tree is build, which closely describes the training set. This tree is called the maximal tree and is grown using a binary split-procedure. In a next step the overgrown tree, which shows overfitting, is pruned. During this procedure a series of less complex trees is derived from the maximal tree. In the final step, the tree with the optimal tree size is selected using a cross-validation (CV) procedure [16].

2.1.1. Building the maximal tree

The maximal tree is build using a binary split procedure, which starts at the tree-root. The tree-root consists of all objects of the training set. At each level, a mother group is considered which is split in two exclusive daughter groups. In the next step, every daughter group becomes a mother group. Every split is described by one value of one descriptor, chosen in such a way that all objects in a daughter group have more similar response variable values. The split for continuous variables is defined by "$x_i < a_j$" where x_i is the selected explanatory variable and a_j its split value.

To choose the most appropriate descriptor for a split and its split value, CART uses an algorithm in which all descriptors and all split values are considered. The split which gives the best reduction in impurity between the mother group (t_p) and the daughter groups (t_L and t_R) is selected. Mathematically this is expressed as:

$$\Delta i(s, t_\mathrm{p}) = i_\mathrm{p}(t_\mathrm{p}) - p_{\mathrm{L}i}(t_\mathrm{L}) - p_{\mathrm{R}i}(t_\mathrm{R})$$

where i is the impurity, s the candidate split value, and p_L and p_R are the fractions of the objects in the left and the right daughter group, respectively.

For regression trees, the impurity i is usually defined as the total sum of squares of the deviations of the individual responses from the mean response of the group in which the considered molecule is classified [16,17]:

$$i(t) = \sum_{n=1}^{n} (y_n - \bar{y}(t))^2$$

with $i(t)$ the impurity of group t, y_n the value of the response variable for object x_n and $\bar{y}(t)$ the mean of the response variable in group t.

This splitting procedure is repeated for each daughter group until the maximal tree is grown. The maximal tree is defined as the tree in which every end node (leaf) consists of one object, or of a predefined number of objects, or of homogeneous groups.

2.1.2. Tree-pruning

The maximal tree usually shows overfitting. As with other modeling techniques it is necessary to find a compromise between tree complexity and its predictive power [19].

During the pruning procedure a series of smaller subtrees derived from the maximal tree is obtained by successively cutting terminal branches. The different subtrees are then compared to find the optimal one. This comparison is based on a cost-complexity measure, in which both tree accuracy and complexity are considered. The cost-complexity parameter $R_\alpha(T)$ is used and for each subtree T it is defined as follow [16]:

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

with $R(T)$ the average within-node sum of squares, $|\tilde{T}|$ the tree complexity, defined as the total number of nodes of the subtree, and α is the complexity parameter, which is a penalty for each additional terminal node. During the pruning procedure, α is gradually increased from 0 to 1 and for each value of α, the tree is selected which minimizes $R_\alpha(T)$. For a value of α equal to zero, $R_\alpha(T)$ is minimized by the maximal tree. By gradually increasing α a series of trees with decreasing complexity is then obtained [16].

2.1.3. Selection of the optimal tree

From the obtained sequence of subtrees, the optimal has to be selected. The selection is based on the evaluation of the predictive error. The predictive error is often evaluated using cross-validation [16]. In cross-validation (CV) a number of objects are randomly removed from the data set, and used as a test set to evaluate the predictive power of the tree, build with the remaining data [20]. The Treeplus® module [21] for Splus® (Mathsoft, Cambridge, Massachussetts, USA) uses 10-fold CV. During 10-fold CV the dataset is divided in 10 subsets, each containing a similar distribution of the response variable. One of these subsets is then used to evaluate the predictive error of the tree build with the other nine subsets. This procedure is repeated ten times using each time another subset as test set. The most accurate tree is the one with the smallest mean CV error, defined as the root mean squared error of cross validation (RMSCEV):

$$\text{RMSECV} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{n}}$$

with y_i is the response value of object i, $\hat{y}_i$ the predicted response value for object i and n the total number of objects. The Treeplus® module identifies the optimal tree as the least complex one with an RMSECV within one standard error of the most accurate tree. The idea here is to choose a less complex tree with a predictive error comparable to that of the most accurate one [16].

2.1.4. Variable ranking

In general it can be stated that tree-structured models are easier to understand than more classical models like PLS or PCR. However, the simple structure of the final tree can be misleading. In fact tree-models are unstable and small changes in the original dataset can lead to the selection of another variable to create a split. To facilitate the interpretation, CART allows evaluating the importance of the different explanatory variables to define a split or more generally to describe the response variable in the selected dataset. The technique used for this purpose is called the "variable ranking method" [16].

A variable which is not selected in the final tree structure could be considered as less important in describing the dataset, but it is possible that this variable is masked. It has been noted, that although a variable x_1 never occurred in the final tree structure, it could be an important variable in a new tree, which is almost as accurate as the original one and which is build after removing the masking variable x_2. In such a situation the variable ranking method is capable to detect the importance of the variable x_1.

The importance of a variable x_m is defined as:

$$M(x_m) = \sum_{t \in T} \Delta I(\tilde{s}_m, t)$$

with $\Delta I(\tilde{s}_m, t) = \max\ \Delta I_{C_1}(s_m, t)$, which equals the maximal decrease in node impurity for the division of a parent node t into daughter nodes C_1 and C_2 guided by a surrogate split $\tilde{s}_m$. This maximal decrease in node impurity is summed for each node t of the optimal subtree T to obtain the importance of a variable. The procedure is repeated for each of the used descriptive variables. A surrogate split is defined by a surrogate variable. This variable is the next most important variable, following the selected variable. This variable gives a similar split of the mother group into daughter groups and gives the second best reduction in impurity of the mother group into the daughter groups. Surrogate splits can also be used to solve the problem of missing values. In this situation, CART will use the surrogate variable to define in which daughter

group an object, with missing primary variable value, will be classified. The importance values allow to rank the different explanatory variables from high to low importance. In this way, the most important variables to describe the response can be identified and CART can be used for feature selection [16].

2.2. Molecular descriptors

A molecular descriptor is the final result of a logical and mathematical procedure, which converts chemical information from a symbolic representation of the molecule into a useful numeric value (theoretical descriptor) or is the result of a standardized experiment (experimental descriptor) [22].

There are different ways by which molecular descriptors are classified. The simplest way is by their origin, i.e. in theoretical and experimental descriptors. The theoretical descriptors can be further classified depending on the molecular representation they are derived from. The simplest representation is the molecular formula. Descriptors derived from it are called zero-dimensional (0D) (e.g. molecular weight, atom-counts, . . .). One-dimensional (1D) descriptors are derived from a substructure list representation of the molecule (e.g. $\log P$ calculated with the method of Rekker [23]). Two-dimensional (2D) (e.g. connectivity indices [22]) and three-dimensional (3D) (e.g. the molecular volume and different geometrical and steric descriptors [22]) descriptors are calculated from a topological and a geometrical molecule representation, respectively. Finally, the descriptors derived from a stereo-electronic or lattice representation, are called four-dimensional (4D). The different descriptors within these classes often are further classified in subclasses. More information can be found in ref. [22].

3. Materials and methods

3.1. Data

The data set used consisted of 141 molecules extracted from Zhao et al. [12]. The drug and drug-like compounds and their percent human intestinal absorption (%HIA) are listed in Table 1. These molecules were selected because they represent absorption data for a high diversity of structures and they cover the whole range of the absorption scale (0–100%)

3.2. Three-dimensional molecular structures

The 3D structures of the molecules were calculated using the Hyperchem® 6.03 professional software (Hypercube, Gainesville, Florida, USA). After the input of the molecule as a topological structure, geometry optimisation was performed by the Molecular Mechanics Force Field method (MM+) using the Polak-Ribière conjugate gradient algorithm with a RMS gradient of 0.1 kcal/(Å mol) as stop criterion. This computational optimisation of the structure results in

Table 1
The absorption data for the 141 molecules (extracted from [12])

Number	Substance	%HIA
1	Acarbose	1.5
2	Acebutolol	89.75
3	Acetaminophen	85
4	Acetylsalicylic acid	100
5	Acrivastine	88
6	Acyclovir	25
7	Adefovir	12
8	Alprenolol	93.75
9	Aminopyrine	100
10	Amoxicillin	93.75
11	Amphotericin B	5
12	Amrinone	93
13	Antipyrine	100
14	Atenolol	51
15	Atropine	90
16	Azithromycin	36
17	Aztreonam	1
20	Benazepril	37
19	Benzylpenicillin	27.5
20	Betaxolol	90
21	Bornaprine	100
22	Bretyliumtosylate	23
23	Bromazepam	84
24	Bromocriptine	28
25	Bumetanide	100
26	Bupropion	87
27	Caffeine	100
28	Camazepam	99
29	Captopril	68
30	Cefatrezine	76
31	Ceftriaxone	1
32	Cefuroxime	5
33	Cefuroximeaxetil	36
34	Cephalexin	98.5
35	Chloramphenicol	90
36	Chlorothiazide	23.75
37	Cimetidine	82.5
38	Ciprofloxacin	84.5
39	Cisapride	100
40	Clonidine	96.25
41	Codein	95
42	Corticosterone	100
43	Cromolynsodium	0.5
44	Cymarin	47
45	Cyproterone acetate	100
46	Dexamethasone	98
47	Diazepam	99.25
48	Doxorubicin	5
49	Enalapril	66
50	Enalaprilat	17.5
51	Erythromycin	35
52	Ethambutol	77.5
53	Ethinylestradiol	100
54	Etoposide	50
55	Felbamate	92.5
56	Fenoterol	60
57	Fluconazole	96.25
58	Foscarnet	17
59	Fosinopril	36
60	Fosmidomycin	30
61	Furosemide	61
62	Gabapentin	50
63	Ganciclovir	3.6

Table 1 (*Continued*)

Number	Substance	%HIA
64	Gentamicin-C1	0
65	Guanabenz	75
66	Guanoxan	50
67	Hydrochlorothiazide	72.75
68	Hydrocortisone	90.25
69	Imipramine	96.25
70	Indomethacin	100
71	Iothalamatesodium	1.9
72	Isoxicam	100
73	Isradipine	92.5
74	Labetalol	93.75
75	Lactulose	0.6
76	Lamotrigine	70
77	Levodopa	85
78	Lincomycin	27.5
79	Lisinopril	25
80	Loracarbef	100
81	Lormetazepam	100
82	Lovastatin	30.5
83	Mannitol	20
84	Meloxicam	90
85	Metaproterenol	44
86	Methotrexate	80
87	Methyldopa	41
88	Methylprednisolone	82
89	Metolazone	63
90	Metoprolol	95
91	Morphine	100
92	Nadolol	31
93	Nefazodone	100
94	Naloxone	91
95	Nordiazepam	99
96	Norfloxacin	35
97	Olsalazine	2.3
98	Ouabain	1.4
99	Oxatomide	100
100	Oxazepam	98.5
101	Oxprenolol	91.75
102	Phenoxymethylpenicillin	45
103	Phenytoin	90
104	Pindolol	91.75
105	Piroxicam	100
106	Practolol	98.75
107	Pravastatin	34
108	Prazosin	100
109	Prednisolone	98.9
110	Progesterone	93.25
111	Propranolol	92.5
112	Propiverine	84
113	Propylthiouracil	75
114	Quinidine	80.25
115	Raffinose	0.3
116	Ranitidine	52.75
117	Reproterol	60
120	Saccharin	88
119	Salicylic acid	100
120	Scopolamine	92.5
121	Sorivudine	82
122	Sotalol	96.25
123	Spironolactone	73
124	Sudoxicam	100
125	Sulfasalazine	38.75
126	Sulindac	90
127	Sulpiride	36

Table 1 (*Continued*)

Number	Substance	%HIA
128	Sumatriptin	70
129	Terazosin	93.25
130	Terbutaline	66.5
131	Testosterone	100
132	Theophylline	96
133	Timolol maleate	85.5
134	Tranexamicacid	55
135	Trimethoprim	97
136	Trovoflaxicin	88
137	Venlafaxine	92
138	Verapamil	95
139	Warfarin	98.5
140	Ximoprofen	100
141	Zidovudine	100

a data matrix consisting of the Cartesian coordinates of the atoms. This data matrix is then used to calculate the molecular descriptors.

3.3. Calculating molecular descriptors

The majority of the molecular descriptors were calculated with Dragon® 4.0 academic version software [24]. It allows calculating 48 constitutional descriptors, 119 topological descriptors, 47 walk and path counts, 33 connectivity indices, 47 information indices, 96 2D autocorrelations, 107 edge adjacency indices, 64 BCUT-descriptors, 21 topological charge indices, 44 eigenvalue-based indices, 41 Randic molecular profiles, 74 geometrical descriptors, 150 RDF descriptors, 160 3D-MoRSE descriptors, 99 WHIM descriptors, 197 GETAWAY descriptors, 121 functional group counts, 120 atom-centered fragments, 14 charge descriptors and 10 molecular properties. The software automatically eliminates constant variables in a given data set. For descriptors with a correlation higher than 0.98, parameters are set that only one is retained in the dataset. For more information about the above descriptors we refer to ref. [22]. The Hyperchem® 6.03 professional software and the ACD-Labs® 6.0 software (Advanced Chemistry Development, Toronto, Ont., Canada) were used to calculate the following additional descriptors: solvent accessible surface area, molecular volume, hydration energy, molar refractivity, molar polarisability, molar mass, parachor index, tension surface, density and the acidity constants. Further the Mc Gowans molecular volume, a parameter applied by Abraham et al. in the linear free energy relationship [5,25], was calculated manually according to Todeschini and Consonni [22].

3.4. Building tree models

The tree models were build using the Treeplus® module [21] in the Splus® software (Mathsoft, Cambridge, Massachussetts, USA). The absorption data were used as response variables and the different molecular descriptors

as explanatory variables. Since the response variable is continuous the resulting tree models are called regression trees.

4. Results and discussion

4.1. Building tree-models

The models are build using the absorption data of all 141 molecules. During the building process the maximal tree is build and pruned. In the next step, 10-fold CV is carried out resulting in a graph of the RMSECV as a function of the tree complexity (Fig. 2). This graph allows selection of the most suited tree. The best tree (with minimal RMSECV) selected by the program has a complexity of three leafs. The leafs in this tree (Fig. 3) cover a wide part of the absorption range, particularly class c. This means that the obtained model cannot be used for our purpose, i.e. the definition of classes with a limited absorption range. Since the response variable is continuous, a regression tree is obtained and the impurity of the nodes is evaluated by the RMSECV, which in fact is the sum of the squared deviation of the absorption values of the molecules in the classes to the mean absorption value of that class. This means that the RMSECV is a good evaluation criterion for regression but not for classification, since an object can be classified correctly but still have a quite high deviation of the mean of the class in which it is predicted. Therefore, it was decided to look at more complex trees, which can be evaluated starting from the same graph (Fig. 2). Each tree was evaluated visually starting with the smallest. As a general rule the smallest tree was selected: (i) in which each leaf represents less than 50% of the absorption range; (ii) where the number of objects in a class is greater than 5 after the outliers, evaluated with the Grubbs test [19], were removed and (iii) the total number of outliers in the model is less than 5% of all molecules used in the model building. A leaf defined by less than five objects is considered as undefined. After selection of the model the absorption range of each class was defined by the lowest and highest value in the class after removing the outliers. Based on the ranges the different absorption classes of the dataset were labelled as follows: class 1, 0–25%; class 2, 26–50%; class 3, 51–70%; class 4, 71–90% and class 5, more than 90% human intestinal absorption. For model building, it was allowed that the range of a leaf covers two consecutive classes. The leaf is then labelled with both class numbers. If the range covers more than two classes, labelling was based on the two most represented consecutive classes.

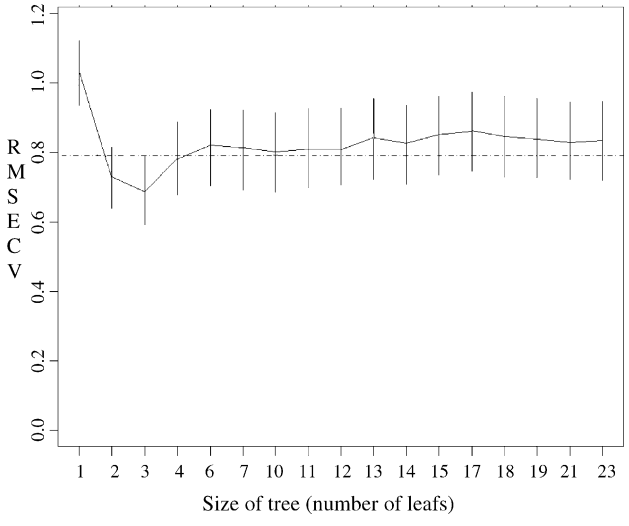

Fig. 2. The tree complexity as a function of the root mean squared error of cross validation (RMSECV).Vertical lines represent the standard deviation around the RMSECV values for the different tree-complexities. The horizontal line represent the minimal RMSECV-value +1 standard error.

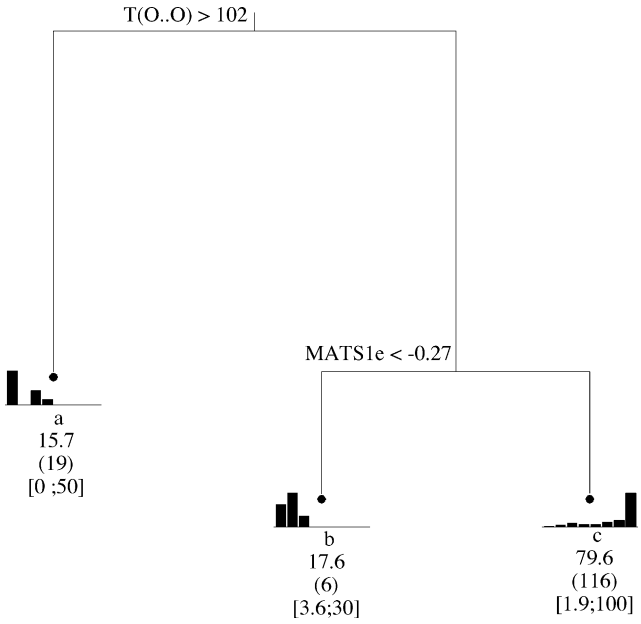

Fig. 3. Optimal regression tree for the absorption values of the 141 molecules, using all calculated descriptors. Classes (a–c) are identified by the mean of the absorption values of its objects, the number of objects and the absorption range of the class, respectivily.

Models were build using three different descriptor sets. The first consists of all calculated descriptors. The second of the 2D descriptors (constitutional descriptors, topological descriptors, walk and path counts, connectivity indices, information indices, 2D autocorrelations, edge adjacence indices, BCUT-descriptors, topological charge indices, eigenvalue based indices, functional group counts and atom centered fragments) and the molecular properties calculated with the Dragon® software, the parameters calculated with Hyperchem® and ACD-Labs® as well as the Mc Gowans volume. The third descriptor set consists of the 3D descriptors (randic molecular profiles, geometrical descriptors, RDF-descriptors, 3D-MoRSE descriptors, WHIM descriptors, GETAWAY descriptors, charge descriptors) and the molecular properties calculated with Dragon®, the parameters calculated with Hyperchem® and ACD-Labs®, and the Mc Gowans volume.

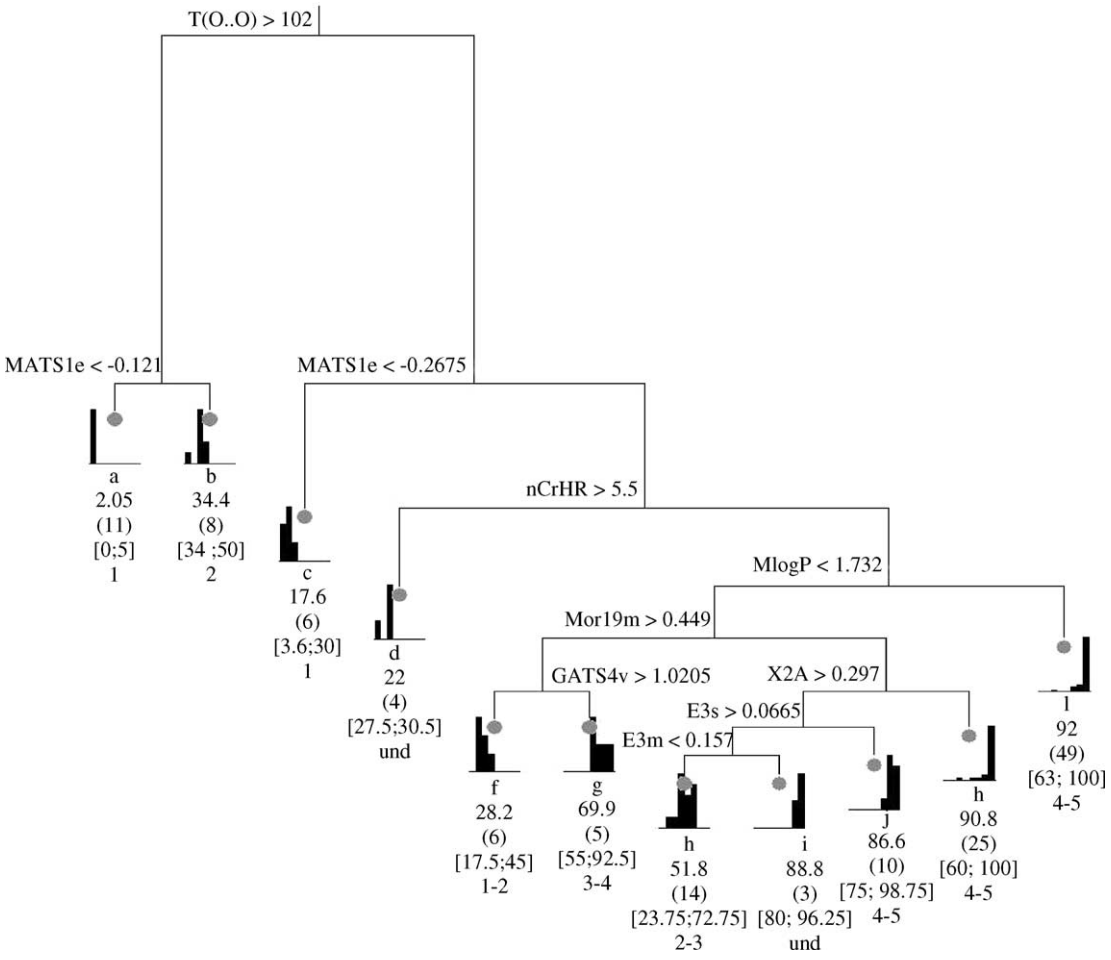

Fig. 4. Model 1 for the dataset of 141 molecules using all descriptors (first descriptor set) Classes (a–l) are identified by the mean of the absorption values of its objects, the number of objects, the absorption range of the class and the class label respectivily.

For the first descriptor set a model with complexity 11 is selected (model 1, Fig. 4). All classes represent less than 50% of the absorption range, only two undefined classes (classes defined by less than five molecules), d and i, are present and only three outliers (aztreonam, iothalamate sodium and norfloxacin) could be detected. Labelling of the different classes was carried out as shown in Fig. 4.

For the second descriptor set also a model with 11 leaves is selected (model 2, Fig. 5). Ten classes cover less than half the absorption range, class g covers a little more (54%). Still the model was selected as a compromise between the classification and the complexity of the tree. Selection of a more complex tree in which class g is split into two smaller classes results in a high number of undefined classes, resulting in a loss of information. Four undefined classes are present in the model (classes d, f, h, and i) and four outliers (aztreonam, iothalamate sodium, reproterol and benazepril) could be detected. Labelling of the classes was carried out as shown in Fig. 5.

The third descriptor set resulted in the selection of a model with complexity 10 (model 3, Fig. 6). All classes except class g cover less than 50% of the absorption range. The model was selected based on the same arguments as for the previous model. Only two undefined classes are present in the model (classes d and f) and two outliers (norfloxacin and gabapentin) could be detected. Labeling of the classes was carried out as shown in Fig. 6.

Comparison of the three models shows that the best model is obtained with all descriptors, since no class covers more than half of the absorption range and the number of undefined classes and outliers is lowest. In the second model it was necessary to come to a compromise between complexity and classification, also the number of undefined classes and outliers is higher. Even though a less complex model was obtained with the third descriptor set, also a compromise had to be found, while the number of undefined classes and outliers is comparable with the first model. This is a first indication of the fact that the combination of 2D and 3D descriptors in CART can be very valuable (see also Section 4.2).

4.2. The selected descriptors

In model 1 (Fig. 4) the first split is defined by the topological descriptor T(O . . O), which represents the sum of

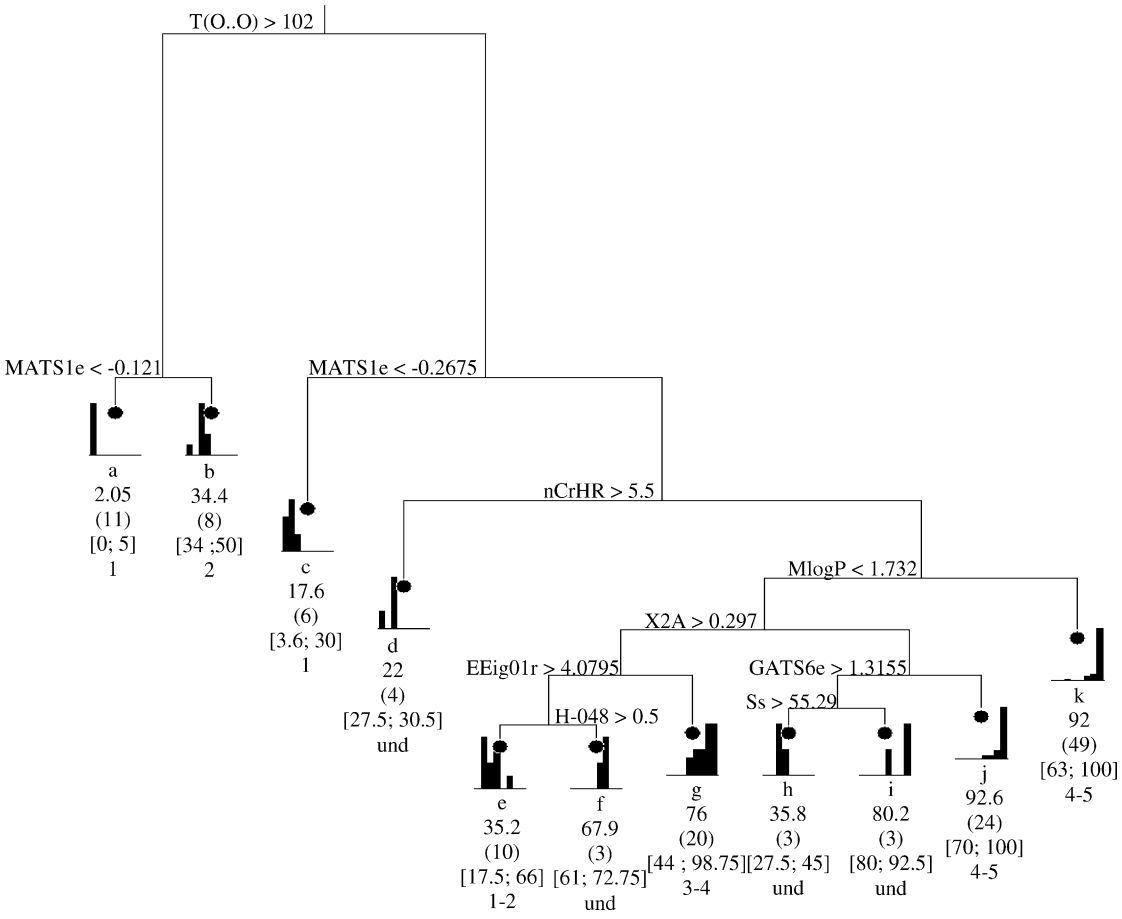

Fig. 5. Model 2 for the dataset of 141 molecules using the second descriptor set. Classes (a–l) are identified as in Fig. 4.

topological distances between oxygen atoms [22]. For the rest of the tree the importance of the 2D-autocorrelation descriptors can be noted (Moran autocorrelation-lag 1/weighted by atomic electronegativities (MATS1e) and Geary autocorrelation-lag 4/weighted by atomic van der Waals volumes (GATS4v)). Further, the number of tertiary C(sp^3) (*n*CrHR) and the average connectivity index chi-2 (X2A) [22] are selected. Interesting is the selection of the *n*-octanol/water partition coefficient log *P*. log *P* is a measure for the hydrofobicity of a molecule and is considered as one of the most important properties of molecules for their passage through biomembranes [6–8]. From the 3D descriptors the 3D-MoRSE descriptor Mor19 m (3D-MoRSE signal 19/weighted by atomic masses) is selected as well as the WHIM descriptors E3s (third component accessibility directional WHIM index/weighted by atomic electrotopological states) and E3m (third component accessibility directional WHIM index/weighted by atomic masses).

In the second model (Fig. 5) the first five splits are identical to these in model 1. The selected descriptors are T(O . . O), MATS1e, *n*CrHR, log *P*, X2A, the Edge Adjacency Index EEig01r (Eigenvalue 01 from edge adjacency matrix weighted by resonance integrals), the 2D-autocorrelation descriptor GATS6e (Geary autocorrelation-lag 6/weighted by atomic Sanderson electronegativities), the atom-centered fragment H-048 (H attached to C2(sp3)/C1(sp2/C0(sp)) and the constitutional descriptor Ss (sum of Kier-Hall electrotopological states).

In the third model (Fig. 6), the selected descriptors are the geometrical descriptors geometrical distance between oxygen atoms (G(O . . O)) and gravitational index G1 (G1), the surface tension, log *P* and the polar surface area (PSA) calculated with ACD-Labs®, the GETAWAY-descriptor R6v (R autocorrelation of lag 6/weighted by atomic van der Waals volumes), the WHIM-descriptor L3s (third component size directional WHIM index/weighted by atomic electrotopological states) and the 3D-MoRSE descriptor Mor24u (3D-MoRSE signal 24/unweighted). More information about these descriptors can be found in ref. [22].

In the previous section, the classification is found worse in models 2 and 3 than in the first model. In the first model, the first five splits are defined by 2D descriptors. The rough classification by these descriptors is then refined by 3D descriptors. This conclusion is confirmed by the fact that in the second model the first five splits are identical to model 1. Replacement of the 3D descriptors by 2D, in model 2, results in less significant splits, resulting in more outliers and broader classes. Therefore, it seems that 3D descriptors

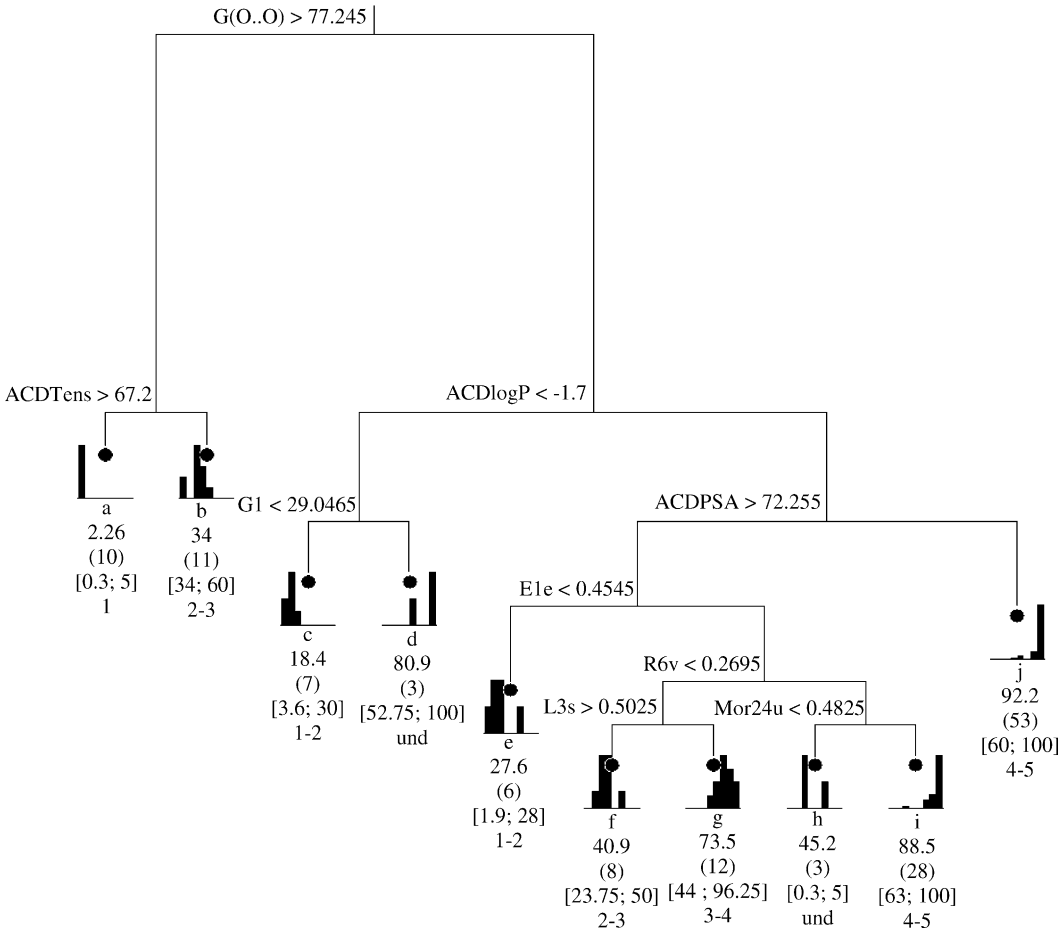

Fig. 6. Model 3 for the dataset of 141 molecules the third descriptor set. Classes (a–j) are identified as in Fig. 4.

can add significant information to the model. If only 3D descriptors are used, as in model 3, also a less good splitting of the molecules is obtained. The above indicates that the combination of the two types of descriptors gives the best description of the used dataset.

4.3. Predictive power

To evaluate the predictive power of the obtained models another subset of the Zhao dataset [12] is used as test set. This test set consists of the HIA-values of 27 molecules (Table 2). It can be noticed that the test set contains only substances with a high %HIA value. Better would have been to be able to create one containing substances for all classes. However, the selection of this test set can be justified as follows. In practice, HIA values are mainly reported for molecules with a high HIA value, because those with a low value are already eliminated during earlier drug molecule screening procedures. Measuring the HIA values is slow, expensive and time-consuming. Therefore, the few molecules with low HIA values in the dataset, were included in the training set. For the molecules of the test set the relevant descriptors were calculated. Based on their HIA-value the molecules were labelled as belonging to one of the five absorption classes defined in previous sections. The calculated descriptors were then used to assign the molecules to one of the classes (nodes) in the different CART models. A molecule was considered correctly classified if the label of the molecule corresponded to the leaf in which it is classified by the model. The results for model 1 are 24 molecules correctly classified (88.9%) and three misclassified (glycine, granisetron, tolmesoxide) (11.1%); for model 2: 23 molecules correctly classified (85.2%), three molecules misclassified (glycine, tolmesoxide, viloxazine) (11.1%) and one molecule classified in an undefined class (granisetron) (3.7%), and for model 3: 21 molecules correctly classified (77.8%) and six misclassified (carfecillin, cicaprost, disulfiram, fluvastatin, gallopamil, sultopride) (22.2%). These results show that the three models have a high predictive power, with model 1 slightly better than the two others. Since the used test set consists mainly of molecules with high absorption values (class 5), the prediction of this test set is not representative for the prediction over the full absorption range. The selection of another test set covering the whole absorption range is not possible. This is due to the fact that too few objects with low absorption are present

Table 2
External test set (extracted from [12])

Number	Substance	%HIA
1	Carfecillin	100
2	Cicaprost	100
3	Clofibrate	96
4	Desipramine	98.75
5	Diclofenac	100
6	Disulfiram	91
7	Felodipine	100
8	Fenclofenac	100
9	Fluvastatin	97.5
10	Gallopamil	100
11	Glycine	100
12	Granisetron	100
13	Ibuprofen	100
14	Ketoprofen	100
15	Ketorolac	100
16	Mexiletine	100
17	Minoxidil	95
20	Naproxen	97.75
19	Nicotine	100
20	Nizatidine	99
21	Ondansetron	100
22	Phenglutarimide	100
23	Praziquantel	100
24	Sultopride	100
25	Tolmesoxide	100
26	Valproic acid	100
27	Viloxazine	100

in the dataset. Deleting these objects would result in bad modeling in the lower part of the absorption range. Therefore, it was decided to carry out also a manual 10-fold CV. The difference with the CV included in the CART methodology is that the misclassification rate was not evaluated by the RMSECV but by the mean number of correctly and misclassified molecules. The dataset of 141 molecules was divided in 10 equal parts of 14 molecules. Each part contains exclusive molecules with absorption values representative for the complete dataset. One part is used to evaluated the predictive abilities of the model build with the other nine parts. This procedure is repeated 10 times, in which each part is used once as test set. The trees were built according to the higher specified criteria. This resulted in trees very similar to those in Figs. 4–6. The prediction results for all molecules from the 10 trees are as follows: model 1: 92 objects correctly classified, 38 objects misclassified and 10 objects misclassified; model 2: 88 objects correctly classified, 39 objects misclassified and 13 objects undefined; model 3: 88 objects correctly classified, 40 objects misclassified and 12 objects undefined. These results show again that the three types of models have a good predictive power with slightly better results for the model using all descriptors.

In general, it can be concluded that the model built with all descriptors shows the best results in descriptive as well as predictive power.

Table 3
The 50 selected molecular descriptors by the variable ranking method. Their abbreviation, name and class

Abbreviation	Name	Class
log *P*	*n*-Octanol/water partition coefficient	Molecular properties
PSA	Polar surface area	Molecular properties
HyHydrat. E	Hydration energy	QSAR-properties calculated with Hyperchem®
R1e	R-autocorrelation of lag 1/weighted by atomic Sanderson electronegativities	GETAWAY descriptors
T(O . . O)	Sum of topological distances between oxygen atoms	Topological descriptors
H-050	H attached to heteroatom	Atom-centered fragments
SEige	Eigenvalue sum from electronegativity weighted distance matrix	Eigenvalue based indices
GATS2v	Geary autocorrelation-lag 2/weighted by atomic van der Waals volumes	2D-autocorrelations
RDF060m	Radial distribution function-6.0/weighted by atomic masses	RDF-descriptors
ACDDens	Density	Macroscopic properties calculated with ACD-labs®
SEigv	Eigenvalue sum from van der Waals weighted distance matrix	Eigenvalue based indices
nO	Number of oxygen atoms	Constitutional descriptors
DELS	Molecular electrotopological variation	Topological descriptors
nCrHR	Number of ring tertiary C(sp3)	Functional group counts
BAC	Balaban centric index	Topological descriptors
GATS2p	Geary autocorrelation-lag 2/weighted by atomic polarizabilities	2D-autocorrelations
IC5	Information content index (neighborhood symmetry of 5-order)	Information indices
MATS1e	Moran autocorrelation-lag 1/weighted by atomic Sanderson electronegativities	2D-autocorrelations
ACDTens	Surface tension	Macroscopic properties calculated with ACD-labs®
RDF065m	Radial distribution function-6.5/weighted by atomic masses	RDF-descriptors
HNar	Narumi harmonic topological index	Topological descriptors
PCR	Ratio of multiple path count over path count	Walk and path counts
SEigZ	Eigenvalue sum from *Z* weighted distance matrix (Barysz matrix)	Eigenvalue-based descriptors
Ms	Mean electrotopological state	Constitutional descriptors
*A*log *P*	Ghose-Crippen octanol–water partition coefficient	Molecular properties
Mor10m	3D-MoRSE signal 10/weighted by atomic masses	3d-MoRSE descriptors
RDF020v	Radial Distribution Function-2.0/weighted by atomic van der Waals volumes	RDF-descriptors
H3u	H autocorrelation of lag 3/unweighted	GETAWAY descriptors

Table 3 (*Continued*)

Abbreviation	Name	Class
RTm	R total index/weighted by atomic masses	GETAWAY-descriptors
ATS1m	Broto-Moreau autocorrelation of a topological structure-lag 1/weighted by atomic masses	2D-autocorrelations
RDF020m	Radial distribution function-2.0/weighted by atomic masses	RDF descriptors
RDF045m	Radial distribution function-4.5/weighted by atomic masses	RDF-descriptors
Mor16m	3D-MoRSE signal 16/weighted by atomic masses	3d-MoRSE descriptors
RDF010m	Radial distribution function-1.0/weighted by atomic masses	RDF descriptors
Mor05m	3D-MoRSE signal 05/weighted by atomic masses	3d-MoRSE descriptors
BLTD48	Verhaar model of Daphnia base-line toxicity from MlogP (mmol/l)	Molecular properties
Ss	Sum of Kier-Hall electrotopological states	Constitutional descriptors
ATS8m	Broto-Moreau autocorrelation of a topological structure-lag 8: weighted by atomic masses	2D-autocorrelations
GGI9	Topological charge index of order 9	Topological charge indices
RDF080m	Radial distribution function-8.0/weighted by atomic masses	RDF descriptors
MAXDN	Maximal electrotopological negative variation	Topological descriptors
RDF050m	Radial distribution function-5.0/weighted by atomic masses	RDF descriptors
RDF035m	Radial distribution function-3.5/weighted by atomic masses	RDF descriptors
RDF075m	Radial distribution function-7.5/weighted by atomic masses	RDF-descriptors
PW4	Path/walk 4-Randic shape index	Topological descriptors
J3D	3D-Balaban index	Geometrical descriptors
Mor03u	3D-MoRSE signal 03/unweighted	3d-MoRSE descriptors
RDF040p	Radial distribution function-4.0/weighted by atomic polarizabilities	RDF descriptors
Mor01m	3D-MoRSE signal 01/weighted by atomic masses	3d-MoRSE descriptors
ATS7m	Broto-Moreau autocorrelation of a topological structure-lag 7: weighted by atomic masses	2D-autocorrelations

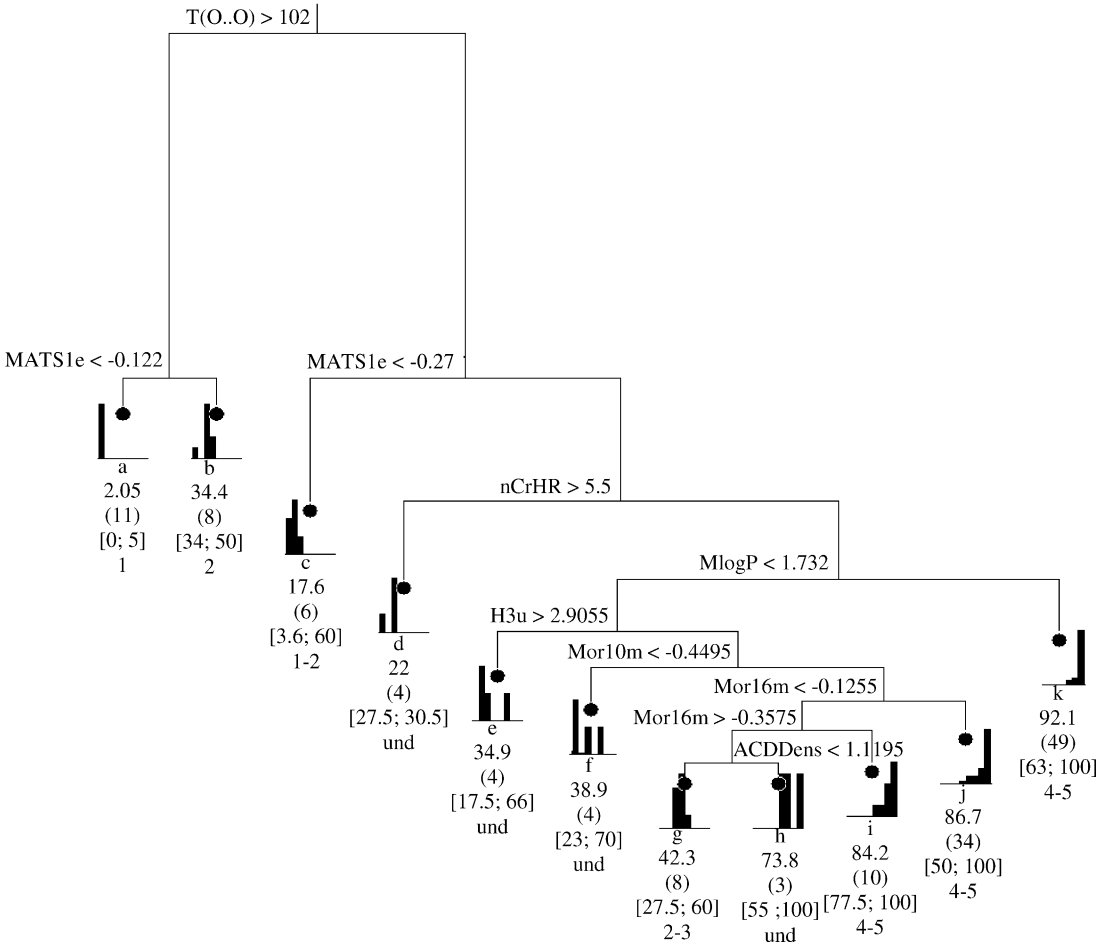

Fig. 7. Model for the dataset of 141 molecules using the 50 descriptors selected through variable ranking. Classes (a–j) are identified as in Fig. 4 (descriptors, see Table 3).

4.4. Variable ranking

In this section the possibility of CART as variable selection method was evaluated. Therefore, the previous build model using all descriptors (model 1) was used as starting point. The variable ranking method was applied to this model. The fifty most important descriptors (Table 3) were used to select a new descriptor set.

In Table 3, the descriptors are ranked in descending order of importance. The two most important descriptors selected by the method are log *P* and the polar surface area (PSA). The importance of log *P* in absorption processes was already mentioned higher [6–8]. The PSA is defined as the part of the surface area of the molecule associated with oxygen, nitrogens, sulfurs and the hydrogens bonded to any of these atoms [22,26]. It is a measure for the H-bonding capacity of a molecule. It has been found that processes involving passive diffusion depend primarily on these H-bonding properties [10]. This shows that CART is capable to select the descriptors, mentioned in the literature to be important in absorption processes.

The thus obtained descriptor set was used to build a new model (Fig. 7), according to the above mentioned rules. A model with complexity 11 was selected. All classes represent less than 50% of the absorption range, four undefined classes (classes d, h, e and f) are present and three outliers (aztreonam, iothalamate sodium and benazepril) could be detected with the Grubbs test [19]. Labelling of the classes was carried out as shown in Fig. 7.

To evaluate the predictive power of this model the test set of 27 molecules was predicted. Twenty-three molecules were correctly classified, one was misclassified (glycin) and three molecules were predicted in undefined classes (granisetron, nizatidine and viloxazine). In analogy with previous sections the manual cross-validation was carried out. The mean results are: 95 molecules correctly classified, 37 molecules misclassified and eight molecules predicted in an undefined class.

Comparing these results with those obtained for the model using all descriptors (model 1), shows that both models perform very comparably. The model obtained after variable selection has the same complexity as model 1 but has two undefined classes more. The results for the test set and the CV can be considered identical for both models.

It can be concluded that the variable ranking method in CART is capable to select the descriptors most useful in describing a given data set. Models based on such variable selection show no significant loss of information.

5. Conclusions

From the three types of models build, it can be concluded that the best results for description of the training set and prediction of the test sets are obtained with models combining 2D and 3D descriptors. The 3D descriptors can add valuable information to the models, which is logical since the geometrical properties of a molecule play a major part in the process of membrane passage. In the models based on all descriptors (model 1 and related models obtained during CV) it is observed that the first splits are always defined by 2D descriptors or molecular properties. The 3D descriptors usually define the latter splits in the models. This indicates that 3D descriptors give a refinement of the model, resulting in better descriptions of datasets and more accurate predictions of test sets.

It is demonstrated that CART can also be used as a variable selection method, resulting in models almost without loss of information.

Generally we can conclude that CART can be a useful tool in QSAR studies. It is capable of selecting the most important descriptors out of hundreds of descriptors and of giving a close description of the used datasets.

Acknowledgment

This research is financed with a specialisation grant from the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT).

References

[1] I.J. Hidalgo, Curr. Top. Med. Chem. 1 (2001) 385–401.
[2] P.V. Balimane, S. Chong, R.A. Morrison, J. Pharmacol. Toxicol. Methods 44 (2000) 301–312.
[3] A. Nasal, A. Bucinski, L. Bober, R. Kaliszan, Int. J. Pharm. 159 (1997) 43–55.
[4] R. Kaliszan, J. Chromatogr. A 656 (1993) 417–435.
[5] A. Detroyer, Y. Vander Heyden, S. Carda-Broch, M.C. García-Alvarez-Coque, D.L. Massart, J. Chromatogr. A 912 (2001) 211–221.
[6] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Adv. Drug Delivery Rev. 46 (2001) 3–26.
[7] S.K. Poole, C.F. Poole, J. Chromatogr. B 797 (2003) 3–19.
[8] S. Yang, J.F. Bumgarner, L.F.R. Kruk, M.G. Khaledi, J. Chromatogr. A 721 (1996) 323–335.
[9] A. Detroyer, S. Stokbroekx, H. Bohets, W. Lorreyne, P. Timmermans, P. Verboven, D.L. Massart, Y. Vander Heyden, Anal. Chem. 76 (2004) 7304–7309.
[10] M.H. Abraham, A. Ibrahim, A.M. Zissimos, Y.H. Zhao, J. Corner, D.P. Reynolds, Drug Discov. Today 7 (2002) 1056–1063.
[11] S. Agatonovic-Kustrin, R. Beresford, A. Pausi, M. Yusof, J. Pharm. Biomed. Anal. 25 (2001) 227–237.
[12] Y.H. Zhao, J. Le, M.H. Abraham, A. Hersey, P.J. Eddershaw, C.N. Luscombe, D. Boutina, G. Beck, B. Sherborne, I. Cooper, J.A. Platts, J. Pharm. Sci. 90 (2001) 749–784.
[13] Q. Shen, Q.Z. Lü, J.H. Jiang, G.L. Shen, R.Q. Yu, Eur. J. Pharm. Sci. 20 (2003) 63–71.
[14] V. Karalis, A. Tsantili-Kakoulidou, P. Macheras, Eur. J. Pharm. Sci. 20 (2003) 115–123.
[15] J.A. Platts, M.H. Abraham, Y.H. Zhao, A. Hersey, L. Ijaz, D. Butina, Eur. J. Med. Chem. 36 (2001) 719–730.
[16] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth, Monterey, 1984.
[17] R. Put, C. Perrin, F. Questier, D. Coomans, D.L. Massart, Y. Vander Heyden, J. Chromatogr. A 988 (2003) 261–276.

[18] R. Kaliszan, M.A. Van Straten, M. Markuszewski, C.A. Cramers, H.A. Claessens, J. Chromatogr. A 855 (1999) 455–486.
[19] 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 Science, Amsterdam, 1997.
[20] 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 Science, Amsterdam, 1997.
[21] G. De'Ath, New statistical methods for modeling species-environment relationships, Ph.D. thesis, James Cook University, Townsville, Australia, 1999.
[22] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, WILEY-UCH, Weinheim, 2000.
[23] R.F. Rekker, The Hydrofobic Fragmental Constant, Pharmacochemistry Library, vol. 1, Elsevier, New York, 1977.
[24] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, Dragon® academic version, software version 4.0 (2003). Milano Chemometrics and QSAR Research Group, Copyright Talete srl© 1997–2003.
[25] M.H. Abraham, H.S. Chadha, R.A.E. Leitao, R.C. Mitchell, W.J. Lambert, R. Kaliszan, A. Nasal, P. Haber, J. Chromatogr. A 766 (1997) 35–47.
[26] D.E. Clark, J. Pharm. Sci. 88 (1999) 807–814.