JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

Aroma Quality Differentiation of Pyrazine Derivatives Using Self-Organizing Molecular Field Analysis and Artificial Neural Network

JOHANNA KLOCKER,† BETTINA WAILZER,† GERHARD BUCHBAUER,‡ AND PETER WOLSCHANN*,†

Institute of Theoretical Chemistry and Molecular Biology, University of Vienna, Waehringer Strasse 17, A-1090 Vienna, Austria, and Institute of Pharmaceutical Chemistry, University of Vienna, Althanstrasse 14, A-1090 Vienna, Austria

The encoding of various aroma impressions and the distinction between different aroma qualities are unsolved problems, as differences between aroma impressions can be described only in a qualitative but not in a quantitative manner. As a consequence, classifications of various aroma qualities cannot easily be performed by standard QSAR methods. To find a proper way to encode aroma impressions for SAR studies, a total of 50 pyrazine-based aroma compounds showing the aroma quality of earthy, green-earthy, or green are analyzed. Special attention is thereby turned on the mixed aroma impression green-earthy. Classifications on the whole data set as well as on smaller subsets are calculated using self-organizing molecular field analysis (SOMFA) and artificial neural networks (ANNs). SOMFA classifies between two or three aroma impressions, leading to models satisfying in predictive power. ANN analysis using multilayer perceptron network architecture with one hidden layer and nominal output as well as genetic regression neural network) with two hidden layers and numerical output both lead to a rather good performance rate of 94%.

KEYWORDS: Artificial neural network; self-organizing molecular field analysis; aroma impression; aroma classification; pyrazines

INTRODUCTION

Aroma compounds play an important role in food chemistry. At the beginning mostly the syntheses of flavor molecules were of interest; later, the main topics of aroma research were the isolation of aroma compounds from foodstuffs and their analytical chemistry as determined by using improved analytical techniques. Nowadays, the flavor science concentrates especially on a better definition of the sensory properties of aroma substances and on the knowledge of the relationships between the molecular structures and the quality as well as the intensity of the aroma impression of these molecules (*1*). Quantitative structure−activity relationships (QSAR) try to correlate the intensity of a biological effect with the molecular structure of the studied compounds. During recent years, various QSAR models have been developed for aroma compounds to give some insight into the parameters, which might influence the aroma intensity (*2*−*8*). Generally, QSAR methods rely on quantitative information of the biological activity of the studied molecules. The activity of aroma compounds can be determined quantitatively by the detection threshold value of the compound dissolved in water. The threshold value is thereby defined as the olfactory detection threshold, which is connected to the ability of a test person to distinguish between water with and without aroma. Studies on the aroma qualities and structural differences between the various aroma impressions (*9*−*11*), which are also important to elucidate the mechanism of aroma recognition, cannot easily be performed by conventional QSAR procedures. This is due to the fact that an aroma impression can be described only in a qualitative but not in a quantitative manner. Moreover, this qualitative specification is not always clearly defined. A widespread problem appears to be the fact that the assignment of some substances to a concrete aroma impression is sometimes arbitrary, as for these compounds two or more aroma descriptions are given. Dominant aroma impressions with weaker tonalities have to be considered as well as real mixtures of different aroma qualities. So the question arises, how could these various aroma impressions and mixtures be encoded in order to investigate them by computational methods and to obtain proper classification models pointing out their structural differences.

In our previous investigations on structure−flavor relationships the application of artificial neural networks (ANNs) with nominal output allowed the discrimination between different classes of aroma impressions of pyrazine-derived flavor compounds (*8*). Pyrazines are one important class of aroma

* Author to whom correspondence should be addressed (e-mail Karl.Peter.Wolschann@univie.ac.at; telephone +43 1 4277 52772; fax 0043 1 4277 9527).
† Institute of Theoretical Chemistry and Molecular Biology.
‡ Institute of Pharmaceutical Chemistry.

10.1021/jf011664a CCC: $22.00
Published on Web 06/04/2002

Table 1. Structures and Aroma Impressions of the 50 Pyrazines

compd	pyrazine	R_1	R_2	R_3	R_4	impression
1	triethyl-	C_2H_5	C_2H_5	C_2H_5	H	earthy[a]
2	2-methyl-3-ethenyl-	CH_3	$CH{=}CH_2$	H	H	earthy[a]
3	2,5-dimethyl-3-ethenyl-	CH_3	$CH{=}CH_2$	CH_3	H	earthy[a]
4	3,5-dimethyl-2-ethenyl-	CH_3	$CH{=}CH_2$	H	CH_3	earthy[a]
5	3,5-dimethyl-2-propyl-	CH_3	C_3H_7	H	CH_3	earthy[a]
6	3,5-dimethyl-2-(*Z*-1-propenyl)-	CH_3	$CH{=}CHCH_3$ (*Z*)	H	CH_3	earthy[a]
7	3,5-dimethyl-2-(*E*-1-propenyl)-	CH_3	$CH{=}CHCH_3$ (*E*)	H	CH_3	earthy[a]
8	3,5-dimethyl-2-(2-propenyl)-	CH_3	$CH_2CH{=}CH_2$	H	CH_3	earthy[a]
9	2-isopropyl-3,5-dimethyl-	CH_3	$CH(CH_3)_2$	H	CH_3	earthy[a]
10	2-butyl-3,5-dimethyl-	CH_3	C_4H_9	H	CH_3	earthy[a]
11	3-ethyl-5-methyl-2-ethenyl-	CH_3	H	$CH{=}CH_2$	C_2H_5	earthy[a]
12	2-ethyl-5-methyl-3-ethenyl-	CH_3	H	C_2H_5	$CH{=}CH_2$	earthy[a]
13	3-ethyl-2-methyl-5-ethenyl-	CH_3	C_2H_5	$CH{=}CH_2$	H	earthy[a]
14	2-isopropyl-3-methyl-	CH_3	$CH(CH_3)_2$	H	H	earthy[b]
15	2,3-diethyl-	C_2H_5	C_2H_5	H	H	green-earthy[c] earthy[a]
16	2-pentyl-	H	C_5H_{11}	H	H	green-earthy[d]
17	2-methylthio-3-methyl-5-(2-methylpentyl)-	SCH_3	CH_3	$CH_2CH(CH_3)C_3H_7$	H	green-earthy[b]
18	2-acetyl-3,6-dimethoxy-5-methyl-	OCH_3	$COCH_3$	OCH_3	CH_3	green-earthy[e]
19	2-methylthio-3-isopropyl-	SCH_3	$CH(CH_3)_2$	H	H	green-earthy[b]
20	2-butyl-	H	C_4H_9	H	H	green-earthy[d]
21	2-ethylthio-3-butyl-	SC_2H_5	C_4H_9	H	H	green-earthy[d]
22	2-methylthio-3-pentyl-	SCH_3	C_5H_{11}	H	H	green-earthy[d]
23	2-methoxy-3-pentyl-	OCH_3	C_5H_{11}	H	H	green-earthy[d]
24	2-ethoxy-3-pentyl-	OC_2H_5	C_5H_{11}	H	H	green-earthy[d]
25	2-methoxy-3-heptyl-	OCH_3	C_7H_{15}	H	H	green-earthy[d]
26	2-methylthio-3-octyl-	SCH_3	C_8H_{17}	H	H	green-earthy[d]
27	2-methoxy-3-octyl-	OCH_3	C_8H_{17}	H	H	green-earthy[d]
28	2-ethylthio-3-octyl-	SC_2H_5	C_8H_{17}	H	H	green-earthy[d]
29	2-methylthio-3-decyl-	SCH_3	$C_{10}H_{21}$	H	H	green-earthy[d]
30	2- methoxy-3-decyl-	OCH_3	$C_{10}H_{21}$	H	H	green-earthy[d]
31	2-ethylthio-3-decyl-	SC_2H_5	$C_{10}H_{21}$	H	H	green-earthy[d]
32	2-phenylthio-3-pentyl-	SC_6H_5	C_5H_{11}	H	H	green-earthy[d]
33	2-phenoxy-3-pentyl-	OC_6H_5	C_5H_{11}	H	H	green-earthy[d]
34	2-methoxy-5-isobutyl-3-methyl-	OCH_3	CH_3	$CH_2CH(CH_3)_2$	H	green[b]
35	2-methoxy-3-methyl-5-(2-methylbutyl)-	OCH_3	CH_3	$CH_2CH(CH_3)C_2H_5$	H	green[b]
36	2-methylthio-3-methyl-5-(2-methylbutyl)-	SCH_3	CH_3	$CH_2CH(CH_3)_2$	H	green[b]
37	2,3-dimethyl-5-pentyl-	CH_3	CH_3	C_5H_{11}	H	green[f]
38	2-phenoxy-5-isopropyl-3-methyl-	OC_6H_5	CH_3	$CH(CH_3)_2$	H	green[b]
39	2-ethylthio-3-methyl-5-(2-methylbutyl)-	SC_2H_5	CH_3	$CH_2CH(CH_3)C_2H_5$	H	green[f]
40	2-ethoxy-5-*sec*-butyl-3-methyl-	OC_2H_5	CH_3	$CH(CH_3)C_2H_5$	H	green[b]
41	3-methoxy-2-isopropyl-5-methyl-	OCH_3	$CH(CH_3)_2$	H	CH_3	green[e]
42	2-dimethylamino-6-isobutyl-	$N(CH_3)_2$	H	H	$CH_2CH(CH_3)_2$	green[e]
43	2-acetyl-3-methoxy-5-methyl-	OCH_3	$COCH_3$	H	CH_3	green[e]
44	2,5-dimethyl-3-(3-methylbutyl)-	CH_3	$C_3H_5(CH_3)_2$	CH_3	H	green[c]
45	2-methoxy-3-isopropyl-5-methyl-	OCH_3	$CH(CH_3)_2$	CH_3	H	green[e]
46	2-propyl-3-methyl-	CH_3	C_3H_7	H	H	green[b]
47	2-methylthio-3-propyl-	SCH_3	C_3H_7	H	H	green[b]
48	2-ethylthio-3-pentyl-	SC_2H_5	C_5H_{11}	H	H	green[d]
49	2-ethoxy-3-octyl-	OC_2H_5	C_8H_{17}	H	H	green[d]
50	2-butoxy-	OC_4H_9	H	H	H	green[g]

[a] Wagner et al. (*6*). [b] Masuda and Mihara (*4*). [c] Boelens and van Gemert (*21*). [d] Masuda and Mihara (*22*). [e] Takken et al. (*3*). [f] Shibamoto (*23*). [g] Pittet and Hruza (*24*).

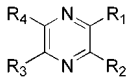

Figure 1. General structure of pyrazines.

compounds. They are potent and characteristic flavorants found in a wide range of raw and processed foods, where they are formed during the Maillard reaction (*12*) or as products of the secondary metabolism (*13*, *14*), respectively. Their general structure is depicted in **Figure 1**.

As an extension of our recent studies, we investigated a series of aroma compounds with two different aroma impressions and compounds that are described to show both of these aroma qualities simultaneously. In particular, pyrazine-derived molecules with earthy, green, and green-earthy odors are analyzed. The aroma impressions are taken from the literature (different panels), whereby 10 compounds are described in the same way by two different authors. Only compound **15** is specified with two different aroma qualities: as green-earthy smelling by Boelens et al. (*21*) and as earthy by Wagner et al. (*6*). In this case we decided to describe compound **15** with a green-earthy aroma on the basis of the fact that the earthy character is included. In total, quality data of seven different panels are used (see **Table 1**). Whereas the group of earthy compounds is mainly investigated by Wagner et al. (*6*) and the green-earthy group by Masuda et al. (*4*), studies of the green-smelling compounds are distributed among several authors. Results show that a correct classification of the various aroma impressions by ANN and self-organizing molecular field analysis (SOMFA) is not related to the amount of panels used describing a particular aroma quality.

Different encoding schemes for the aroma qualities are considered in order to compare the predictivity of the classification models obtained by using SOMFA and ANNs.

The SOMFA method is a technique for 3D-QSAR studies; it has been developed just recently by Robinson et al. (*15*). The method has similarities to both comparative molecular field analysis (CoMFA) and molecular similarity studies. Related to CoMFA, a grid-based approach is used, but no probe interaction energies have to be evaluated. As in similarity methods intrinsic molecular properties, such as the molecular shape and electrostatic potential, are used to develop the QSAR models. Three-dimensional grids are created as in other QSAR techniques with values at the grid points representing shape or electrostatic potentials. Crucial for SOMFA is the notion of the "mean centered activity", which is derived by subtracting the mean activity of the training set from each molecule's activity. The value of the shape or the electrostatic potential at every grid point for a given molecule is then multiplied by the mean centered activity for this molecule. This weights the grid points in such a way that the most active and least active molecules have higher values and, therefore, have more influence on the final model than the molecules with an activity close to the mean value.

QSAR techniques such as SOMFA often do not lead to satisfactory results, a fact that may be due to nonlinearities. One possibility to overcome such problems is the application of ANNs, which are able to handle nonlinear relationships sufficiently. The importance of ANNs in the field of drug design is strongly increasing (*16*), and these methods have already been applied in different structure−activity relationship studies of aroma compounds (*9*−*11*, *17*−*19*). As ANNs are originally motivated by attempts to describe the working principles of individual neurons or networks of neurons in the human brain mathematically (*20*), their texture can be seen in analogy to it. In an ANN a processing element plays the role of a biological neuron. It receives inputs from other processing elements through the input connectors, which represent the dendrites. The incoming data are interpreted in the processing element, and the resulting output is sent to other processing elements through the output connectors, simulating the function of the axons. The pattern of connections between the neurons and the strength of these connections, the so-called weights, play an important role in the training of ANNs. This involves iterative changes of the weights to minimize the error in the predictions on the training set. If the network is properly trained, it learns to model the function which relates the input variables to the output variables and can subsequently be used to make predictions where the output is unknown. Usually, the studied data set is divided into training and verification sets. The training set is used to train the network, and the verification set is applied to check the network's error performance. Finally, it is common practice to reserve a third set of cases (test set) for external prediction, to ensure that the results on the training and verification set are real and no artifacts of the training process.

COMPUTATIONAL METHODS

Fifty pyrazine-based aroma compounds (**Table 1**), with earthy or green flavor as well as with mixtures of both aroma qualities, are selected from the literature (given in the table) in order to find some molecular descriptors and to elucidate convenient models for the predictive discrimination between these various flavor impressions.

The three-dimensional structures of the pyrazines are built using the Hyperchem 5.0 program package (*25*). They are minimized with the MM+ force field implemented in this software. The obtained geometries are then optimized in the Gaussian 98 program on the basis of the ab initio Hartree−Fock method at the 3-21G level (*26*). Afterward, the structures are superimposed in such a manner that R_1 is always defined as the substituent possessing a heteroatom. For compounds without a heteroatom, the methyl group is considered to be located at position R_1. In the case of the other parts of the pyrazines, showing neither a heteroatom nor a methyl group, the longest side chain is decided to be placed at position R_2. With the TSAR 3.21 (Tools for Structure−Activity Relationships) software different properties for all structures are calculated, among them steric descriptors (e.g., molecular mass, molecular surface, molecular volume, and Verloop parameters), molecular refractivity as a measure of polarizability, atom counts (carbon atoms for each substituent and different heteroatoms for substituent R_1), and the sum of electrotopological indices, which gives information about the electronic and topological state of the atoms in the molecule (*27*). Furthermore, the Hartree−Fock-derived dipole moments and point charges on the atoms of the heterocycle and the first atoms of the four substituents R_1−R_4 are used as electrostatic descriptors.

Table 2. Encoding Scheme for the Data Sets Used for the SOMFA Investigations

model	earthy	green-earthy	green	test compounds
1	0.0	0.5	1.0	**2**, **8**, **13**, **19**, **22**, **24**, **31**, **35**, **43**, **49**
2	1.0	X[a]	0.0	**2**, **5**, **8**, **10**, **13**, **35**, **41**, **43**, **46**, **49**
3	0.0	X	1.0	**2**, **5**, **8**, **10**, **13**, **35**, **41**, **43**, **46**, **49**
4	1.0	0.0	X	**2**, **5**, **8**, **10**, **13**, **19**, **22**, **24**, **25**, **31**
5	X	0.0	1.0	**19**, **22**, **24**, **25**, **31**, **35**, **41**, **43**, **46**, **49**
6	1.0	1.0	0.0	**2**, **8**, **13**, **19**, **22**, **24**, **31**, **35**, **43**, **49**
7	0.0	1.0	1.0	**2**, **8**, **13**, **19**, **22**, **24**, **31**, **35**, **43**, **49**

[a] X = not included in this calculation.

SOMFA calculations are performed using seven different subsets of the pyrazines as data set. The encoding schemes for the various calculations are depicted in **Table 2**.

The first investigation includes all 50 pyrazines and aims at a classification of all structures into the three studied aroma impressions. The activity is, therefore, encoded as 0.0 for earthy-smelling described structures, 1.0 for green flavor, and 0.5 for the mixture of those aroma qualities. As SOMFA distinguishes between active and less active molecules, a value of 0.5 for green-earthy means that this flavor could be described as half-active for earthy and green aroma qualities. For most cases this is not exactly true, as a mixed aroma impression on the one hand can result from dominant aroma impressions with weaker tonalities or, on the other hand, may represent a real mixture of aroma qualities of equal intensity. To verify if an application of SOMFA on smaller subsets of the data set allows additional information to be gained, six more models are calculated. Each of these models includes only two different aroma classes, whereby in all cases one of the aroma impressions is set to an activity value of 1.0 ("active") and the other aroma quality to 0.0 ("not active"). For all of the studies, shape and electrostatic potential based SOMFA models are calculated. To sum up the predictive power of these two properties into one final model, we combine their individual predictions using a mixing coefficient (c_1) as illustrated in eq 1 (*15*).

$$\text{activity} = c_1(\text{activity}_{\text{shape}}) + (1 - c_1)(\text{activity}_{\text{ESP}}) \quad (1)$$

The quality of the resulting models is proven by calculation of the correlation coefficient (r), the standard deviation (s), and the F value (F). Moreover, the predictive power of the models is checked by comparing the forecasted aroma impressions with the aroma descriptions from literature.

Table 3. Statistics of the Various SOMFA Models[a]

model	r[a]	s	F	c_1
1	0.734	0.161	44.51	0.7
2	0.882	0.197	66.90	0.8
3	0.882	0.197	66.90	0.8
4	0.887	0.180	77.53	0.6
5	0.732	0.214	27.72	0.5
6	0.658	0.180	28.98	0.6
7	0.831	0.176	84.50	0.7

[a] r = correlation coefficient; s = standard deviation; F = F value; c_1 = steric contribution.

ANN analyses are performed with the TRAJAN software package (*28*). Two different types to encode the target values are used. A multilayer perceptron (MLP) network architecture is trained on a nominal output variable. In this case, the network has to learn to distinguish among the three aroma classes, whereby a structure can have only one of these defined impressions. On the other hand, a general regression neural network (GRNN) is trained on a numerically defined output, where one of the two aroma impressions is set to 1.0 ("present") and the other one is set to 0.0 ("absent"). In the case of the mixed aroma impression green-earthy, the output values for green as well as for earthy are set to 1.0. The trained GRNN has to decide for every molecule if it shows (1) an earthy aroma impression and (2) a green aroma quality. As the output values of the GRNN can be interpreted as probabilities of showing a distinct aroma impression, this type of network also allows a quantitative interpretation of the observed results. For both network types, statistics such as the percentage of correctly predicted cases or correlation coefficients are calculated. The predictive quality of the ANN is evaluated by comparison of the known aroma impression from the literature with the predicted ones received from ANN.

RESULTS

SOMFA Studies. SOMFA calculations for both shape and electrostatic potentials are performed and combined according to eq 1. Results and statistics for the various models are summed in **Table 3**. Generally, it turns out that the steric contribution is of higher importance ($c_1 > 0.5$). To get some idea about the external predictive ability of the different SOMFA models, each of the studied subsets is divided into a training and a test set by excluding 10 pyrazines as test compounds. The compounds of the test sets are given in **Table 2**. The number of misclassified cases within the different SOMFA models is depicted in **Table 4**.

The best statistics for the discrimination of the whole data set into the three aroma impressions show a correlation coefficient, r, of 0.734, a standard deviation, s, of 0.161, and an overall F value of 44.51. By convention we decide that activity predictions up to a value of 0.4 belong to the earthy aroma impression, activities between 0.4 and 0.6 characterize a mixed aroma impression, and values >0.6 are typical for green-smelling pyrazines. Using this decision rule, the activity predictions obtained from the model are compared to the aroma descriptions from the literature. As can be seen from **Table 4** the misclassification within model 1 is rather high, as 16 of 50 pyrazines are not correctly predicted. To get some idea if the used type of output encoding (0.0−0.5−1.0) is perhaps not useful for SOMFA, the data set is divided into smaller subsets, which are then investigated by the same procedure. The difference to model 1 is that the remaining models include only two aroma impressions each and that SOMFA therefore should distinguish between active and not active. The limit between the two classes is set to 0.5, meaning that activity predictions <0.5 are decided to be not active, whereas values >0.5 stand for active. SOMFA calculations 2 and 3 include the earthy and the green-smelling described pyrazines. Both models show a correlation coefficient r of 0.882 and a standard deviation s of 0.197. Investigations 4 and 5 gain some differentiation between earthy and green-earthy as well as between green and green-earthy, respectively. As can be seen from **Tables 3** and **4**, model 4 results in 100% correct classification combined with rather good statistical values ($r = 0.887$, $s = 0.180$, $F = 77.53$). On the other hand, the distinction between green and green-earthy-smelling pyrazines (model 5) is rather poor. Six of the studied pyrazines are not correctly forecasted by this investigation. A rather low correlation coefficient r of 0.658 combined with a high misclassification rate is obtained by model 6, within which the earthy and green-earthy groups are regarded as active, whereas green is described as not active. Finally, combination of the green and green-earthy classes into one aroma impression and comparison to the earthy one (model 7) results in a correlation coefficient r of 0.831 and a standard deviation s of 0.176. The misclassification rate of this model is low, as only one of the structures is not correctly predicted.

ANNs. Nominal classification is performed by comparison of the three output neuron activation levels to two threshold values, namely, the accept and the reject thresholds. Values above the accept threshold are classified as positive (case belongs to the class represented by this output neuron), whereas activations below the reject threshold are classified as negative (case does not belong to the class represented by the output neuron). If the activation value lies between the accept and reject thresholds, the case is not classified. As for our studies the accept threshold is set to 0.0 and the reject threshold to 1.0; unclassified cases do not appear. The 50 structures are split randomly into three sets: 30 pyrazines are used for training, 10 compounds (compounds **6**, **8**, **9**, **15**, **27**, **30**, **39**, **42**, **44**, and **50**) for verification, and 10 structures (compounds **12**, **13**, **24**, **25**, **33**, **36**, **38**, **40**, **43**, **48**) for testing the neural network. The best classification of the 50 pyrazines (94%) into the three different groups of aroma impressions is obtained by back-propagation training of a multilayer perceptron network (MLP) architecture with seven input neurons, one hidden layer containing two neurons, and the three defined output neurons, one for each aroma impression. The MLP is composed by interconnecting these neurons, whereby a weight is associated with each connection. These weights are randomly initialized and iteratively optimized during the learning phase by back-propagating the error function from the outputs to the inputs (*29*).

The seven input neurons of the trained neural network contain the values of the following descriptors: charge of the first atom of the substitutent R_1, sum of electrotopological indices, number of heteroatoms at substituent R_1, dipole moment of the whole structure, number of carbon atoms at substituent R_2, molecular surface of the substituent R_1, and finally the charge of the carbon atom C_4 within the heteroaromatic ring. These seven inputs are obtained by sensitivity analysis on the various descriptors. The sensitivity analysis gives information about the relative importance of the variables. Therefore, the data set is submitted to the network repeatedly, with each variable in turn treated as missing, and the resulting network error is recorded. Removal of an important variable results in a significant increase of the network error. For the determination of the number of neurons in the hidden layer, the empirical rule mentioned by So and Richards is taken into account (*30*). Therefore, the ratio of the

Table 4. Number of Misclassified Cases (Code As Given in Table 1 in Parentheses) within the Various SOMFA Models[a]

	training			test		
model	earthy	green-earthy	green	earthy	green-earthy	green
1	1 (**1**)	6 (**16**, **17**, **21**, **26**, **29**, **33**)	6 (**41**, **42**, **44**, **46**, **47**, **50**)	1 (**13**)	2 (**22**, **24**)	0
2	0	X[a]	0	0	X	1 (**46**)
3	0	X	0	0	X	1 (**46**)
4	0	0	X	0	0	X
5	X	1 (**17**)	2 (**47**, **48**)	X	0	3 (**41**, **46**, **49**)
6	0	1 (**17**)	5 (**41**, **44**, **46**, **47**, **48**)	0	0	1 (**49**)
7	0	0	1 (**46**)	0	0	0

[a] X = not included in this calculation.

Table 5. Classification Statistics of ANN with Nominal Output

	training			verification			test		
	earthy	green-earthy	green	earthy	green-earthy	green	earthy	green-earthy	green
total	9	13	8	3	3	4	2	3	5
correct	9	13	7	3	2	4	2	3	4
wrong	0	0	1	0	1	0	0	0	1

number of input samples to the number of adjustable weights should result in a value for ρ in the range of $1.8 < \rho < 2.2$. If $\rho < 1.0$, the network simply memorizes the data, whereas for $\rho > 3.0$ the network is not able to generalize. Using an MLP architecture with seven input neurons, two hidden neurons, and three output neurons on our data set (30 training and 10 verification compounds), a value of 2.0 is obtained for ρ. Training of this network results in a test error of 0.166, a verification error showing a value of 0.187, and a test error of 0.258, whereby the errors are defined as the sum of the squared differences between the predicted and actual output on each output unit. As there are only 3 (compounds **15**, **48**, and **49**) of 50 pyrazines misclassified (**Table 5**), the performance of the network is rather high (94%). With regard to the "wrong" prediction of compound **15**, it should be mentioned that this compound is described as earthy by Wagner et al. (*6*). and as green-earthy by Boelens et al. (*21*). The ANN prediction model determines an earthy odor impression for this compound according to the information given by the second author.

Numerical classification is performed by calculation of correlation coefficients between the descriptors used as input for the neural network and the desired output values. For each compound two decisions have to be made: (1) Does this compound show an earthy odor? (2) Does the molecule possess a green odor? The best result is obtained by training of a GRNN. This network type achieves the estimation of the probability density function for each unknown pattern and predicts the most probable value of the dependent feature based on a finite number of measurements (*31*). The data set of 50 pyrazines is divided into 40 training and 10 test compounds (compounds **1**, **7**, **13**, **22**, **24**, **25**, **31**, **36**, **40**, and **44**); no verification set is needed for a GRNN. The GRNN used for this study is trained with the following four inputs, which are observed from sensitivity analysis: the number of heteroatoms at substituent R_1, the charge of the first atom of the substituent R_4, the shape flexibility of the whole molecule, and the number of carbon atoms at substituent R_3. As is usual for GRNNs, the first hidden layer contains one neuron for each training case (pattern layer). Furthermore, the network consists of three neurons in the second hidden layer (summation layer) and two output neurons, one for green and one for earthy aroma impression. By running the data set, we observe a final model with correlation coefficients for the earthy odor of 0.795 (training set) and 0.741 (test set), whereas correlations between the inputs and the green aroma impression are 0.946 and 0.973, respectively. The training as well as the test error of this model both show a value of 0.226. If we drag a limit between the "presence" and "absence" of an aroma impression at a value of 0.5 (no impression < 0.5 > impression), only a small misclassification rate can be observed. Six of the 100 decisions performed by GRNN are not correct (one quality character for each of the following structures: compounds **14**, **19**, **44**, **46**, **48**, and **49**). This results in a quite impressive correct classification rate of 94%. The prediction of the aroma impression of the test set is depicted in **Table 6**.

Table 6. Prediction of the Aroma Impression of the Test Set by GRNN

	earthy		green	
compd	actual	pred	actual	pred
1	1	0.943	0	0.286
7	1	1	0	0
13	1	0.966	0	0.107
22	1	0.776	1	1
24	1	0.772	1	1
25	1	0.705	1	1
31	1	0.976	1	1
36	0	0.002	1	1
40	0	0.003	1	1
44	0	0.870	1	0.782

DISCUSSION

Comparison of the results obtained by the different encoding types for the aroma impression and the two different calculation methods shows that for the given classification problem models of different predictive power are obtained. In general, statistics and classification performance of the SOMFA models are not as good as the ones obtained from ANN. SOMFA model 1 and the ANN model with nominal output consider the same problem. Both aim at a classification of the studied structures into the three aroma impressions earthy, green-earthy, and green. As can be seen from the results, the ANN works definitely better on this problem than SOMFA. This is an indication for nonlinear relationships within the given classification problem. Moreover, the ANN with numerically encoded output performs well on the aroma differentiation. This application additionally provides some quantitative information about the relative part of the aroma qualities green and earthy to the mixed impression. Application of SOMFA on smaller subsets allows more insight

into the appropriation of this and other two-dimensional methods on a classification problem. As for SOMFA models 2 and 3, totally equal statistical parameters are obtained, so it can be concluded that the final result of a model is not influenced by the decision of which one of the two studied aroma qualities is defined as active and which one as not active. Closer inspection of the remaining results of the SOMFA applications indicates that misclassification occurs especially within the green aroma quality. This seems to be due to the relatively high electrostatic similarity between the pyrazines assigned to the green and the green-earthy aroma impression, as—in contrast to the earthy-smelling structures—these two groups mostly contain a heteroatom at position R_1. If structurally similar compounds are described by strongly varying activities and are, therefore, assigned to different aroma impressions, a clear distinction between them is not possible by means of SOMFA.

Comparison of the correlation coefficients and classification performances of the different models points out that in general ANNs work better on the given classification problem. However, it has to be considered that the two different methods use different calculation procedures and variables to obtain their final models. In the case of SOMFA the three-dimensional structures and the electrostatic potentials of the studied molecules are used as input. On the other hand, ANNs build their decision on various two-dimensional descriptors and the Hartree−Fock-derived point charges. It is evident that the nonlinear handling of the data set leads to better results than the linear calculation procedure. A further advantage of ANNs is that their results allow more detailed insight into the structural differences between the studied aroma qualities. One important electrostatic descriptor used for the ANNs presented here is the charge of the first atom of the substituent R_1. This feature displays a characteristic difference between the studied aroma impressions. Substituent R_1 of the earthy-smelling pyrazines is usually a methyl group, whereas structures belonging to the green class mostly show a substituent R_1 containing a heteroatom. As a consequence, molecules of the green and earthy aroma qualities have different charges at this position. Moreover, the molecular surface of the substituent R_1 varies within these two groups. Additional information about the differences between green and earthy aroma qualities of pyrazines can be obtained from the charge of the carbon atom C_4 within the ring. This charge depends on the substituent R_4. If this substituent represents a short carbon side chain, the substances have a strong tendency to show an earthy aroma impression, whereas the green-smelling structures mostly contain a hydrogen atom at this position. Furthermore, the number of carbon atoms at substituent R_3 influences the differentiation between earthy and green aroma impressions. Generally, green-smelling pyrazines show a long side chain at this position, whereas a small substituent indicates an earthy impression. The mixed aroma impression seems to be a result of a combination of the described characteristics of the green and earthy aroma qualities, as green-earthy-smelling pyrazines usually contain a heteroatom at substituent R_1 (green), a hydrogen atom at position R_4 (green), and a small substituent at R_3 (earthy).

LITERATURE CITED

(1) Maga, J. A. Pyrazine Update. *Food Rev. Int.* **1992**, *8*, 479−558.

(2) Parliament, T. H.; Epstein, M. F. Organoleptic Properties of Some Alkyl-Substituted Alkoxy- and Alkylthiopyrazines. *J. Agric. Food Chem.* **1973**, *21*, 714−716.

(3) Takken, H. J.; van der Linde, L. M.; Boelens, M.; van Dort, J. M. Olfactive Properties of a Number of Polysubstituted Pyrazines. *J. Agric. Food Chem.* **1975**, *23*, 638−642.

(4) Masuda, H.; Mihara, S. Synthesis of Alkoxy-, (Alkylthio)-, Phenoxy-, and (Phenylthio)pyrazines and Their Olfactive Properties. *J. Agric. Food Chem.* **1986**, *34*, 377−381.

(5) Yoshii, F.; Hirono, S. Construction of a Quantitative Three-dimensional Model for Odor Quality using Comparative Molecular Field Analysis (CoMFA). *Chem. Senses* **1996**, *21*, 201−210.

(6) Wagner, R.; Czerny, M.; Bielohradsky, J.; Grosch, W. Structure-odour-activity relationships of alkylpyrazines. *Z. Lebensm. Unters. Forsch.* **1999**, *208*, 308−316.

(7) Buchbauer, G.; Klein, C. Th.; Wailzer, B.; Wolschann, P. Threshold-Based Structure−Activity Relationships of Pyrazines with Bell-Pepper Flavor. *J. Agric. Food Chem.* **2000**, *48*, 4273−4278.

(8) Wailzer, B.; Klocker, J.; Buchbauer, G.; Ecker, G.; Wolschann, P. Prediction of the Aroma Quality and Threshold Values of Some Pyrazines Using Artificial Neural Networks. *J. Med. Chem.* **2001**, *44*, 2805−2813.

(9) Chastrette, M.; El Aidi, C. In *Neural Networks in QSAR and Drug Design;* Devillers, J., Ed.; Academic Press: London, U.K., 1996; pp 83−96.

(10) Zakarya, D.; Cherqaoui, D.; Esseffar, M.; Villemin, D.; Cense, J. M. Application of neural networks to structure sandalwood odour relationships. *J. Phys. Org. Chem.* **1997**, *10*, 612−622.

(11) Zakarya, D.; Chastrette, M.; Tollabi, M.; Fkih-Tetouani, S. Structure-camphour odour relationships using the Generation and Selection of Pertinent Descriptors Approach. *Chemom. Intell. Lab.* **1999**, *48*, 35−46.

(12) Ho, C. T. Thermal Generation of Maillard Aromas. In *The Maillard Reaction, Consequences for the Chemical and Life Science*; Ikan, R., Ed.; Wiley: New York, 1996; pp 27−53.

(13) Buttery, R. G.; Seifert, R. M.; Guadagni, D. G.; Ling, L. C. Characterization of Some Volatile Constituents of Bellpeppers. *J. Agric. Food Chem.* **1969**, *17*, 1322−1327.

(14) Murray, K. E.; Whitfield, F. The Occurrence of 3-Alkyl-2-methoxypyrazines in Raw Vegetables. *J. Sci. Food Agric.* **1975**, *26*, 937−986.

(15) Robinson, D. D.; Winn, P. J.; Lyne, P. D.; Richards, W. G. Self-Organizing Molecular Field Analysis: A Tool for Structure−Activity Studies. *J. Med. Chem.* **1999**, *42*, 573−583.

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

(17) Chastrette, M.; de Saint Laumer, J. Y. Adapting the structure of a neural network to extract chemical information. Application to structure odor relationships. *SAR QSAR Environ. Res.* **1992**, *1*, 221−231.

(18) Chastrette, M.; Zakarya, D.; Peyraud, J. F. Structure musk odour relationship studies for tetralins and indans using neural networks. *Eur. J. Med. Chem.* **1994**, *29*, 343−348.

(19) Cherqaoui, D.; Esseffar, M.; Villemin, D.; Cense, J. M.; Chastrette, M.; Zakarya, D. Structure musk odour relationship studies of tetralins and indan compounds using neural networks. *New. J. Chem.* **1998**, *22*, 839−843.

(20) Peterson, K. L. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds; Wiley-VCH: New York, 2000; Vol. 16, pp 52−140.

(21) Boelens, M. H.; van Gemert, L. J. Structure−Activity Relationships of Natural Volatile Nitrogen Compounds. *Perfum. Flavor.* **1995**, *20* (Sept/Oct), 63−76.

(22) Masuda, H.; Mihara, S. Olfactive Properties of Alkylpyrazines and 3-Substituted 2-Alkylpyrazines. *J. Agric. Food Chem.* **1988**, *36*, 584−587.

(23) Shibamoto, T. Odor Threshold of Some Pyrazines. *J. Food Sci.* **1986**, *51*, 1098−1099.

(24) Pittet, A. O.; Hruza, D. E. Comparative Study of Flavor Properties of Thiazole Derivatives. *J. Agric. Food Chem.* **1974**, *22*, 264−269.

(25) Hyperchem 5.0, Hypercube Inc., 1997.

(26) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J.

M.; Daniels, D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98*, revision A.6; Gaussian, Inc.: Pittsburgh, PA, 1998.

(27) TSAR 3.2, Oxford Molecular, Ltd., 1999.

(28) Trajan Neural Networks 4.0, Trajan Software Ltd., 1999.

(29) Chtioui, Y.; Panigrahi, S.; Marsh, R. Conjugate gradient and approximate Newton methods for an optimal probabilistic neural network for food color classification. *Opt. Eng.* **1998**, *37*, 3015−3023.

(30) So, S.; Richards, W. G. Application of neural networks: Quantitative structure−activity relationships of the derivatives of 2,4-diamino-5-(substituted-benzyl)-pyrimidines as DHFR inhibitors. *J. Med. Chem.* **1992**, *35*, 3201−3207.

(31) Chtioui, Y.; Panigrahi, S.; Francl, L. A. Generalized regression neural network and its application for leaf wetness prediction to forecast plant disease. *Chemom. Intell. Lab.* **1999**, *48*, 47−58.

Received for review December 17, 2001. Revised manuscript received April 9, 2002. Accepted April 12, 2002.

JF011664A