SIMO 2006 – systèmes d'information, modélisation, optimisation et commande en Génie des Procédés
Le virtuel dans la réalité quotidienne.
11-12 octobre 2006 - Toulouse (France)

Relation Structure moléculaire - Odeur
Utilisation des Réseaux de Neurones pour l'estimation de l'Odeur Balsamique

Mourad KORICHI[1,2], Vincent GERBAUD[1], Thierry TALOU[3], Christine RAYNAUD[3], Pascal FLOQUET[1]

- (1) Laboratoire de Génie Chimique
  5 rue Paulin Talabot
  F-3110 Toulouse, FRANCE
  e-mail : Vincent.Gerbaud@ensiacet.fr ; France
- (2) Laboratoire de Valorisation et Promotion des Ressources Sahariennes
  Université de Ouargla, BP 511
  30000, Ouargla, ALGERIE
  e-mail : M.Korichi@arn.dz ; Algérie
- (3) Laboratoire de Chimie Agro-Industrielles,
  118 route de Narbonne,
  F-31077 Toulouse FRANCE

Abstract: Structure – odor relationships (SOR) are key issues for the synthesis of new odorant molecules. But, this relation is hard to model, due to limited understanding of olfaction phenomena and the subjectivity of odor quantity and quality as stated in Rossitier's review (1996). Many molecular descriptors are used to correlate molecule's odor, but no universal rules emerge in this field. In this paper, we focus on the use of molecular descriptors as an alternative approach in the prediction of odors, by the mean of regression techniques. Principal Component Analysis (PCA) and Stepwise Collinearity Diagnosis (SCD) techniques are used to reduce the dimensionality of data, by the identification of significant molecular descriptors. Then, the chosen molecular descriptors are used with a neural networks algorithm to correlate the structure to molecular odor quality. The results are validated on balsamic flavor..

Keywords: Molecular graph; Group contribution; Property prediction; CAMD

Résumé : Les molécules odorantes (parfums ou flaveurs) sont utilisées dans une grande variété de produits de consommation, pour inciter les consommateurs à associer les impressions favorables à un produit donné. La Relation Structure moléculaire-Odeur (SOR) est cruciale pour la synthèse de ces molécules mais est très difficile à établir due à la subjectivité de l'odeur. Ce travail présente une approche de prédiction de l'odeur des molécules basée sur les descripteurs moléculaires. Les techniques d'analyse en composantes principales (PCA) et de d'analyse de colinéarité permettent d'identifier les descripteurs les plus pertinents. un réseau de neurones supervisé5 à deux couches (cachée et sortie) est employé pour corréler la structure moléculaire à l'odeur. La base de données décrite précédemment est utilisée pour l'apprentissage. Un ensemble de paramètres est modifié jusqu'à la satisfaction de la meilleure régression.
Les résultats obtenus sont encouragent, ainsi les descripteurs moléculaires convenables corrèlent efficacement l'odeur des molécules. C'est la première étape d'un modèle générique en développement pour corréler l'odeur avec les structures moléculaires

MOTS-CLÉS : Graphe moléculaire ; Contribution de groupes ; Prédiction des propriétés ; CPAO.

INTRODUCTION

Odorant compounds are found in a wide variety of products ranging from foods, perfumes, health care products and medicines. Either combined or alone, flavor and fragrance compounds are used to induce consumers to associate favorable impressions with a given product. In some cases, products have one predominant component which provides the characteristic odor. However, in most cases, products containing odors include a complex mixture of fragrant compounds. Some of them are classified within REACH, a European Community document, regulating the use of chemicals in terms of environment and toxicity. Structure – Odor relationships (SOR) are very important for the synthesis of new odorant molecules. This relation is difficult to model due to the subjectivity of the odor quantity and quality. Olfaction phenomenon is not yet completely understood and odor measurements are often inaccurate (Amboni *et al.*, 2000). Research has been oriented to the use of structural, topological, geometrical, electronic, and physicochemical parameters as descriptors, to generate odor predictive equations. Therefore, a number of computational techniques have been used successfully. Artificial Neural Networks (ANN's) are one of these promising techniques readily suited for the assessment of poorly understood properties like odor. In this paper, we aim to use molecular descriptors as an alternative approach in the prediction of molecule's odor by the mean of regression techniques. Principal Component Analysis (PCA) and Pairwise Collinearity Diagnosis techniques are used to reduce the dimensionality of data, by the identification of significant molecular descriptors.

Then, the chosen molecular descriptors are used with a neural networks algorithm to correlate the structure to molecular odor quality. Figure 1 summarizes the methodology.

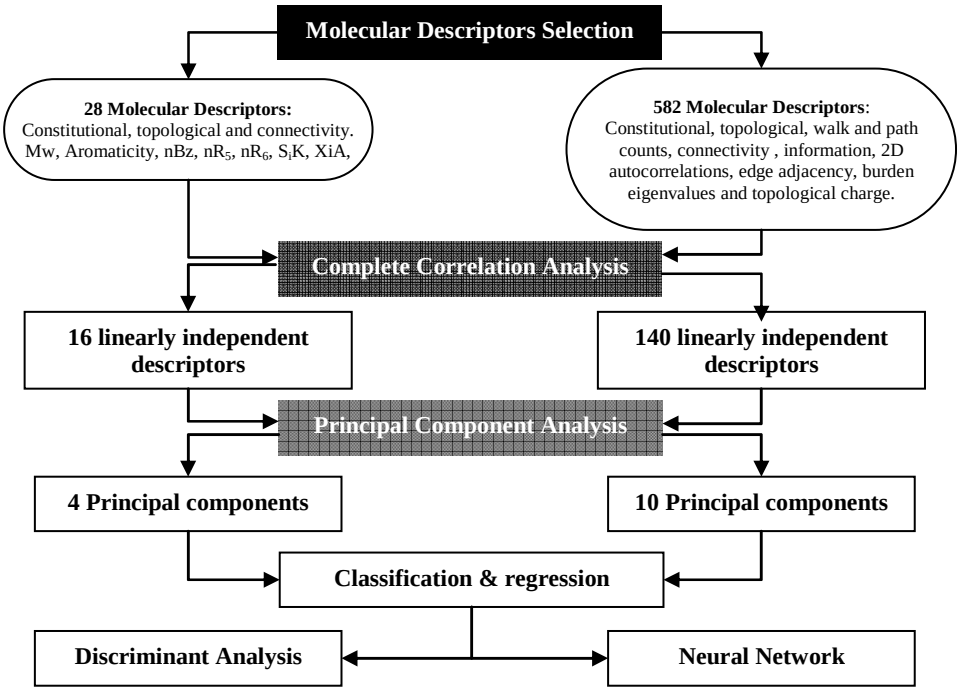

Figure 1. Schematic representation of methodology on structure – odor relationships.

MOLECULAR DESCRIPTOR SELECTION

Molecular descriptors accounts for a particular aspect of the molecule structure. As examples, *simply count of atoms*, *functional groups* and *characteristic fragments* are some of the *constitutional descriptors* family of the studied structure. *Topological descriptors* are related to the two-dimensional representation of the molecular structure. Molecular descriptors are the most significant common features of molecular structure that can be used to develop Structure - property relationships. In our case, the property is the odor of a molecule. Our input data set contains 121 molecules of balsamic odor splited in 5 sub-notes of typical odors (see Table 1): anise, balsam, honey, vanilla and sweet (Aldrich Flavors and Fragrances catalog, 2005). The dragon software (TALETE, 2005) is used to calculate up to 582 molecular descriptors of the input data set. According to figure 1, two cases are explored, first, 34 simple descriptors (molecular weight, aromatic ratio, number of benzene-like rings, connectivity indices, Kier symmetry index and n-path Kier alpha-modified shape index) used in the different relations described in Rossitier's review on structure – odor works since 1930 (1996) are calculated. In the second case, 582 molecular descriptors are considered: constitutional descriptors, topological indices, walk and path counts, connectivity and information indices, 2D autocorrelations, edge adjacency indices, burden eigenvalues and topological charge indices. All descriptors are calculated from the 2D molecular representation.

Table 1. Input data set of molecular structure

Odor type	Number of compounds	Arbitrary continuous Odor codification	Arbitrary discontinuous Odor codification
Anise	10	0 to 0.15	0.15
Balsam	18	0.25 to 0.35	0.35
Honey	21	0.45 to 0.55	0.55
Vanilla	15	0.60 to 0.75	0.75
Sweet	58	0.85 to 0.95	0.95

Complete Correlation Analysis

The complete correlation analysis is used to select a subset of linearly independent descriptors. Descriptor dependency is evaluated using the Dragon software by setting a predefined value R_{max} (In this work, $R_{max} = 0.97$) below which descriptors are considered linearly independent.

Principal component analysis

Principal component analysis is one of the oldest, but still most widely used techniques of multivariate analysis. The basic idea of the method is to try to describe the variation of the variables in a set of multivariate data, as parsimoniously as possible using a set of uncorrelated variables, each of which is a particular linear combination of those in the original data. This enables us to reduce the molecular descriptors dimensionality, by the identification of the principal components that can be used in the structure - odor relationship. All eigenvalues greater than 1 are retained to describe the principal axes. In the first case, four principal components are kept to describe the 16 molecular descriptors, widely used in the correlation of the structure – odor. In the second case, ten principal components are retained to represent the 140 molecular descriptors.

ARTIFICIAL NEURAL NETWORKS (ANN) APPROACH

The ANN (Dreyfus *et al.*, 2004) trained by back-propagation (BP) network algorithm has a two layers architecture: the first layer is called the hidden layer. The number of hidden neuron is a variable X between 7 and 9. The output layer consists of one neuron, namely the odor quality. The network configuration is m-X-1, where m represents the number of principal components encapsulating the maximum of information of the linearly independent molecular descriptors (§ 2.2). The ANN has a feed forward layered structure with connections allowed only between adjacent layers. The balsamic odor sub-notes as output, are represented by arbitrary continuous codification described in table 1. Input and output data are normalized, and hyperbolic tangent sigmoid transfer function is used in the hidden layer. The output layer is a linear function. In this work, the training and the validation sets are generated randomly, corresponding respectively to 70% and 30% of the input data set of 121 molecules. After several training sessions, an optimal number of hidden neurons X equal 8 and 7 is retained for case one and two respectively. The network is trained for 500 epochs with a gradient algorithm. The performance goal is 0.001.

DISCRIMINANT ANALYSIS

Discriminant analysis is an analytical technique, whereby a multivariate data set containing m variables is separated into a number (k) of pre-defined groups, using discriminant functions (Z) which are linear combinations of the variables. Two cases are studied to discriminate the molecules into different odors based on the results of PCA described on the paragraph (§2.2).

Discriminant Analysis based on the first PCA study (four principal components)

In the first case there are four principal components. Results are presented in the table 3 with an overall 69.4% of the molecules in the data set, are well classified. The molecules of vanilla odor have the highest correctly classification per cent.

Table 3. Discriminant analysis based on the first PCA study.

Groups	Predicted groups						Correctly classified
	Anise	Balsam	Honey	Vanilla	Sweet	molecules	
Anise	8	0	0	2	0	10	0.800
Balsam	1	14	1	1	1	18	0.778
Honey	0	3	15	0	3	21	0.714
Vanilla	0	2	0	12	0	14	0.857
Sweet	3	8	8	4	35	58	0.603

Discriminant Analysis based on the second PCA study (ten principal components)

In the second case, there are ten principal components. 83.4% of 121 molecules in the data set are well classified, with the honey molecules having the highest classification, and the sweet molecules having the lowest, like in the first case.

Table 4. Discriminant analysis based on the second PCA study

Groups	Predicted groups						Correctly classified
	Anise	Balsam	Honey	Vanilla	Sweet	molecules	
Anise	8	0	0	2	0	10	0.800
Balsam	0	17	0	0	1	18	0.944
Honey	0	0	20	0	1	21	0.952
Vanilla	0	1	0	13	0	14	0.929
Sweet	3	2	5	4	44	58	0.759

RESULTS AND DISCUSSIONS

Artificial Neural Networks Approach

It is well-known that ANN performance depends on many variables, as the number of hidden neurons, the degree of homology between the training and the validation sets and the input variables (Principal components in this case). In figure 2, results are represented as the variation of the odor quality versus the molecule identification code, (a) for the first case and (b) for the second case. In the first case (a), the ANN does not converge as shown by the similarity of the response for all molecules despite their initial differences. In the second case (b), the training set is well represented, but almost all the validation set is not. This clearly shows the non predictive capacity of the ANN approach. Kovatcheva *et al.* (2004) on a *k*NN approach for modeling structure - ambergris odor relationship suggest to use division procedures based on sphere-exclusion algorithms and demonstrate a predictive capacity. But the ambergris odor is due to well known chemical structures, unlike the balsamic odor where molecule structure is more heterogeneous with several odor sub-notes. Also, Chastrette *et al.*, (1995 & 1996), Cherqaoui *et al.* (1998) and Zakarya *et al.* (1999) do not consider sub-notes, only, requesting a discrete response.

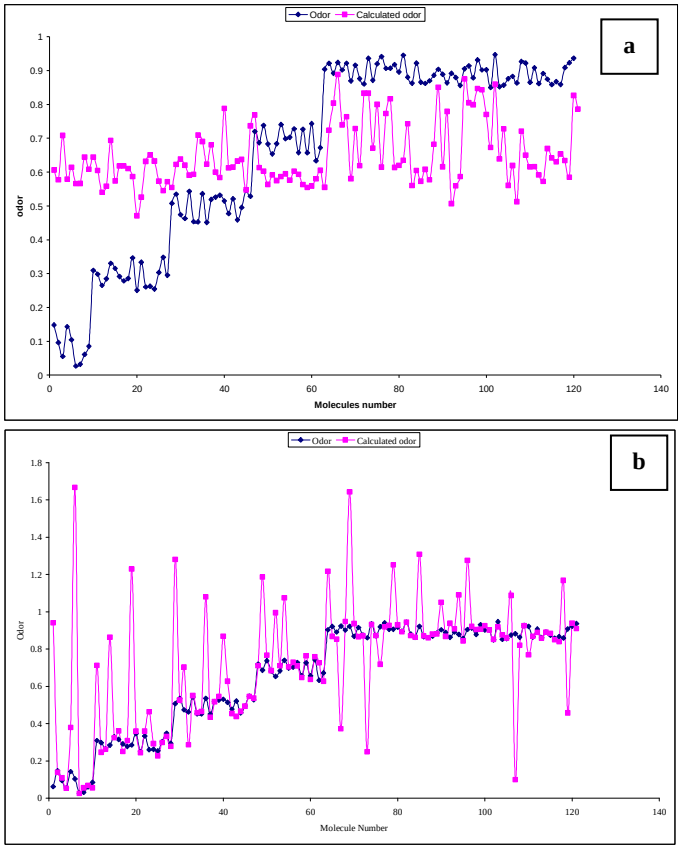

Figure 2. Reference odor (Aldrich, 2005) and calculated odor versus molecular identification.

Discriminant Analysis Approach

69.4% and 83.4% of molecules are well discriminated in the two cases respectively, especially anise and vanilla molecules groups. This is better than the ANN approach. Case 2 has more odor discriminant, because it incorporates more numerous and diverse molecular descriptors. Among the molecules that are not discriminated, the two molecules from anise, classified in the vanilla group bear similar molecular structure with vanilla type molecules, which have three oxygen atoms, high Kier symmetry index and n-path Kier alpha-modified shape index. For case 2, in balsam and honey sub-note odors, the molecule wrongly classified is considered differently depending on the referential nomenclature (we used Aldrich's) that both belong to balsamic and/or rose main odors. In the group of vanilla, one molecule is discriminated as balsam sub-note. From the referential nomenclature, we can say that the molecule has different odor types. In the sweet sub-note, fourteen molecules are distributed into other sub-notes. The low discrimination of the sweet odor may be attributed to the subjectivity of this sub-note, unlike vanilla or anise. Indeed, sweet is not considered as a typical odor type in the reputed referential chart "the field of odors" of Jaubert et al. (1995).

CONCLUSION AND PERSPECTIVES

In this work we present different ways to estimate and discriminate odors of molecules, based on molecular descriptors using multidimensional data analysis, and neural network applied to balsamic odors. The multidimensional data analysis is a powerful tool to reduce data sets and encapsulate the maximum of molecule's structure information. Discriminant analysis results using only 2D molecular representation are encouraging. Further work using 3D representation molecular descriptors may improve the results. The neural network satisfactorily correlates the molecules with their assigned odor, based on sufficiently numerous and diverse molecular descriptors. But it is unable to predict balsamic odor and its sub-notes. Compared with literature, successful results in ANN approach are due to the well known families of odor. The heterogeneous nature of the molecules assigned to balsamic odor and the absence of evident structure – odor relationship, forces us to request a continuous discrimination between sub-notes.

REFERENCES

Aldrich Inc., Flavors and Fragrances catalog, , http://www.sigmaaldrich.com/, 2005.

Amboni, R. D. C., Junkes, B., Yunes R. A. and Heinzen, V. E. F., ***J. Agric. Food Chem.,*** 48 (2000) 3517-3521.

Chastrette, M., Aïdi, C. E. and Peyraud, J. F., ***Eur. J. Med. Chem.***, 30 (1995), 679-686.

Chastrette, M., D. Cretin, D. and Aïdi, C. E., ***J. Chem. Inf. Comput. Sci.***, 36 (1996), 108-113.

Cherqaoui, D., Essefar M., Villemin, D., Cense, J.-M., Chastrette, M., and Zakarya, D., ***New J. Chem***, 1998, 839- 843.

Dreyfus, G., Martinez, J.-M., Samuelides, M., Gordon, A. B., Badran, F., Thiria, S, Hérault, L, *Réseaux de neurones :: méthodologie et applications,* Paris , Eyrolles, 2004.

Jaubert, J. N., Tapiero, C. And Dore, J. C. ***Perfumer Flavorist***, 20 (1995), 1-16.

Kovatcheva, A., Golbraikh, A., Oloff, S. Xiao, Y_D., Zheng, W., Wolschann, P. Buchbauer, G., and Tropsha, A., ***J. Chem. Inf. Comput. Sci.***, 44 (2004), 582-595.

Rossitier, K. **J., *Chem. Rev.***, 96 (1996), 3201 - 3240.

Talete srl, *Dragon Profesional Software*, Via V.Pisani, 13-20124 Milano(ITALY), 2005.

Zakarya, D., Chastrette, M., Tollabi, M. and Fkih-Tetouani, S., ***Chemometrics and Intelligent Laboratory Systems***, 48 (1999), 35–46.