

Quantitative Relationships Between Structure and the Fruity Odor of Esters

By Karen Rossiter, Quest International, Ashford, Kent, England

Over the last ten years an explosive growth in the development of computer hardware has been matched by the appearance of many commercial and academic molecular modeling and structure-activity relationship (SAR and QSAR) packages. (The "Q" is used in QSAR when describing the structure of the compound in a quantitative way, the simplest examples of quantitative descriptors being the mass of the compound or the number of atoms present.) The molecular modeling and SAR packages have been primarily developed within the arena of the drug and agrochemical industries in an attempt to ease the search for relationships between structure and activity. The fragrance chemist has followed in the footsteps of the drug designer and used similar techniques in his own SAR studies. However, there have been very few studies comparing the usefulness of different SAR approaches to the field of olfaction. In addition, the rapid developments currently being made within the SAR field mean there is a continuing need to evaluate the use of new molecular modeling and SAR techniques in the study of structure-odor relationships.

This article describes the evaluation of three QSAR approaches (CoMFA, Hansch and Principal Component Analysis) which were used to investigate the correlation between chemical structure and the fruitiness of esters. It includes the first published study of the use of comparative molecular field analysis (CoMFA) in the formulation of a structure-odor relationship.

Purpose of the Study

The objective of this work was to evaluate the relative usefulness of various QSAR approaches in understanding and predicting the odor properties of chemicals. The odor property chosen for this initial assessment was the perceived intensity of the fruity character of 27 aliphatic esters. This data set was considered to be ideal for the following reasons:

- Good reproducible odor data was already available in-house.
- The data set is of a suitable size for QSAR work.
- The structural variation exhibited by these compounds is limited (position of the ester group, the pattern of substitution and molecular weight).

If useful QSAR models could not be obtained for this data

set, it is highly unlikely that the same techniques could be applied successfully to more complicated problems. Thus the results from this work provide a good initial indication as to the usefulness of various QSAR approaches in the field of olfaction.

Materials and Methods

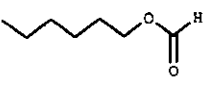
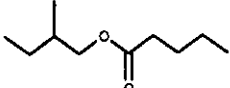
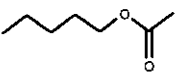
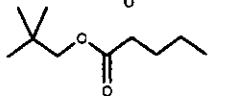
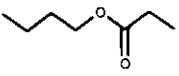
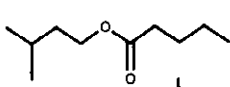
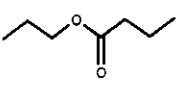
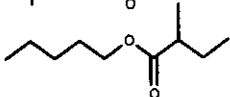
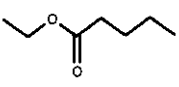
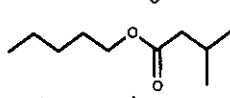
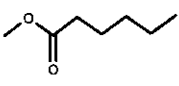
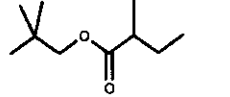
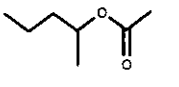
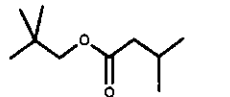
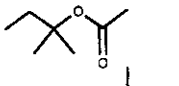
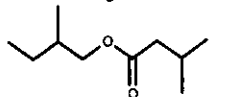
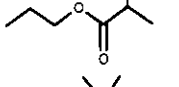
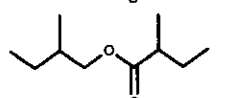
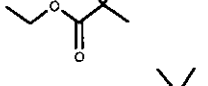
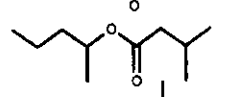
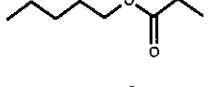
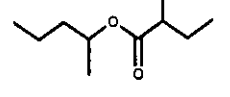

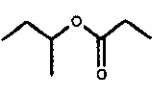
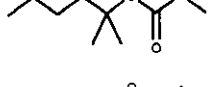
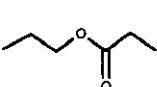
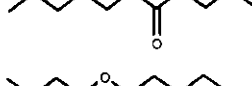
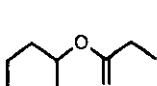
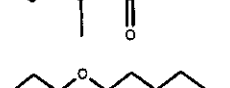
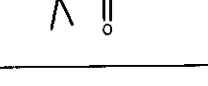
The materials for study were prepared by standard synthetic procedures. Each product's structure and purity (>99%) were confirmed by GC, GC/MS, NMR and IR spectroscopy.

The odor of each ester was profiled by Quest's sensory analysis team. The sensory panel consisted of a pool of 28 assessors, all of whom were trained to be able to identify standard odors, both individually and in complex mixtures, and to score their perceived intensity using a ratio scoring technique known as magnitude estimation.¹ The materials were assessed as 10% solutions in diethyl phthalate. At each session, six to eight panelists profiled the odor of six to eight esters by scoring the perceived intensity of 51 standard descriptors. The assessments were replicated until a minimum of 20 had been obtained for each ester. The panelists' scores were normalized and averaged to give a consensus profile for each ester. Assessments across the complete set of esters were standardized by including pentyl ethanoate in each test.

The fruity descriptors consisted of pear, peardrop, pineapple, apple and lactone. The combined fruit score, which in all cases was dominated by the pear or peardrop score, was used in this SAR study. Sensory data of this kind, which uses human subjects to quantitatively describe differences in odor character, is innately variable. However, by using highly trained and experienced panel members and carefully designed experiments, this variability can be minimized. Over the range of data obtained in this study, care should be taken not to interpret differences of less than 10 as anything more than variability in the data. The fruit score, structure and reference code for each ester in the study are listed in Table I.

Several compounds in the data set contain one or even two asymmetric carbon atoms and, as such, exist as isomeric mixtures. Ideally these isomers should be separated and their individual odor properties ascertained. However, this would be very time consuming, and extreme care would

Table I. Structures and fruit scores of the 27 esters and the structures of the four test esters

Structure	Reference	Fruit Score %	Structure	Reference	Fruit Score %
	A	24		Q	56
	B	100		R	72
	C	92		S	80
	D	47		T	79
	E	85		U	66
	F	81		V	19
	G	34		W	37
	H	0		X	55
	I	44		Y	54
	J	32		Z	36
	K	31		AA	32
	L	25		AB	?
	M	11		AC	?
	N	65		AD	?
	O	40		AE	?
	P	22			

Key to Table I.

A = hexyl methanoate
 B = pentyl ethanoate
 C = butyl propanoate
 D = propyl butanoate
 E = ethyl pentanoate
 F = methyl hexanoate
 G = pent-2-yl ethanoate
 H = 2-methylbut-2-yl ethanoate
 I = propyl 2-methylpropanoate
 J = ethyl 2,2-dimethylpropanoate

K = pentyl 2,2-dimethylpropanoate
 L = 2-methylhex-2-yl ethanoate
 M = 2-methylhex-2-yl 2,2-dimethylpropanoate
 N = pentyl pentanoate
 O = pent-2-yl pentanoate
 P = 2-methylbut-2-yl pentanoate
 Q = 2-methylbutyl pentanoate
 R = 2,2-dimethylpropyl pentanoate
 S = 3-methylbutyl pentanoate
 T = pentyl 2-methylbutanoate
 U = pentyl 3-methylbutanoate

V = 2,2-dimethylpropyl 2-methylbutanoate
 W = 2,2-dimethylpropyl 3-methylbutanoate
 X = 2-methylbutyl 3-methylbutanoate
 Y = 2-methylbutyl 2-methylbutanoate
 Z = pent-2-yl 3-methylbutanoate
 AA = pent-2-yl 2-methylbutanoate
 AB = but-2-yl propanoate
 AC = propyl propanoate
 AD = cyclohexyl propanoate
 AE = cyclopentyl propanoate

have to be taken to make sure that any observed differences in odor were not due to the presence of trace impurities. This is often the case with flexible molecules, where the reported differences tend to be so small that they have often been explained away as due to trace impurities. The most striking differences in odor tend to occur in rigid molecules such as carvone and menthol. For this data set of very flexible molecules it was assumed that the stereoisomers exhibited identical odor properties; thus, the fruit odor score of the isomeric mixture was used.

The implications of chirality in the development of a QSAR are as follows. If physicochemical parameters or 2-D structural descriptors are used to describe structural variation, the resulting model will not be able to discriminate between enantiomers even if they do exhibit different biological activity. The best way of discriminating between enantiomers is by using three-dimensional QSAR techniques such as CoMFA or conformational analysis.

The molecular modeling, CoMFA analysis and partial least squares statistical analysis were carried out using the Tripos SYBYL software version 6.03 and 6.1 on an IRIS INDIGO ELAN workstation. The operating system was IRIX version 4.05F. Forward stepwise regression, backward elimination regression and principal component analysis were carried out using the SAS software.

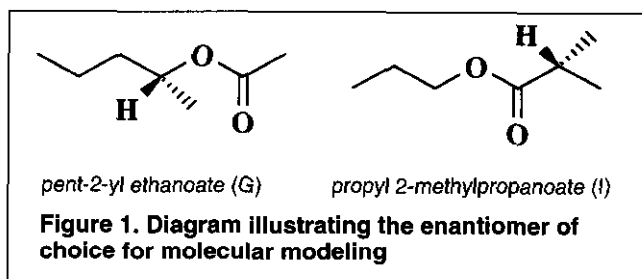
The CoMFA Approach

CoMFA, Comparative Molecular Field Analysis,² is a relatively new QSAR approach. It was introduced by Tripos Associates in 1988. To the best of my knowledge, this is its first use in the field of olfaction.

The idea underlying CoMFA is that differences in biological activity are often related to differences in the shapes of the fields surrounding the molecules. Thus, as its name suggests, CoMFA compares the steric and/or electrostatic fields of a set of molecules. These fields are measured by placing a hypothetical probe atom, usually an SP^3 hybridized C^+ atom, at regular positions around the molecule and, at these various locations, calculating the energy of interaction between the probe atom and the molecule. When the probe atom is close to the molecule the energy of steric interaction is high, and when it is close to an electron-rich moiety, such as an oxygen atom or p electrons, the energy of electronic interaction is high. Thousands of energy terms are calculated and become the structural descriptors in a QSAR table. These are then analyzed using partial least squares (PLS) analysis to see whether there is a correlation between the molecular fields and the biological activity. Partial least squares is one of the recommended statistical techniques for data sets which contain a much larger number of explanatory variables than compounds.³ It produces an equation relating biological activity (y) to the explanatory variables (x terms), which can be used to predict the activity of untested compounds. In this case y is the fruit score and the x terms are electrostatic and steric interaction energies. Since the equation resulting from a CoMFA procedure is very large, the model is also displayed graphically showing

the regions of desired and undesired steric bulk and the areas where negative potential is favorable or unfavorable.

The most important parameter in a CoMFA study is the relative alignment of the individual molecules when their fields are computed. Properly aligned molecules have a comparable conformation and a similar orientation in Cartesian space. For a series of molecules exhibiting the same biological activity via the same mechanism, it is assumed that there is a common conformational arrangement of key structural features. The so-called "active" conformational arrangement of fruity esters, or for that matter any odiferous molecule, is not known. It is therefore up to the organic chemist or molecular modeler to explore possible energetically favorable conformations that all of the molecules can adopt. In this study the chosen conformation was that in which the longest carbon chain backbone formed a staggered straight-chain arrangement. The esters were aligned using pentyl pentanoate as the template by superimposing the carbonyl carbon atom and the two ester oxygen atoms. The ester group was chosen for the alignment based upon the assumption that fruity esters may interact with olfactory receptors through hydrogen bonding with either or both of the ester oxygen atoms acting as hydrogen bond acceptors. In the case of the chiral esters, only one enantiomer was modeled, and it was assigned the fruity score of the mixture. Thus, for esters such as pent-2-yl ethanoate (C) and propyl



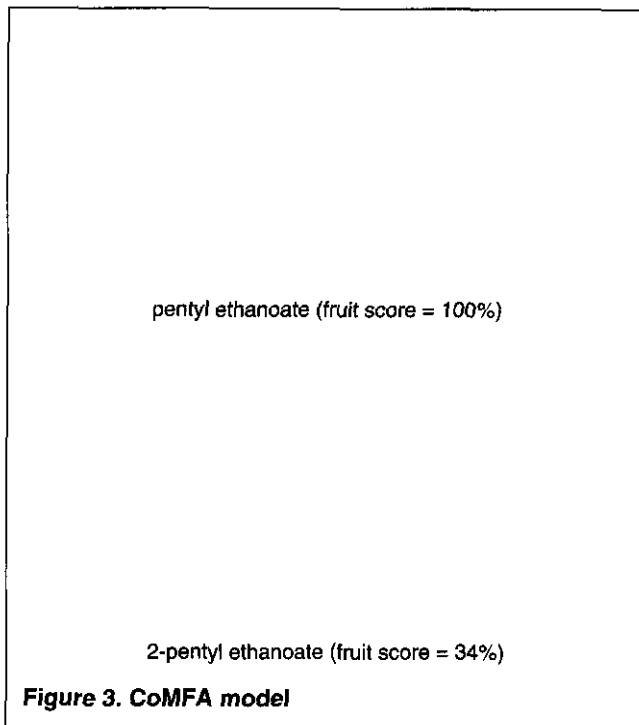
2-methylpropanoate (I), the ester chain was arranged as in Figure 1 and the methyl substituents placed behind the plane of the molecule. The 27 mutually aligned molecules are shown in Figure 2.

The first CoMFA model was poor. It could explain only 59% of variation in the observed fruit score (as indicated by the correlation coefficient R^2 of 0.59) and was expected, from cross-validation experiments, to be of poor predictive value (cross-validated R^2 , using the leave-one-out technique, was 0.27). The low value for the cross-validated R^2 suggests that either the assumptions underlying the model are incorrect or one or more of the compounds included in the model are outliers. The compounds that were poorly modeled were identified from a plot of the predicted vs. the actual fruit score. The worst outlier was hexyl methanoate (A) with a predicted fruity score of 83% and an actual score of 24%. Since the steric and electrostatic properties of methanoate esters are quite different from those of higher esters, this compound was omitted from the data set. Another poorly modeled compound was pentyl 3-methylbutanoate (U). The predicted fruit score for this compound was much higher than the observed score (124% predicted vs. 66% actual). However, this material also had a high sweaty score, which is probably due to the presence of trace quantities of isovaleric acid.* This sweaty note is likely to mask some of the fruity character of pentyl 3-methylbutanoate. Therefore, it was decided that this compound should also be omitted during rederivatization of the QSAR.

Omission of hexyl methanoate and pentyl 3-methylbutanoate resulted in an improvement in the model. The cross-validated R^2 , using the leave-one-out technique, was 0.52; the R^2 measure of fit was 0.84; the F value was 36.9; and the optimum number of components was 3. The fraction of the 84% variance explained by the electrostatic fields was 9%, and that explained by the steric fields was 91%. This suggests that the fruitiness of aliphatic esters is predominantly governed by steric effects, and that electrostatic effects appear not to be important. This is not too surprising since all of the compounds in this data will have similar electronic properties (they all contain only one functional group, the ester group, and this group is superimposed).

* The sample of pentyl 3-methylbutanoate was washed with base and subsequently determined to be >99% pure. However, a compound that appears to be pure may contain impurities present at levels below the detection limit of even today's sophisticated analytical techniques. If these are strongly odoriferous, as is the case with isovaleric acid, they will significantly affect the overall odor profile of a sample. This illustrates the importance of olfactory purity as opposed to chemical purity in the measurement of odors.

Figure 2. Alignment of 27 esters



The results from this QSAR analysis are shown graphically in Figures 3 and 4. Figure 3 shows the spatial distribution of important steric and electrostatic properties affecting the fruit score. Negative potential is favorable in the red areas (corresponding to the high electronic density of the ester group) and unfavorable in the blue areas. Bulky substituents are desirable in the green areas and undesirable in the yellow areas. To relate these regions to actual molecules, both pentyl ethanoate and pent-2-yl ethanoate have been placed within the CoMFA model. From the latter it can be seen that substituents close to the ethereal oxygen atom fall into the forbidden yellow region, resulting in a decrease in the fruit score. The location of the yellow region on one side of the molecule reflects the input of only one enantiomer for the chiral esters.

The graphs in Figure 4 are plots of the predicted vs. actual fruit score. For a perfect model the points would fall on a diagonal line from left to right. Points which deviate from this line correspond to compounds that are poorly modeled by the QSAR. This graph visually shows that the model is good. In fact, an R^2 of 0.84 is extremely good for a correlation between odor and structure. The difficulties associated with odor measurement generally give rise to a 10-15% variation in the individual odor scores. Thus, even if a perfect correlation between structure and odor was

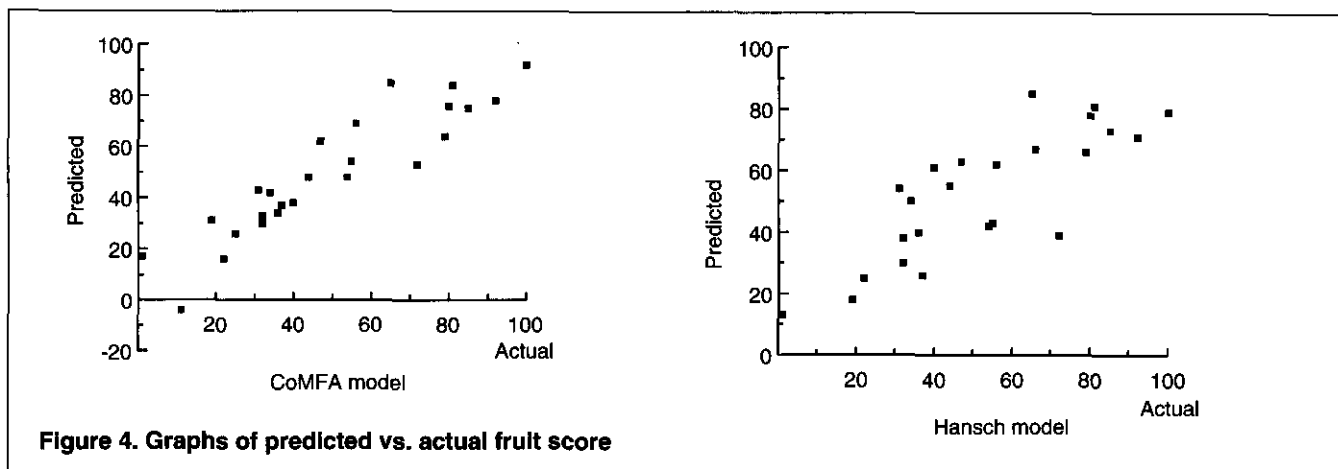


Figure 4. Graphs of predicted vs. actual fruit score

Table II. The CoMFA-predicted fruit scores and the observed fruitiness of the four test esters

Test compound	Predicted fruit score (%)	Observed fruitiness
but-2-yl propanoate (AB)	31	moderate
propyl propanoate (AC)	66	strong
cyclohexyl propanoate (AD)	38	strong
cyclopentyl propanoate (AE)	34	strong

obtained, one would not expect the R^2 to be much greater than 0.90. However, the cross-validated R^2 of 0.52 is indicative of a model of only moderate predictive ability.

The CoMFA procedure was repeated using only the achiral esters to determine whether or not a more robust model was obtained. Because chiral esters were excluded, we could remove the assumptions about the effect of chirality on the organoleptic properties of aliphatic esters. The resulting CoMFA model was very similar to that obtained using the larger data set ($R_2 = 0.84$, cross-validated $R_2 = 0.47$, optimum number of components = 3, $F = 23.3$, fraction of variance explained by the steric and electrostatic fields = 82% and 18% respectively). The spatial distribution of the important steric fields was also very similar except that there were now two areas where substituents were undesirable, one on either side of the molecule. This reflects the input of compounds such as 2-methylbut-2-yl ethanoate (H), 2-methylhex-2-yl ethanoate (L), 2-methylhex-2-yl 2,2-dimethylpropanoate (M) and 2-methylbut-2-yl pentanoate (P), which all have very low fruit scores. The CoMFA model using the larger data set had only one region of undesired steric bulk because the analysis was dominated by the presence of single enantiomers.

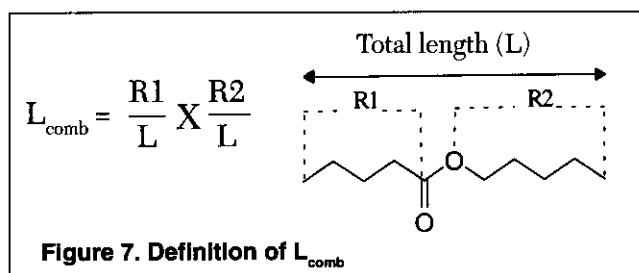
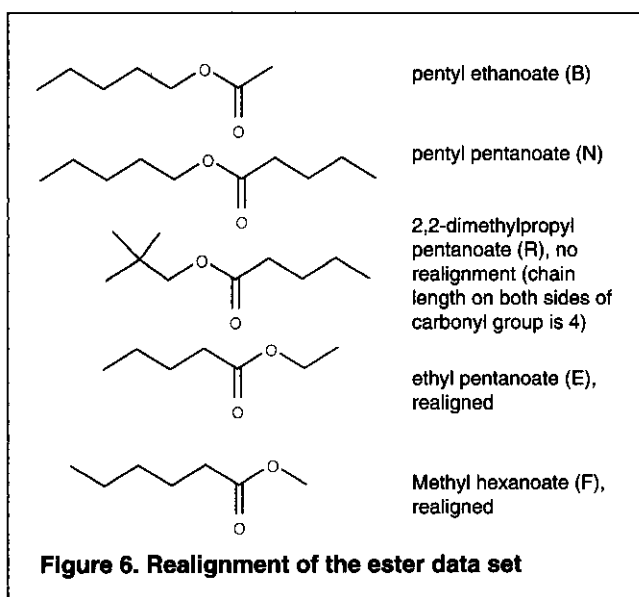
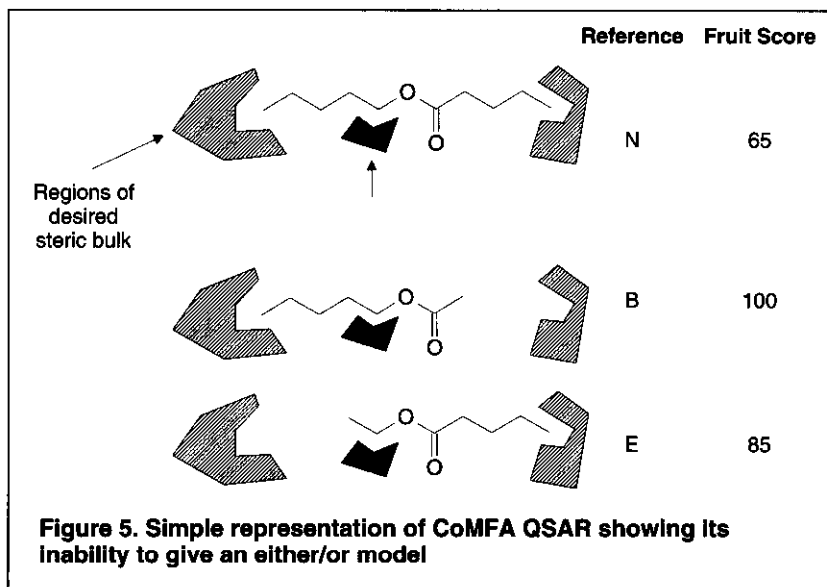
The CoMFA model was used to predict the activity of four test esters (Table II). These esters were prepared as part of a local school's chemistry project and as such have not been odor profiled by the Quest expert panel. Instead, the four esters were assessed by a panel of fragrance chemists and ranked as weak, moderate or strong. All of the esters were strongly fruity except for but-2-yl propanoate,

which was of only moderate fruit intensity. It was felt that this qualitative data would be adequate for testing the predictive ability of QSAR models.

In the case of the CoMFA model, it was concluded that accurate predictions were restricted to compounds (such as propyl propanoate and but-2-yl propanoate) which were very similar in structure to those used to derive the model. In contrast, extrapolation of the QSAR to predict the fruit score of compounds (such as cyclopentyl propanoate and cyclohexyl propanoate) that structurally fell outside the scope of the data set, gave poor results. Reasons for this limitation include the sensitivity of the CoMFA approach to changes in conformational arrangement and molecular alignment. For example, the predicted fruit scores for five different low energy conformations of cyclohexylpropanoate were in the range of 18-48%.

The CoMFA model has regions of desired steric bulk (green regions) on both sides of the ester function. Since both these areas are approximately five carbon atoms away from the functional group, the CoMFA model predicts that pentyl pentanoate (N) will have a very high fruit score. However, this is not the case. Pentyl pentanoate has a lower observed fruit score than, for example, either pentyl ethanoate (B) or ethyl pentanoate (E) (65% vs. 100% and 85%, respectively). From this it is clear that one of the limitations of the CoMFA approach for this data set is that it cannot give an either/or model. It cannot graphically represent that the most active esters have a long chain on one side of the ester group and a short chain on the other, and that it doesn't matter which way the ester group is oriented between the two chains. That is, as shown in Figure 5, the short chain can be attached either to the ether oxygen atom (-O-) (as in compound E) or to the carbonyl group (C=O) (as in compound B); the same is true for the long chain. Sell,^{4,5} during his studies into the SAR of fruit odor, also concluded that for an ester to be strongly fruity the ester group is best placed one or two carbons in from the end of the chain.

The above observations suggest that perhaps a better alignment of these esters is one in which the longest alkyl chain and the carbonyl group are superimposed. However, when this alignment was investigated, only two out of the 27



esters in the data set had to be realigned. They were ethyl pentanoate (E) and methyl hexanoate (F). This stems from the fact that the data set had originally been chosen to study the effect of steric hindrance while maintaining a constant molecular weight.^{4,5} One of the reasons for this is that increasing molecular weight is believed to be associated with a decrease in fruitiness.^{6,7} Consequently, the data set contains a large number of C10 isomeric esters derived from various C5 alcohols and C5 carboxylic acids. Since

alignment is based upon superimposition of the carbonyl group, the number of atoms, including oxygen, on the alcohol side is six and on the other side four. Thus the alcohol side of these C10 esters is always longer than the chain derived from the acid moiety, even when there is extensive branching as in the case of 2,2-dimethylpropyl pentanoate (R) (Figure 6). The CoMFA model derived from the data set containing the two realigned molecules was very poor.

The Hansch Approach

This approach is named after the founder of modern QSAR, Corwin Hansch, who suggested that the biological activity of a molecule was a function of its electronic, steric and hydrophobic properties. The biological activity is correlated to a range of physical and structural parameters in the form of a

regression equation. In the literature there are several examples in which the Hansch approach has been used to correlate odor with molecular properties. These include the odor threshold values of a series of homologous compounds,⁸ the odor quality of fruity and floral odorants,⁷ the odor similarities of ethereal, floral and benzaldehyde-like odorants and anosmia to fatty acids.⁹

Since the CoMFA approach had highlighted the importance of the position of the ester group in the chain, a parameter (L_{comb}) was invented to quantitatively describe this structural feature. L_{comb} was derived as follows. The length of the acid moiety (R_1) was defined as the distance between the carbonyl carbon atom and the terminus carbon atom, and the length of the alcohol moiety (R_2) as the distance between the ether oxygen atom and the end carbon atom (Figure 7). These lengths were expressed as a fraction of the total lengths (L) so comparisons could be made between compounds of varying chain length. The two fractions were then multiplied together to give L_{comb} . L_{comb} is low when the ester group is toward the end of the chain, irrespective of which end, and higher when the ester group is in the middle. A simple 2-D plot of L_{comb} against fruit score for six unsubstituted esters (B,C,D,E,F,N) shows, as is expected, that L_{comb} increases as the fruitiness of the ester decreases (Figure 8).

We evaluated combinations of L_{comb} with a variety of other parameters. One of the combinations which best accounted for the observed differences in the fruit score of aliphatic esters was equation 1 (Figure 9). This equation contains a molecular length term (L), L_{comb} and Charton steric substituent constants (ν). The Charton parameter is related to measured rates of hydrolysis of esters and, as such, is a measure of intramolecular steric effects around nearby reaction centers.¹⁰ The negative coefficient of the Charton parameters means that as the value of these constants increases, that is, as steric hindrance increases, the fruit score goes down. The larger coefficient of ν_{OR} means

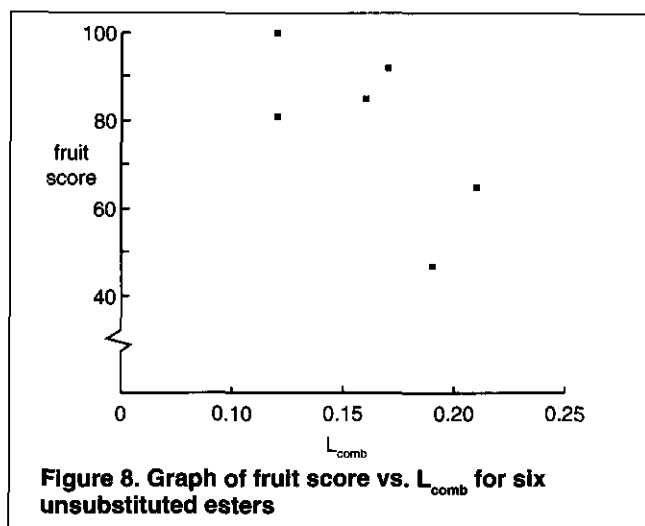


Figure 8. Graph of fruit score vs. L_{comb} for six unsubstituted esters

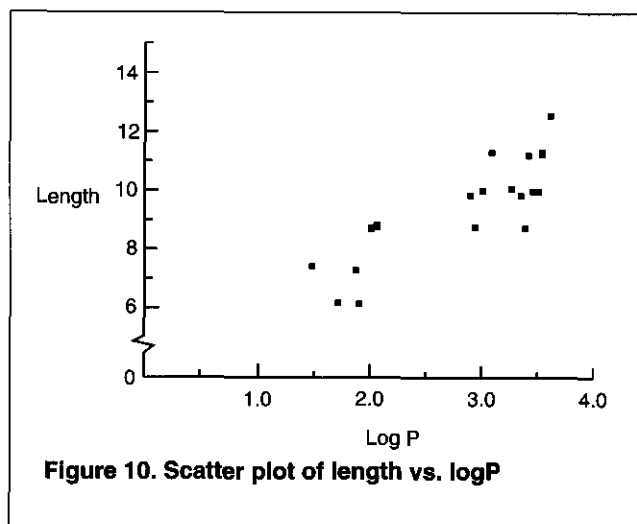


Figure 10. Scatter plot of length vs. $\log P$

$$\text{Fruit score} = 94.4 - 37.8v_R - 48.6v_{OR} + 6.1L - 135L_{comb}$$

$$(R^2 = 0.67, \text{ number of components} = 2, \text{ cross-validated } R^2 = 0.50, F = 21.34, n = 24)$$

where: v_R = Charton substituent constant for the alkyl group attached to C=O

v_{OR} = Charton substituent constant for the alkyl group attached to ether oxygen atom

L = molecular length

L_{comb} = descriptor for position of ester group in the chain

Figure 9. QSAR equation 1 obtained by partial least squares analysis

that steric hindrance around the ether oxygen atom has a greater effect than steric bulk around the carbonyl group. The positive coefficient for length (L) means that fruitiness increases as the molecule becomes longer. The negative coefficient of L_{comb} means that as L_{comb} increases, that is, as the ester group is placed toward the center of the chain, the fruity character is reduced.

The relative importance of the four variables was estimated using forward selection regression and backward elimination regression. In forward selection, variables are successively added to the model and retained if the fit of the model is significantly improved. The first variable entered will be the one that is most highly correlated, on its own, with the y variable (in this case, the fruit score). The second variable entered will be the one that causes the greatest maximization of R^2 , and so on. The backward elimination regression method begins with all the variables in the model and proceeds by eliminating the least useful variables one at a time. Both regression techniques showed that the most important parameter out of those evaluated was v_{OR} , the second most important v_R , the third either molecular length or the octanol-water partition coefficient ($\log P$), and the fourth L_{comb} .

$\log P$ can replace the role of length in the correlation QSAR because these two parameters are highly correlated (Figure 10). However, their physical interpretation in understanding structure odor correlations is quite different. Molecular length could be associated with shape requirements for optimum interaction with the receptor system,

whereas the balance between hydrophobicity and lipophilicity will affect the transport properties of a molecule across aqueous-lipid interfaces.

The correlation between molecular length and $\log P$ is clearly seen in the corresponding 2-D plot (Figure 10). However, a better estimate of the correlation between the two parameters can be obtained from the parameter correlation matrix (Table III). The elements of the correlation matrix can take on a value from -1 to $+1$, where $+1$ indicates perfect positive correlation between two parameters, -1 perfect negative correlation and a value of zero indicates no correlation. Thus the $+1.00$ values on the diagonal of Table III simply tell us that a parameter is perfectly correlated with itself. The pairs of parameters which are significantly correlated with each other are length and $\log P$ and—because molecular length is used in the derivation of L_{comb} —length and L_{comb} and $\log P$ and L_{comb} . It is interesting that the correlations between the fruit score and the parameters suggest a slightly different order of relative importance for the parameters than the forward selection and backward elimination regression analyses. Once again v_{OR} comes out as being the most important. However, there is little difference between the second two most important parameters (v_R and length) and it is somewhat surprising, based upon the observation that $\log P$ can replace the role of length in the QSAR equations, that $\log P$ is not correlated to the fruit score at all.

At first sight, the Hansch correlation QSAR (equation 1, Figure 9) does not appear to be as good as the CoMFA

Table III. Correlation matrix

	fruit score	v_R	v_{OR}	length	L_{comb}	logP
fruit score	+1.00	-0.30	-0.67	+0.36	-0.16	-0.01
v_R	-0.30	+1.00	-0.08	+0.08	+0.31	+0.36
v_{OR}	-0.67	-0.08	+1.00	-0.15	+0.25	+0.21
length	+0.35	+0.08	-0.15	+1.00	+0.62	+0.83
L_{comb}	-0.16	+0.31	+0.25	+0.62	+1.00	+0.72
logP	-0.01	+0.36	+0.21	+0.83	+0.72	+1.00

Table IV. The Hansch-predicted fruit scores and the observed fruitiness of the four test esters

Test compound	Predicted fruit score (%)	Observed fruitiness
but-2-yl propanoate (AB)	47	moderate
propyl propanoate (AC)	62	strong
cyclohexyl propanoate (AD)	58	strong
cyclopentyl propanoate (AE)	58	strong

model. Although they are expected to have similar predictive abilities, as indicated by their cross-validated R^2 values of 0.50 and 0.52 respectively, the Hansch model accounts for only 67% of the observed variation in fruit score, as opposed to the 84% explained by the CoMFA model. The same conclusion is drawn by comparing the corresponding graphs of the actual vs. predicted fruit scores (Figure 4). However, when the two models were put to the real test and used to predict the activity of the four test esters, it was found that the Hansch model, in contrast to the CoMFA model, was capable of both interpolative and extrapolative prediction (Table IV). This difference in predictive ability is believed to be due to the different methods used to quantify steric hindrance. Advantages of the Charton steric substituent constants include their relationship to measured rates of hydrolysis of substituted esters and their independence of the postulation of a single active conformation.

Other steric parameters, notably the Kier and Hall connectivity indices,¹¹ have also been used successfully in the Hansch approach to correlate odor with molecular structure.^{6,9,12} These indices are simply calculated from the topology of the molecule and, as such, represent the size and extent of branching in a molecule. Data sets such as these are ideal for further comparisons of the Hansch and CoMFA approach. Preliminary investigations at Quest International into the use of CoMFA to correlate Amoore's¹³ bitter almond data set were disappointing. On the other hand, Kier⁹ and Dearden¹² both obtained good correlations using various connectivity indices in the Hansch approach.

Principal Component Analysis

Principal component analysis (PCA) is a technique used to reduce an n-dimensional data set to either a two- or three-dimensional data set. This allows the display of n

Table V. Variance explained by the principal components

PC	Eigenvalue	Proportion of variance	Cumulative variance
1	1.779	44.4%	44.4%
2	1.076	26.9%	71.4%
3	0.808	20.1%	91.6%
4	0.337	8.4%	100.0%

molecular properties on a simple 2-D or 3-D plot, where each point represents a compound which can be coded according to either its chemical name or biological activity. If the chosen molecular properties are indeed important in determining the biological activity of interest, then distinct clusters will be formed for compounds of different activity. This technique is thus well suited for classified biological data such as active/inactive, strong/moderate/weak and odor descriptors. Since the majority of odor data is of a classified nature, the evaluation of QSAR techniques in the field of olfaction would be incomplete without the inclusion of a classification method, such as PCA.

Reduction in dimensionality is achieved by the creation of principal components (PCs), each being a linear combination of all the original explanatory variables. The first principal component explains the greatest variation in the data cloud, the second the next maximum variation and so on. The principal components become the new axes in the 2-D or 3-D plot.

The fruit score for each ester was converted to a qualitative odor measurement by ranking the ester as weak, moderate or strong according to the following rules:

<u>% fruit score</u>	<u>ranking</u>
< 33	weak
33-66	moderate
> 66	strong

Four principal components were created from the four variables used in the Hansch approach: the two Charton parameters v_{OR} and v_R , molecular length and L_{comb} . The results are summarized in Tables V and VI. Each principal component (PC) has an associated eigenvalue which is related to the proportion of variance explained by the PC. Each new PC explains successively less of the total variance. As a rule of thumb, a PC which has an eigenvalue greater than 1 contains more information than a single variable and, as such, should be included in the final PC analysis. Using this criterion, this data set can be reduced from four dimensions to two dimensions; the two PCs explain 71.4% of the original variance.

The new principal components can be interpreted using the eigenvectors (Table VI). These measure the contribution of each original property to each principal component. Thus, PC1 is predominantly made up of L_{comb} , length and v_R ; PC2 is predominantly v_{OR} ; PC3 is predominantly v_R .

Interestingly, a 2-D plot of PC1 and PC2, which in

Table VI. Eigenvectors

Variable	PC1	PC2	PC3	PC4
v_R	0.435	-0.235	0.861	0.118
v_{OR}	0.121	0.924	0.146	0.331
length	0.605	-0.228	-0.453	0.614
L_{comb}	0.655	0.195	-0.181	-0.707

combination explain 71.4% of the variation in the molecular property cloud, resulted in no well-defined clusters of different activity classes. This suggests that the variables that make the greatest contribution to PC1, namely length and L_{comb} , are not the molecular properties which primarily determine the fruitiness of the esters. These findings are in agreement with the results from the forward selection and backward elimination regression analyses. However, a plot of PC2 vs. PC3 produced some clustering of esters with strong, moderate and weak fruity odor characteristics (Figure 11). The strongly fruity esters form a tight cluster between PC2 values of -1.5 and -0.5 and PC3 values of -0.75 to +0.25. The moderately fruity esters radiate out from the "strong" cluster, and the weakly fruity esters are on the periphery of the graph.

In Figure 11, the boundaries between each class are somewhat "fuzzy" because of the arbitrary nature of the

classification. For example, compound G with a fruit score of 34% is ranked as moderate and compound J with a fruit score of 32% as weak. However, from the fruit scores it is clear that there is no significant difference between the fruitiness of these two esters. Consequently, the principal component analysis was repeated using only the strong and weak compounds to see whether or not this gave better discrimination. The results from this analysis were very similar to those obtained using all three classes of esters. The eigenvalues and eigenvectors were virtually identical. The PC2 vs. PC3 plot was also very similar to that obtained from the original analysis, with, of course, just the moderate points missing. Both models poorly modeled 2,2-dimethylpropyl pentanoate (R). This is an ester which has a strong fruity odor but which falls amongst the weak esters in the PC plot. Interestingly, it is also the main outlier in the Hansch model where, again, its activity is significantly under predicted (predicted fruit score = 35, observed score = 72).

The eigenvectors for PC2 and PC3 show that the Charton substituent constant for the alcohol moiety makes the greatest contribution to PC2, and that the Charton substituent constant for the acid moiety makes the greatest contribution to PC3. From the above results, one might expect that a simple 2-D plot of the two different Charton steric substituent constants would produce an equally good discriminating map. However, this is not the case. The boundaries between the different activity classes are less well defined, and there are examples of esters from different

activity classes with exactly the same x,y coordinates. Thus the contribution made by the other properties in PC2 and PC3 improves the discriminating ability of the resulting model and also gives unique coordinates for every compound in the data set.

The activity of the four test esters can be estimated from their position on the PC2 vs. PC3 plot (Figure 11). Although all four compounds lie within the moderate activity class, but-2-yl propanoate (AB) lies toward the moderate-weak boundary line and thus would be predicted, out of the four test esters, to be the one with the weakest fruity character, while propyl propanoate (AC), which is close to the strong-moderate boundary, would be predicted to be the strongest fruit ester. The two cyclic esters (AD and AE) would be predicted to have a fruit odor of moderate intensity. Thus the PCA model, although it uses the same explanatory variables as the Hansch model, appears to be poorer at predicting the activity of the two esters which structurally fall outside the scope of the test data set.

In order to make a better comparison between the principal component analysis and the Hansch model, the scores for the second and third principal components were used parameters in a linear regression. The resulting QSAR equation (equation 2, Figure 12) is comparable to equation 1. They both have similar R^2 values (0.64 and 0.67, respectively) and give similar predicted fruit score values for the four test esters. The predicted fruit scores obtained using equation 2 are 46% for but-2-yl propanoate (AB), 63% for propyl propanoate (AC), 55% for cyclohexyl propanoate (AD) and 57% for cyclopentyl propanoate (AE).

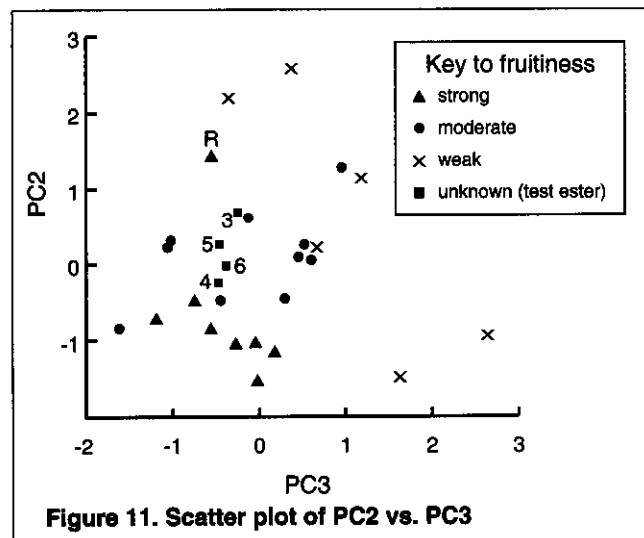


Figure 11. Scatter plot of PC2 vs. PC3

Conclusion

Each QSAR technique produced a model which was capable of relating the fruitiness of aliphatic esters to a limited number of molecular properties. Steric effects around the ester group were found to be the most important determining property in all three cases. The Hansch and CoMFA approaches also showed that the introduction of substituents on the alcohol side had a greater detrimental effect on the fruity character than substituents on the acid side. All of these results were in agreement with the findings of Sell,^{4,5} who in 1986 used his chemical expertise to draw qualitative conclusions about the effect of substitution on the intensity of fruit odor. It was thus concluded that these techniques can be used to identify sensible relationships between structure and odor, and that they can therefore be applied, with some degree of confidence, to more compli-

$$\text{Fruit score} = 52.97 - 14.26 \text{ PC2} - 13.40 \text{ PC3}$$

$$(R^2 = 0.64, F = 18.90, n = 24)$$

where: ν_R = Charton substituent constant for the alkyl group attached to C=O
 ν_{OR} = Charton substituent constant for the alkyl group attached to ether oxygen atom
 L = molecular length
 L_{comb} = descriptor for position of ester group in the chain

Figure 12. QSAR equation 2

cated problems where the relationship between structure and activity is not so obvious. One advantage of all of these quantitative techniques over empirical SAR rules is that they can be used to predict the activity score or activity class of a new compound. In addition, computer-assisted SAR allows the simultaneous comparison of hundreds of compounds, whereas the organic chemist is restricted to comparing only a few compounds at a time.

The CoMFA model is quick and easy to use once the molecules have been aligned. Its use in prediction is also particularly straightforward. The new compound only needs to be drawn, minimized and aligned, then "at a press of a button" its activity is predicted. The most important adjustable parameter in CoMFA is the relative alignment of the individual molecules. However, it is difficult to decide how to superimpose conformationally flexible molecules, or molecules which appear to have little in common in terms of obvious chemical functionality. Molecular modeling software companies have developed programs that are capable of identifying common conformational arrangements of a set of active compounds. However, these have been designed primarily for compounds that possess a high degree of chemical functionality, such as drug molecules. Odoriferous molecules, on the other hand, tend to have only one or two functional groups and a relatively large hydrophobic group. It is thus anticipated that the use of the CoMFA approach in olfaction will be restricted to the study of rigid molecules and closely related analogues.

The success of an SAR derived from statistical techniques such as the Hansch approach or principal component analysis is heavily dependent upon the number and quality of the selected descriptors. If the QSAR is going to be used to design new materials, one important criterion is that the parameters are understandable. Ideally, values of the explanatory properties should be available for every compound in the data set. However, this is often not the case, and the SAR worker is forced to omit from the analysis compounds which contain valuable information. The fruity ester study was no exception. The unavailability of three Charton steric substituent constants meant that the Hansch and PCA models were derived from a reduced data set of 24 compounds. This scenario is best avoided by the use of calculated properties, many of which can be obtained from molecular modeling systems. One property that is particularly difficult to quantify and thus include in a statistical QSAR approach is shape. Therefore, biological activity that is believed to be strongly dependent on shape, such as odor quality, is best studied using 3-D QSAR tools such as CoMFA or conformational analysis. Odor intensity, on the other hand, particularly when restricted to a specific odor type and chemical class, can be related to molecular properties using the statistical QSAR approaches reviewed here. It is believed that in these cases one can use a general QSAR equation, which includes a volatility term (such as vapor pressure or gas chromatography retention data), a hydrophobicity term (such as $\log P$) and a parameter related to steric hindrance around the osmophore group.

The potential for SAR in the field of olfaction is exemplified by a few studies where it is claimed that structure-odor correlations have led to the discovery of a new fragrance ingredient.¹⁴⁻¹⁸ As progress in the biological sciences leads to an increased understanding of the mechanism of olfaction, and as more sophisticated SAR tools are developed, the search for such correlations should become easier. This challenge, coupled with the potential predictive ability of this approach, will entice chemists and molecular modelers to continue research in this area.

Acknowledgements: I would like to thank all my colleagues at Quest, especially Dr. Anne Richardson for the odor profiling, Christopher Newman for the synthetic work, Steven Bociek for software management support and Martin Till for his assistance and teaching in the use of the SAS software. I am also grateful to my PhD supervisors, Dr. Charles Sell from Quest International and Dr. John Mitchell from University of Kent, England, for their helpful discussions and for their continuing support and encouragement.

References

Address correspondence to Mrs. Karen Rossiter, Quest International, Ashford, Kent, TN24 OLT, England.

1. HR Moskowitz, *Lebens-Wiss Technol* **8** 249 (1975)
2. RD Cramer III, DE Patterson and JD Bunce, *J Amer Chem Soc* **110** 5959 (1988)
3. LStahle and S Wold, in *Progress in Medicinal Chemistry*, GP Ellis and GB West, eds, Amsterdam, Netherlands: Elsevier (1988) pp 292-338
4. CS Sell, *Seifen Ole Fette Wachse* **112**(8) 267 (1986)
5. CS Sell, in *Flavors and Fragrances: A World Perspective*. Proceedings of the 10th International Congress of Essential Oils, Fragrances and Flavors, Washington D.C., Nov 1986, BM Lawrence, BD Mookherjee and BJ Willis, eds, Amsterdam: Elsevier Science Publishers BV (1988) pp 777-795
6. H Boelens, The influence of molecular structure on olfactive quality: A quantitative approach, in Proceedings 36th Tobacco Chemists, Research Conference, Raleigh NC, USA, October (1982)
7. H Boelens, HG Haring and HJ Takken, *Chem Ind* 26-30 (1983)
8. MJ Greenberg, *J Agric Food Chem* **27** 347 (1979)
9. LB Kier et al, *J Theor Biol* **67** 585-595 (1977)
10. M Charton, *JACS* **91** 615 (1967); *JOC* **44** 2284 (1979)
11. LB Kier and LH Hall, in *Molecular Connectivity in Chemistry and Drug Research*, G De Stevens, ed, New York: Academic Press (1976)
12. JC Dearden, *Food Quality and Preference* **5** 81 (1994)
13. JE Amoore, *Nature* **233** 270 (1971)
14. C Anselmi, M Centini, M Mariani, A Segal and P Pelosi, *J Agric Food Chem* **40** 853 (1992)
15. R Pelzer, U Harder, A Krömpel, H Sommer, H Surburg and P Hoever, in *Recent Developments in Flavor and Fragrance Chemistry*, Proceedings of the 3rd Int. Haarmann & Reimer Symposium, Apr 12-15, 1992, Kyoto, Japan, R Hopp and M Kenji, eds, Weinheim, Germany: VCH Publishers (1993) pp 29-67
16. H Boelens, PC Traas and HJ Takken, *Perf & Flav* **5**(1) 39 (1980)
17. C Fehr, J Galindo, R Haubrichs and R Pernet, *Helv Chim Acta* **72** 1537 (1989)
18. JA Bajgrowicz and C Broger, in *Flavours, Fragrances and Essential Oils*, Proceedings of the 13th International Congress of Flavours, Fragrances and Essential Oils, Istanbul, Turkey, 15-19 October 1995, vol 3, KHC Baser, ed, Istanbul: AREP Publ (1995) pp 1-15

