Comparing the Information Content of Two Large Olfactory Databases

Marco Pintore,[†] Christophe Wechman,[‡] Gilles Sicard,[§] Maurice Chastrette,[||] Nicolas Amaury,[†] and Jacques R. Chretien*[,†]

BioChemics Consulting, 16 Rue L. de Vinci, F-45074 Orléans, France, University of Orléans, LBLGC/CBI, UPRES EA 1207, F-45067 Orléans, France, University Claude Bernard, Lab. Neurosciences & Systèmes Sensoriels, 50 Av. Tony Garnier, 69366 Lyon Cedex 07, France, and Fondation Edmond Roudnitska, 40 Avenue Hoche, 75008 Paris

Received June 20, 2005

The expert's subjectivity in establishing an olfactory description can produce wide discrepancies in different databases listing the odor profile of identical compounds. A representative example is obtained by comparing the odorous compounds included in the "Perfumery Materials and Performance 2001" (PMP2001) database and in Arctander's books (1960 and 1969). To better assess this problem, classification models obtained by using the adaptive fuzzy partition method were established on subsets of these databases distributed into the same olfactory classes. The robustness and the prediction power of these models give a powerful criterion for evaluating the "quality" of their information content and for deciding which is the most trustable database.

1. INTRODUCTION

Olfaction remains a permanent challenge in academic and industrial research, and its economic impact is proportional to its complexity. The latter is related to different factors: (1) a huge number of receptors is involved in the olfactory processes,[1−3] (2) knowledge related to the 3D structure of these receptors is still missing, (3) different types of chemical compounds can affect the same receptor, and (4) one compound can simultaneously exhibit different odors.

All these critical issues make it quite difficult to transpose to the olfaction field the progress achieved in the past few years by computer-aided molecular design, above all in medicinal chemistry.[4−6] Nevertheless, the use of data mining (DM) approaches can play an important part in better knowing the role of chemical and physical parameters in olfaction and, then, in the implementation of robust predictive models.[7,8] These methods analyze the molecular diversity of biochemical databases and obtain automated classification by establishing structure−activity relationships (SARs) with the help of multivariate analysis algorithms. Traditional DM procedures, like principal component analysis,[9] discriminant and cluster analysis,[10,11] and methods pertaining to the field of artificial neural networks,[12] have been widely used in the development of several electronic noses[13−15] and in the data analysis of olfactory data sets.[16,17] These approaches offer different possibilities and objectives, but also many limitations.

Recently, we have shown that methods derived from fuzzy logic[18] (FL) deliver very interesting results in the classification of olfactory data sets.[19,21] Actually, FL, based on the possibility of handling the "concept of partial truth", provides solutions to problems within the context of imprecise categories, which olfaction can be included in. The main ability of a fuzzy classification consists of representing the boundaries between neighboring activity classes as continuous, assigning to compounds a degree of membership of each class within a 0−1 range. More particularly, the best models in predicting olfactory notes were obtained by using the adaptive fuzzy partition[21,22] (AFP) method, a recursive partitioning method derived from FL concepts. This method was successfully applied on a data set of 412 olfactory molecules, divided into animal, camphoraceous, ethereal, and fatty compounds. The ability of the proposed tool to model the four 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.[21] Furthermore, the method showed its ability to lead to generalist models and simple rules describing SAR relationships.

A main problem in establishing predictive models from an olfactory data set, regardless of the DM procedure used, is represented by the expert's subjectivity in assessing odors. Much progress has been made in the knowledge of physiological and psychological factors influencing the evaluation of olfaction,[23,24] but it is not sufficient to clearly discriminate between objectivity and subjectivity in the expert's characterization. This could result in differences between trustable databases listing the odor profiles of identical compounds. An example of this problem is represented by the comparison of two large databases, "Perfumery Materials and Performance 2001" from Bacis, for one,[25] and that derived from the works of Arctander;[26,27] these data sets will be hereafter called "PMP2001" and "Arctander", respectively. After suppressing all oils, mixtures, polymers, unknown structures, and molecules associated with unspecific notes (fruity, floral, etc.), each of the databases included about 2600 compounds. A straight comparison between them isolated 923 common

* Author to whom correspondence should be addressed. Tel: (33) 2 38 41 70 76. Fax: (33) 2 38 41 72 21. E-mail: jacques.chretien@biochemics-consulting.com.

[†] BioChemics Consulting.

[‡] University of Orléans.

[§] University Claude Bernard.

[||] Fondation Edmond Roudnitska.

10.1021/ci0502505 CCC: $33.50 © 2006 American Chemical Society
Published on Web 09/16/2005

Table 1. Simplified Explanation of the Olfactory Notes Included in the PMP2001 and Arctander Data Sets

animal	odors recalling the animal kingdom (so as their excrements), but also fragrances related to organic musk
camphoraceous	compounds characterized by the very typical odor of camphor
hesperide	fragrances associated with citrus fruits, i.e., orange, lemon, mandarin, bergamot, etc.
humus	odors concerning wood and undergrowth, such as moss, mushrooms, molds.
balsamic	sickly sweet smells regrouping incense, resin, vanilla, etc. The word comes from the balsam tree, characterized by a very distinct smelling resin
spicy	odors evoking spices such as pepper, clove, nutmeg
ethereal	compounds smelling like ether

compounds; 40% of these compounds were associated with totally different olfactory profiles, in which no odor included in the Arctander description could be found in the PMP2001 one. Only 2% of them exhibited the same notes, whereas 58% had at least one same odor present in each database.

This comparison indicates how different the databases are but gives no criterion for evaluating which is the better one. The objective of this work consisted, then, in applying the DM strategy based on the AFP method to compare the quality of the databases with the help of robust statistical databases, even if the collected raw information issued from the expert panels are based on subjective data. The comparison was performed by selecting two pairs of subsets, derived from Arctander and PMP2001, which regrouped a few hundred compounds. Each pair of subsets included four classes that associated the same olfactory descriptors for both Arctander and PMP2001 descriptions. A large set of molecular descriptors was computed on the 2D structures, and the most relevant parameters were selected by a procedure combining the genetic algorithm concepts and a stepwise technique.[28] Then, structure−odor relationships were established with help of AFP; the robustness and the prediction power of the classification models established on the Arctander and PMP2001 data sets give a powerful criterion for evaluating the "quality" of their information content.

2. MATERIALS AND METHODS

2.1. Compound Selection. The databases here called Arctander and PMP2001, derived respectively from Arctander's books[26,27] and "Perfumery Materials and Performance 2001",[25] include 2589 and 2610 compounds subdivided into 29 and 35 olfactory notes. A first pair of data sets (Selection A) was selected from these databases by isolating all compounds that were associated with one of the following odors as the main olfactory note: animal, camphoraceous, hesperide, and humus. More detail about the meaning of these notes can be found in Table 1. These odors were selected to include a similar number of compounds in each class for each of the two data sets. The total number of compounds (262 and 207); their distribution in the four classes; and the subdivision into training, validation, and test sets are reported in Table 2. The test set included molecules that were never used for developing the model. The validation set was used during the development of the model, based on the training set, to optimize the parameters and to validate the model.

Table 2. Compound Repartition of the Two Pairs of Olfactory Data Sets, Derived from the PMP2001 and Arctander Databases in the Training, Validation, and Test Sets

	PMP2001			Arctander		
odor	training	validation	test	training	validation	test
	Selection A (207 and 262 compounds)					
animal	30	11	11	50	16	16
camphoraceous	19	7	7	32	10	10
hesperide	40	14	14	28	9	9
humus	32	11	11	50	16	16
total	**121**	**43**	**43**	**160**	**51**	**51**
	Selection B (436 and 403 compounds)					
animal	32	10	10	52	15	15
balsamic	85	36	36	91	27	27
spicy	69	20	20	40	11	11
ethereal	74	22	22	76	19	19
total	**260**	**88**	**88**	**259**	**72**	**72**

Later, a second pair of data sets (Selection B) was selected from the PMP2001 and Arctander databases, including a larger number of compounds, 436 and 403, respectively. This second selection associates the compounds with the following odors: animal, balsamic, spicy, and ethereal (Table 1). The compounds' subdivision into training, validation, and test sets is also reported in Table 2.

2.2. Molecular Descriptor Selection. General molecular descriptors have proved to be a good compromise for data mining in large databases, as they are able to account for the main structural feature of each molecule. All olfactory data sets were then distributed within a hyperspace defined by 167 descriptors, computed on 2D structures by the MDL QSAR software.[29] These descriptors included constitutional, informational, topological, physicochemical, and electronic parameters. More details about the different molecular descriptors used can be found in ref 22.

To select the best parameters, amongst the 167 descriptors, for classifying the data set compounds, a hybrid selection algorithm (HSA) was used, on the basis of genetic algorithm (GA) concepts.[28,30] GA, inspired by population genetics, is very effective for exploratory searches, applicable to problems where little information is available, but it is not particularly suitable for local searches. So, a stepwise approach was also implemented in combination with GA in order to reach local convergence,[28,31] as it is quick and adapted to find solutions in "promising" areas already identified.

Finally, a specific classification index was derived by the fuzzy clustering method[32] to evaluate the fitness function of HSA. This index has the advantage of being calculated quite quickly and giving an estimate of the descriptor relevance also by analyzing complex molecular distributions, in which finding separating edges between the different categories is difficult.

Furthermore, to prevent overfitting and a poor generalization, a cross validation procedure was included in the algorithm during the selection procedure, randomly dividing the data set into training and test sets. The fitness score of each set of descriptors derives from the combination of the scores of both training and test sets.

All details about the strategy proposed for molecular descriptor selection and the proprietary software used can be found in ref 28.

The following parameters were used to process the olfactory data set by HSA: (i) Fuzzy parameters: weighting coefficient = 1.5; tolerance convergence = 0.001; number of iterations = 30; cluster number = 6. (ii) Genetic parameters: chromosome number = 10; chromosome size = 167 (number of descriptors used); initial active descriptors in each chromosome = 8; crossover point number = 1, percentage of rejections = 0.1, percentage of crossover = 0.8, percentage of mutation = 0.05, number of generations = 10. (iii) Stepwise parameters: ascending coefficient = 0.02; descending coefficient = −0.02.

2.3. Self-Organizing Maps (SOMs). SOM[33] is a nonlinear mapping technique that gives a 2D space representation of a given set of points from a multidimensional space derived from a series of molecular descriptors. Each point of this set is related to a SOM node, which is characterized by *N* weighted connections varying between 0 and 1.

Training SOM consists of rearranging the layer nodes by gradually adjusting their weights. After selecting a first hyperspace point, the distances between its coordinates and each node of the SOM layer are calculated. The nearest node is called "winner" and the hyperspace point is "projected" on this node of the map. Then, the weights of the winning node and its neighbors are modified according to the equation

$$w_{ij}(t+1) = w_{ji}(t) + \alpha(t)\,\gamma(t,r)\,[x_j - w_{ij}(t)] \qquad (1)$$

where x_j is the component *j* of input vector *x*, w_{ij} represents the weight vector of the node *i* for the descriptor *j*, *t* and $\alpha(t)$ are respectively the iteration number and the learning rate, and $\gamma(t,r)$ is the triangular neighborhood function depending on the iteration number and the distance *r* between the node *i* and the winning unit.

The learning rate $\alpha(t)$ is linearly decreased during the training process from $\alpha(0)$ to zero. The triangular function $\gamma(t,r)$ works on the whole map, and it is discretely decreased with an increase in the distance and the number of iterations.

The same procedure is successively repeated for all the hyperspace vectors, and each point is associated with a node in the SOM layer. The points that are close in the descriptor hyperspace remain close in the SOM layer, occupying the same nodes or the neighboring ones. When SOM is applied on a chemical data set, the maps can then reveal similar compounds or natural regroupings, if the Euclidean distance is accepted as a similarity measure.

The calculations were performed using proprietary software and the following parameters: number of columns = 10; number of rows = 10; coefficient γ for bias calculations = 0.01; number of iterations for the training phase = 50 000; coefficient β for frequency calculations = 0.001.

2.4. Adaptive Fuzzy Partition (AFP). AFP is a supervised classification method implementing a fuzzy partition algorithm,[34] and it was already fully presented[22] and validated elsewhere.[21,22,35−37] It models relations between molecular descriptors and activities by dynamically dividing the descriptor space into a set of fuzzy partitioned subspaces defined by fuzzy rules. The aim of the algorithm is then to select the descriptor and the cut position, which allow retrieval of the maximal difference between the two fuzzy rule scores generated by the new subspaces. The score is determined by the weighted average of the activity values in an active subspace A and in its neighboring subspaces.

Let us assume that the working space is an *n*-dimension hyperspace defined by *n* molecular descriptors; each dimension *i* can be partitioned into *L* intervals I_{ij}, where *j* represents an interval in the partition selected. Indicating with $P(x_1, ..., x_n)$ a molecular vector in an *n*-dimensional hyperspace, a *rule* for a subspace S_k, derived by combining *n* intervals I_{ij}, is defined by[38]

$$\begin{gathered}\text{if } x_1 \text{ is associated with } \mu_{1k}\\ (x_1)\ \textbf{and}\ x_2 \text{ is associated with } \mu_{2k}\\ (x_2)\ \textbf{and}\ x_N \text{ is associated with } \mu_{Nk}\\ (x_N) \Longrightarrow \text{then the score of the activity } O \text{ for } P \text{ is } O_{kP}\end{gathered} \qquad (2)$$

where x_i represents the value of the *i*th descriptor for the molecule *P*, μ_{ik} is a trapezoidal membership function related to the descriptor *i* for the subspace *k*, and O_{kP} is the activity value related to the subspace S_k. The "and" of the fuzzy rule is represented by the *Min operator*,[39] which selects the minimal value amongst all the μ_{ik} components.

All the rules created during the fuzzy procedure are considered to establish the model between the descriptor hyperspace and activities. The global score in the subspace S_k can be represented by

$$O_k = \frac{\sum_{j=1}^{M}[\operatorname*{Min}_{i}^{N} \mu_{ik}(x_i)_{P_i}](A_{P_j})}{\sum_{j=1}^{M}[\operatorname*{Min}_{i}^{N} \mu_{ik}(x_i)_{P_j}]} \qquad (3)$$

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_{P_j} is the experimental activity of the compound P_j. A classic centroid defuzzification procedure[40] 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 P_j is

$$O(P_j) = \frac{\sum_{k=1}^{N\_\text{subsp}}[\operatorname*{Min}_{i}^{N} \mu_{ik}(x_i)_{P_j}](O_k)}{\sum_{k=1}^{N\_\text{subsp}}[\operatorname*{Min}_{i}^{N} \mu_{ik}(x_i)_{P_j}]} \qquad (4)$$

where *N*_subsp represents the total number of subspaces.

The following AFP parameters were used to process the olfactory data sets:

maximal number of rules for each olfactory note = 30; minimal number of compounds for a given rule = 5; maximal number of cuts for each axis = 9.

2.5. Validation Tools. The robustness of the AFP models was evaluated on the training set by two main techniques, leave-several-out (LSO) and the *Y*-randomization test.[4,41] LSO is a cross-validation method consisting in leaving out a given number of compounds from the training set and rebuilding the model, which is then used to predict the compounds left out. A classification LSO coefficient is

Table 3. List of the Most Relevant Descriptors Selected by HSA for Classifying the PMP2001 and Arctander Data Sets Associated with Selection A (see Table 2)

symbol	definition	descriptor family
	PMP2001	
SdssC	sum of all (═C<) E-state values	electrotopological
SaaN	sum of all (⋯N⋯) E-state values	electrotopological
SsOH	sum of all (−OH) E-state values	electrotopological
X0	simple 0 order χ index	topological
Xp9	simple ninth-order path χ index	topological
Nxp6	number of paths of length 6 (number of edges)	topological
Nelem	number of chemical elements	constitutional
log P	lipophilicity at pH = 7	physicochemical
	Arctander	
SaaCH_acnt	count of all (⋯CH⋯) groups	constitutional
Xp8	simple eighth-order path χ index	topological
Dxv1	difference valence first-order χ index	topological
Dxvp8	difference valence eighth-order path χ index	topological
Hmin	smallest atom hydrogen E-state value	electrotopological
Idcbar	Bonchev−Trinajsti mean information content	information
NumHBa	number of hydrogen bond acceptors	constitutional

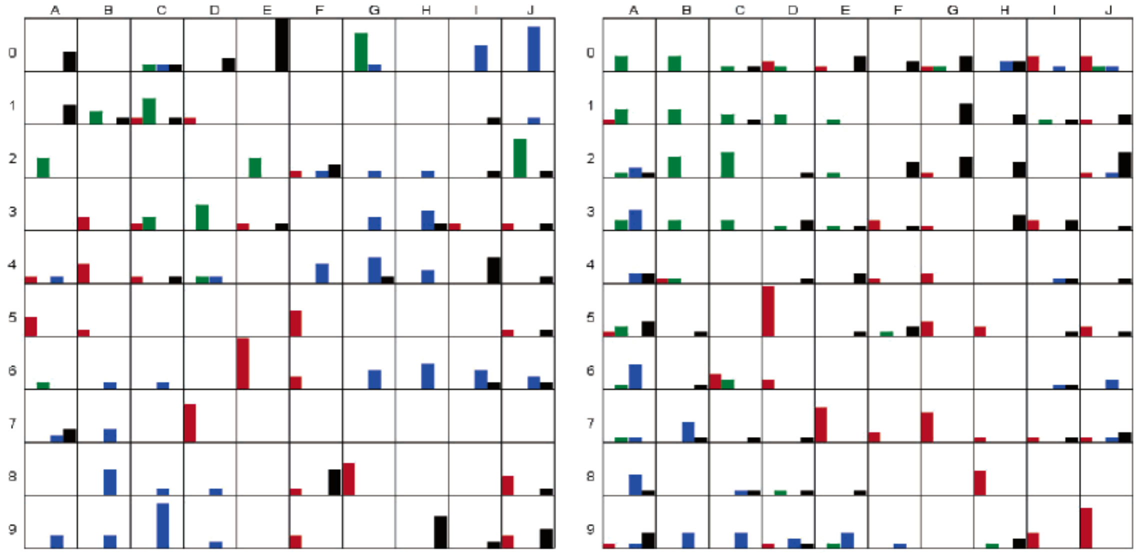

Figure 1. SOM charts derived from the descriptor hyperspaces in which the PMP2001 (left side) and Arctander (right side) data sets coming from Section A were distributed. The descriptors used are listed in Table 3. 10 × 10 maps were employed, including 100 cells. The histogram heights represent the number of compounds of each class in any cell of the map. These same charts were also used to define the training and test sets, selecting the compounds according to the molecular frequency and the olfactory note. Red = animal; green = camphoraceous; blue = hesperide; black = humus.

computed by evaluating the percentage of these "test" compounds rightly predicted. The procedure is iterated many times, and the related model should be reasonably robust if a high average LSO coefficient is obtained. In this work, the training set was subdivided into 10 groups and, then, the LSO procedure was performed 10 times before computing the final LSO value.

The *Y*-randomization test is another widely used method in which the dependent-variable vector, *Y* vector, is randomly shuffled and a new model is established by using the same original independent-variable matrix. After repeating this test several times, the average LSO value is expected to be low. If a high score is obtained, the original model is not acceptable, as the result of a chance correlation or a structural redundancy in the training set.[41]

3. RESULTS AND DISCUSSION

3.1. Modeling Selection A. *3.1.1. Descriptor and Training Set Selection.* The first step of the DM strategy consisted of selecting by HSA, amongst the global set of 167 parameters, the most relevant molecular descriptors for classifying both the PMP2001 and Arctander data sets (Table 3). The two sets of descriptors selected show that the properties necessary to discriminate the same odors in the two data sets are totally different. The PMP2001 series is characterized by lipophilicity, electrotopological, and simple topological descriptors, whereas the Arctander data set is mainly represented by H-bond and complex topological parameters. Globally, the two descriptor sets cover a wide diversity and, in the PMP2001 case, a descriptor (log *P*) is selected that plays a fundamental role in developing classification models also in many other biological fields, for example, cancer;[36] central nervous system mechanisms;[22] absorption, distribution, metabolism, and excretion properties;[35] and environmental toxicity.[37]

These descriptor sets were first used to derive two SOM charts, reported in Figure 1, which represents the 2D distributions of the four classes included in both olfactory

Table 4. Statistical Validation Scores Derived from the Best AFP Model Developed on the PMP2001 and Arctander Data Sets Associated with Selection A (see Table 2)

	PMP2001			Arctander		
odor	training (%)	validation (%)	test (%)	training (%)	validation (%)	test (%)
animal	97	82	82	84	87	56
camphoraceous	95	100	71	66	90	70
hesperide	98	93	93	89	44	56
humus	81	91	73	80	87	69
total	**93**	**91**	**81**	**80**	**80**	**63**

data sets. A 10 × 10 representation was used, including 100 cells. The histogram heights define the number of compounds of each class in any cell of the map. Several regions of these charts can be clearly associated with specific odors, and globally, they suggest the AFP method should be able to define robust models by working directly on the hyperspace.

Besides giving a very useful representation of the compounds' distribution, SOM is also a useful tool for selecting the training and test sets, maximizing the molecular diversity for each activity. The compounds were selected in each of the 100 cells of the map, according to the molecular frequency and the mechanism of action. The molecular subdivision of these three sets, within the four odor classes, is summarized in Table 2.

3.1.2. AFP Classification. The AFP models related to the PMP2001 and Arctander data sets were established on the training set compounds distributed in the descriptor hyperspaces associated with the parameters represented in Table 3. These models allowed the definition of four descriptor−odor relationships, one for each olfactory note, and the parameters used for developing them were tuned with help of the validation set.

The AFP method allowed the retrieval of the degrees of membership of the different activities for each compound, within a 0−1 range. Then, a compound was attributed to a given olfactory class if its degree of membership was the highest amongst all four values and was superior to 0.3. The detailed validation results for the best model are shown in Table 4, for both the PMP2001 model and the Arctander one. The results are satisfactory for both models, but the global and class-by-class scores are clearly better for the PMP2001 model, with differences ranging between 10 and 20% in the three sets.

By focusing on the PMP2001 scores, the "experimental" odor was predicted correctly for 91% of the validation compounds and a similar score was obtained when testing the training compounds, showing the model developed has a general application. The good robustness of this AFP model was chiefly confirmed by the LSO score that reached, quite impressively, 80%! Even more important, this value was not too dissimilar from the prediction scores associated with the training and validation sets; this means the model was not significantly perturbed by eliminating 10% of the training set information and well-represented all different structure−odor relationships associated with the PMP2001 data set. Once more, this score was superior to the Actander one, that is, 68%.

All the models were also submitted to a *Y*-randomization test, by shuffling 50% of the training set compounds; this process was repeated five times. In all cases, the final average LSO score fell down to about 25% and underlined that no chance correlation or compound redundancy was present in the structure−odor relationships developed.

Finally, the prediction ability of the models was assessed only with the help of an external test set never used to build or validate the model. The excellent average score of 81% indicated that the AFP model derived from the PMP2001 data set was undoubtedly predictive, and its performance was highly superior to the Arctander one, with an average score of 63%.

3.2. Modeling Selection B. The same DM procedure was applied to the second pair of data sets, regrouping 436 PMP2001 and 403 Arctander compounds (see Selection B in Table 2). This further test allowed the verification of the previous results on different olfactory notes and on molecular series that seemed more "robust" from a statistical point of view.

The most relevant molecular descriptors selected by the HSA method to model the new data sets are represented in Table 5. As in Selection A, the two descriptor sets covered a wide diversity and represented quite different properties,

Table 5. List of the Most Relevant Descriptors Selected by HSA for Classifying the PMP2001 and Arctander Data Sets Associated with Selection B (see Table 2)

symbol	definition	descriptor family
	PMP2001	
SaaCH	sum of all (∸(-−)CH∸) E-state values	electrotopological
SaasC	sum of all (∸CH∸) E-state values	electrotopological
SaaN	sum of all (∸N∸) E-state values	electrotopological
Shother	sum of all [other] E-state values	electrotopological
Gmax	largest atom E-state value in molecule	electrotopological
k2	second order κ shape index	topological
Nvx	number of graph vertices	topological
Totop	total topological index	topological
log P	lipophilicity at pH = 7	physicochemical
	Arctander	
SdssC	sum of all (=C<) E-state values	electrotopological
SaasC	sum of all (∸CH∸) E-state values	electrotopological
SsssNHp	sum of all (>NH+−) E-state values	electrotopological
Shtvin	sum of H E-state on sp^2 C and terminal double bonds	electrotopological
dx0	difference simple 0-order χ index	topological
dxp4	difference simple fourth-order path χ index	topological
xvp7	valence seventh-order path χ index	topological
Nrings	number of rings in a molecular graph	constitutional

Table 6. Statistical Validation Scores Derived from the Best AFP Model Developed on the PMP2001 and Arctander Data Sets Associated with Selection B (see Table 2)

	PMP2001			Arctander		
odor	training (%)	validation (%)	test (%)	training (%)	validation (%)	test (%)
animal	94	60	80	89	73	73
balsamic	77	86	78	96	93	82
spicy	88	80	75	60	36	46
ethereal	95	100	96	86	100	90
total	**87**	**85**	**82**	**86**	**82**	**76**

although in both cases, there is the presence of a cluster of electrotopological parameters related to sp^2 carbons and nitrogens. It is also interesting to observe that the PMP2001 set still included the lipophilicity parameter. These descriptor sets were used to generate SOM charts and, then, to select training, validation, and test sets according to the same procedure used for Selection A. The main trends represented by these maps were very similar to those included in Figure 1, so they will not be reported here.

The validation scores associated with the best AFP models established on the PMP2001 and Arctander training sets are shown in Table 6. These results show that very satisfactory structure−odor relationships were established in both cases. Training and validation scores all ranged from 82 to 87%, and the prediction powers evaluated on the test set associated with good values of about 80%. Moreover, the LSO procedure gave scores of 80 and 73% for the PMP2001 and Arctander data sets, respectively, whereas the *Y*-randomization test provided good results identical to those of the Selection A cases. Therefore, all the models generated could be considered to be robust, general, and predictive.

However, the PMP2001 model is preferable also in this second compound selection, for two main reasons: (1) its global scores, associated with the training and test sets, are superior to the Arctander ones, with a gap included within a 5−8% interval, and (2) its class-by-class predictions are always satisfactory, whereas the Arctander model shows weak scores when assessing the spicy compounds.

4. CONCLUSION

Besides the intrinsic biochemical factors that make olfaction one of the most complex properties to be modeled, the difficulty of this research field is amplified by the expert's subjectivity in the characterization of the odorous notes. So, the olfactory description of the same compounds in two different and, apparently, trustable databases can produce wide discrepancies. A representative example is obtained by comparing the odorous compounds included in the PMP2001 database and those in Arctander's books. A systematic analysis allows the isolation of about 900 shared molecules, and amongst them, only 2% recover the same olfactory descriptors, whereas 40% of them have a totally different profile, in which no odor included in a description can be found in the other one.

This comparison underlines how different the databases are but gives no information on their absolute and reciprocal quality. The objective of this paper consisted, then, in defining a criterion able to compare the information content of two or more databases. This task was achieved by using a data mining procedure based on the AFP method, a recursive partitioning technique derived from fuzzy logic concepts. The latter are particularly suitable to classify olfactory databases, as they can represent the "fuzziness" linked to the expert's subjectivity in the characterization of odors, by computing intermediate values between absolutely true and absolutely false for each olfactory category. These values are named degrees of membership and range between 0.0 and 1.0.

Two pairs of data sets were derived from the PMP2001 and Arctander databases, including respectively 207 and 262 compounds in the first case study, and 436 and 403 molecules in the second one. For each couple of data sets, the PMP2001 and Arctander compounds were distributed into the same olfactory classes. After selecting the most relevant descriptors by a procedure based on genetic algorithms, the AFP models were established on the training compounds selected rationally by SOM.

The robustness and the prediction power of the classification models established on these data sets gave a useful criterion for evaluating the "quality" of their information content. All the models established, on both PMP2001 descriptors and the Arctander ones, delivered satisfactory results, underlining that the newly developed AFP models are general, robust, and predictive. The best structure−odor relationships show excellent cross-validated scores of about 80% and are able to predict correctly the "experimental" odor of more than 80% of the test molecules.

But, more important for the aim of this work, all validation and prediction scores associated with the PMP2001 data sets were clearly superior to the Arctander ones, with differences ranging between 5 and 20%. *Then, these results seem to indicate the information content included in the PMP2001 database is more trustable and it should be adopted as reference database for deeper studies on more extended olfactory data sets.*

To conclude, the data mining procedure here presented can be used as an efficient strategy to better evaluate and compare *the quality of two databases with the help of robust statistical databases, even if the collected raw information issued from the expert panels are based on subjective data*. More generally, the above strategy of comparing the information content of public or corporate databases is not limited to olfactory properties but can be extended to any biological property for which enough information is collected.

ACKNOWLEDGMENT

Our gratitude goes to the Edmond Roudnitska Foundation for their support and grant to C.W. This paper is dedicated in memoriam to Mrs. Thérèse Roudnitska (1919−2004), Past President of the Edmond Roudnitska Fondation.

REFERENCES AND NOTES

(1) Chastrette, M.; Zakarya, D. Molecular structure and smell. In *The Human Sense of Smell*; Laing, D. G., Doty, R. L., Breipohl, W., Eds.; Springer-Verlag: New York, 1991.

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

(3) Malnic, B.; Hirono, J.; Sato, T.; Buck, L. B. Combinatorial receptor codes for odors. *Cell* **1999**, *96*, 713−723.

(4) *Chemometric Methods in Molecular Design*; Van de Waterbeemd, H., Ed.; Wiley-VCH: Weinheim, Germany, 1995.

(5) *Practical Application of Computer-Aided Drug Design*; Charifson, P. S., Ed.; Marcel Dekker: New York, 1997.

(6) *3D QSAR in Drug Design. Recent Advances*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer Escom: Dordrecht, Netherlands, 1998.

(7) Rossiter, K. J. Structure−Odor Relationships. *Chem. Rev.* **1996**, *96*, 3201−3240.

(8) Chastrette, M. Data management in olfaction studies. *SAR QSAR Environ. Res.* **1998**, *8*, 157−181.

(9) Niemi, G. J. Multivariate analysis and QSAR: Applications of principal component analysis. In *Practical Applications of Quantitative Structure−Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*; Karcher, W., Devillers, J., Eds.; Kluwer Academic: Dordrecht, Netherlands, 1990.

(10) Hubert, C. J. *Applied Discriminant Analysis*; Wiley-Interscience: New York, 1994.

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

(12) Zupan, J.; Johann Gasteiger, J. *Neural Networks in Chemistry and Drug Design*; Wiley-VCH: Weinheim, Germany, 1999.

(13) Gardner, J. W.; Hines, E. L.; Wilkinson, M. Application of Artificial Neural Networks to an Electronic Olfactory System. *Meas. Sci. Technol.* **1990**, *1*, 446−451.

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

(15) Keller, P. Physiologically Inspired Pattern Recognition for Electronic Noses. *Proc. SPIE* **1999**, *3722*, 144−153.

(16) Ham, C. L.; Jurs, P. C. Structure−activity studies of musk odorants using pattern recognition: monocyclic nitrobenzenes. *Chem. Senses* **1985**, *10*, 491−505.

(17) Chastrette, M.; Cretin, D.; Aidi, C. E. Structure−Odor Relationships: Using Neural Networks in the Estimation of Camphoraceous or Fruity Odors and Olfactory Thresholds of Aliphatic Alcohol. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 108−113.

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

(19) Ros, F.; Audouze, K.; Pintore, M.; Chretien, J. R. Hybrid System for Virtual Screening: Interest of Fuzzy Clustering Applied to Olfaction. *SAR QSAR Environ. Res.* **2000**, *11*, 281−300.

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

(21) Pintore, M.; Audouze, K.; Ros, F.; Chrétien, J. R. Adaptive Fuzzy Partition in data base mining: application to olfaction. *Data Sci. J.* **2002**, *1*, 99−110.

(22) Ros, F.; Taboureau, O.; Pintore, M.; Chrétien, J. R. Development of CNS predictive models by Adaptive Fuzzy Partitioning. *Chemom. Intell. Lab. Syst.* **2003**, *67*, 29−50.

(23) Manley, C. H. Psychophysiological effect of odor. *Crit. Rev. Food Sci. Nutr.* **1993**, *33*, 57−62.

(24) Qureshy, A.; Kawashima, R.; Imran, M. B.; Sugiura, M.; Goto, R.; Okada, K.; Inoue, K.; Itoh, M.; Schormann, T.; Zilles, K.; Fukuda, H. Functional mapping of human brain in olfactory processing: a PET study. *J. Neurophysiol.* **2000**, *84*, 1656−1666.

(25) *PMP 2001*, database of perfumery materials and performance; BACIS: The Netherlands, 2001.

(26) *Perfume and Flavor Materials of Natural Origin*; Arctander, S., Ed.; Allured Pub Corp: Wheaton, IL, 1960.

(27) *Perfume and Flavor Chemicals*; Arctander, S., Ed.; Allured Pub Corp: Wheaton, IL, 1969.

(28) Ros, F.; Pintore, M.; Chrétien, J. R. Molecular descriptor selection combining genetic algorithms and fuzzy logic: application to data base mining procedure. *Chemom. Intell. Lab. Syst.* **2001**, *63*, 15−26.

(29) *MDL QSAR*, version 2.2; MDL Information Systems Inc.: San Leandro, CA, 2003.

(30) Haupt, R. L.; Haupt, S. E. *Practical Genetic Algorithms*; Wiley-Interscience: New York, 1998.

(31) Leardi, R.; Gonzales, A. L. Genetic algorithms applied to feature selection in PLS regression: how and when to use them. *Chemom. Intell. Lab. Syst.* **1998**, *41*, 195−207.

(32) Pintore, M.; Taboureau, O.; Ros, F.; Chretien, J. R. Database mining applied to central nervous system (CNS) activity. *Eur. J. Med. Chem.* **2001**, *36*, 349−359.

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

(34) Lin, Y.; Cunninghan, G. J. Building a Fuzzy System from Input−Output Data. *J. Intell. Fuzzy Syst.* **1994**, *2*, 243−250.

(35) Pintore, M.; van de Waterbeemd, H.; Piclin, N.; Chrétien, J. R. Prediction of oral bioavailability by Adaptive Fuzzy Partitioning. *Eur. J. Med. Chem.* **2003**, *38*, 427−431.

(36) Piclin, N.; Pintore, M.; Wechman, C.; Chretien, J. R. Classification of a large anticancer data set by Adaptive Fuzzy Partition. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 577−586.

(37) Pintore, M.; Piclin, N.; Benfenati, E.; Gini, G.; Chretien, J. R. Predicting toxicity against the fathead minnow by Adaptive Fuzzy Partition. *QSAR Comb. Sci.* **2003**, *22*, 210−219.

(38) Sugeno, M.; Yasukawa, T. A fuzzy-logic-based approach to qualitative modeling. *IEEE Trans. Fuzzy Syst.* **1993**, *1I*, 7−31.

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

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

(41) Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.* **2003**, *22*, 69−77.

CI0502505