ADAPTIVE FUZZY PARTITION IN DATA BASE MINING: APPLICATION TO OLFACTION

M Pintore, K Audouze, F Ros and JR Chrétien*

Laboratory of Chemometrics & BioInformatics, University of Orléans, BP 6759, 45067 Orléans Cedex2, France.
Email: Jacques.Chretien@univ-orleans.fr

ABSTRACT

A data set of 412 olfactory compounds, divided into animal, camphoraceous, ethereal and fatty olfaction classes, was submitted to an analysis by a Fuzzy Logic procedure called Adaptive Fuzzy Partition (AFP). This method aims to establish molecular descriptor/chemical activity relationships by dynamically dividing the descriptor space into a set of fuzzily partitioned subspaces. The ability of these AFP models to classify the four olfactory notes was validated after dividing the data set compounds into training and test sets, including 310 and 102 molecules, respectively. The main olfactory note was correctly predicted for 83 % of the test set compounds.

Keywords: Fuzzy Logic, Structure-Activity Relationship, Genetic Algorithm, olfactory compounds.

1 INTRODUCTION

Flavor and odor remain permanent challenges in academic and industrial research. The economic impact of the olfactory field explains the large number of papers involving data analysis methods to process sensorial and experimental measurements (Rossiter, 1996; Chastrette, 1998; Kermani, Schiffman & Nagle, 1999; de Mello Castanho Amboni, da Silva Junkes, Yunes & Heinzen, 2000). However, odor evaluation by man represents a special field of research, whose specific difficulties need to be overcome to lead to robust results. The multiplicity of factors involved in the olfaction biological process prevents the derivation of efficient predictive mathematical models. Four points mainly define this complexity (Chastrette & Zakarya, 1991; Buck & Axel, 1991; Malnic, Hirono, Sato & Buck, 1999):

(i) a huge number of receptors is involved in olfaction;
(ii) knowledge related to the 3D structure of these receptors is still missing;
(iii) different types of chemical compounds can affect the same receptor;
(iv) one compound can exhibit simultaneously different odors.

Furthermore, the importance of fuzziness linked to the expert's subjectivity has to be considered. Much progress has been made in the knowledge of physiological and psychological factors influencing the expert's olfaction evaluation (Manley, 1993; Qureshy, Kawashima, Imran, Sugiura, Goto, Okada et al., 2000), but it is not sufficient to clearly discriminate between objectivity and subjectivity in the characterization exhibited by panels of experts.

All these factors prevent the direct transposition of advances in Chemometrics and Molecular Modeling in Medicinal Chemistry (Van de Waterbeemd, 1995; Kubinyi, Folkers & Martin, 1998) into the field of olfaction. Nevertheless, the use of multivariate data analysis approaches can play an important part to improve the knowledge of the molecular descriptor role in olfaction and, then, the implementation of robust mathematical models. Traditional pattern recognition procedures, like Principal Component Analysis (PCA) (Niemi, 1990), Discriminant Analysis (DA) (Hubert, 1994), and Cluster Analysis (Kaufman & Rousseeuw, 1990), and methods pertaining to the field of Artificial Neural Networks, like

Back Propagation Neural Networks (BPNN) (Hecht-Nielsen, 1989) or Kohonen Self-Organizing Maps (SOM) (Kohonen, 2001), are been widely used in the development of several electronic noses (Gardner, Hines & Wilkinson, 1990; Moriizumi, Nakamoto & Sakuraba, 1992; Keller, 1999) and in data analysis of olfactory data sets (Ham & Jurs, 1985; Chastrette, Cretin & Aidi, 1996). These approaches offer different possibilities and objectives. PCA can be considered as being only a projective technique. It is worth using this method when clusters or classes can be visually delineated. DA is really a discriminant technique as it aims to find linear relations in the molecular descriptor hyperspace able to separate different compound categories included in the data set. Both methods, PCA and DA, work correctly if the compounds, belonging to different classes, are grouped in well separated regions, but, in more complex distributions, their classification power becomes poor.

Cluster Analysis offers a first solution to this problem. It consists of obtaining self partitioning of the data, in which each cluster can be identified as a set of compounds clearly delineated regarding the molecular descriptor set involved. Instead of trying to inspect all the compounds in the database to understand and analyze their chemical properties, it is only required to select typical compounds representing each cluster to get a deeper knowledge of the structure of the data base, i.e. of the distribution of the compounds in the derived hyperspace. The main problems related to this method are that:

(i) the number of clusters and the initial positions of the cluster centers can influence the final classification results;
(ii) compound separation is based on a binary notion of belonging, for which a compound located between two clusters is included in only one cluster.

SOM has been considered as an alternative method to overcome the above limitations. It integrates non linearity into the data set, so as to project the molecular descriptor hyperspace onto a two-dimensional map and to preserve the original topology, as the points located near each other in the original space remain neighbors in SOM. This technique has been used to process huge amounts of data in a high-dimensional space (Varfis & Versino, 1992), but, like PCA, it remains an unsupervised projective method. Then, for predictive objectives, SOM has to be combined with another technique, generating a hybrid system that offers an automatic objective map interpretation (Ros, Audouze, Pintore & Chretien, 2000; Audouze, Ros, Pintore & Chretien, 2000).

Contrary to SOM, BPNN is a supervised predictive method. It is able to discriminate any non linearly separable class, relating continuous input and output spaces with an arbitrary degree of accuracy. This method, applied to several fields of chemical database analysis (Zupan & Gasteiger, 1993; Devillers, 1996), has proved to be very efficient in modeling complex data set relationships. However, as in other Artificial Neural Networks techniques, the complexity of the modeling function often prevents extraction of relevant information suitable to explain the model and, therefore, to deliver a better understanding of biological mechanism.

Fuzzy concepts introduced by Zadeh (1977) provide interesting alternative solutions to the classification problems within the context of imprecise categories, in which olfaction can be included. In fact, fuzzy classification represents the boundaries between neighboring classes as a continuous, assigning to compounds a degree of membership of each class. It has been widely used in the field of process control, where the idea is to convert human expert knowledge into fuzzy rules (Hathaway, Bezdek & Pedrycz, 1996), and it should be able to extract relevant structure-activity relationships (SAR) from a database, without *a priori* knowledge.

The aim of this work is to apply a Fuzzy Logic procedure, that we called Adaptive Fuzzy Partition (AFP), to a chemical database derived from olfactory studies (Arctander, 1960; Arctander, 1969), in order to develop a predictive SAR model. The database included 412 compounds associated with an odor appreciation defining the presence or the absence of 4 different olfactory notes. A set of 61 molecular descriptors was examined and the most relevant descriptors were selected by a procedure derived from the Genetic Algorithm concepts (Haupt & Haupt, 1998).

2 MATERIALS AND METHODS

2.1 Compound selection

A database derived from the Arctander's books (Arctander, 1960; Arctander, 1969), including 2620 compounds and 81 olfactory notes, was submitted to a PCA analysis, in order to determine a reduced subset of compounds representing very weakly correlated odors. The relative results allowed to determine a data set of 412 olfactory compounds homogeneously distributed in four classes: animal, camphoraceous, ethereal and fatty odors.

2.2 Molecular descriptors

The reduced data set was distributed in a 61 multidimensional hyperspace derived from a selected set of 61 molecular descriptors. This descriptor set includes topological (Kier & Hall, 1986; Sabljic, 1990), physico-chemical and electronic parameters (Dearden, 1990). In virtual screening, general descriptors have proved a good compromise, from an efficiency point of view, for data mining in large databases. The advantage of these descriptors is their ability to take into account not only the main structural features of each molecule, but also their global behaviors. Then, they should be able to take simultaneously into account the complexity of the olfaction mechanism and the approximation of the odor scale.

Molar refractivity (MR), molar volume (MV) molecular weight (MW) and Van Der Waals volume (VdWV) were used as size descriptors. The shape features of the molecules were characterized by topological indices which account for the ramification degree, the oblong character, etc. The following molecular descriptors were used: 20 molecular connectivity indices ($^{0}\chi$, $^{1}\chi$, $^{2}\chi$, $^{3}\chi_{C}$, $^{3}\chi_{P}$, $^{4}\chi_{P}$, $^{4}\chi_{PC}$, $^{5}\chi_{P}$, $^{5}\chi_{C}$, $^{6}\chi_{P}$, $^{0}\chi^{v}$, $^{1}\chi^{v}$, $^{2}\chi^{v}$, $^{3}\chi^{v}_{C}$, $^{3}\chi^{v}_{P}$, $^{4}\chi^{v}_{P}$, $^{4}\chi^{v}_{PC}$, $^{5}\chi^{v}_{P}$, $^{5}\chi^{v}_{C}$, $^{6}\chi^{v}_{P}$), a series of information content descriptors (IC^{0}, SIC^{0}, CIC^{0}, IC^{1}, SIC^{1},CIC^{1}, IDW), Wiener index (W), centric index (C), Balaban index (J), Gutman index (M2), Platt number (F), counts of paths of lengths 1-4, counts of vertices with 1-4 nearest neighbors. The number of N, O and S atoms in a molecule was also considered. A lipophilicity descriptor represented by the octanol/water partition coefficient (log $P_{oct/water}$) was calculated using the Hansch and Leo method (Hansch & Leo, 1979). Another descriptor was derived from the electronegativity of molecules (E_{M}^{S}) by the Sanderson method (Sanderson, 1976).

2.3 Descriptor selection

To select, amidst the 61 descriptors, the best parameters for classifying the data set compounds, a method based on Genetic Algorithm (GA) (Haupt & Haupt, 1998; Ros, Pintore & Chretien, 2001) concepts was used.

GA, inspired by population genetics, consists of a population of individuals competing on the basis of natural selection concepts. Each individual, or chromosome, represents a trial solution to the problem to be solved. In the context of descriptor selection, the structure of the chromosome is very simple. Each descriptor is coded by a bit (0 or 1) and represents a component of the chromosome. 0 defines the absence of the descriptor, 1 defines its presence. The algorithm proceeds in successive steps called generations. During each generation, the population of chromosomes evolves by means of a “fitness” function (Davis, 1991), which selects them by standard crossover and mutation operators (Kinnear, 1994). The crossover phase takes two chromosomes and produces two new individuals, by swapping segments of genetic material, i.e. bits in this case. Within the population, mutation removes the bits affecting a small probability.

Genetic algorithms are very effective for exploratory search, applicable to problems where little knowledge is available, but it is not particularly suitable for local searches. In the latter case, it is combined with a stepwise approach in order to reach local convergence (Ros, Guillaume, Rabatel & Sevila, 1995). Stepwise approaches are quick and are adapted to find solutions in “promising” areas that have been already identified.

To evaluate the fitness function, a specific index was derived by using a fuzzy clustering method (Ros, Audouze, Pintore & Chretien, 2000). Furthermore, to prevent over-fitting and a poor generalization, a cross validation procedure was included in the algorithm during the selection procedure, by randomly dividing the database into training and test sets. The fitness score of each chromosome is derived from the combination of the scores of the training and test sets.

The following parameters were used in the data processing of the data set of 412 olfactory compounds:
(i) Fuzzy parameters - weighting coefficient = 1.5, tolerance convergence = 0.001, number of iterations = 50, number of clusters = 10.
(ii) Genetic parameters - number of chromosomes = 10, chromosome size = 60 (number of descriptors used), number of crossover points = 1, percentage of rejections = 0.1, percentage of crossovers = 0.8, percentage of mutations = 0.05, time off (10,100), number of generations = 10, ascendant coefficient = 0.02, descendant coefficient = -0.02.

Calculations were performed using proprietary software (Ros, Pintore & Chretien, 2001).

2.4 Adaptive Fuzzy Partition

AFP is a supervised classification method implementing a fuzzy partition algorithm (Lin & Cunninghan, 1994). It models relations between molecular descriptors and chemical activities by dynamically dividing the descriptor space into a set of fuzzy partitioned subspaces. In a first phase, the global descriptor hyperspace is considered and cut into two subspaces where the fuzzy rules are derived. These two subspaces are divided step by step into smaller subspaces until certain conditions are satisfied, namely when:

(i) the number of molecular vectors within a subspace attains a minimum threshold number;
(ii) the difference between two generated subspaces is negligible in terms of chemical activities represented;
(iii) the number of subspaces exceeds a maximum threshold number.

The aim of the algorithm is to select the descriptor and the cut position which allows the maximal difference between the two fuzzy rule scores generated by the new subspaces to be determined. The score is defined by the weighted average of the chemical activity values in an active subspace A and in its neighboring subspaces. If the number of trial cuts per descriptor is defined by N_cut, the number of trial partitions equals (N_cut + 1)N. Only the best cut is selected to subdivide the original subspace.

All the rules created during the fuzzy procedure are considered to establish the model between descriptor hyperspace and biochemical activities. Indicating with $P(x_1, \ldots x_n)$ a molecular vector in a n dimensional descriptor hyperspace, a *rule* for a subspace S_k is defined by (Sugeno & Yasukawa, 1993):

if x_1 is associated with $\mu_{1k}(x_1)$ **and** x_2 with associated to $\mu_{2k}(x_2)$ … **and** x_N is associated with $\mu_{Nk}(x_N)$
$\Rightarrow$ the score of the activity O for P is O_{kP}, (1)

where x_i represents the value of the i^{th} descriptor for the molecule P, μ_{ik} is the membership function related to the descriptor i for the subspace k, and O_{kP} is the biochemical activity value related to the subspace S_k. The "and" of the fuzzy rule is represented by the *Min operator* (Dubois & Prade, 1990) and the membership functions are defined by trapezoidal shapes. These latter functions are based on the boundaries of the subspaces. If the width of a subspace S_k on the i^{th} dimension, after each cut, is represented by w_i, the p and q parameters defining the shape of the trapezoid are calculated by

$$p = \lambda_i w_i \text{ and } q = \nu_i w_i \quad (2)$$

where the parameters λ_i and ν_i vary so that $p \geq 1$ and $q \leq 1$. If $p = 1$ and $q = 1$, the membership function becomes a rectangle.

The global score in the subspace S_k can be represented by

$$O_k = \frac{\sum_{j=1}^{M} (\mathrm{Min}_i^N \mu_{ik}(x_i)_{P_j}).(A_{P_j})}{\sum_{j=1}^{M} (\mathrm{Min}_i^N \mu_{ik}(x_i)_{P_j})} \quad (3)$$

M is the number of molecular vectors in a given subspace, N is the total number of descriptors, $\mu_{ik}(x_i)_{P_j}$ is the fuzzy membership function related to the descriptor i for the molecular vector P_j, and A_{Pj} is the experimental activity of the compound P_j. A classic centroid defuzzification procedure (Gupta & Qi, 1991) is implemented to determine the chemical activity of a new test molecule. All the subspaces k are considered and the general formula to compute the score of the activity O for a generic molecule Pj is

$$O(P_j) = \frac{\sum_{k=1}^{N\_subsp} (\mathrm{Min}_i^N \mu_{ik}(x_i)_{P_j}).(O_k)}{\sum_{k=1}^{N\_subsp} (\mathrm{Min}_i^N \mu_{ik}(x_i)_{P_j})} \quad (4)$$

where N_subsp represents the total number of subspaces.

The following parameters were used to process the data set of 165 pesticide compounds:
maximal number of rules for each chemical activity = 35; minimal number of compounds for a given rule = 4; number of cutting for each axis = 4; p = 1.2 and q = 0.8.

3 RESULTS AND DISCUSSION

3.1 Descriptor selection

Four relevant descriptors were selected by the GA procedure: ${}^5\chi_p$, ${}^3\chi^v_c$, NV_2 and VES. The first three descriptors correspond to topological indices encoding information about molecular structure (Kier & Hall, 1986). Descriptor ${}^5\chi_p$ is a simple molecular connectivity χ index of order 5, in which 5 bonds in the fragment molecule are considered in a path arrangement. All the atoms are considered to be carbon atoms. ${}^3\chi^v_c$ is a valence molecular connectivity χ index of order 3, in which 3 bonds are considered in a cluster arrangement. The values for non-carbon heteroatoms are computed differently regarding the values for identically connected carbon atoms. NV_2, number of vertices of degree 2, corresponds to the number of hybrid groups with two bonded neighbors.

Finally, VES, an electronic index, represents the variance of electronegativity computed by the Sanderson method (Sanderson, 1976).

3.2 AFP model

The AFP model was established on the training set compounds, defining four molecular descriptor - odor relationships, one for each olfactory note. The number of rules implemented in each relationship was dependent on the complexity of the compound distribution regarding a given odor. The animal, camphoraceous, ethereal and fatty odors were respectively represented by seventeen, eighteen, fourteen and twenty-four rules. The number of rules concerning the fatty odor shows that the corresponding relationship was the most difficult to establish. A possible explanation could be found in the fact that only complex combinations of molecular descriptors can represent the distribution of the ethereal compounds, so requiring a high number of rules. Another one can be related to the cutting procedure performed by the algorithm. But this hypothesis is less probable as a different number of cuts, 3, 4 and 5 per axis, leads to similar results.

An example of a rule defining a subspace for the ethereal class is represented by the following definition:

if $0 < x(^5\chi_p) < 0.26$ and $0 < x(^3\chi^v_c) < 0.51$ and $0 < x(NV_2) < 10.8 \Rightarrow$ the ethereal score for a given compound is 1.
This relation demonstrates that the subspace is specially devoted to define the ethereal class.

Figure 1 shows the descriptor composition in each class for each rule. In the ethereal and animal classes, descriptor $^5\chi_p$ plays a major role, being included in all the rules. In the camphoraceous class the most relevant descriptor is represented by $^3\chi^v_c$, while in the fatty category descriptors NV_2 and VES fill the same important role. However, all four descriptors selected seem to be important in modeling the olfactory notes, as they occur in each class.

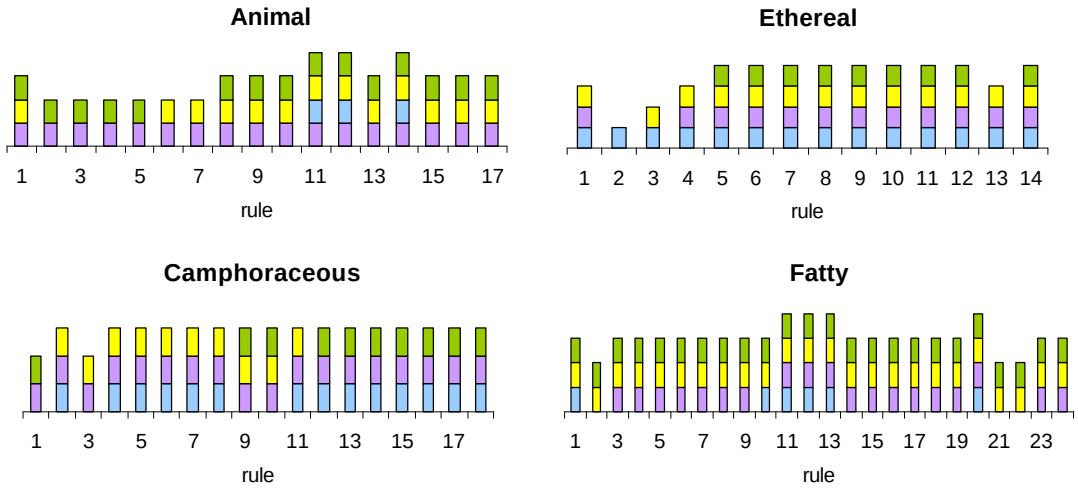

Figure 1. Representation of the descriptors included in each rule for all four categories.
■ = VES; ■ = NV_2; ■ = $^5\chi_p$; ■ = $^3\chi^v_c$.

These first results underline a most important ability of the AFP method, its capacity to solve such complex problems as olfaction, transcribing the molecular descriptor - activity relationships into simple rules that are directly related to the selected descriptors. The contribution of the GA procedure is obviously fundamental: it reduces the amount of information in the input step, making it easier to determine and interpret the model in the following steps.

3.3 Model validation

The AFP model was validated by attempting to predict the olfactory notes of the 102 test compounds. For each compound, the method allows the degrees of membership of the different odors to be determined within a 0 to 1 range. The comparison between predicted and experimental values for all the test set compounds is reported in Table 1. A very good agreement between membership degrees and experimental olfactory scores was found for several compounds. The main olfactory note was correctly predicted for 85 % of the compounds. Amidst the 15 wrong predictions, typed out in bold in Table 1, only 11 were a complete miss, the remaining 4 molecules were weakly predicted.

Table 1. Comparison between predicted and experimental scores for the test set compounds.
A= animal; C = camphoraceous; E = ethereal; F = fatty. The wrong predictions are represented in bold.

ID	Compound	Experimental score				Predicted score			
		A	C	E	F	A	C	E	F
19	Acetone diethylketal	0	0	1	0	0.05	0.17	0.75	0.00
34	Acetyl carbinol	0	0	1	0	0.05	0.01	1.00	0.27
40	5-acetyl-1,1,2,3,3,6-hexamethyl indan	1	0	0	0	0.71	0.24	0.05	0.01
70	Allyl hexanoate	0	0	0	1	0.05	0.01	0.06	0.29
81	Allyl formate	0	0	1	0	0.05	0.01	0.97	0.00
88	Allyl nonanoate	0	0	0	1	0.05	0.01	0.20	0.75

Table 1. (continued)

ID	Compound	Experimental score				Predicted score			
		A	C	E	F	A	C	E	F
109	o-aminoacetophenone	1	0	0	0	0.88	0.11	0.31	0.01
167	o-tertiary-amyl cyclohexyl acetate	0	1	0	0	0.05	1.00	0.05	0.00
169	**Amyl ether**	0	0	1	0	0.05	0.01	0.55	1.00
199	Iso-amyl nitrite	0	0	1	0	0.05	0.07	0.93	0.00
215	Alpha-iso-amyl pyridine	1	0	0	0	0.86	0.41	0.05	0.04
365	Bornyl valerate	0	1	0	0	0.20	1.00	0.05	0.00
405	Tertiary –butyl benzol	0	1	0	0	0.37	1.00	0.17	0.03
435	o-tertiary-butyl cyclohexanone	0	1	0	0	0.05	1.00	0.05	0.01
453	Iso-butyl formate	0	0	1	0	0.05	0.00	0.93	0.00
498	Butyl myristate	0	0	0	1	0.21	0.01	0.05	0.61
515	Iso-butyl phenyl propionate	1	0	0	0	0.57	0.10	0.05	0.07
519	Butyl propionate	0	0	1	0	0.05	0.01	0.93	0.00
545	**Iso-butyl-iso-valerate**	0	0	1	0	0.05	0.95	0.84	0.00
593	Cedrene	0	1	0	0	0.71	1.00	0.05	0.00
609	**Chloroform**	0	0	1	0	0.05	0.95	0.79	0.00
668	Alpha-citronellidene cyclopentanone	0	1	0	0	0.20	0.42	0.05	0.01
723	p-cresol propyl ether	1	0	0	0	0.64	0.04	0.05	0.13
728	p-cresyl butyrate	1	0	0	0	0.76	0.08	0.05	0.14
737	m-cresyl phenylacetate	1	0	0	0	0.95	0.01	0.05	0.08
743	p-cresyl valerate	1	0	0	0	0.72	0.00	0.05	0.36
788	2-cyclohexyl cyclohexanol	0	1	0	0	0.08	1.00	0.05	0.21
817	Cyclyl acetate	0	1	0	0	0.05	1.00	0.05	0.00
839	Decane nitrile	0	0	0	1	0.02	0.01	0.17	1.00
845	7-decenolactone	0	0	0	1	0.05	0.01	0.05	0.29
891	Di-n-butyryl	0	0	0	1	0.18	0.01	0.00	0.97
894	**Dicrotyl sulfide**	1	0	0	0	0.00	0.01	0.78	0.58
901	Diethyl citronellol	0	1	0	0	0.15	0.96	0.16	0.08
911	Diethyl methylmalonate	0	0	1	0	0.05	0.14	0.85	0.00
976	Dill-iso-apiol	0	1	0	0	0.71	0.94	0.05	0.31
989	**Dimethyl benzyl carbinol**	1	0	0	0	0.30	0.95	0.05	0.01
998	Dimethyl carbonate	0	0	1	0	0.05	0.00	1.00	0.00
1023	3,5-dimethyl-2-iso-hexylcyclohexanone	0	1	0	0	0.05	0.96	0.05	0.01
1070	**Dioxane**	0	0	1	0	0.05	0.45	0.41	0.00
1117	Alpha-dodecyl-gamma-butyrolactone	0	0	0	1	0.05	0.01	0.05	0.40
1138	Ethyl acetoacetate	0	0	1	0	0.05	0.01	0.93	0.00
1178	Ethyl butyrate	0	0	1	0	0.05	0.01	0.93	0.00
1209	Ethyl decyl ether	0	0	0	1	0.05	0.01	0.04	1.00
1226	Ethylene oxide	0	0	1	0	0.00	0.00	1.00	0.00
1273	Beta-ethyl indole	1	0	0	0	0.89	0.01	0.05	0.17
1277	Ethyl laurate	0	0	0	1	0.05	0.01	0.04	1.00
1278	Ethyl levulinate	0	0	1	0	0.05	0.01	0.91	0.00
1356	Ethyl tiglate	0	0	1	0	0.05	0.00	0.70	0.00
1385	Fenchone	0	1	0	0	0.05	0.87	0.05	0.00
1396	Formaldehyde diethyl acetal	0	0	1	0	0.05	0.21	0.41	0.00
1415	Furfuryl acetate	0	0	1	0	0.05	0.01	0.71	0.00
1473	**Guaiacol allylether**	0	0	0	1	0.61	0.01	0.05	0.30
1505	Cis-4-hepten-1-al	0	0	0	1	0.00	0.01	0.15	0.79
1532	p-heptyl cyclohexanone	0	1	0	0	0.05	0.38	0.05	0.36
1540	**Heptyl formate**	0	0	0	1	0.05	0.01	0.93	0.85

Table 1. (continued)

ID	Compound	Experimental score				Predicted score			
		A	C	E	F	A	C	E	F
1550	**Heptyl laurate**	0	0	0	1	0.66	0.01	0.05	0.61
1592	Hexanal	0	0	0	1	0.02	0.01	0.00	0.94
1625	Cis-3-hexenyl propionate	0	0	0	1	0.05	0.01	0.06	0.37
1654	n-hexyl-4-cyclohexanone	0	1	0	0	0.05	0.38	0.05	0.36
1743	2-hydroxy-4,4-dimethyl-4-cyclohexyl butane	0	1	0	0	0.05	1.00	0.05	0.00
1746	4-hydroxy-2-hexenylacetate	0	0	1	0	0.05	0.01	0.60	0.04
1787	Jasmine lactone	0	0	0	1	0.05	0.01	0.05	0.29
1855	Mesitylene	0	0	1	0	0.27	0.00	0.62	0.15
1872	6-methoxy dicyclopentadiene aldehyde	0	1	0	0	0.71	1.00	0.05	0.00
1896	Methyl acetyl cyclopentane	0	0	1	0	0.05	0.01	0.72	0.01
1939	Alpha-methyl butyraldehyde	0	0	1	0	0.02	0.03	1.00	0.00
1966	7-methyl coumarin	1	0	0	0	0.79	0.00	0.05	0.01
2020	Methyl ethyl ketone	0	0	1	0	0.02	0.01	1.00	0.00
2052	Nonan-2-one	0	0	0	1	0.05	0.01	0.00	0.99
2079	Methyl-3-hydroxyhexanoate	0	0	1	0	0.05	0.01	0.93	0.00
2102	**Methyl-3-methoxy-2-methyl aminobenzoate**	1	0	0	0	0.06	0.11	0.05	0.00
2114	Methyl-5-methyl-n-hex-1-yne carbonate	0	0	0	1	0.06	0.00	0.00	0.91
2139	Methyl nonylenate	0	0	0	1	0.05	0.01	0.07	0.89
2152	Methyl-7-octynoate	0	0	0	1	0.05	0.01	0.00	0.93
2194	2-methyl-2-phenyl hexanone-4	0	1	0	0	0.25	0.81	0.05	0.01
2226	Methyl propyl ketone	0	0	1	0	0.02	0.01	1.00	0.00
2236	7-methyl quinoline	1	0	0	0	0.94	0.00	0.05	0.00
2263	Methyl valerate	0	0	1	0	0.05	0.01	0.93	0.00
2303	Beta-naphtyl phenylether	1	0	0	0	1.00	0.00	0.05	0.00
2307	**Beta-naphtyl ethylalcohol**	0	1	0	0	1.00	0.00	0.05	0.00
2343	Nonanal	0	0	0	1	0.03	0.01	0.00	1.00
2352	Nonan-3-one-1-yl acetate	0	0	1	0	0.05	0.01	0.60	0.30
2362	Nonenyl nitrile	0	0	0	1	0.00	0.01	0.53	1.00
2408	**n-octyl acetate**	0	0	0	1	0.05	0.01	0.78	0.43
2419	p-octyl cyclohexanone	0	1	0	0	0.05	0.38	0.05	0.36
2542	**Phenylethyl methyl ethyl carbonyl acetate**	1	0	0	0	0.16	0.42	0.05	0.01
2603	3-phenyl propyl undecylenate	0	0	0	1	0.50	0.01	0.05	0.61
2622	Pinocamphone	0	1	0	0	0.05	0.72	0.05	0.00
2653	Propione	0	0	1	0	0.02	0.01	1.00	0.00
2676	Propyl butyrate	0	0	1	0	0.05	0.01	0.93	0.00
2688	Alpha-propyl cinnamic aldehyde	1	0	0	0	0.88	0.01	0.05	0.00
2690	p-iso-propyl cyclohexane ethanol	0	1	0	0	0.05	0.63	0.05	0.00
2714	Iso-propyl heptyl ether	0	0	1	0	0.05	0.07	0.89	0.14
2760	Iso-propyl-iso-valerate	0	0	1	0	0.05	0.16	0.93	0.00
2833	Santene	0	1	0	0	0.05	0.54	0.05	0.00
2847	**Skatolene**	1	0	0	0	0.00	0.42	0.05	0.01
2923	**Tetrahydronaphtyl ethyl alcohol**	0	1	0	0	0.58	0.00	0.05	0.21
2981	Tricyclopentadiene	0	1	0	0	0.71	0.80	0.05	0.00
3026	Undecamethylene carbonate	0	1	0	0	0.05	0.45	0.05	0.19
3041	Undecenyl acetate	0	0	0	1	0.05	0.01	0.08	1.00
3090	Vinyl acetate	0	0	1	0	0.05	0.01	1.00	0.00
3101	Zingerone	1	0	0	0	0.84	0.00	0.05	0.01

In Table 2, the validation ratios, the Root Mean Square (RMS) errors and the correlation coefficients for each class and for all the molecules are reported. RMS error is calculated by the formula:

$$\text{Error} = \sqrt{\left[\frac{1}{N}\sum_{i=1}^{N}(\text{exsc(i)} - \text{prsc(i)})^2\right]} \quad (5)$$

where N is the total number of molecules, and exsc(i) and prsc (i) represent the experimental and predicted scores for molecule i, respectively.

Considering the complexity of the olfactory field, the prediction power of the model is good. Furthermore, the errors and the correlation coefficients could probably be improved if the experimental scores were not limited to values 0 and 1, but included in the range [0,1].

Table 2. Statistical values defining the robustness of the AFP model. Error is represented by RMS score.

	Training set			Test set		
	Validation Ratio (%)	Error	Correlation Coefficient	Validation ratio (%)	Error	Correlation coefficient
Animal	77.6	0.25	0.79	73.7	0.31	0.63
Camphoraceous	92.2	0.21	0.88	92.0	0.27	0.79
Ethereal	88.8	0.25	0.84	88.0	0.21	0.89
Fatty	86.6	0.27	0.79	84.0	0.24	0.82
Total value	88.7	0.25	0.83	85.2	0.26	0.80

The statistical criteria concerning the training set are reported in Table 2. They show that the prediction values for the training and test sets are comparable, demonstrating the robustness and generalized behavior of the proposed model. In fact, the AFP method includes the possibility of errors for training set compounds, as it memorizes the main characteristics of the compound distribution without considering all the details. Using the same original experimental data, the only way to increase the classification power consists in adding new relevant descriptors; improving model performance for the training set by adding further rules is useless, as the test set compounds are not better predicted.

4 CONCLUSION

Data Base Mining (DBM) algorithms, based upon molecular diversity analysis, are becoming a must for pharmaceutical companies in the search for new leads. They allow the automated classification of chemical databases, but the huge amounts of information provided by the large number of molecular descriptors tested is difficult to exploit. Then, new tools have to be developed to give a user-friendly representation of the compound distribution in the descriptor hyperspace.

Furthermore, the difficulty of data mining in olfaction databases is amplified by the fact that one compound can have different odors and its activity is usually expressed in a qualitative way. Another source of complexity derives from the fact that one receptor can recognize different chemical determinants and the same compound can be active on different receptors.

Fuzzy logic methods, developed to mimic human reasoning in its ability to produce correct judgements from ambiguous and uncertain information, can provide interesting solutions in the classification of olfactory databases. In fact, these techniques should be able to represent the “fuzziness” linked to an expert’s subjectivity in the characterization of the odorous notes, computing intermediate values between absolutely true and absolutely false for each olfactory category. These values are named degrees of membership and are ranged between 0.0 and 1.0.

In this work, a new procedure, the Adaptive Fuzzy Partition (AFP) algorithm, was applied to a data set of 412 olfactory molecules, divided into animal, camphoraceous, ethereal and fatty compounds. This method consists of modeling molecular descriptor – activity relationships by dynamically dividing the descriptor hyperspace into a set of fuzzy subspaces. A large number of molecular descriptors was tested and the best

ones were selected with help of an innovative procedure based on Genetic Algorithm concepts. The ability of the proposed tool to model the 4 olfactory classes was validated after separating the 412 compounds into a training set and a test set, including 310 and 102 molecules, respectively. The main experimental olfactory notes were predicted correctly for 83% of the test set compounds. Furthermore, the method showed its ability to lead to generalist models and simple rules describing SAR relationships. These preliminary results show that the proposed methods are worth investigating more thoroughly and testing with a large chemical database. Work is underway to entirely exploit the database derived from the Arctander's books, which involves a sensible increase in the number of compounds and complex olfactory notes to be treated.

5 ACKNOWLEDGEMENTS

Our gratitude goes to Mrs Roudnitska and the E. Roudnitska Foundation for their support and the grant to K.A.. Our thanks also go to Professor M. Chastrette for his kind comments and encouragement for this work.

6 REFERENCES

Arctander, S. (1960) *Perfume and Flavor Materials of Natural Origin*, Elizabeth, USA: Steffen Arctander.

Arctander, S. (1969) *Perfume and Flavor Chemicals: Aroma Chemicals*, Montclair, USA: Steffen Arctander.

Audouze, K., Ros, F., Pintore, M., & Chretien, J.R. (2000) Prediction of odours of aliphatic alcohols and carbonylated compounds using fuzzy partition and self organising maps (SOM). *Analusis 28*(7), 625-632.

Buck, L., & Axel, R. (1991) A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell 65*(1), 175-187.

Chastrette, M. (1998) Data management in olfaction studies. *SAR and QSAR in Environmental Research 8*(3-4), 157-181.

Chastrette, M., Cretin, D., & Aidi, C.E. (1996) Structure-Odor Relationships: Using Neural Networks in the Estimation of Camphoraceous or Fruity Odors and Olfactory Thresholds of Aliphatic Alcohol. *Journal of Chemical Information and Computer Sciences 36*(1), 108-113.

Chastrette, M. & Zakarya, D. (1991) Molecular structure and smell. In Laing, D.G, Doty R.L & Breipohl W. (Eds.), *The human sense of smell*, New York: Springer-Verlag.

Davis, L. (1991) *Handbook of Genetic Algorithms*, New York: Van Nostrand Reinhold.

Dearden, J.C (1990) Physico-chemical descriptors. In Karcher, W. & Devillers, J. (Eds.), *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*, Dordrecht: Kluwer Academic.

de Mello Castanho Amboni, R.D., da Silva Junkes, B., Yunes, R.A., & Heinzen V.E. (2000) Quantitative structure-odor relationships of aliphatic esters using topological indices. *Journal of Agricultural and Food Chemistry 48*(8), 3517-3521.

Devillers, J. (1996) *Neural Networks in QSAR and Drug Design*, New York: Academic Press.

Dubois, D. & Prade, H. (1990) An introduction to possibilistic and fuzzy logics. In Shafer, G. & Pearl, J. (Eds.), *Readings in Uncertain Reasoning*, San Francisco: Morgan Kaufmann.

Gardner, J.W., Hines, E.L., & Wilkinson, M. (1990) Application of Artificial Neural Networks to an Electronic Olfactory System. *Measurement Science and Technology 1*(5), 446-451.

Gupta, M.M., & Qi, J. (1991) Theory of T-norms and fuzzy inference methods. *Fuzzy Sets and Systems 40*(3), 431-450.

Ham, C.L., & Jurs P.C. (1985) Structure-activity studies of musk odorants using pattern recognition: monocyclic nitrobenzenes. *Chemical Senses 10*(4), 491-505.

Hansch, C. & Leo, A.W (1979) *Substituent constants for correlation analysis in chemistry and biology*, New York: Wiley.

Hathaway, R.J., Bezdek, J.C., & Pedrycz, W. (1996) A Parametric Model for Fusing Heterogeneous Fuzzy Data. *IEEE Transactions on Fuzzy Systems 4*(3), 270-281.

Haupt, R.L & Haupt, S.E (1998) *Practical Genetic Algorithms*, New York: Wiley Inter-science.

Hecht-Nielsen, R. (1989) Theory of the backpropagation neural network. *Proceedings of the International Joint Conference on Neural Networks* (pp. 593-605). Washington DC, USA.

Hubert, C.J (1994) *Applied Discriminant analysis*, New York: Wiley Interscience.

Kaufman, L. & Rousseeuw, P.J (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*, New York: Wiley Interscience.

Keller, P. (1999) Physiologically Inspired Pattern Recognition for Electronic Noses. *Proceedings of the SPIE 3722*(13), 144-153.

Kermani, B.G., Schiffman, S.S., & Nagle, H.T. (1999) Using neural networks and genetic algorithms to enhance performance in an electronic nose. *IEEE Transactions on Biomedical Engineering 46*(4), 429-39.

Kier, L.B & Hall, L. H (1986) *Molecular Connectivity in structure analysis*, New York: Wiley.

Kinnear, K.E (1994) *Advances in Genetic Programming*, Cambridge: MIT Press.

Kohonen, T. (2001) *Self-Organizing Maps*, Berlin: Springer-Verlag.

Kubinyi, H., Folkers, G. & Martin, Y.C (Eds.) (1998) *3D QSAR in Drug Design. Recent Advances*, Dordrecht: Kluwer Escom.

Lin, Y., & Cunninghan, G.J. (1994) Building a Fuzzy System from Input-Output Data. *Journal of Intelligent and Fuzzy Systems 2*(3), 243-250.

Malnic, B., Hirono J., Sato T., & Buck L.B. (1999) Combinatorial receptor codes for odors. *Cell 96*(7), 713-723.

Manley, C.H. (1993) Psychophysiological effect of odor. *Critical Reviews in Food Science and Nutrition 33*(1), 57-62.

Moriizumi, T., Nakamoto T., & Sakuraba Y. (1992) Pattern Recognition in Electronic Noses by Artificial Neural Network Models. In Gardner, J.W & Bartlett, P.N (Eds.), *Sensors and Sensory Systems for an Electronic Nose*, Amsterdam: Kluweer Academic.

Niemi, G.J (1990) Multivariate analysis and QSAR: Applications of principal component analysis. In

Karcher, W. & Devillers, J. (Eds.), *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*, Dordrecht: Kluwer Academic.

Qureshy, A., Kawashima, R., Imran, M.B., Sugiura, M., Goto, R., Okada, K., Inoue, K., Itoh, M., Schormann, T., Zilles, K., & Fukuda, H. (2000) Functional mapping of human brain in olfactory processing: a PET study. *Journal of Neurophysiology 84*(3), 1656-1666.

Ros, F., Audouze, K., Pintore, M., & Chretien, J.R. (2000) Hybrid System for Virtual Screening: Interest of Fuzzy Clustering Applied to Olfaction. *SAR and QSAR in Environmental Research 11*(3-4), 281-300.

Ros, F., Guillaume, S., Rabatel, G., & Sevila F. (1995) Recognition of overlapping particles in granular product images using statistics and neural networks. *Food Control 6*(1), 37-43.

Ros, F., Pintore, M., & Chrétien, J.R. (2001) Molecular descriptor selection combining genetic algorithms and fuzzy logic: application to data base mining procedure. *Chemometrics and Intelligent Laboratory Systems* (in press).

Rossiter, K.J. (1996) Structure-Odor Relationships. *Chemical Review 96*(8), 3201-3240.

Sabljic, A. (1990) Topological indices and environmental chemistry. In Karcher, W. & Devillers, J. (Eds.), *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*, Dordrecht: Kluwer Academic.

Sanderson R. (1976) *Chemical bonds and bond energy*, New York: Academic Press.

Sugeno, M., & Yasukawa, T.(1993) A fuzzy-logic-based approach to qualitative modeling. *IEEE Transactions On Fuzzy Systems 1I*(1), 7-31.

Van de Waterbeemd, H. (Ed.) (1995) *Chemometric methods in molecular design*. Weinheim: VCH.
Varfis, A., & Versino, C. (1992) Clustering of socio-economic data with kohonen maps. *Neural Network World 2*(6), 813-834.

Zadeh, L.A (1977) Fuzzy sets and their applications to classification and clustering. In J. Van Ryzin J. (Ed.), *Classification and Clustering* (pp. 251-299). New York: Academic Press.

Zupan, J. & Gasteiger, J. (1993) *Neural networks for Chemists: An Introduction*, Weinheim: VCH.