



ELSEVIER

Journal of Chromatography A, 906 (2001) 443–458

JOURNAL OF  
CHROMATOGRAPHY A

www.elsevier.com/locate/chroma

## Review

# Reviewing mobile phases used on Chiralcel OD through an application of data mining tools to CHIRBASE database

Patrick Piras, Christian Roussel\*, Johanna Pierrot-Sanders

Université Aix-Marseille III, ENSSPICAM, CNRS-UMR6516, Avenue Escadrille Normandie-Niemen, 13397 Marseille Cedex 20, France

### Abstract

During the past decade, thousands of compounds have been resolved on Chiralcel OD (a cellulose-based chiral stationary phase) under diverse eluting conditions. Many researches have documented the effects of mobile phase on enantioselectivity for a given family of samples but today no comprehensive study aimed at identifying the associations between the structural features present on solute and appropriate mobile phase conditions has yet been proposed. In this review of mobile phases used on Chiralcel OD, we try to go far beyond a simple enumeration of eluting conditions and an effort is made to explore the utility of data mining tools for assessing the knowledge contained in CHIRBASE database. We have extracted from CHIRBASE the chemical features of 2363 chiral compounds separated on Chiralcel OD and their corresponding mobile phases. This data set was submitted to data mining programs for molecular pattern recognition and mobile phase predictions for new cases. Some substructural characteristics of solutes were related to the efficient use of some specific mobile phases. For example, the application of CH<sub>3</sub>CN/salt buffer at pH 6–7 was found convenient for reversed-phase separation of compounds bearing a tertiary amine functional group. Furthermore, a cluster analysis allowed the arrangement of the mobile phases according to similarity found in molecular patterns of solutes. A decision tree, which may lead to a more rational choice of the mobile phase under reversed-phase conditions, is also proposed. © 2001 Elsevier Science B.V. All rights reserved.

**Keywords:** Reviews; Mobile phases; Chiral stationary phases, LC; Data mining; CHIRBASE database

### Contents

1. Introduction .....	444
2. Data source .....	446
3. Data analysis .....	447
4. Discussion .....	447
4.1. Bayesian classifier approach .....	448
4.2. Decision tree approach .....	454
5. Conclusion .....	456
6. Nomenclature .....	456
References .....	457

\*Corresponding author. Tel.: +33-491-288-257; fax: +33-491-027-776.

E-mail address: roussel@u3pic105.u-3mrs.fr (C. Roussel).

## 1. Introduction

Since the beginning of its development in 1989, the main goal of CHIRBASE database project has not changed [1–3]. Its goals are to provide together comprehensive structural, experimental and bibliographic information on successful or unsuccessful chiral separations which have been obtained on chiral stationary phases (CSPs) in liquid chromatography (HPLC). Today, the enormous increase in the number of groups working on chiral chromatography [4–8] has led to a fast and impressive accumulation of data in CHIRBASE (Table 1). A result from all these developments is that the use of CSPs for analysis or preparation of enantiomers is today a routine task. This gives rise to one important consequence for the application of chiral technology: an increasing number of end-users of these tools who have different scientific knowledge and experiences as they are involved in a broad range of applications and activities including pharmacokinetics, asymmetric synthesis, enzymatic resolution, simulated moving bed technology and so on in the pharmaceutical, agro-chemical, as well as in the food and biotechnology industries.

This situation is well depicted in Fig. 1, which shows an almost regular increase of the total number of articles stored in CHIRBASE, whereas we can note since 1994 a relative stability of the number of references provided by the more fundamental chromatographic literature. Hence, with the rapidly growing accumulation of data and the interdisciplinary use of chiral technologies, accordingly the need to systematise and analyse the data becomes ever more demanding. Then, today, the challenge does not only remain the management of this huge quantity of information, but also its investigation in order to disclose the knowledge hidden in the data.

The extraction of hidden and useful patterns from

experimental data is facilitated in CHIRBASE because the database can be molecularly interrogated with ISIS software [9]. ISIS is a powerful chemical information system that provides both storage and retrieval of two- and three-dimensional chemical structures of both solutes and CSPs. Especially the ability of this information system to export the key structural descriptors contained in thousands of compounds is one of the most important features for data mining approaches because these molecular descriptors incorporate a remarkable amount of pertinent molecular arrangements covering each type of interaction involved in solute-CSP bindings [10].

In some recent studies from our group (to be published elsewhere), these structural key descriptors were found valuable to estimate the molecular diversity within a set of molecules resolved on a given CSP. More precisely, we have calculated the similarity indices (ranging from 0 to 100) between all possible molecule pairs using the Tanimoto method [11]. A similarity value of 0 means that the two molecules are totally dissimilar whereas a value of 100 will be obtained when the two molecules are 100% identical. If we display these indices in a dot plot we obtain a similarity map. Fig. 2 illustrates some results of dot plots. The similarity measures are displayed here according to a grey value gradient (white for 0 to black for 100). In this figure, the axis have no meaning. Dots are put in the maps at random positions by the algorithm in order to provide a good statistical repartition and thus facilitating the ability to visually distinguish the global diversity of samples. Then the application range of a given CSP can be immediately estimated from the average luminance of the full image: the higher diversity between molecules, the higher brightness in the picture will be found.

From the comparison of these maps, we could establish a scoring scheme for the classification of

Table 1  
CHIRBASE current statistics (April 2000)

Number of entries (unique sample–CSP combinations)	44 479
Number of experiments (different chiral separations)	69 106
Number of solutes	19 102
Number of CSPs	1157
Number of solvents or modifiers	219
Number of new chiral separations per update (each 4 months)	3000–5000

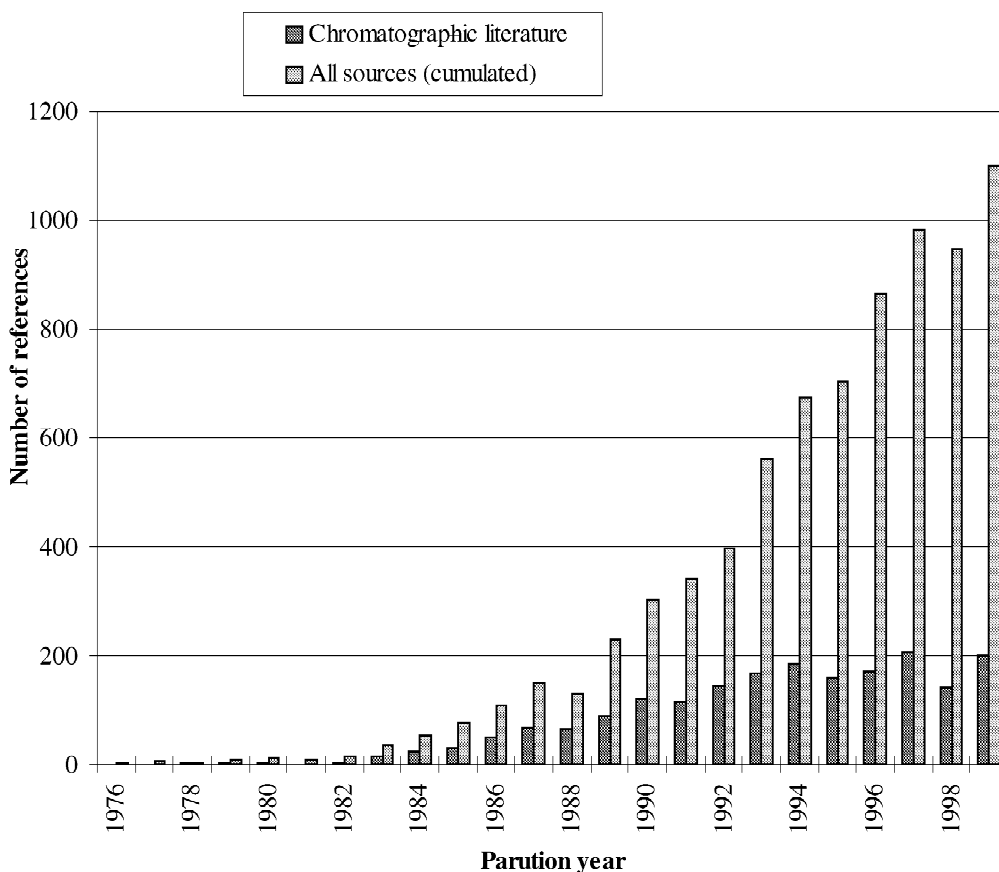


Fig. 1. Yearly evolution of the number of articles mentioning the use of a chiral stationary phase for the separation of enantiomers (source: CHIRBASE).

CSPs from specific to broad application range. In this scheme, Chiralcel OD (Fig. 3), the most commonly used CSP today was confirmed to resolve the broadest class of samples.

Furthermore, the growing number of reverse mode applications on Chiralcel OD, mostly since the availability of Chiralcel OD-R (see some applications in Refs. [12–16]), brings up a new challenge: the need for expertise in the choice of mobile phase conditions. Today, the most commonly used mobile phases are mixtures of hexane/2-propanol (63% of the Chiralcel OD separations found in CHIRBASE). In practice, analysts usually employ 10–20% 2-propanol in hexane (recommended by the manufacturer) to separate a diverse range of structurally and functionally unrelated chiral drugs. Despite the wide use of this elution system, a broader diversity of

mobile phases including pure methanol or acetonitrile have been tested on Chiralcel OD, producing a plethora of behavioural effects with highly variable enantioselective responses. For instance, there is evidence that separation of acidic or basic samples are affected by the addition of specific modifiers as trifluoroacetic acid [17] or diethylamine [18] in the mobile phase. However many other rules still remain obscure to the analysts and the solution to this problem is rather difficult, because the enantioresolution mechanisms of Chiralcel OD are still not well understood. Obviously, any attempt at identifying the relationships between the solute structure and choice of mobile phase is useful. Indeed, such relationship could also help in our understanding of the chiral recognition mechanisms of this CSP. Several attempts at studying the effect of mobile phase con-

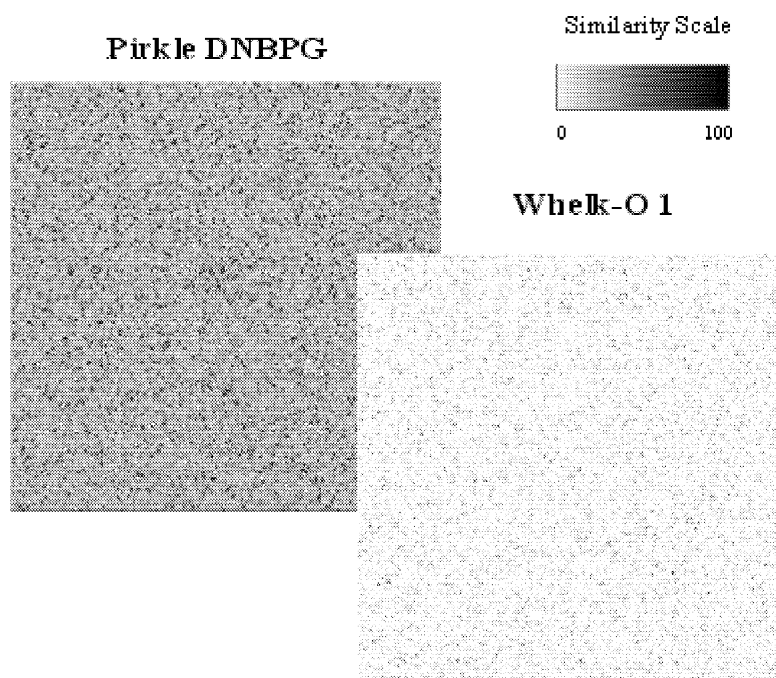


Fig. 2. Similarity map comparison of Whelk-O 1 and Pirkle DNBPG CSPs. Here, the well-recognised ability of Whelk-O 1 to resolve a broader range of samples than standard Pirkle-like CSPs is confirmed. Whelk-O 1: (3*R*,4*S*)-4-(3,5-dinitrobenzamido)-3-[3-(dimethylsilyloxy)propyl]-1,2,3,4-tetrahydrophenanthrene. Pirkle DNBPG: (*R*)-*N*-3,5-dinitrobenzoyl-phenylglycine covalently bonded to aminopropyl silica.

stituents on separation have been reported in the literature [19,20]. It should be noted that most of these studies were qualitative observations based on

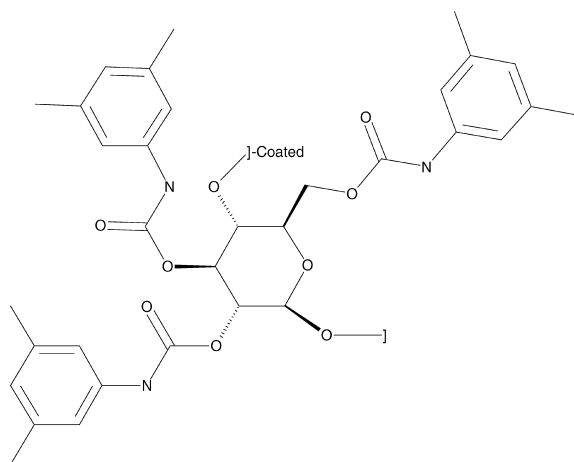


Fig. 3. Molecular structure of Chiralcel OD CSP as found in CHIRBASE. CSP name: cellulose tris(3,5-dimethylphenylcarbamate) coated on silica. CSP supplier: Daicel (Japan).

the examination of relatively small molecular data sets.

In this work, we have analysed a large data set of experimental material obtained on Chiralcel OD that is exported from CHIRBASE. The main purpose of this study is to evaluate different data analysis tools for mobile phase prediction using as attributes the molecular keys readily available in ISIS. Because of the large diversity of chemical structures in this data set, conventional quantitative structure activity relationship (QSAR) programs were not used, as they are more appropriate for the analysis of a given class of compounds. The data set was therefore submitted to data mining programs that are suitable for the analysis of highly diverse data sets.

## 2. Data source

The review of the successful separations reported in literature ( $\alpha$  found in CHIRBASE and  $\alpha \geq 1.1$ ) for Chiralcel OD resulted in an experimental data set of

2363 molecular structures. This data set has been extracted from CHIRBASE by an in-house program developed using the application development kit of ISIS. For each solute, we have exported from CHIRBASE 166 molecular key descriptors (available in ISIS) and all the corresponding evaluated mobile phases. This procedure leads to a matrix data file containing a total of 2494 chiral separations (some of the 2363 solutes may be resolved using different eluting conditions). In this file, the molecular descriptors are coded with 0 for “absence” and 1 for “presence”. Calculations have been performed on a Pentium PC (Windows 95, 32 Mbytes RAM) and on a Silicon Graphics O2 RX-10000 (IRIX 6.3).

### 3. Data analysis

The RoC (Robust Bayesian Classifier) package has been used for class membership prediction of molecular structures. RoC is a Bayesian supervised classifier that includes an easy-to-use wizard interface. RoC is developed by the Bayesian Knowledge Discovery Project of The Open University (UK) [21]. Decision trees have been built with the MLC++ package (Machine Learning Library in C++) initially developed at Stanford University [22]. Clustering analysis was performed using Ward reciprocal nearest neighbour algorithm as available in TSAR, a QSAR package available from Oxford Molecular [23].

### 4. Discussion

The first step in any data mining studies is the preparation of data.

First, in our Chiralcel OD data set file, many of the 166 structural key descriptors (such as actinide, lanthanide, isotope, etc.) were found irrelevant for our study and removed from the data set. Some of the 48 descriptors that have been retained are reported in Table 2.

In addition, it was also required to transform the reviewed mobile phases into classes. This conversion provided 20 different categories of mobile phases as seen in Table 3. In these categories:

Table 2

Some examples of the 48 molecular descriptors retained for the analysis of Chiralcel OD experimental data set extracted from CHIRBASE

Symbol assigned by ISIS	Designation
NCOO	Carbamate
NCON	Urea
C%N	Aromatic amine
CH3CH2CH2AA	Alkyl chain
CCCC	<i>tert.</i> -Butyl
NH2	Primary amine
NH	Secondary amine
CNCC	Tertiary amine
OCNC	Amide
C–O–C	Ether
X	Halogen
OCCC	Secondary alcohol
C=O	Carbonyl
AROMATIC	Aromatic group
COOH	Carboxylic acid
CTN	Cyano
NHETERO	N Heterocycle

Table 3

Class frequency distribution of the Chiralcel OD data set<sup>a</sup>

Class	Frequency distribution
ALKANE/CHCl3	9
ALKANE/ALCOHOL/AMINE/H2O	7
CH3CN/ACID	39
ALKANE/AMINE	5
ALKANE/MTBE	8
ALCOHOL/H2O	20
ALCOHOL	21
ALKANE/ALCOHOL/H2O	28
CH3CN	7
CH3CN/H2O	14
ALKANE/CHCl3	11
ALKANE/THF	15
ALKANE	31
ALKANE/ALCOHOL/AMINE	235
CO2 (supercritical)/MEOH	31
ALKANE/ETOH	126
CH3CN/SALT	69
ALKANE/2-PROH	1577
ALKANE/ALCOHOL/ACID	206
ALKANE/ALCOHOL/ACID/AMINE	35

<sup>a</sup> Number of classes (mobile phases): 20; number of attributes (molecular descriptors): 47; number of cases: 2494.

- ALKANE represents any hexane, heptane, pentane, etc., solvents;
- AMINE represents any basic modifiers as diethyl- or triethylamine;
- ACID represents any acid modifiers such as formic, trifluoroacetic, acetic acid, etc.;
- CH3CN/ACID includes all the mobile phases with CH<sub>3</sub>CN in a pH acid buffer;
- CH3CN/SALT corresponds to CH<sub>3</sub>CN in a non-acid salt buffer (pH 6–7);
- ALCOHOL alone means a pure alcoholic mobile phase (as MeOH or EtOH);
- MTBE stands for methyl-*tert.*-butyl-ether.

Some comments about this table are required here. As we needed to fit the different mobile phases into a small number of discrete categories, we could not take in consideration the different proportions of solvents or modifiers. In addition, the frequency distribution reported in this table reveals that the number of items (separations) corresponding to each class (mobile phases) varies significantly between the classes. Then, even if the data mining tools chosen in this study solve the problem of rare items in data set, we can already anticipate here that badly represented classes as ALKANE/AMINE (five items) or CH3CN (seven items) will often be ignored by classification studies (no samples may be classified in these classes). One can note that some supercritical applications (CO<sub>2</sub>/MEOH) of Chiralcel OD have been included in our study. We can also note that the manufacturer of Chiralcel OD does not recommend some solvents found in CHIRBASE like THF or CHCl<sub>3</sub>. In consideration of the theoretical purpose of our study, we decided to keep them in our analysis. Moreover, these solvents may gain importance with the advances in linking the Chiralcel OD-type selector onto silica [24].

#### 4.1. Bayesian classifier approach

Bayesian analysis are interesting alternative to “classical” statistics based on linear models (linear regression, discriminant analysis or factor analysis). Bayesian methods provide a strictly probabilistic system, which is often well appropriate to a various of diagnostic domains where uncertainty must be taken into account (health-care, industrial processes,

economical or biological sciences) [25–27]. Bayesian systems apply the Bayes theorem:

$$p(A|B) = [p(A) \cdot p(B|A)]/p(B)p(A|B)$$

is the probability of *A* for a given *B*. It indicates the probability that *A* is true given that *B* is true.

A Bayesian classifier is first trained by estimating the conditional probabilities distribution (relative frequency of relevant cases) of each attribute. Then, using Bayes theorem, given a set of attribute values (*A*<sub>1</sub>, . . . , *A*<sub>*n*</sub>) of a new case, the Bayesian classifier will find the class *C*<sub>*i*</sub> for which *p*(*C*<sub>*i*</sub>|*A*<sub>1</sub> & . . . . . & *A*<sub>*n*</sub>) is the greatest.

In this section, we trained the RoC Bayesian classifier on the whole Chiralcel OD data set file and then used the resulting classifier to predict the classes of the same data set. The resulting accuracy (predictive quality) was 49%. The low predictivity may seem to be a weakness of the method. The difference here, however, is that in chiral chromatography many mobile phases are interchangeable for a given solute. As there are a number of elution system that can satisfy the same constraints, the selection of a particular one is unimportant, and often determined by what appeared to be available in the laboratory. Moreover, this result may also denote some conflicts with the experience of analytical chemists in setting up an elution system for particular samples. As discussed above, it is also reasonable to assume a weakness in the user’s knowledge about the rules for the most appropriate choices. In this perspective we can hypothesise that when the classifier makes an incorrect conclusion for a given sample, it may mean that another appropriate mobile phase could be chosen.

This assessment was found very pertinent when we manually looked at the misclassified samples and identified the molecular features, which lead the classifier model to choose a particular mobile phase. This was not a difficult task because RoC also provided the distribution of the conditional probabilities (comprise between 0 and 1) assigned to the presence and absence of each attribute with each mobile phases. As an example, we report in Fig. 4 the probability distribution of the “CNCC” (tertiary amine) attribute. The highest probability values of the presence of “CNCC” are associated with the

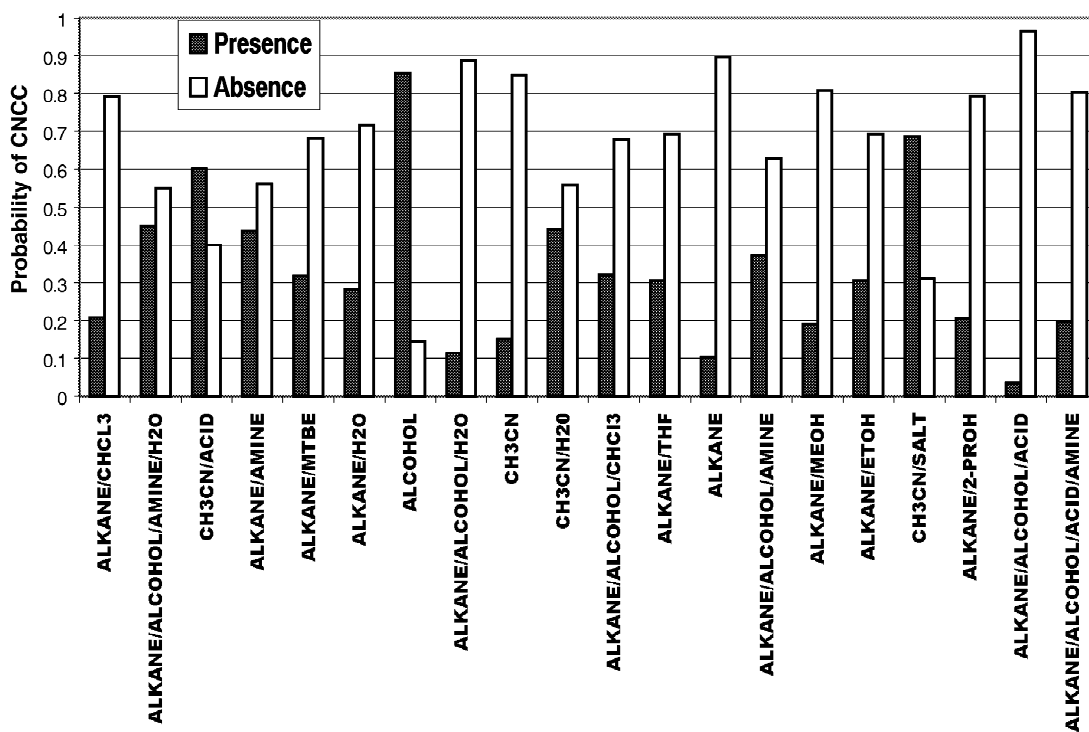


Fig. 4. Conditional probability distribution of the attribute CNCC (tertiary amine).

classes ALCOHOL and CH3CN/SALT. For absence of “CNCC” we retrieve ALKANE/ALCOHOL/ACID, ALKANE/ALCOHOL/H2O and ALKANE.

When we look at the probability distribution of the “COOH” (carboxylic acid) attribute, the two highest values for its presence are found for ALKANE/ALCOHOL/ACID ( $P=0.6$ ) and CH3CN/ACID ( $P=0.3$ ). If we combine these results with the probabilities of presence of “CNCC” with ALKANE/ALCOHOL/ACID ( $P=0.03$ ) and CH3CN/ACID ( $P=0.7$ ), we may predict that classifier will often propose using a CH3CN/ACID mobile phase for samples bearing both carboxylic acid and tertiary amine groups.

Because we can relate the misclassified samples to probability values, we can easily find many other simple rules. Some of them explain well the misclassified mobile phase choice for samples reported in Table 4:

- ALKANE is associated with the lowest probabilities ( $P=0.036$ ) of presence of hydrogen-bond donors and high probabilities of lipophilic groups

(alkyl chain, *tert.*-butyl). It is often proposed for very lipophilic samples (entries 1, 2 and 3).

- ALKANE/ALCOHOL/CHCl3 is proposed for non-aromatic bicyclic compounds (highest probability for O-heterocycle and non-aromatic attributes, high probability for bicycle).
  - If ALKANE/ALCOHOL/ACID is the mobile phase of choice for acids (entry 20), CH3CN/ACID is also frequently suggested for acids. This result confirms that acids can be resolved under normal as well reversed-phase mode.
- More surprising are the following results:
- CH3CN/ACID is associated with the highest probability for presence of the attribute amide ( $P=0.7$ ) as confirmed in entries 13 and 14.
  - ALCOHOL (entry 18) or CH3CN/SALT (entries 15 and 16) with the presence of “CNCC”.
  - ALCOHOL hold the highest probability for presence of *N*-heterocycle (entry 19) and the attribute “HETEROCYC-ATOM>1” (more than one heterocyclic atom).

It is also interesting to note that ALKANE/AL-

Table 4  
Some typical structures misclassified by RoC Bayesian Classifier

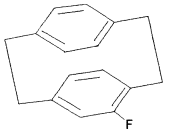
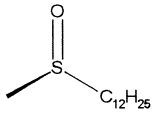
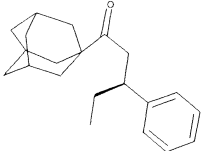
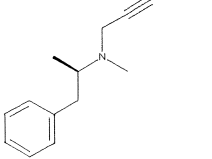
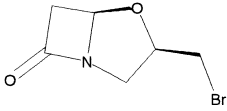
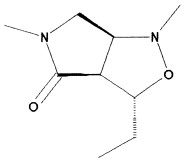
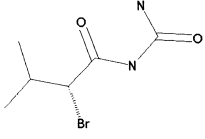
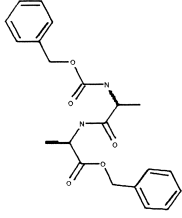
Structure	Mobile phase proposed by the classifier	Mobile phase quoted in literature	$\alpha$	Ref.
1 	Alkane	Hexane/2-PrOH	1.04	[28]
2 	Alkane	Hexane/2-PrOH	1.07	[29]
3 	Alkane	Hexane/2-PrOH	1.09	[30]
4 	Alkane/MTBE	Hexane/2-PrOH	1.20	- <sup>a</sup>
5 	Alkane/Alcohol/ CHCl <sub>3</sub>	Hexane/2-PrOH	1.09	[31]
6 	Alkane/Alcohol/ CHCl <sub>3</sub>	Hexane/2-PrOH	1.10	[32]
7 	Alkane/Alcohol/ CHCl <sub>3</sub>	Hexane/2-PrOH	1.06	[33]
8 	Alkane/Alcohol/ Amine	Hexane/EtOH/ H <sub>2</sub> O	1.05	[34]



Table 4. Continued

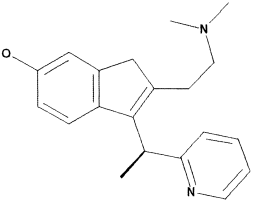
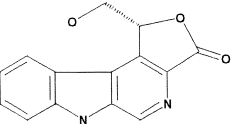
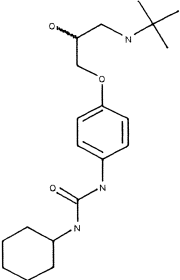
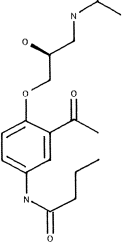
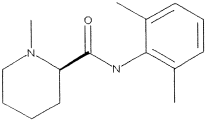
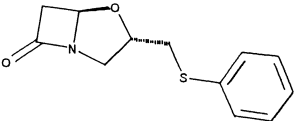
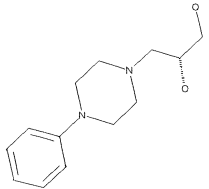
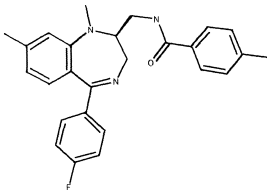
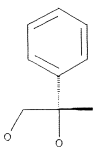
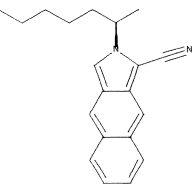
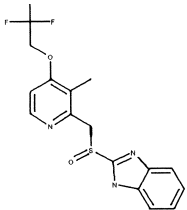
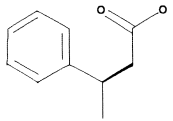
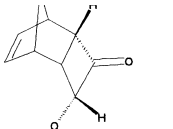
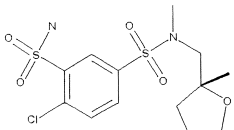
Structure	Mobile phase proposed by the classifier	Mobile phase quoted in literature	$\alpha$	Ref.
	Alkane/Alcohol/ Amine	Hexane/2-PrOH	1.06	[35]
	Alkane/Alcohol/ Amine	Heptane/2-PrOH/ H <sub>2</sub> O	1.09	[36]
	Alkane/Alcohol/ Acid/Amine	Heptane/2-PrOH/ Et <sub>2</sub> NH	1.20	[37]
	Alkane/Alcohol/ Acid/Amine	Heptane/EtOH/ Et <sub>2</sub> NH	1.10	[37]
	CH <sub>3</sub> CN/buffer pH acid	Hexane/EtOH	1.07	[38]
	CH <sub>3</sub> CN/buffer pH acid	Hexane/2-PrOH	1.10	[31]

Table 4. Continued

Structure	Mobile phase proposed by the classifier	Mobile phase quoted in literature	$\alpha$	Ref.
	CH <sub>3</sub> CN/Salt	Hexane/EtOH	1.08	[39]
	CH <sub>3</sub> CN/Salt	Hexane/2-PrOH	1.09	[40]
	Alkane/Alcohol/ H <sub>2</sub> O	Hexane/2-PrOH	1.09	[41]
	Alcohol	Hexane/2-PrOH	1.05	[42]
	Alcohol	MeOH/H <sub>2</sub> O 50 mM NaClO <sub>4</sub>	1.16	[43]
	Alkane/Alcohol/ Acid	CH <sub>3</sub> CN/pH 2.0, NaClO <sub>4</sub> aq.	1.10	[44]
	Alkane/CHCl <sub>3</sub>	Hexane/2-PrOH	1.42	[45]
	Alcohol/H <sub>2</sub> O	Hexane/2-PrOH	1.10	[46]

<sup>a</sup> R. Stradi, G. Celentano, personal Communication in CHIRBASE, 1991.

COHOL/AMINE has significant (highest mean probability values) but not the highest probabilities for the presence of most amino groups.

It, however, holds the highest probabilities for “N” and “NH” attributes.

Some other interesting results and often difficult to interpret are:

- Addition of amine in mobile phase (ALKANE/AMINE class) or MTBE is suggested in presence of carbamate or urea groups.
- Addition of H<sub>2</sub>O (ALCOHOL/H<sub>2</sub>O or ALKANE/ALCOHOL/H<sub>2</sub>O) is often proposed in presence of hydroxyl groups (alcohols).
- ALCOHOL/H<sub>2</sub>O is often proposed in presence of “NH<sub>2</sub>” attribute (primary amine).
- CH<sub>3</sub>CN/SALT hold the highest probability ( $P = 0.5$ ) for presence of “C=N” (imine) attribute.
- ALKANE/ALCOHOL/ACID/AMINE and ALKANE/ALCOHOL/AMINE/H<sub>2</sub>O provide higher probabilities than ALKANE/ALCOHOL/AMINE for alcohol groups or presence of many hydrogen-bond donors.
- A significant probability of the “ESTER” attribute is assigned to ALKANE/ALCOHOL/ACID and ALKANE/MTBE mobile phases.
- CH<sub>3</sub>CN and CH<sub>3</sub>CN/H<sub>2</sub>O mobile phases hold the highest probability for presence of a cyano group.
- Pure alcoholic mobile phases are chosen before mixture of ALKANE/ALCOHOL for basic compounds and for compounds bearing several aromatic rings and halogens.
- ALKANE/ETOH is preferred to ALKANE/2-PROH for basic compounds. We observe no significant difference between these two mobile phases for amide, carbamate or urea groups.
- ALKANE/ALCOHOL/ACID provides the lowest probability values for urea, amide, aromatic amine, imine, primary and tertiary amine groups ( $0.001 > P > 0.03$ ). Therefore, one should prefer CH<sub>3</sub>CN/ACID in the presence of these substructural features.

As we have already pointed out, the first objective of this approach is the production of a model based on existing data. It helps well to discover some hidden patterns contained in data and allows the prediction of future results (supervised learning). However it does not clearly reveal the groups of

items (mobile phases) that are similar. For this purpose, cluster analysis techniques are well recommended.

Cluster analysis algorithms are based on the division of a data set so that entries with similar content are in the same group, and groups are as different as possible from each other.

The results from a cluster analysis are best viewed in a dendrogram as displayed in Fig. 5.

The dendrogram of this figure has been built using the knowledge (probabilities) that the program gained during the training process. We chose the Ward reciprocal nearest neighbour method [47] because it computes the distance between two subgroups as the minimum distance between any two members of opposite groups. Ward’s method is a favourable default linkage because it produces condensed groups of well-distributed size. As shown in this dendrogram, some mobile phase groups are in good agreement with some of the patterns outlined above. The cluster analysis differentiated well, in one cluster, three mobile phases used for reversed-phase application of Chiralcel OD: ALCOHOL, CH<sub>3</sub>CN/ACID and CH<sub>3</sub>CN/SALT. It is quite interesting to find CH<sub>3</sub>CN/ACID and CH<sub>3</sub>CN/SALT, mainly used on Chiralcel OD-R in the same sub-cluster.

In another cluster, unusual mobile phases (using THF, CHCl<sub>3</sub> or MTBE) are well distinguished from “classical” mobile phases (as alkane/alcohol). This result may reveal singular effects of these solvents. It was noteworthy to see that closely related eluents (CH<sub>3</sub>CN and CH<sub>3</sub>CN/H<sub>2</sub>O) are grouped in the same cluster. A far more striking result is found when the largest group of clustered mobile phases is examined where we observe closely related ALKANE/2-PROH and ALKANE/ETOH in the same cluster. In this group, one cluster is dominated by conditions often applied for separation of basic compounds. Inside this cluster, it may be surprising to find that ALKANE/ALCOHOL/AMINE and supercritical CO<sub>2</sub>/MEOH are grouped together. This relationship is clearest in data from experiments. For instance, a large number of  $\beta$ -blockers (amino-alcohols) are well resolved under both conditions [48,49]. On a finer level we may note that alcoholic mobile phases containing H<sub>2</sub>O are clustered next to, or in the immediate vicinity of, each other.

Some other unexpected observations are:

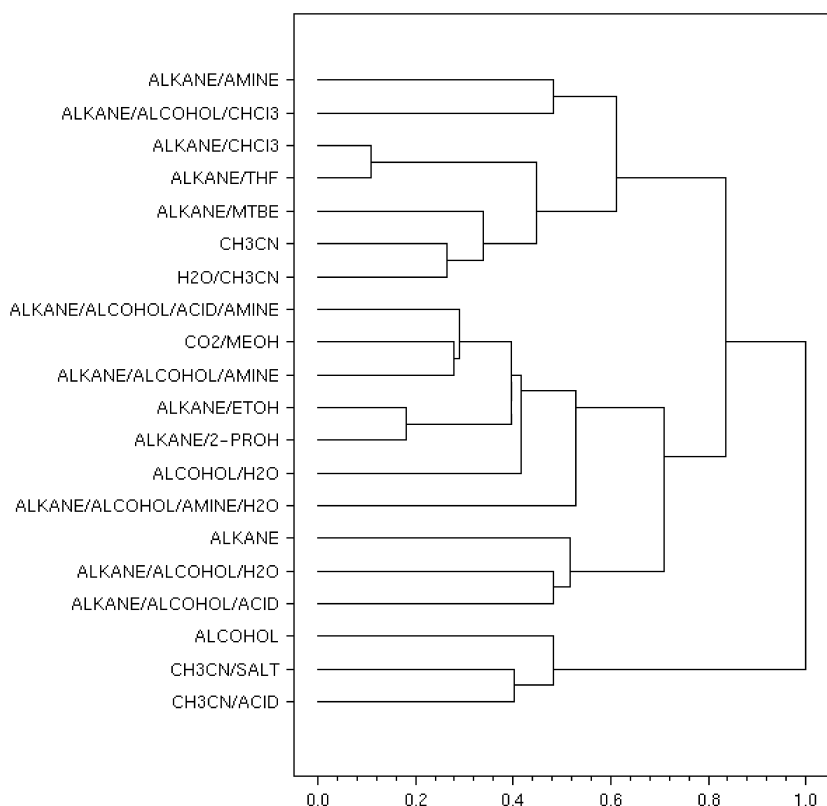


Fig. 5. Dendrogram revealing similarity of mobile phases according to the Bayesian analysis results. Agglomeration method: Ward reciprocal nearest neighbour.

- ALKANE/ALCOHOL/AMINE/ACID is closer to ALKANE/ALCOHOL/AMINE than to ALKANE/ALCOHOL/AMINE/H2O.
- The proximity of ALCOHOL and CH3CN/SALT (probably due to the high frequency of the amine tertiary group in samples as seen above).
- ALKANE/ALCOHOL/H2O and ALKANE/ALCOHOL/ACID are found in one cluster suggesting that acid may sometimes be replaced by water.

Notwithstanding the probabilistic character of the data analysed here and thus the possibility of errors, it seems that some interchangeable mobile phases often cluster tightly together.

#### 4.2. Decision tree approach

A decision tree contains two types of nodes: parents and leaves. Each parent node corresponds to

a question or an attribute; each leaf node designates a single class. The branches connected to a parent node correspond to a split of the population node according to the answers to the question or the value of the attribute. Each subset of the population is split again, recursively, using different questions or attributes until a subset belongs to a single class. In this case, the branch of the tree stops with a leaf node labelled with a single class.

A tree is read from root to leaves. We begin at the root of the tree that contains all the population. Then, following the relevant branches according to the question asked at each branch node, we finally reach a leaf node. The label on that leaf node provides the class, which is the resulting conclusion induced from the tree.

In this study, attributes are the molecular key features already described above (1=presence, 0=absence) and leaves are mobile phases. We used a

decision tree algorithm originally developed by Quinlan [50] as it is available in the MLC++ package. The purpose was to evaluate the decision tree approach for prediction of Chiralcel OD-R mobile phases (a reversed-phase version of Chiralcel OD). This assay was carried out on a CHIRBASE file of 124 molecular structures corresponding to a list of samples tested on Chiralcel OD-R. For each solute, we have picked in CHIRBASE all the mobile phases with their corresponding enantioselectivities. The transformation of the various mobile phases into classes leads to the three following categories:

- CH3CN/SALT (usually CH<sub>3</sub>CN, pH 6.0, NaClO<sub>4</sub> aqueous)
- CH3CN/ACID (usually CH<sub>3</sub>CN, pH 2.0, NaClO<sub>4</sub> aqueous)
- CH3CN/H2O

A first attempt using the 48 molecular descriptors applied previously showed that many kinds of “noise” exist in the data set because some attributes were found irrelevant to the decision-making process. Principal component analysis (PCA), well

```

NH = 0
...CN(C)C = 1 → CH3CN/SALT
: CN(C)C = 0
: ...COOH = 0 → CH3CN/H2O
:   COOH = 1
:     ...alpha <= 1.04 → CH3CN/SALT
:     alpha > 1.04 → CH3CN/ACID
NH = 1
...OC(N)C = 1
...CN(C)C = 1 → CH3CN/ACID
: CN(C)C = 0
: ...alpha <= 2.02 → CH3CN/H2O
:   alpha > 2.02 → CH3CN/ACID
OC(N)C = 0
...NC(O)N = 1 → CH3CN/ACID
NC(O)N = 0
...NC(O)O = 1 → CH3CN/ACID
NC(O)O = 0
...alpha <= 1.04 → CH3CN/ACID
alpha > 1.04 → CH3CN/SALT

```

known in practice to reduce the dimensionality of problems and to transform interdependent coordinates into significant and independent ones, was successfully used here. PCA reduced significantly the number of variables of our Chiralcel OD-R data set to 10 molecular descriptors. This new data set finally yields to the full tree shown in Fig. 6.

In this tree, each path, from the root to a leaf, corresponds to a rule. All of the decisions about the presence or absence of a molecular feature leading to a mobile phase choice at the leaf node define the conditions of the rules. For example, from the tree above we can generate the following rules:

*Rule 1:*  
 NH=0  
 CN(C)C=1  
 →CH3CN/SALT

*Rule 2:*  
 NH=0  
 CN(C)C=0  
 COOH=0  
 →CH3CN/H2O

*Rule 3:*  
 NH=0  
 CN(C)C=0  
 COOH=1  
 →CH3CN/ACID ( $\alpha > 1.04$ )  
 →CH3CN/SALT ( $\alpha \leq 1.04$ )

*Rule 4:*  
 NH=1  
 OC(N)C=1  
 CN(C)C=1  
 →CH3CN/ACID

*Rule 5:*  
 NH=1  
 OC(N)C=1  
 CN(C)C=0  
 →CH3CN/ACID ( $\alpha > 2.02$ )  
 →CH3CN/H2O ( $\alpha \leq 2.02$ )

*Rule 6:*  
 NH=1  
 OC(N)C=0  
 NC(O)N=1  
 →CH3CN/ACID

Fig. 6. Decision tree built from the analysis of 124 molecular structures using CH3CN/ACID, CH3CN/SALT or CH3CN/H2O mobile phases (on Chiralcel OD-R).

**Rule 7:**

NH=1  
 OC(N)C=0  
 NC(O)N=0  
 NC(O)O=1  
 →CH3CN/ACID

**Rule 8:**

NH=1  
 OC(N)C=0  
 NC(O)N=0  
 NC(O)O=0  
 →CH3CN/ACID ( $\alpha \leq 1.04$ )  
 →CH3CN/SALT ( $\alpha > 1.04$ )

As previously noted, CH3CN/SALT is indicative for all samples bearing a tertiary amine; when the sample does not contain a tertiary amine then in presence of a carboxylic acid, CH3CN/ACID should be chosen. If a solute contains an amide, urea or carbamate functional group, then CH3CN/ACID should be preferred to CH3CN/SALT. CH3CN/H2O is proposed when none of these structural features is found on the solute.

## 5. Conclusion

From the results derived with these preliminary studies, we can conclude the following:

Data mining and molecular pattern recognition can be used in HPLC in order to rapidly classify mobile phase applications. We believe that algorithms for automatic classification like the Bayes classifier applied in our study are in principle able to perform such classification task.

The advantage of the Bayes classification algorithm is the relevance of the predictions achieved if a set of sufficiently discriminating features is determined. The examples presented in this review of mobile phases on Chiralcel OD demonstrate that it is particularly useful when it is difficult to discover complicated non-linear relationships in data. The main disadvantage is that it often produces a huge amount of quantitative results in large tables. Consequently, we needed to develop our skill to detect information among all these quantitative measures.

We must also point out that some findings highlighted in this work could be enhanced and com-

pleted if we succeed to gather a better representative set of samples in the training set prior to performing predictions. It seems likely that the future addition of more and diverse analytical conditions in CHIRBASE will help us gain more accurate and numerous results. On the basis of this observation, it is probable that our new research project which makes use of an automated CSP screening equipment to supply more original data in CHIRBASE, will help us to address this issue. More precisely, we also aim to use this new technology for the rational design of large collections of experimental data. Indeed, we expect to obtain from such rational studies more discriminating experimental data, and therefore more fruitful and reliable prediction models.

Finally, as our objective is also to implement an expert system in CHIRBASE, these developments will certainly play an important role in this challenge aiming to build an information system that provides not only data collection but also rule sets for each CSP and knowledge about the processes of chiral separations.

## 6. Nomenclature

ACID	Descriptor: any acid modifier
ALCOHOL	Descriptor: any alcoholic modifier
ALKANE	Descriptor: any alkane solvent (heptane, hexane, pentane, etc.)
AMINE	Descriptor: any basic modifier
CH3CN/ACID	Descriptor: all the mobile phases with CH <sub>3</sub> CN in a pH acid buffer
CH3CN/H2O	Descriptor: all the mobile phases with CH <sub>3</sub> CN in water (no salt)
CH3CN/SALT	Descriptor: all the mobile phases with CH <sub>3</sub> CN in a non-acid salt buffer
CNCC	Descriptor: tertiary amine functional group
COOH	Descriptor: carboxylic acid functional group
CSP	Chiral stationary phase
HPLC	High-performance liquid chromatography
MLC	Machine Learning Library in C++
MTBE	Descriptor: methyl- <i>tert.</i> -butyl-ether
NC(O)N	Descriptor: urea functional group

NC(O)O	Descriptor: carbamate functional group
NH	Descriptor: secondary amine functional group
OC(N)C	Descriptor: amide functional group
PCA	Principal component analysis
QSAR	Quantitative structure activity relationship
RoC	Robust Bayesian classifier

## References

- [1] C. Roussel, P. Piras, *Pure Appl. Chem.* 65 (1993) 235.
- [2] C. Roussel, P. Piras, in: *Proceedings CHIRAL'94, USA, 1994*.
- [3] C. Roussel, P. Piras, in: J.E. Dubois, N. Gershon (Eds.), *Data and Knowledge in a Changing World: Modeling Complex Data for Creating Information: Real and Virtual Objects*, Springer, Berlin, 1996.
- [4] Y. Okamoto, E. Yashima, *Angew. Chem. Int. Ed. Engl.* 37 (1998) 1021.
- [5] E. Francotte, *Chem. Anal. Ser.* 142 (1997) 633.
- [6] S. Ahuja (Ed.), *Chiral Separations: Application and Technology*, American Chemical Society, Washington, DC, 1996.
- [7] P. Schreier, A. Bernreuther, M. Huffer (Eds.), *Analysis of Chiral Organic Molecules. Methodology and Applications*, Walter de Gruyter, Berlin, 1995.
- [8] G. Subramanian (Ed.), *A Practical Approach to Chiral Separations by Liquid Chromatography*, VCH, Weinheim, New York, 1994.
- [9] MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577.
- [10] R.D. Brown, Y.C. Martin, *J. Chem. Inf. Comput. Sci.* 37 (1997) 1.
- [11] G.M. Downs, P. Willett, in: K.B. Lipkowitz, D.B. Boyd (Eds.), *Reviews in Computational Chemistry*, Vol. 7, VCH, Weinheim, New York, 1996, Chapter 1.
- [12] J.G. Hou, Y.L. Wang, C.X. Li, X.Q. Han, J.Z. Gao, J.W. Kang, *Chromatographia* 50 (1999) 89.
- [13] E.B. Asafu-Adjaye, G.K. Shiu, *J. Chromatogr. B* 707 (1998) 161.
- [14] T. Santa, A. Takeda, S. Uchiyama, T. Fukushima, H. Homma, S. Suzuki, H. Yokusu, C.K. Lim, K. Imai, *J. Pharm. Biomed. Anal.* 17 (1998) 1065.
- [15] M.G. Quaglia, A. Farina, E. Bossu, V. Cotichini, *J. Pharm. Biomed. Anal.* 18 (1998) 171.
- [16] T. Fukushima, T. Santa, H. Homma, S.M. Al Kindy, K. Imai, *Anal. Chem.* 69 (1997) 1793.
- [17] R.J. Bopp, F. Geiser, *Presentation IB2-2 ISCD'97 Nagoya, 1997*.
- [18] A.M. Krstulovic, M.H. Fouchet, J.T. Burke, G. Gillet, A. Durand, *J. Chromatogr.* 452 (1988) 477.
- [19] K.M. Kirkland, *J. Chromatogr. A* 718 (1995) 9.
- [20] K. Balmer, P.-O. Lagerström, B.-A. Persson, G. Schill, *J. Chromatogr.* 592 (1992) 331.
- [21] Bayesian Knowledge Discovery Project, Knowledge Media Institute Top Floor, Berrill Building, The Open University, Milton Keynes MK7 6AA UK (URL: <http://kmi.open.ac.uk/projects/bkd>).
- [22] MLC++ (Machine Learning Library in C++) was initially developed at Stanford University (R. Kohavi, D. Sommerfeld, J. Dougherty) and is public domain. The new version 2.0 is freely distributed by Silicon Graphics, Inc.
- [23] Oxford Molecular Ltd., the Medawar Centre, Oxford Science Park, Sandford on Thames, Oxford OX4 4GA, UK.
- [24] E.R. Francotte, Patent WO 9749733 (1997).
- [25] D. Heckermann, *Data Mining Knowledge Discovery 1* (1997) 79.
- [26] P. Kontkanen, P. Myllymäki, H. Tirri, in: D.L. Dowe, K.B. Korb, J.J. Oliver (Eds.), *Comparing Bayesian Model Class Selection Criteria by Discrete Finite Mixtures. Information, Statistics and Induction in Science, Proceedings of the ISIS'96 Conference in Melbourne, Australia, August 1996*, World Scientific, Singapore, 1996, p. 364.
- [27] P. Kontkanen, P. Myllymäki, H. Tirri, *Predictive Data Mining with Finite Mixtures, Proceedings of The Second International Conference on Knowledge Discovery and Data Mining, 1996*, 176.
- [28] H. Hopf, W. Grahm, D.G. Barrett, A. Gerdes, J. Hilmer, J. Hucker, Y. Okamoto, Y. Kaida, *Chem. Ber.* 123 (1990) 841.
- [29] E. Küsters, V. Loux, E. Schmid, Ph. Floersheim, *J. Chromatogr. A* 666 (1994) 421.
- [30] K. Soai, M. Okudo, M. Okamoto, *Tetrahedron Lett.* 32 (1991) 95.
- [31] Y. Okamoto, T. Senoh, H. Nakane, K. Hatada, *Chirality* 1 (1989) 216.
- [32] R. Ficarra, M.L. Calabro, S. Tommasini, D. Costantino, M. Carulli, S. Melardi, M.R. Di Bella, F. Casuscelli, R. Romeo, P. Ficarra, *Chromatographia* 43 (1996) 365.
- [33] V. Venizelos, H. Irth, U.R. Tjaden, J. van der Greef, D.D. Breimer, T.M.T. Mulders, G.J. Mulder, *J. Chromatogr. B* 573 (1992) 259.
- [34] S.-H. Wu, S.-L. Lin, S.-L. Lai, T.-H. Chou, *J. Chromatogr.* 514 (1990) 325.
- [35] D. Prien, G. Blaschke, *Poster: 3rd International Symposium on Chiral Discrimination, Tübingen, 1992*.
- [36] L. Dubois, G. Dorey, P. Potier, R.H. Dodd, *Tetrahedron: Asymmetry* 6 (1995) 455.
- [37] C. Vandenbosch, D.L. Massart, W. Lindner, *J. Pharm. Biomed. Anal.* 10 (1992) 895.
- [38] M. Siluveru, J.T. Stewart, *Anal. Lett.* 30 (1997) 1167.
- [39] R. Stradi, D. Pitre, G. Celentano, E. Speroni, M. Gentile, E. Marinone, *Poster P149, Chiral Discrimination Symposium, Rome, 1991*.
- [40] R.L. Meurisse, C.J. De Ranter, *Chromatographia* 38 (1994) 629.
- [41] E. Höft, H.J. Hamann, A. Kunath, W. Adam, U. Hoch, C.R. Saha-Möller, P. Schreier, *Tetrahedron: Asymmetry* 6 (1995) 603.

- [42] A.L.L. Duchateau, M.G. Hillemans, I. Hindriks, *Enantiomer* 2 (1997) 61.
- [43] M. Tanaka, H. Yamazaki, H. Hokusui, *Chirality* 7 (1995) 612.
- [44] Daicel Chem. Ind., Ltd., Chiralcel OD-R Brochure, 1992.
- [45] T. Taniguchi, R.M. Kanada, K. Ogasawara, *Tetrahedron: Asymmetry* 8 (1997) 2773.
- [46] Daicel Chem. Ind., Ltd., Application Guide, 1989.
- [47] J.H. Ward, *J. Am. Stat. Assoc.* 58 (1963) 236.
- [48] N. Bargmann-Leyder, A. Tambuté, M. Caude, *Chirality* 7 (1995) 311.
- [49] C.R. Lee, J.-P. Porziemsky, M.-C. Aubert, A.M. Krstulovic, *J. Chromatogr.* 539 (1991) 55.
- [50] J.R. Quinlan, in: *C4.5: Programs for Machine Learning*, Morgan Kaufmann, Los Altos, CA, 1993.