ARTICLE

OBC
www.rsc.org/obc

A structure–odour relationship study using EVA descriptors and hierarchical clustering†

Shin-ya Takane‡[a,b] and John B. O. Mitchell[a]

[a] *Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Cambridge, UK CB2 1EW*

[b] *Department of Information Systems Engineering, Faculty of Engineering, Osaka Sangyo University, Daito, Osaka 574-8530, Japan*

Received 30th June 2004, Accepted 10th September 2004
First published as an Advance Article on the web 28th September 2004

Structure–odour relationship analyses using hierarchical clustering were carried out on a diverse dataset of 47 molecules. These molecules were divided into seven odour categories: ambergris, bitter almond, camphoraceous, rose, jasmine, muguet, and musk. The alignment-independent descriptor EVA (EigenVAlue) was used as the molecular descriptor. The results were compared with those of another kind of descriptor, the UNITY 2D fingerprint. The dendrograms obtained with these descriptors were compared with the seven odour categories using the adjusted Rand index. The dendrograms produced by EVA consistently outperformed those from UNITY 2D in reproducing the experimental odour classifications of these 47 molecules.

Introduction

Olfaction is the least well understood of our five senses. Since it is a chemical sense, initiated by the interaction between odour molecules and receptors, it is of interest to know how the chemical properties of odorants determine the odour we perceive. An understanding of olfaction at the molecular level would also facilitate odour prediction, and hence is of interest to the fragrance industry. However, although in the case of taste there are only five characteristics (sweet, salty, sour, bitter, and umami) each corresponding to an independent receptor, the mechanism of olfaction is much more complicated. The number of types of olfactory receptors (ORs) expressed is thought to be about 1000 and 350 in the proteomes of mice and men, respectively.[1,2] The number of human proteins associated with olfaction is thus second in number only to the immune system. In addition, it has become apparent that one OR recognizes multiple odour molecules, that one odour molecule is recognized by multiple ORs, and that different odour molecules are recognized by different combinations of ORs. Thus, the olfactory system is believed to use a combinatorial receptor coding scheme to encode odour identities.[3] However, it is not yet clear which combinations of receptors correspond to specific odours or which characteristics of odour molecules are involved in recognition by ORs.

To elucidate the detailed mechanism of olfaction, one must investigate the interaction between odour molecules and the olfactory receptors at the molecular level. Unfortunately, no 3D structures of olfactory receptors are available yet. Thus, in the fields of fragrance and food science, empirical methods such as quantitative structure activity relationships (QSAR) are extensively used to design novel molecules from known datasets. A common approach within the field of fragrance SAR is 3D-QSAR (*e.g.*, conformational analysis and the search for olfactophores). It has, however, the problem that it requires molecular alignment (structural superposition), and thus cannot be applied to diverse datasets.

Most studies using 3D-QSAR methods like CoMFA are, at least implicitly, based on the "binding theory", which is widely accepted as the mechanism of interaction between odour molecules and the olfactory receptors. Under this theory, odorants are recognized and bound by the receptors, while molecules sharing an olfactophore—a three dimensional arrangement of functional groups associated with a particular odour—produce very similar responses from the receptors because they bind in similar ways. There is a competing theory that is very different from this. The "vibrational theory" goes back to Dyson's paper[4] in the 1930s, later extended by Wright,[5] and in 1996 Turin[6] revived it, in a modified form. The vibrational theory claims that the molecular vibrations of odour molecules are the direct cause of odour. The vibrational theory is controversial, to say the least, and is widely considered to be inconsistent with recent experimental evidence.[7]

Although we do not think that the evidence justifies belief in a causal link between molecular vibrations and odour, vibrational information does provide a convenient alignment-independent way to generate descriptors for odour molecules. Practically, the EVA descriptor based on the vibrational eigenvalues is available and has been applied to some diverse datasets in general structure–activity relationship studies. In this paper, we tackle the classification problem for a dataset containing 47 structurally diverse molecules (Fig. 1), by applying the alignment-independent QSAR descriptor EVA (Fig. 2). The results are also compared with those of another descriptor, the UNITY 2D fingerprint.

Results and discussion

The dendrograms obtained from EVA descriptors using the complete linkage method at various σ values (σ = 100, 50, 20 cm^{-1}) are shown in Fig. 3 (a)–(c). In Fig. 4 (a)–(c), the corresponding dendrograms obtained using the modified Ward's method at the same σ values are shown. For both these methods, as σ increases the shape of the dendrogram becomes slightly more compact because of smoothing by the Gaussian convolution. As a consequence of this effect, the difference of two EVA spectra is less affected by the peak shift caused by small changes or differences in molecular conformation than other 3D QSAR methods. Furthermore, the dendrogram from the modified Ward's method is more compact than that from the complete linkage method at the same σ. The dendrograms from UNITY 2D (Fig. 3 (d) and Fig. 4 (d)) are markedly bulkier and thus more diverse than those from EVA descriptors. This gives the visual impression that UNITY 2D generates a less clear-cut partition into appropriate clusters than does EVA, and this was subsequently confirmed by the numerical data from the adjusted Rand indices (see below).

† This is one of a number of contributions on the theme of molecular informatics, published to coincide with the RSC Symposium "New Horizons in Molecular Informatics", December 7th 2004, Cambridge UK.

‡ Permanent address: Department of Information Systems Engineering, Faculty of Engineering, Osaka Sangyo University, Daito, Osaka 574-8530, Japan

DOI: 10.1039/b409802a

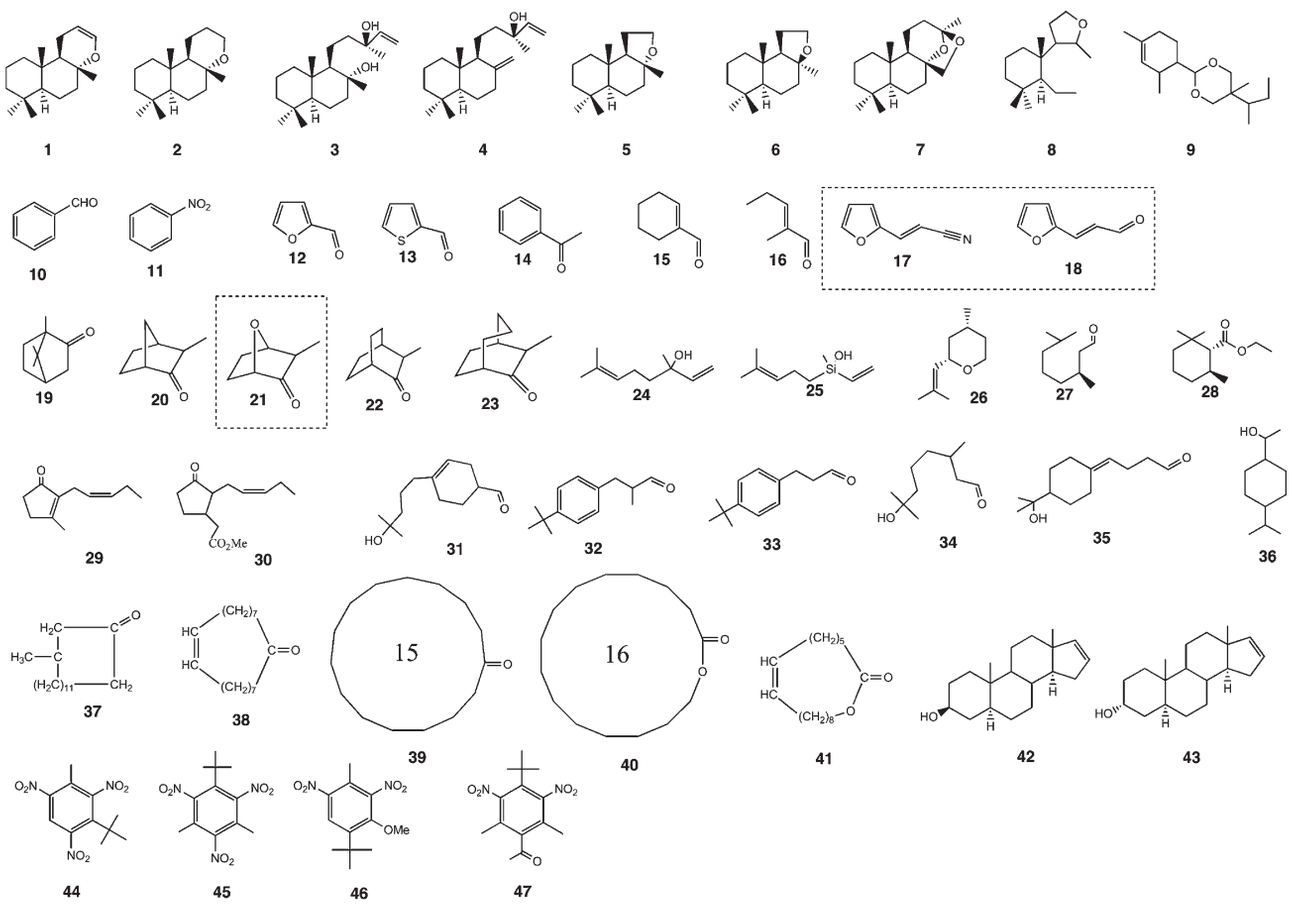

Fig. 1 Dataset used in the present study.

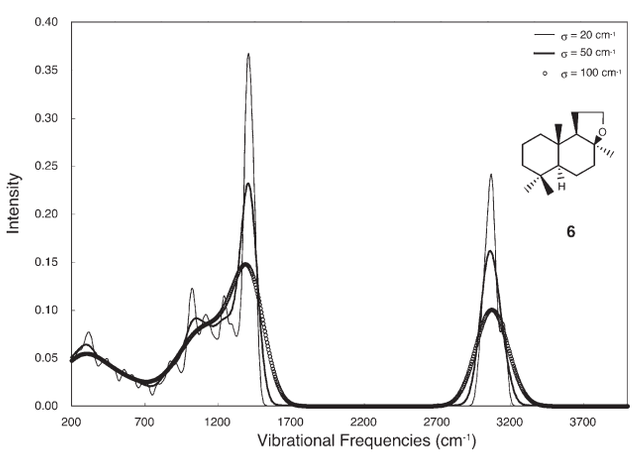

Fig. 2 Example of EVA spectra.

In Fig. 4 (a)–(c), the molecules appearing at the far right of the dendrograms are well separated from other clusters (**10**, **11**, **12**, **13**, **14**, **17**, and **18**). These are all bitter almond molecules, except for **17** and **18** which are structurally similar to others but their actual odour characteristics remain debatable, as described in the review by Rossiter.[8]

From the dendrograms obtained, we next determined the number of clusters. A feature of cluster analysis is that the determination of the number of clusters is somewhat subjective. We examined two kinds of methods for this. The first was to use Mojena's stopping rule. In this study, we adopted the significance probability p of $p < 0.05$. As shown in Table 2, the number of clusters was typically three to six for all methods in this partition. The second approach is to fix the number of clusters manually. Since the number of odour categories for this dataset is seven (Table 1), we fixed the numbers of clusters to be seven or ten for comparison (considering that the three inactive molecules might each constitute a further odour category).

Some methods for comparing partitions have been proposed in the literature. In this study, we adopted the adjusted Rand index (I_{HA}) to compare our clustering results with the reference experimental classification into seven odour categories. The index values I_{HA}, obtained by comparing the results from each of the various clustering methods with the reference classification into seven experimental odour categories, are listed in Table 2. The value of the adjusted Rand index would be 1 in the ideal case of comparing two identical partitions. For all the partitions, the adjusted Rand index values obtained from EVA descriptors were higher than those from UNITY 2D. That is, the results from EVA were consistently closer to the experimental odour classification than were the results from UNITY 2D. It can be seen that I_{HA} is not affected by σ when the modified Ward's method and Mojena's cut-off probability ($p < 0.05$) are used.

Table 3 shows the contingency table when the number of partitions is ten at $\sigma = 100$ cm^{-1}, clustering with the modified Ward's method. The adjusted Rand index I_{HA} is 0.480, which is the highest value in this study. The molecules in each cluster are listed in Table 4. C2, C3, C4, C6, C9, and C10 are successfully categorized. Cluster 2 contains only five bitter almond molecules (**10**, **11**, **14**, **17**, and **18**), although the odour characteristics of **17** and **18** are debatable,[8] as discussed above. Both cluster 3 (**12** and **13**) and cluster 4 (**15** and **16**) also contain bitter almond molecules. Clusters 9 and 10 contain musk molecules and the partition between them structurally discriminates aromatic nitromusks from the structurally unrelated macrocyclic ketone and lactone musks. Other clusters mix odour qualities, but cluster 1 contains all nine ambergris molecules and the remaining two musks, which are both benzenoid. Inspection of Fig. 1 suggests structural similarity between the ambergris molecules and the benzenoid musks.

In the present study, we have used only the EVA descriptor to classify a diverse set of 47 molecules, without a training set. We do not claim that the EVA descriptor alone can predict odour qualities, but it is a potential descriptor as a zeroth-order model

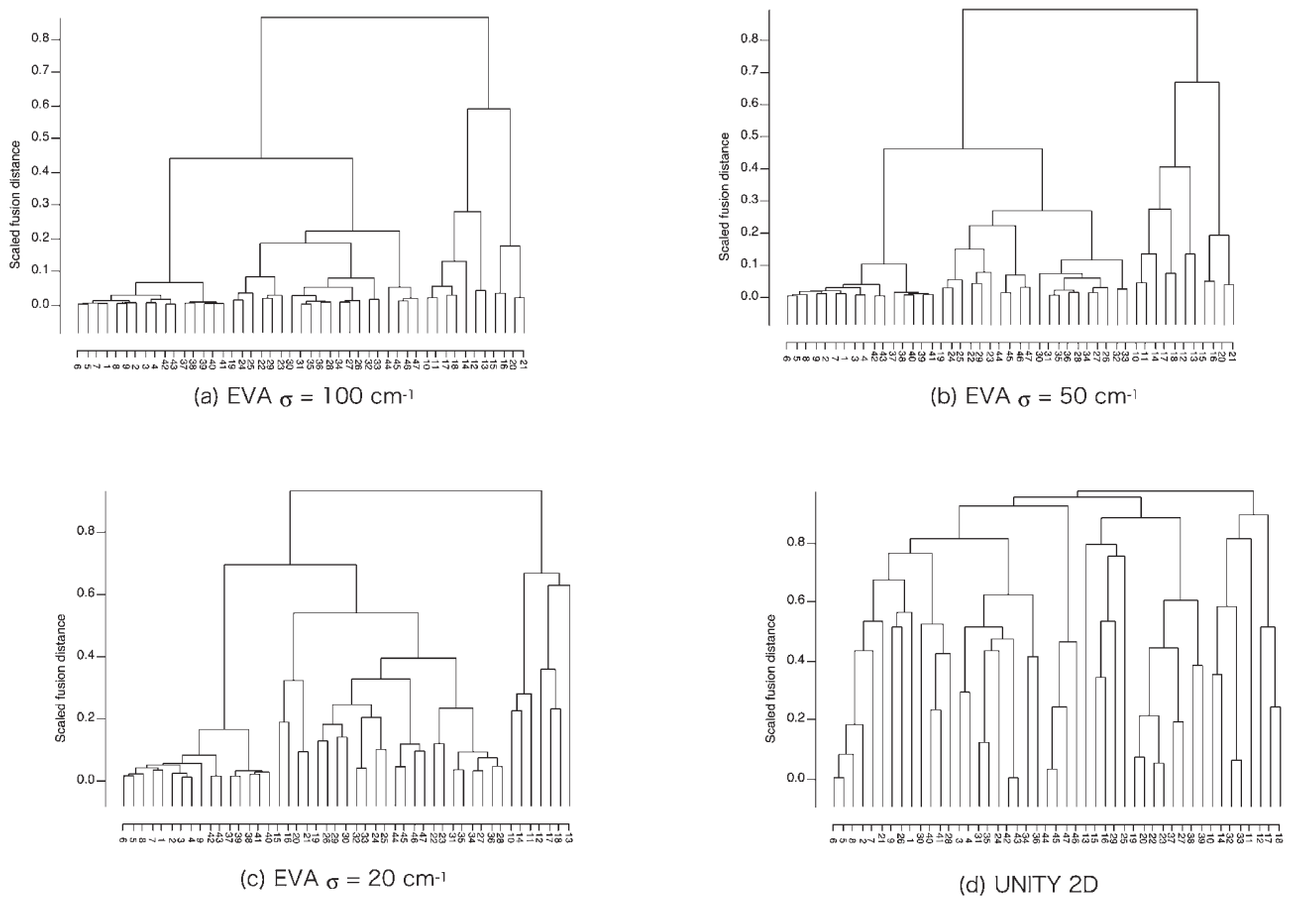

Fig. 3 Dendrograms using complete linkage method.

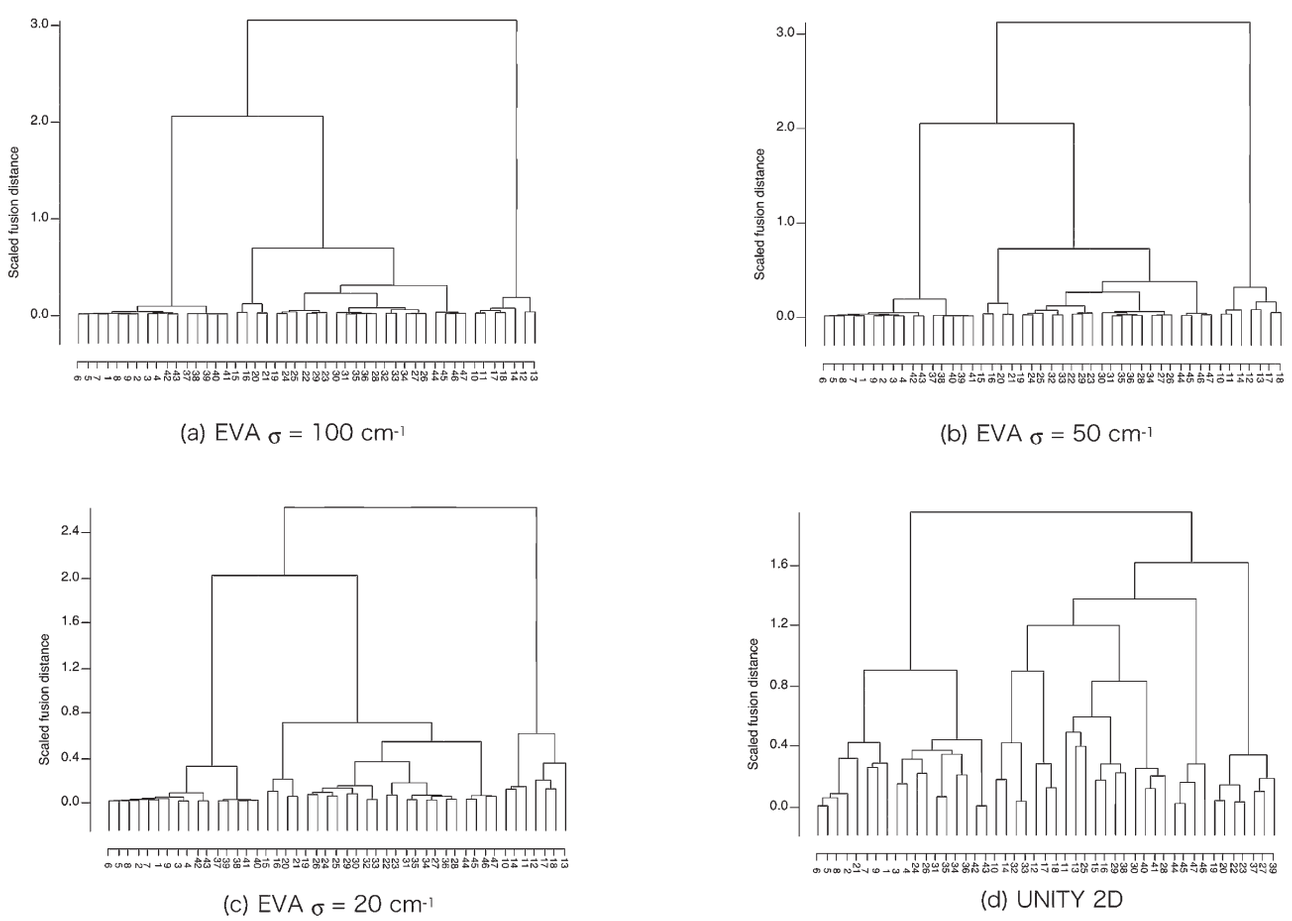

Fig. 4 Dendrograms using modified Ward's method.

Table 1 Odour quality of the molecules[a] in the dataset

Odour quality	Molecules [b]
Ambergris	**1**–**9**
Bitter almond	**10**–**16**, **17**[c], **18**[c]
Camphoraceous	**19**, **20**, **21**[d], **22**, **23**
Rose	**24**–**28**
Jasmine	**29**, **30**
Muguet	**31**–**36**
Musk	**37**–**47**

[a] From ref. 8. [b] See Fig. 1. [c] Similar structure to other molecules, but their odour characters are debatable. See ref. 8. [d] Similar structure to other molecules, but non-camphoraceous odour.

to predict molecular similarities, in particular for structure–odour relationship problems.

Turin has in recent years revived the decades-old[4] idea that the smell of a molecule is related to its vibrational spectrum, the "vibrational theory". His ideas have found little support, as they appear to contradict the established "binding theory" that olfaction results from the specific binding of an odour molecule to a small number of the 350 or so human olfactory receptors (ORs),[2] which are the largest family of G-protein-coupled receptors (GPCRs).[9] Any theory of olfaction must be consistent with the currently available experimental data. It is very desirable that such a theory should both provide explanations of the available data and accurately predict the results of future experiments. A good theory should also be able to suggest a credible biochemical mechanism by which olfaction operates. If two theories do approximately equally well in predicting and explaining data, then the one with the greater *a priori* plausibility is likely to prevail.

In the first[6] of three publications,[6,10,11] Turin put forward his theory that olfaction operates *via* a biological version of Inelastic Electron Tunnelling Spectroscopy (IETS), with different ORs tuned to different regions of the vibrational spectrum. Enantiomers and isotopomers present two possible tests of the theory. Enantiomers must have identical vibrational spectra, but in some cases have quite different odours; a classical example is *R*- and *S*-carvone smelling of mint and of caraway, respectively. In addressing this, Turin distanced himself from a purely vibrational model by invoking the notion that the binding of enantiomers to a chiral receptor leads to differences in the IETS spectra sampled by the receptor in each case. He justified this in terms of differing relative orientations of dipolar groups. Turin's theory is thus effectively one of vibration plus binding, rather than vibration alone. Turin made the point that there are other cases where enantiomers apparently smell the same to humans, which he saw as a challenge to the conventional binding theory. While this may indicate that the binding is not always highly specific, we note that Rubin and Katz[12] have found that rats (with roughly three times as many functional ORs) can distinguish enantiomers even in cases where humans cannot and that enantiomeric pairs produce different spatial patterns of activity on the olfactory bulb. This suggests that the mammalian olfactory system is, in principle, sensitive to the differences between enantiomers.

The other suggested "Turin test" involves the odours of isotopomers, molecules differing only in the masses of some of their nuclei. The replacement of ^{1}H with ^{2}H (deuterium) will reduce the corresponding X–H vibrational frequency by a factor of about $\sqrt{2}$. Under a vibrational theory of olfaction, such a change would be expected to have a significant impact on the odour impression. Turin claimed to have observed such an effect for acetophenone,[6] dimethyl sulfide[11] and decaborane.[11] The first independent experimental study of this effect, by Haffenden *et al.*,[13] was inconclusive, with about half (once guessing had been accounted for) of a trained panel able to distinguish deuterated benzaldehyde from its normal cousin. This indicates to us that any difference in odour is very small, and we are somewhat sceptical about its authors' interpretation that their results suggested a putative vibrationally dependent receptor in the range 2500 to 3000 cm^{-1}. In fact, small isotope dependencies have been observed in many molecular properties, both biological and physical,[14] as might be expected from consideration of zero point energy effects. Thus, the receptor binding energies of deuterated and normal molecules will be unequal and small differences in odour between isotopomers, even if they did exist, need not require the vibrational theory for their explanation. In any case, Keller and Vosshall[7] have recently published a study on volunteer human subjects showing no difference in odour between isotopomers.

Our view is that the relationship between vibrational frequencies and odour is not causal (as in Turin's theory), but may come about indirectly as a consequence of similar molecules having similar properties. Irrespective of whether Turin's IETS theory is true or not, it seems that his method of calculation[6,10] is a kind of extended version of EVA descriptor analysis, with a different weighting scheme used. Our work uses the EVA descriptor to predict odour characteristics of molecules within small regions of chemical space. At least within such localised regions, empirically there are some relationships between the vibrational spectrum of an odour molecule and its smell. In this study, we chose to investigate these structure–odour relationships using cluster analysis. We applied (non-weighted) EVA descriptors to cluster the molecules in an unbiased way on the basis of their structural similarity and used the adjusted Rand index to determine whether these clusters reflected the odour characteristics of the molecules.

The most expensive part of the EVA procedure is calculating the vibrational frequencies of the molecules in the dataset. However, the molecules considered in structure–odour relationship problems are typically rather small (molecular weight up to 300), and the vibrational frequencies need only be calculated once per molecule. Furthermore, once the vibrational frequencies of molecules have been obtained, the EVA spectrum can be easily redrawn and can be examined in subsequent analyses. Along these lines, we are now considering extended versions of EVA, including combining it with other descriptors, to investigate problems involving structure–odour relationships, and to construct an odour molecule database. This would contain not only molecular names, formulae, properties, and odour qualities, as in other existing databases, but also structural information and a data management system, allowing the calculation of EVA descriptors. We believe that such databases will prove useful in the future for structure-based studies in this field.

Methods

Dataset

Fig. 1 shows the 47 molecules in our dataset, comprising nine ambergris, nine bitter almond and similar structures (seven active and two inactive), five camphoraceous and similar structures (four active and one inactive), 13 floral and 11 musk odour molecules. The musk molecules contain five macrocyclic musks, four nitro musks, and two non-nitro aromatic benzenoids. The floral molecules are further subcategorised to two jasmine, six muguet (lily of the valley) and five rose odour molecules. All of these molecules were extracted from a review by Rossiter.[8]

EigenVAlue (EVA) descriptor

EVA is a vector descriptor based on eigenvalues corresponding to individual calculated normal modes, originally developed by Ferguson *et al.*[15] and extensively studied by Turner *et al.*[16–19] It has been successfully applied to some diverse datasets[20,21] in structure–activity relationships.

Since the EVA descriptor requires a geometry optimisation followed by a normal coordinate analysis, we used the GAMESS program package[22] for this purpose. The geometries were

Table 2 Adjusted Rand indices at the various partitions

Stopping rule	Descriptor	I_{HA} Complete Linkage	I_{HA} Modified Ward's
$p < 0.05$[a]	EVA, σ = 100 cm^{-1}	0.372 (4)[c]	0.323 (3)
	EVA, σ = 50 cm^{-1}	0.372 (4)	0.323 (3)
	EVA, σ = 20 cm^{-1}	0.267 (6)	0.323 (3)
	UNITY 2D	0.225 (3)	0.265 (5)
7 Clusters[b]	EVA, σ = 100 cm^{-1}	0.442	0.442
	EVA, σ = 50 cm^{-1}	0.370	0.388
	EVA, σ = 20 cm^{-1}	0.311	0.381
	UNITY 2D	0.173	0.247
10 Clusters[b]	EVA, σ = 100 cm^{-1}	0.437	0.480
	EVA, σ = 50 cm^{-1}	0.427	0.419
	EVA, σ = 20 cm^{-1}	0.365	0.417
	UNITY 2D	0.253	0.255

[a] Mojena's cut-off probability. [b] Fixed number of clusters. [c] Number of clusters in parentheses.

Table 3 Contingency table of the dendrogram using EVA descriptor [a]

Odour quality	Cluster C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Ambergris	9	0	0	0	0	0	0	0	0	0
Bitter almond	0	5	2	2	0	0	0	0	0	0
Camphoraceous	0	0	0	0	3	2	0	0	0	0
Rose	0	0	0	0	2	0	1	2	0	0
Jasmine	0	0	0	0	1	0	1	0	0	0
Muguet	0	0	0	0	0	0	3	3	0	0
Musk	2	0	0	0	0	0	0	0	5	4

[a] At n = 10, σ = 100 cm^{-1}. I_{HA} = 0.480.

Table 4 Molecules in the 10 clusters by modified Ward's method using EVA descriptor at σ = 100 cm^{-1}

Cluster	Molecules [a]
C1	**1–9, 42, 43**
C2	**10, 11, 14, 17, 18**
C3	**12, 13**
C4	**15, 16**
C5	**19, 22, 23, 24, 25, 29**
C6	**20, 21**
C7	**28, 30, 31, 35, 36**
C8	**26, 27, 32–34**
C9	**37–41**
C10	**44–47**

[a] See Fig. 1.

initially generated by using CORINA[23] to construct 3D structures. These were then subjected to geometry optimisations and normal coordinate analyses. The Hamiltonian used here was the semi-empirical AM1.[24,25] This procedure generated a single low energy conformation for each molecule. We took this as the representative conformation, although we are aware that the EVA descriptor has some dependence on conformation.

The resulting vibrational frequencies were then convolved using a sum of Gaussian functions to generate a pseudo-spectrum I(x):

$$\mathrm{I}(x) = \sum_{i=1}^{3N-6} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-f_i)^2}{2\sigma^2}} \qquad (1)$$

where N is the number of atoms, σ is a parameter describing the standard deviation (approximately equal to the half-width at half-height) of the Gaussian function, and f_i is the i-th vibrational frequency of the molecule. It should be noted that the spectrum generated in this way does not simulate actual IR or Raman spectra. Fig. 2 shows an example of an EVA spectra. As shown in Fig. 2, the spectrum becomes smoother and some peaks are merged together as the σ value increases (from 20 to 100 cm^{-1}). The intensity of the EVA spectrum does not correspond to the IR or Raman intensity, but it shows a kind of density of vibrational eigenvalues. Finally, the element of the vector descriptor was obtained by sampling at each step size δ (= 5 cm^{-1} in this study) in the range of 200–4000 cm^{-1}. In this case, the total number of elements for each molecule was 761. The similarity between each pair of molecules was calculated by Tanimoto association coefficients for continuous variables,[26] which are given by the formula

$$S_{\mathrm{A,B}} = [\sum_{j=1}^{j=n} x_{j\mathrm{A}} x_{j\mathrm{B}}] / [\sum_{j=1}^{j=n} (x_{j\mathrm{A}})^2 + \sum_{j=1}^{j=n} (x_{j\mathrm{B}})^2 - \sum_{j=1}^{j=n} x_{j\mathrm{A}} x_{j\mathrm{B}}] \qquad (2)$$

where $x_{j\mathrm{A}}$ and $x_{j\mathrm{B}}$ are elements of the EVA descriptors for molecules A and B, respectively. In the present calculations, the value of n is 761 as above. The calculations of EVA were performed by an in-house program written in Java.

UNITY 2D fingerprint

In this study, we have treated the EVA descriptor as something like a 3D fingerprint of each molecule, not as a descriptor usually used as the basis for a logistic regression study. Therefore, we tried to compare the results of the EVA descriptors with those of other molecular fingerprints. A 2D fingerprint of a compound is an array of binary variables, which have the value of 1 if the compound contains a particular fragment and 0 otherwise. Two molecules having similar structures will have many of the corresponding bits set, and so comparison of their fingerprints will generate a high similarity score. In this study, the fingerprint was generated for each of the molecules, using the default general-purpose UNITY screen definition file, as a 988-member bit-string. The similarity between each pair of molecules was calculated in the UNITY system, which is part of SYBYL® 6.8,[27] by means of the Tanimoto association coefficient.

Hierarchical clustering

For both the EVA descriptor and the UNITY 2D fingerprint, the similarity matrices obtained are then subjected to hierarchical clustering. It is an unsupervised classification method that does not require a training set. The initial N clusters are reduced in number one at a time until all N objects are in one cluster.

In this study, we have used two methods for clustering. One is the complete linkage method and the other is a modified Ward's method. Although the original Ward's clustering method[28] requires the use of the Euclidean distance, we used the Tanimoto coefficient in this study, as described above. To determine the final partition for the dendrogram, Mojena's stopping rule one[29] was used. These clustering calculations were carried out with the CEOPS program written by Smith (Department of Chemistry, Cambridge).

To compare the results of these two clustering methods, we used the adjusted Rand index.[30] This index can be applied even if the numbers of clusters differ between the two partitions. The adjusted Rand index (I_{HA}) is given by N/D, where

$$N = \sum_{i=1}^{c_1}\sum_{j=1}^{c_2}\binom{n_{ij}}{2} - \sum_{i=1}^{c_1}\binom{n_{i\bullet}}{2}\sum_{j=1}^{c_2}\binom{n_{\bullet j}}{2}\Big/\binom{n}{2}$$

$$D = \left[\sum_{i=1}^{c_1}\binom{n_{i\bullet}}{2} + \sum_{j=1}^{c_2}\binom{n_{\bullet j}}{2}\right]\Big/2 - \sum_{i=1}^{c_1}\binom{n_{i\bullet}}{2}\sum_{j=1}^{c_2}\binom{n_{\bullet j}}{2}\Big/\binom{n}{2} \qquad (3)$$

in which n_{ij} is the number of objects in group i of partition 1 ($i = 1, 2, \ldots, c_1$) and group j of partition 2 ($j = 1, 2, \ldots, c_2$),

$$n_{i\bullet} = \sum_{i=1}^{c_1} n_{ij} \qquad (4)$$

and

$$n_{\bullet j} = \sum_{j=1}^{c_2} n_{ij} \ . \qquad (5)$$

The table summarizing n_{ij} is called a contingency table (for example, see Table 3) and is useful to calculate I_{HA}. The index is within a range between 0 and 1. The adjusted Rand indices quoted in Table 2 describe comparisons of our clustering results with the experimental classification into seven odour categories (ambergris, bitter almond, camphoraceous, jasmine, rose, muguet, and musk).

Conclusions

We find that the dendrograms produced by the EVA method consistently outperform those from UNITY 2D in reproducing the reference experimental odour classifications of these 47 molecules. The highest adjusted Rand index I_{HA} is 0.480, obtained when the number of partitions is 10 at σ = 100 cm^{-1}, clustering with the modified Ward's method.

Acknowledgements

The authors thank Dr James Smith of the Department of Chemistry in Cambridge for the use of his CEOPS clustering software. They also thank Unilever for their support of the Centre for Molecular Science Informatics. This work was also supported by the overseas research programme of Osaka Sangyo University, Japan.

References

1 X. Zhang and S. Firestein, *Nat Neurosci.*, 2002, **5**, 124–133.
2 S. Zozulya, F. Echeverri and T. Nguyen, *GenomeBiology*, 2001, **2**, DOI: 10.1186/gb-2001-2-6-research0018.
3 B. Malnic, J. Hirono, T. Sato and L. B. Buck, *Cell*, 1999, **96**, 713–723.
4 G. M. Dyson, *Perfum. Essent. Oil Rec.*, 1937, **28**, 13–19.
5 R. H. Wright, *J. Appl. Chem.*, 1954, **4**, 611–615.
6 L. Turin, *Chem. Sens.*, 1996, **21**, 773–791.
7 A. Keller and L. B. Vosshall, *Nat. Neurosci.*, 2004, **7**, 337–338.
8 K. J. Rossiter, *Chem. Rev.*, 1996, **96**, 3201–3240.
9 L. Buck and R. Axel, *Cell*, 1991, **65**, 175–187.
10 L. Turin, *J. Theor. Biol.*, 2002, **216**, 367–385.
11 L. Turin and F. Yoshii, in *Handbook of Olfaction and Gustation*, ed. R. Doty, Marcel Dekker, New York, 2003, p. 275–294.
12 B. D. Rubin and L. C. Katz, *Nat. Neurosci.*, 2001, **4**, 355–356.
13 L. J. W. Haffenden, V. A. Yaylayan and J. Fortin, *Food Chem.*, 2001, **73**, 67–72.
14 D. Wade, *Chem.-Biol. Interact.*, 1999, **117**, 191–217.
15 A. M. Ferguson, T. Heritage, P. Jonathon, S. E. Pack, L. Phillips, J. Rogan and P. J. Snaith, *J. Comput.-Aided Mol. Des.*, 1997, **11**, 143–152.
16 D. B. Turner, P. Willett, A. M. Ferguson and T. Heritage, *J. Comput.-Aided Mol. Des.*, 1997, **11**, 409–422.
17 D. B. Turner, P. Willett, A. M. Ferguson and T. Heritage, *J. Comput.-Aided Mol. Des.*, 1999, **13**, 271–296.
18 D. B. Turner and P. Willett, *J. Comput.-Aided Mol. Des.*, 2000, **14**, 1–21.
19 D. B. Turner and P. Willett, *Eur. J. Med. Chem.*, 2000, **35**, 367–375.
20 M. T. Makhija and V. M. Kulkarni, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1569–1577.
21 D. J. Livingstone, R. Greenwood, R. Rees and M. D. Smith, *SAR QSAR Environ. Res.*, 2002, **13**, 21–33.
22 M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. J. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. Su, T. L. Windus, M. Dupuis and J. A. Montgomery, *J. Comput. Chem.*, 1993, **14**, 1347–1363.
23 http://www2.chemie.uni-erlangen.de/software/corina/.
24 M. J. S. Dewar, E. G. Zoebisch, E. F. Healy and J. J. P. Stewart, *J. Am. Chem. Soc.*, 1985, **107**, 3902–3909.
25 M. J. S. Dewar and Y. C. Yuan, *Inorg. Chem.*, 1990, **29**, 3881–3890.
26 P. Willett, J. M. Barnard and G. M. Downs, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 983–996.
27 SYBYL 6.8, Tripos Inc., 1699 South Hanley Road, St. Louis, Missouri, 63144, USA; http://www.tripos.com/.
28 J. H. Ward, *J. Am. Stat. Assoc.*, 1963, **58**, 236–244.
29 R. Mojena, *Comp J.*, 1977, **20**, 359–363.
30 L. Hubert and P. Arabie, *J. Classification*, 1985, **2**, 193–218.