Chem. Senses 37: 723–736, 2012

doi:10.1093/chemse/bjs058
Advance Access publication July 23, 2012

Odor–Structure Relationship Studies of Tetralin and Indan Musks

Barry K. Lavine[1], Collin White[1], Nikhil Mirjankar[1], C. Matthew Sundling[2] and Curt M. Breneman[2]

[1]Department of Chemistry, Oklahoma State University, Stillwater, OK 74078, USA and [2]Department of Chemistry, Rensselaer Polytechnic Institute, Cogswell Laboratory, 110 8th Street, Troy, NY 12180, USA

Correspondence to be sent to: Barry K. Lavine, Department of Chemistry, Oklahoma State University, Stillwater, OK 74078, USA. e-mail: bklab@chem.okstate.edu

Accepted April 15, 2012

Abstract

A list of 147 tetralin- and indan-like compounds was compiled from the literature for investigating the relationship between molecular structure and musk odor. Each compound in the data set was represented by 374 CODESSA and 970 TAE descriptors. A genetic algorithm (GA) for pattern recognition analysis was used to identify a subset of molecular descriptors that could differentiate musks from nonmusks in a plot of the two largest principal components (PCs) of the data. A PC map of the 110 compounds in the training set using 45 molecular descriptors identified by the pattern recognition GA revealed an asymmetric data structure. Tetralin and indan musks were found to occupy a small, but well-defined region of the PC (descriptor) space, with the nonmusks randomly distributed in the PC plot. A three-layer feed-forward neural network trained by back propagation was used to develop a discriminant that correctly classified all the compounds in the training set as musk or nonmusk. The neural network was successfully validated using an external prediction of 37 compounds.

Key words: asymmetric classification, genetic algorithms, musks, odor–structure relationships, olfaction, SAR, TAE descriptors

Introduction

Compounds with musk odor have been studied intensively for many years (Amoore 1970; Jennings-White 1985). Musks have aroused considerable interest because of their characteristic odor and fixative properties. Almost all fragrances sold commercially contain musk. Musks are also interesting from a structural point of view because they span a variety of structural manifolds thereby making musk odor prediction challenging. A wealth of information about musks is available in the scientific literature.

Originally, commercially available musks were derived from the glands of the Asian musk deer (Diete 1982). The first synthetic musk was developed by Baur (1891) who discovered that many nitrated derivatives of benzene imitated the odor of the naturally occurring musks. At present, indan and tetralin musks are an important group of artificial musks and include some of the most powerful musks known (Sell 2006). Phantolid, the first indan musk, was synthesized in 1951 (Fuchs 1956). Ever since, indan and tetralin musks have been extensively studied and many analogues have been synthesized. Indan and tetralin musks that are important commercially include Phantolid, versalide, Tonalid, and structural variants of them. The commercial success of indan and tetralin musks has been attributed to their chemical stability in comparison with nitrobenzene musks and to their similarity in odor to naturally occurring macrocyclic musks.

The study of relationships between structure and odor has facilitated the development of new synthetic musks. Odor–structure relationships (OSRs) of musks have been the subject of reviews by Rossiter (1996), Kraft et al. (2000), Kraft (2004), and Frater et al. (1998). Analysis of OSR using computational and pattern recognition methods can provide an approach to the analysis of indan and tetralin musks. The heart of this approach is finding a set of molecular descriptors from which a discriminating relationship can be found. Since the ultimate goal of any structure–activity relationship (SAR) study is the design of new compounds, the molecular descriptors must contain information related to the olfactory process. Using computer-assisted SAR techniques, Narvaez et al. (1986) studied a set of 148 nitro-free bicyclo- and tricyclo- benzenoids (67 musks

and 81 nonmusks). In that work, 14 molecular descriptors were identified that correctly classified every compound in the training set and 15 of 16 compounds in an external validation set. These 14 descriptors contained information about the number of rings, the number of quaternary centers, distances between the polar heteroatom and quaternary centers and the nearest methyl group, and the degree of branching of substituents attached to the nonaromatic ring. Klopman and Ptscelintsev (1992) analyzed a set of 152 non-nitroaromatic musks and their odorless structural analogues using CASE methodology. A hierarchical multi-case analysis identified 23 descriptors of which 9 were structural fragments responsible for musk odor and 7 were fragments encountered exclusively in the nonmusks. The 23 descriptors correctly classified 18 of the 20 set compounds. Charastrette et al. (1994) and Cherqaoui et al. (1998) employed a three-layer neural network to classify homogeneous set of tetralin and indan musks and their nonmusk analogues using linear free–energy relationship parameters as descriptors for well-characterized aromatic substituents. The SAR methodology in these studies employed either 2D fragment-based descriptors, which are not highly correlated to biological responses, or some variation of Hansch analysis, which necessarily limited the analysis to a homologous set of compounds.

In a previous study, Lavine and coworkers (2003) showed that Breneman's *T*ransferable *A*tom *E*quivalent (TAE)–based RECON methodology (Breneman et al. 1995; Song et al. 2002; Whitehead et al. 2003) could be used to map a molecule to a large set of spatially resolved property descriptors of a type that correlates with key modes of intermolecular interactions. Additionally, the use of spatially resolved shape/property hybrid electron density–based property-encoded surface translation (PEST) descriptors (Breneman et al. 2003) removed many of the limitations brought about by using descriptors based on 2D fragments, molecular surface properties, or other whole-molecule descriptors. The SAR of aromatic nitro musks, which was previously not well understood because of the complex substitution pattern and the varied polyfunctional character of the nitro group, was successfully modeled using six TAE-derived descriptors (Lavine et al. 2003). These six descriptors were found to contain information about molecular interactions important in olfaction.

Using Breneman's RECON and PEST methodologies, a set of 147 indan- and tetralin-like compounds was investigated for better understanding the relationship between molecular structure and musk odor for this class of polycyclic musks. Previous studies undertaken in our laboratory using fragment-based descriptors to develop an SAR for this set of compounds could not differentiate these tetralin and indan musks from their nonmusk structural analogues. A genetic algorithm (GA) for pattern recognition (Lavine et al. 2002, 2004a, 2004b, 2011; Karasinski 2005; Eiceman 2006) was used to identify molecular descriptors that could differentiate musks from nonmusks in a plot of the two largest principal components (PCs) of the data. Because PCs maximize variance, the descriptors selected by the pattern recognition GA contained information primarily about differences between the musks and the nonmusks in the data set. Due to the risk of model over-determination when performing feature selection, a variety of "best practices" methods for model cross-validation were used in this study, the most important of which is blind testing. Therefore, the discriminating relationships developed during this study were validated using a separate external validation set. The molecular structural descriptors that were employed in discrimination contained information about the size, shape, electronic nature, and intermolecular interaction capabilities of these compounds.

Materials and methods

Data set

An olfaction database of 147 compounds was compiled from literature reports (Wood 1970; Tenahsi 1971; Beets 1973, 1974, 1978; Theimer 1982; Fehr et al. 1989; Bersuker et al. 1991; Ohloff 1994; Pybus and Sell 1999) of chemical structure and

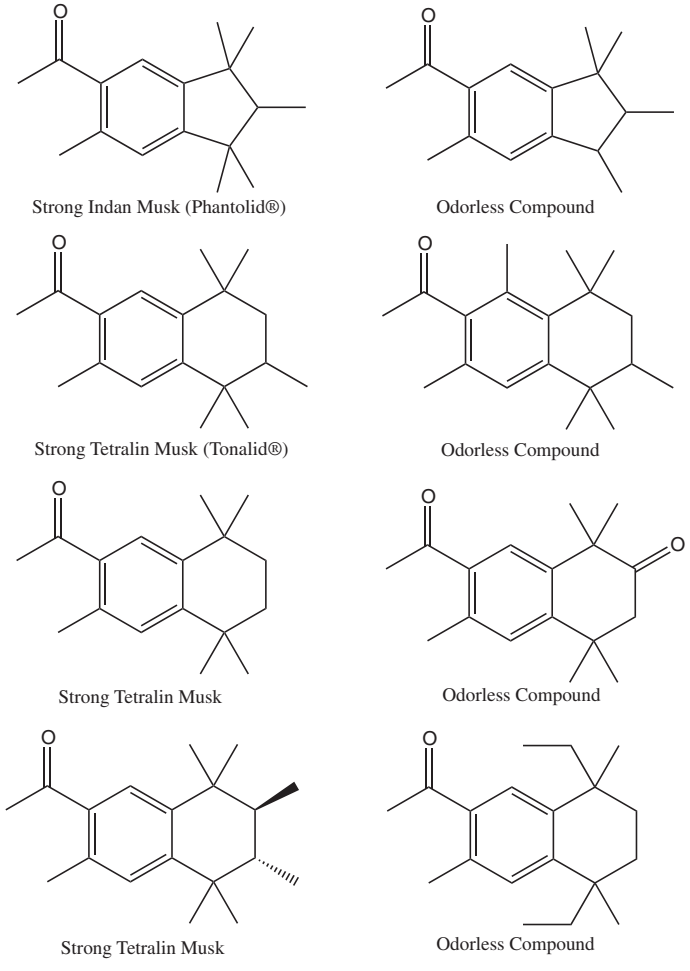

Figure 1 Examples of musks and nonmusks present in the data set.

Table 1 List of training set compounds

Compound name	Odor quality
6-(1,1-Dimethylethyl)-4-ethyl-2,3-dihydro-1,1-dimethyl-1H-indene	OLES
6-Acetyl-3,4-dihydro-1,1,4,4,5-pentamethyl-2(1H)-naphthalenone	OLES
7-Acetyl-3,4-dihydro-1,1,4,4,5-pentamethyl-2(1H)-naphthalenone	OLES
8-Acetyl-3,4-dihydro-1,1,4,4,5-pentamethyl-2(1H)-naphthalenone	OLES
5-Acetyl-3,4-dihydro-1,1,4,4,6-pentamethyl-2(1H)-naphthalenone	OLES
7-Acetyl-3,4-dihydro-1,1,4,4,6-pentamethyl-2(1H)-naphthalenone	OLES
8-Acetyl-3,4-dihydro-1,1,4,4,6-pentamethyl-2(1H)-naphthalenone	OLES
5-Acetyl-3,4-dihydro-1,1,4,4,7-pentamethyl-2(1H)-naphthalenone	OLES
6-Acetyl-3,4-dihydro-1,1,4,4,7-pentamethyl-2(1H)-naphthalenone	OLES
5-Acetyl-3,4-dihydro-1,1,4,4,8-pentamethyl-2(1H)-naphthalenone	OLES
6-Acetyl-3,4-dihydro-1,1,4,4,8-pentamethyl-2(1H)-naphthalenone	OLES
7-Acetyl-3,4-dihydro-1,1,4,4,8-pentamethyl-2(1H)-naphthalenone	OLES
1-(2,3-Dihydro-1,1-dimethyl-1H-inden-4-yl)-ethanone	OLES
1-(2,3-Dihydro-3,3-dimethyl-1H-inden-4-yl)-ethanone	OLES
1-[3-(1,1-Dimethylethyl)-5,6,7,8-tetrahydro-1-naphthalenyl]-ethanone	OLES
1-(2,3-Dihydro-1,1-dimethyl-1H-inden-5-yl)-ethanone	OLES
1-(2,3-Dihydro-3,3-dimethyl-1H-inden-5-yl)-ethanone	OLES
1-(2,3-Dihydro-1,2,3,3,6-pentamethyl-1H-inden-5-yl)-ethanone	OLES
5,6,7,8-Tetrahydro-3-methoxy-5,5,8,8-tetramethyl-2-naphthalenecarboxaldehyde	OLES
1-(5,6,7,8-Tetrahydro-3-methoxy-5,5,8,8-tetramethyl-5-naphthalenyl)-ethanone	OLES
5,6,7,8-Tetrahydro-4-methoxy-5,5,8,8-tetramethyl-2-naphthalenecarboxaldehyde	OLES
3-Acetyl-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenecarbonitrile	OLES
2-Methyl-1-(5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)-1-propanone	OLES
1-(5,8-Diethyl-5,6,7,8-tetrahydro-3,5,8-trimethyl-2-naphthalenyl)-ethanone	OLES
1-(3-Ethyl-5,6,7,8-tetrahydro-8,8-dimethyl-2-naphthalenyl)-ethanone	OLES
1-[2,3-Dihydro-1,3,3,6-tetramethyl-1-(2-methylpropyl)-1H-inden-5-yl]-ethanone	OLES
1-[5,6,7,8-Tetrahydro-5,5,8,8-tetramethyl-3-(1-methylethyl)-2-naphthalenyl]-propanone	OLES
1-(5,8-Dihydro-5,5,7,8,8-pentamethyl-2-naphthalenyl)-ethanone	OLES
1-(5,8-Dihydro-5,5,6,8,8-pentamethyl-2-naphthalenyl)-ethanone	OLES
1-(5,8-Dihydro-5,5,6,8,8-pentamethyl-1-naphthalenyl)-ethanone	OLES
1-[5,6,7,8-Tetrahydro-3,8,8-trimethyl-5-(1-methylethyl)-2-naphthalenyl)-ethanone	OLES
1-[5,6,7,8-Tetrahydro-3,5,5-trimethyl-8-(1-methylethyl)-2-naphthalenyl)-ethanone	OLES
4-(5,6,7,8-Tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)-3-buten-2-one	OLES
1,2,3,4-Tetrahydro-5-methoxy-1,1,4,7,7-pentaethylnaphthalene	OLES
5,6,7,8-Tetrahydro-1-methoxy-4,5,5,8,8-pentaethylnaphthalene	OLES
5,6,7,8-Tetrahydro-1,3-dimethoxy-5,5,8,8-tetraethylnaphthalene	OLES
3-Amino-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenecarboxylic acid methyl ester	OLES
5,6,7,8-Tetrahydro-5,5,8,8-tetramethyl-2-naphthalenol	OLES
5,6,7,8-Tetrahydro-2-methoxy-5,5,8,8-tetramethylnaphthalene	OLES

Table 1 Continued

Compound name	Odor quality
2-Ethoxy-5,6,7,8-tetrahydro-5,5,8,8-tetramethylnaphthalene	OLES
5,6,7,8-Tetrahydro-3-methylbutoxy-5,5,8,8-tetramethylnaphthalene	OLES
5,6,7,8-Tetrahydro-3,5,5,8,8-pentamethyl-2-naphthalenol	OLES
5,6,7,8-Tetrahydro-2-methoxy-3,5,5,8,8-pentamethylnaphthalene	OLES
5,6,7,8-Tetrahydro-2,3-dimethoxy-5,5,8,8-tetramethylnaphthalene	OLES
5,6,7,8-Tetrahydro-2-methoxymethyl-3,5,5,8,8-pentamethylnaphthalene	OLES
3-Ethyl-5,6,7,8-tetrahydro-2-methoxy-5,5,8,8-tetramethylnaphthalene	OLES
2-Methoxy-1-(5,6,7,8-tetrahydro-3-methoxy-5,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	OLES
3-Acetyl-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenecarboxylic acid methyl ester	NONM
1-[5,6,7,8-Tetrahydro-3-(methoxymethyl)-5,5,8,8-tetramethyl-2-naphthalenyl]-ethanone	NONM
1-(5,6,7,8-Tetrahydro-3-hydroxy-5,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	NONM
5,6,7,8-Tetrahydro-5,5,8,8-tetramethyl-2,3-naphthalenedicarboxylic acid	NONM
5,6,7,8-Tetrahydro-5,5,8,8-tetramethyl-2,3-naphthalenedicarboxylic acid diethyl ester	NONM
1-(3,5,8-Triethyl-5,6,7,8-tetrahydro-5,8-dimethyl-2-naphthalenyl)-ethanone	NONM
1-{5,8-Diethyl-5,6,7,8-tetrahydro-5,8-dimethyl-3-(1-methylethyl)-2-naphthalenyl]-ethanone	NONM
1-(5,6,7,8-Tetrahydro-3,8,8-trimethyl-2-naphthalenyl)-ethanone	NONM
1,3,3-Trimethyl-1-propyl-1H-indene	NONM
1-(2,3-Dihydro-1,1,2,3,3-pentamethyl-1H-inden-5-yl)-1-propanol	MMUS
1-[2,3-Dihydro-1,1,3,6-tetramethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	MMUS
1-[3-(1,1-Dimethylethyl)-2,3-dihydro-1,1,2,6-tetramethyl-1H-inden-5-yl]-ethanone	MSTR
1-[2,3-Dihydro-1,1,2,2,6-pentamethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	MUSK
1-[6-Ethyl-2,3-dihydro-1,1,2,2-tetramethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	MUSK
1-[6-Ethyl-3-dihydro-1,1,2,2-tetramethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	MUSK
1-[2,3-Dihydro-1,1,2,6-tetramethyl-3-(1-methylethyl)-1H-inden-5-yl]-1-propanone	MUSK
1-Ethyl-2,3-dihydro-1,3,3,6-tetramethyl-1H-indene-5-carboxyaldehyde	MSTR
2,3-Dihydro-1,1,3,3,6-pentamethyl-1H-indene-5-carboxyaldehyde	MSTR
1-(2,3-Dihydro-1,1,2,3,3,6-hexamethyl-1H-inden-5-yl)-ethanone ("Phantolid")	MSTR
1-(2,3-Dihydro-1,1,2,3,6-pentamethyl-1H-inden-5-yl)-ethanone	MWEA
1-(2-Ethyl-2,3-dihydro-1,1,3,3,6-pentamethyl-1H-inden-5-yl)-ethanone	MMUS
1-(6-Ethyl-2,3-dihydro-1,1,3,3-tetramethyl-1H-inden-5-yl)-ethanone	MMUS
1-(6-Ethyl-2,3-dihydro-1,1,2,3,3-tetramethyl-1H-inden-5-yl)-ethanone	MMUS
1-(2,3-Dihydro-1,1,3,3,5,6-hexamethyl-1H-inden-4-yl)-ethanone	MMUS
1-(2,3-Dihydro-1,1,2,3,3,-pentamethyl-1H-inden-5-yl)-ethanone	MMUS
1-[2,3-Dihydro-1,1,2,3,3-pentamethyl-6-(1-methylethyl)-1H-inden-5-yl]-ethanone	MUSK
1-[2,3-Dihydro-1,1,3,3-pentamethyl-6-(1-methylethyl)-1H-inden-5-yl]-ethanone	MUSK
1-[2,3-Dihydro-1,1,2,6-tetramethyl-3-(trimethylsilyl)-1H-inden-5-yl]-ethanone	MSTR
2,3-Dihydro-1,1,2,6-tetramethyl-3-(1-methylethyl)-1H-indene-5-carboxaldehyde	MSTR
1-[2-Ethyl-2,3-dihydro-1,1,6-trimethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	MSTR
1-[2,3-Dihydro-1,1,2,6-tetramethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	MSTR
2,3-Dihydro-1,1,2,3,3,6-hexamethyl-1H-indene-5-carbonitrile	MSTR

Table 1 Continued

Compound name	Odor quality
3-Ethyl-5,6,7,8-tetrahydro-1-methoxy-5,5,8,8-tetramethyl-2-naphthalenecarboxaldehyde	MMED
5,6,7,8-Tetrahydro-1-methoxy-3,3,5,8,8-pentamethyl-2-naphthalenecarboxaldehyde	MSTR
5,6,7,8-Tetrahydro-1-hydroxy-3,3,5,8,8-pentamethyl-2-naphthalenecarboxaldehyde	MSTR
5,6,7,8-Tetrahydro-5,5,8,8-tetramethyl-2,3-naphthalenedicarboxaldehyde	MMED
5′,6′,7′,8′-Tetrahydro-5′,5′,8′,8′-tetramethyl-3′-(methylthio)-2′-acetonaphthone	MUSK
5′,6′,7′,8′-Tetrahydro-5′,5′,6′,8′,8′-pentamethyl-3′-(methylthio)-2′-acetonaphthone	MUSK
5′,6′,7′,8′-Tetrahydro-5′,5′,6′,7′,8′,8′-pentamethyl-3′-(methylthio)-2′-acetonaphthone	MUSK
3′-(Ethylthio)-5′,6′,7′,8′-tetrahydro-5′,5′,8′,8′-tetramethyl-2′-acetonaphthone	MUSK
1-(3-Ethyl-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)-propanone	MWEA
1-(3-Ethyl-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)-ethanone ("Versalide")	MSTR
1-(5,6,7,8-Tetrahydro-3,5,5,6,8,8-hexamethyl-2-naphthalenyl)-ethanone ("Tonalid")	MSTR
1-(5,6,7,8-Tetrahydro-2-naphthalenyl)-ethanone	MSTR
1-(5,6,7,8-Tetrahydro-3,5,5,6,8,8-hexamethyl-2-naphthalenyl)-carbonitrile	MMED
5,6,7,8-Tetrahydro-3,5,5,8,8-pentamethyl-2-naphthalenecarboxaldehyde	MSTR
3-Ethyl-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenecarboxaldehyde	MSTR
1-(5,8-Dihydro-5,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	MMUS
1-(5,6,7,8-Tetrahydro-5,5,8,8-tetramethyl-3-(1-methylethyl)-2-naphthalenyl)-ethanone	MWEA
1-(5,6,7,8-Tetrahydro-3,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	MMED
1-(5-Ethyl-5,6,7,8-tetrahydro-3,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	MWEA
1-(3-Ethyl-5,6,7,8-tetrahydro-5,5,6,8,8-pentamethyl-2-naphthalenyl)-ethanone	MSTR
1-(5,6,7,8-Tetrahydro-5,5,6,8,8-pentamethyl-2-naphthalenyl)-ethanone	MMUS
5,6,7,8-Tetrahydro-3,5,5,8,8-pentamethyl-2-naphthalenecarbonitrile	MMUS
3-Ethyl-5,6,7,8-tetrahydro-5,5,8,8-pentamethyl-2-naphthalenecarbonitrile	MMUS
1-(1-Ethyl-2,3-dihydro-1,3,3,5,6-pentamethyl-1H-inden-4-yl)-ethanone	MWEA
1-(3-Ethyl-2,3-dihydro-1,1,3,5,6-pentamethyl-1H-inden-4-yl)-ethanone	MWEA
1-(1-Ethyl-2,3-dihydro-1,3,3-trimethyl-1H-inden-4-yl)-ethanone	MWEA
1-(3-Ethyl-2,3-dihydro-1,1,3,6-tetramethyl-1H-inden-5-yl)-ethanone	MWEA
1-(2,3-Dihydro-1,1,2,3,3,5,6-heptamethyl-1H-inden-4-yl)-ethanone	MWEA
1-(2,3-Dihydro-1,1,3,3-tetramethyl-1H-inden-5-yl)-ethanone	MWEA
1-[3-(1,1-Dimethylethyl)-5,6,7,8-tetrahydro-5,5-dimethyl-1-naphthalenyl]-ethanone	MWEA
1-(5,8-Dihydro-3,5,5,8,8-pentamethyl-2-naphthalenyl)-ethanone	MWEA

OLES, odorless; NONM, nonmusk; MSTR, strong musk; MMUS, musk of medium or strong odor intensity; MMED, medium musk; MUSK, musk of unspecified odor intensity; MWEA, weak musk.

olfactory quality for investigating the relationship between chemical structure and musk odor. Nonmusks were selected to be similar in structure to the musks. This not only contributed to the additional challenge of separating very similar structures according to odor quality, but this also increased our understanding of how small structural changes affect odor quality. Examples of musks and nonmusks present in the data set are shown in Figure 1. Of the 147 compounds, 70 are musks and 77 are nonmusks. The musks are strong, medium, or weak odor intensity, whereas the nonmusks are odorless or have an odor other than musk. The training set consisted of 110 of these cases, and 37 compounds comprised the validation set. Compounds included in the validation set were chosen by random lot. A list of the 110 training set compounds is given in Table 1, and a list of the 37 compounds comprising the validation set is given in Table 2.

Table 2 List of validation set compounds

Compound name	Odor quality
8-Acetyl-3,4-dihydro-1,1,4,4,7-pentamethyl-2(1H)-naphthalenone	OLES
1-(2,3-Dihydro-1,3,3,6-tetramethyl-1-propyl-1H-inden-5-yl)-ethanone	OLES
1-(5,8-Dihydro-5,5,7,8,8-pentamethyl-1-naphthalenyl)-ethanone	OLES
1-(5,6,7,8-Tetrahydro-1-methoxy-3,5,5,8,8-pentamethyl-2-naphthalenyl)-ethanone	OLES
5,6,7,8-Tetrahydro-3,5,5,8,8-pentamethyl-2-(3-methylbutoxymethyl)-naphthalene	OLES
2,3-Dihydro-1,1,2,3,3,6-hexamethyl-1H-indene-5-carboxaldehyde	MSTR
1-(2,3-Dihydro-1,1,2,3,3,6-hexamethyl-1H-inden-5-yl)-1-propanone	MSTR
5,6,7,8-Tetrahydro-5,5,8,8-tetramethyl-3-(1-methylethyl)-2-naphthalenecarboxaldehyde	MMUS
1-(5-Ethyl-5,6,7,8-tetrahydro-3,5,8-trimethyl-2-naphthalenyl)-ethanone	MWEA
5,6,7,8-Tetrahydro-5,5,8,8-tetramethyl-2,3-naphthalenedicarboxylic acid dimethyl ester	NONM
1-(3-Ethyl-2,3-dihydro-1,1,3-trimethyl-1H-inden-4-yl)-ethanone	MWEA
5,6,7,8-Tetrahydro-3,5,5,6,8,8-hexamethyl-2-naphthalenylcarboxaldehyde	MSTR
(5,6,7,8-Tetrahydro-1,3,5,5,8,8-hexamethyl-2-naphthalenylcarboxaldehyde	MSTR
1-(5,6,7,8-Tetrahydro-1,3,5,5,8,8-hexamethyl-2-naphthalenyl)-ethanone	OLES
1-(5,6,7,8-Tetrahydro-4,5,5,8,8-pentamethyl-2-naphthalenyl)-ethanone	MWEA
1-(5,6,7,8-Tetrahydro-4,5,5,8,8-pentamethyl)-2-naphthalenecarboxaldehyde	MWEA
1-(5,6,7,8-Tetrahydro-1,4,5,5,8,8-hexamethyl-2-naphthalenyl)-carboxaldehyde	MWEA
1-(5,6,7,8-Tetrahydro-1,3,4,5,5,8,8-heptamethyl)-2-naphthalenecarboxaldehyde	MWEA
1-(5,6,7,8-Tetrahydro-1,3,5,5,6,8,8-heptamethyl)-2-naphthalenecarboxaldehyde	MSTR
1-(5,6,7,8-Tetrahydro-1,3,5,5,7,8,8-heptamethyl)-2-naphthalenecarboxaldehyde	MSTR
1-(5,6,7,8-Tetrahydro-1,3,5,5,6,8,8-heptamethyl-2-naphthalenyl)-ethanone	OLES
Trans-1-(5,6,7,8-tetrahydro-3,5,5,6,7,8,8-heptamethyl-2-naphthalenyl)-ethanone	MSTR
Trans-1-(5,6,7,8-tetrahydro-3,5,5,6,7,8,8-heptamethyl-2-naphthalenyl)-carboxaldehyde	MSTR
Cis-1-(5,6,7,8-tetrahydro-3,5,5,6,7,8,8-heptamethyl-2-naphthalenyl)-carboxaldehyde	MSTR
1-(3-Chloro-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	NONM
1-(3-Bromo-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	NONM
1-(3-Iodo-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	NONM
1-[2,3-Dihydro-1,1,6-trimethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	NONM
1-[6-Ethyl-2,3-dihydro-1,1-dimethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	NONM
1-[2,6-Diethyl-2,3-dihydro-1,1-dimethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	NONM
1-(1,2,3,4,5,6,7,8-Octahydro-2,3,8,8-tetramethyl-2-naphthalenyl)-ethanone	AMBER
1-(1,2,3,4,5,6,7,8-Octahydro-3,8,8-trimethyl-2-naphthalenyl)-ethanone	WOODY
1-(1,2,3,4,6,7,8,8a-Octahydro-4,8,8-trimethyl-2-naphthalenyl)-ethanone	WOODY
1-(1,2,3,4,5,6,7,8-Octahydro-4,8,8-trimethyl-2-naphthalenyl)-ethanone	WOODY
1-(1,2,3,5,6,7,8,8a-Octahydro-4,8,8-trimethyl-2-naphthalenyl)-ethanone	WOODY
2,3-Dihydro-β,1,1,2,3,3-hexamethyl-1H-indene-5-ethanol	WOODY
1-[5,6,7,8-Tetrahydro-5,5-dimethyl-3-(1-methylethyl)-2-naphthalenyl)-ethanone	WOODY

OLES, odorless; NONM, nonmusk; MSTR, strong musk; MMUS, musk of medium or strong odor intensity; MMED, medium musk; MUSK, musk of unspecified odor intensity; MWEA, weak musk; AMBER, Amber odor; WOODY, woody odor.

Molecular descriptors

For each compound, a. mol file was generated using ChemDraw (Cambridge Soft), which then served as input for a 3D molecular mechanics model building routine that used the CHARMm force field in the modeling program Quanta (Molecular Simulations). In this study, traditional molecular property descriptors, for example, specific atom counts in a molecule, partial positive and partial negative surface area, and principal moments of inertia, were computed using CODESSA (CompuDrug International, Sedona, AZ) for each compound. In addition, TAE molecular surface property reconstructions were generated for each compound using RECON methods. The Property Encoded Surface Translator (PEST) algorithm was used to generate wavelet and hybrid shape/property descriptors. Molecular electron density properties for all compounds in the study were represented using TAE surface histogram descriptors, wavelet coefficient descriptors (WCD), and PEST hybrid shape/property descriptors. Advantages of using TAE, WCD, and PEST descriptors to study biological and nonbiological molecular behavior have been previously reported (Breneman 2003). Descriptors based on experimental data were not used in this study because it would be extremely difficult to obtain this type of data from the literature for a large set of compounds, and it would not be possible to use the SAR developed in this study as a screening tool to identify new musks since the compounds would have to be synthesized in order to obtain the necessary data.

Genetic algorithm for feature selection and pattern recognition

After the removal of invariant descriptors, each compound in the data set was represented by 1344 molecular descriptors: 374 CODESSA and 970 TAE and PEST descriptors. The descriptors were autoscaled to zero mean and unit standard deviation. A GA with a pattern recognition loss function was used on the training set to identify subsets of molecular descriptors that optimized the separation of the musks from the nonmusks in a plot of the two or three largest PCs of the data. The principal component analysis (PCA) routine embedded in the fitness function of the GA served as an information filter, reducing the size of the search space, since it restricts the search to descriptors whose PC plots show clustering on the basis of class. In addition, the algorithm focuses on those classes and/or compounds difficult to classify as it trains by boosting the class and sample (compound) weights. Compounds that consistently classify correctly are not as heavily weighted as compounds that are difficult to classify. Over time, the algorithm learns its optimal parameters in a manner similar to a neural network.

The fitness function of the pattern recognition GA uses machine learning to score the PC plots and thereby identify descriptors that optimize the separation of the odorants in a plot of the two or three largest PCs of the data. To facilitate the tracking and scoring of the PC plots, class and sample weights, which are an integral part of the fitness function, are computed (see Equations (1) and (2)), where CW(c) is the weight of class c (with c varying from 1 to the total number of classes in the data set). $SW_c(s)$ is the weight of sample (compound) s in class c. The class weights sum to 100, and the sample weights for compounds comprising a particular class (musks or odorless) sum to a value equal to the weight of that class.

$$\mathrm{CW(c)} = 100\frac{\mathrm{CW(c)}}{\sum_{\mathrm{c}}\mathrm{CW(c)}}. \tag{1}$$

$$F(d) = \sum_{\mathrm{c}}\sum_{\mathrm{s}\in\mathrm{c}}\frac{1}{K_{\mathrm{c}}}\times\mathrm{SHC(s)}\times\mathrm{SW(s)}. \tag{2}$$

Each PC plot generated for each descriptor subset is scored using the K-nearest neighbor (K-NN) classification algorithm. For a given compound that is treated as a data point, Euclidean distances are computed between the compound and every other data point (compound) in the PC plot. These distances are arranged from smallest to largest. A poll is then taken of the compound's K_c-nearest neighbors. For the most rigorous classification, K_c is equal to the number of compounds comprising the class to which the point belongs. The number of K_c-nearest neighbors with the same class label as the data point in question, the so-called "sample hit count" (SHC(s)), is computed ($0 < \mathrm{SHC(s)} < K_c$) for each compound. Each PC plot is scored using Equation (3), where $F(d)$ is the fitness function score. The contribution of each compound in class 1 to the overall fitness is computed, with the scores of the compounds comprising this class summed to yield the contribution of class 1 to the overall fitness. This same calculation is repeated for each class with their scores summed to yield the overall fitness, $F(d)$.

$$F(d) = \sum_{\mathrm{c}}\sum_{\mathrm{s}\in\mathrm{c}}\frac{1}{K_{\mathrm{c}}}\times\mathrm{SHC(s)}\times\mathrm{SW(s)} \tag{3}$$

The fitness function of the pattern recognition GA is able to focus on compounds and odor classes that are difficult to classify by adjusting (i.e., boosting) sample and/or class weights over successive generations. To boost class and sample weights, it is necessary to compute the sample-hit rate (SHR), which is the average value of SHC/K_c over all descriptor subsets investigated in a generation (see Equation (4)), and the class-hit rate (CHR), which is the average value of SHR over all compounds in a class (see Equation (5)). The ϕ in Equation (4) is the number of chromosomes in the population, and AVG in Equation (5) refers to the average value. During each generation, class and sample weights are adjusted by a perceptron (see Equations (6) and (7)) where P is the momentum set by the user, g + 1 is the current generation, and g is the previous

generation. Classes with a lower CHR will be boosted more heavily than those classes that score well.

$$\mathrm{SHR(s)} = \frac{1}{\phi}\sum_{i=1}^{\phi}\frac{\mathrm{SHC}_i\mathrm{(s)}}{K_\mathrm{c}} \quad (4)$$

$$\mathrm{CHR_g(c)} = \mathrm{AVG(SHR_g(s) : \forall_{s\,\in\,c})} \quad (5)$$

$$\mathrm{CW_{g+1}(s)} = \mathrm{CW_g(s)} + P(1-\mathrm{CHR_g(s)}) \quad (6)$$

$$\mathrm{SW_{g+1}(s)} = \mathrm{SW_g(s)} + P(1-\mathrm{SHR_g(s)}) \quad (7)$$

Boosting is crucial to the successful operation of the pattern recognition GA as it modifies the fitness landscape by adjusting the values of the class and sample weights. The fitness function of the pattern recognition GA is changing as the population evolves toward a solution. This helps to minimize the problem of convergence to a local optimum.

Pattern recognition analysis

For pattern recognition analysis, each compound was represented as a data vector X = (x_1, x_2, x_3, x_4, x_5 ... x_j ... x_{1344}) where x_j is a CODESSA or TAE molecular descriptor. The tetralin and indan musk data set was also analyzed using the Advanced Data Analysis and Pattern Recognition Toolkit (ADAPT), which was written in MATLAB 7.6.0.324 (R2008a) using the graphical user interface development environment (GUIDE). The toolkit consists of a collection of MATLAB M-files and MATLAB Figure files that control the graphical user interface's (GUI) computational and graphical components. The M-file provides both a code to initialize the GUI and a framework for the GUI routines that execute in response to user-generated events. The four main types of pattern recognition methods are mapping and display, discriminant development, clustering, and modeling. ADAPT has routines in all four areas, for example, PCA, canonical variate analysis, hierarchical and fuzzy c-varieties clustering, linear, regularized, and quadratic discriminant analysis, K-NN, and back propagation neural networks with one or two hidden layers and many were used in this study.

Results

Figure 2 shows a plot of the two largest PCs of the 110 compounds and 1344 descriptors that comprise the training set. The PC plot explains 26.8% of the total cumulative variance. Each compound in the training set is represented as a point in the plot (1, nonmusks and 2, musks). The overlap of musks and nonmusks in the PC plot of the 1344 descriptors is not surprising in view of the similarity of their chemical structures.

The pattern recognition GA was used to identify molecular descriptors from which a discriminating relationship could be developed for the tetralin and indan musks. Sixteen runs with each run requiring 200 generations were performed using different initial populations and mutation rates. A frequency histogram of the best descriptors selected by the pattern recognition GA during each generation is shown in Figure 3. The histogram details the fact that approximately 33% of the descriptors were never selected by the pattern recognition GA for inclusion in the chromosomes with the highest fitness, and approximately 75% of the descriptors were selected less than 1% of the time for inclusion in the best chromosomes.

One-hundred descriptors that were most frequently selected by the pattern recognition GA for inclusion in the

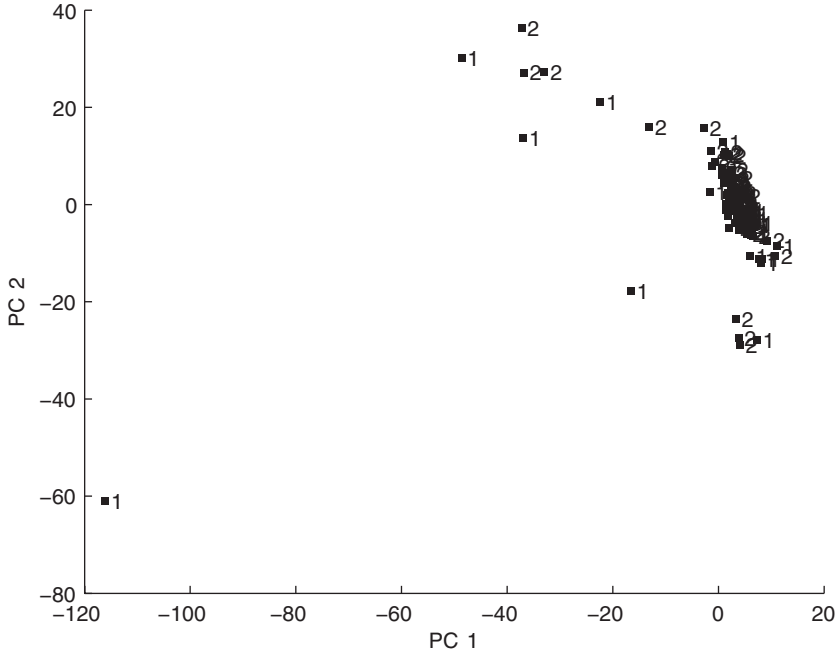

Figure 2 Plot of the two largest PCs of the 110 compounds and 1344 descriptors that comprise the training set. 1, nonmusks and 2, musks.

best chromosomes were retained for further study. The pattern recognition GA was used to identify the most informative of these 100 descriptors by sampling key descriptor subsets, scoring their PC plots, and tracking those classes and/or compounds that were difficult to classify. The boosting routine used this information to steer the population to an optimal solution. After 200 generations, the pattern recognition GA identified 48 molecular descriptors whose PC plot showed clustering of the compounds on the basis of odor. Figure 4 shows a PC plot of the 110 training set compounds and 48 descriptors identified by the pattern recognition GA.

The data structure of the tetralin and indan musk classification problem depicted in Figure 4 is asymmetric (Dunn and Wold 1980, 1995; Rose et al. 1992). Tetralin and indan musks occupy a small but well-defined region of the PC (descriptor) space, whereas the nonmusks are randomly distributed in this space. The musks form a class of structurally similar compounds whose odor is a well-behaved function of the molecular descriptors used to characterize the SAR, whereas the change in the structure of the nonmusks cannot be modeled because the compounds are inactive for a variety of reasons (see Figure 1). This is analogous to problems in process

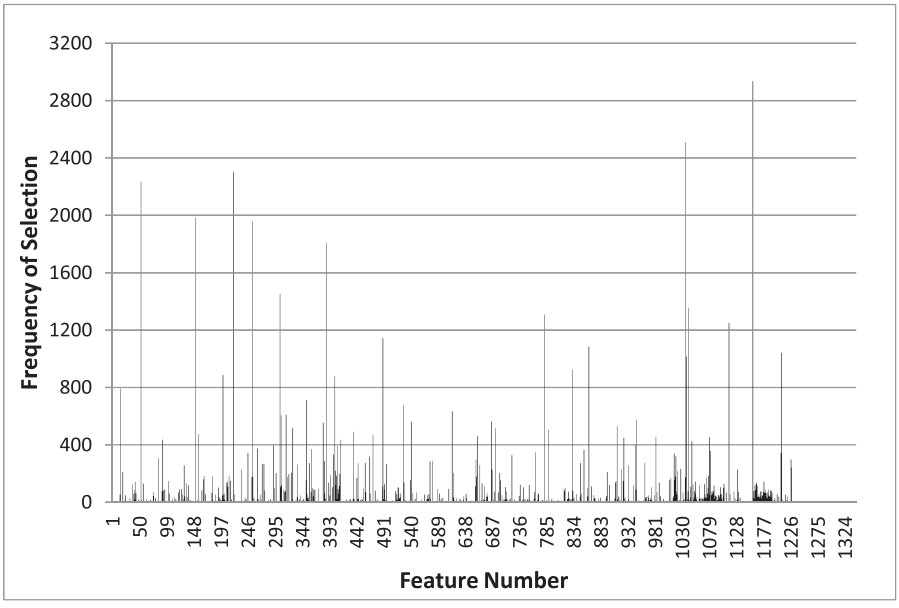

Figure 3 Frequency histogram of the best descriptors selected by the pattern recognition GA during each generation.

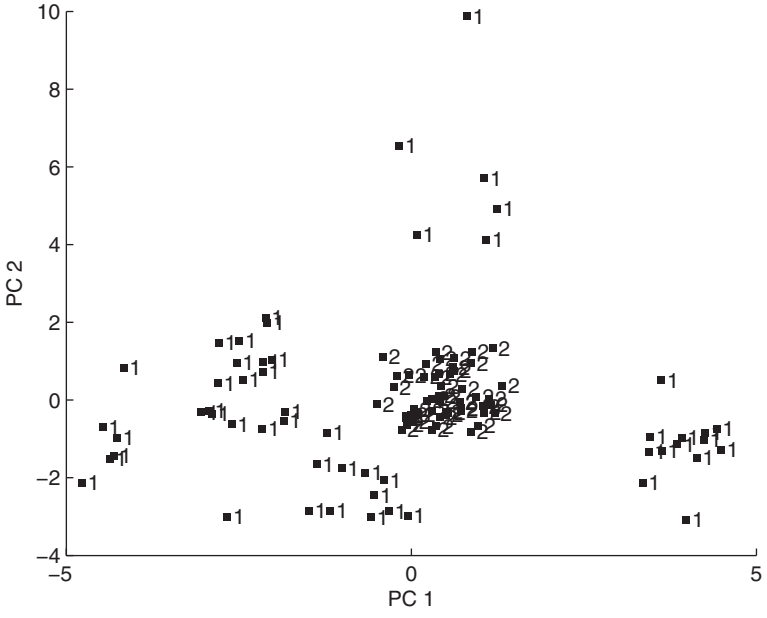

Figure 4 PC score plot of the 110 training set compounds and 48 descriptors identified by the pattern recognition GA. 1, nonmusks and 2, musks.

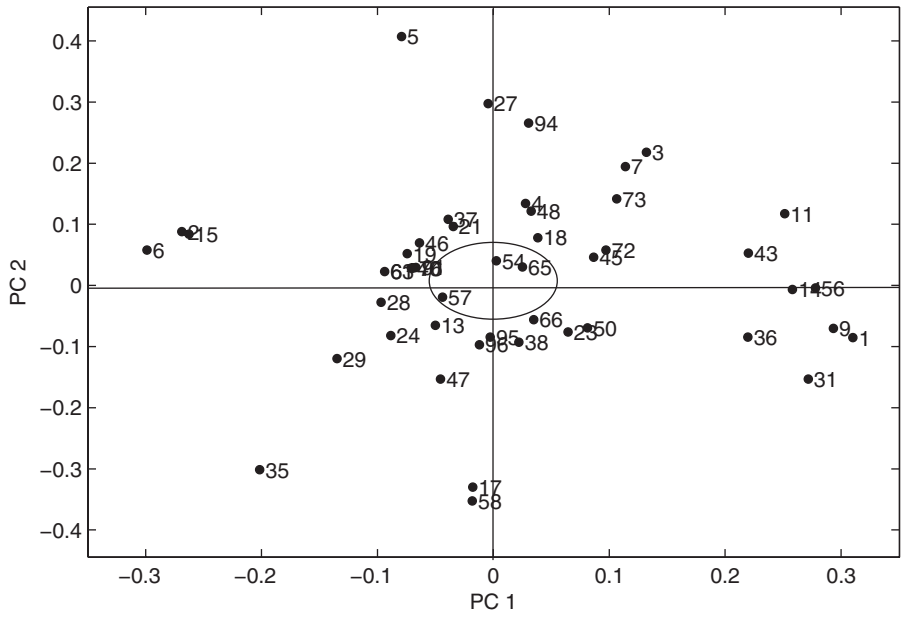

Figure 5 Loading plot for the PC score plot developed from the 110 training set samples and 48 descriptors. Three descriptors (circled) have values near zero for both PCs.

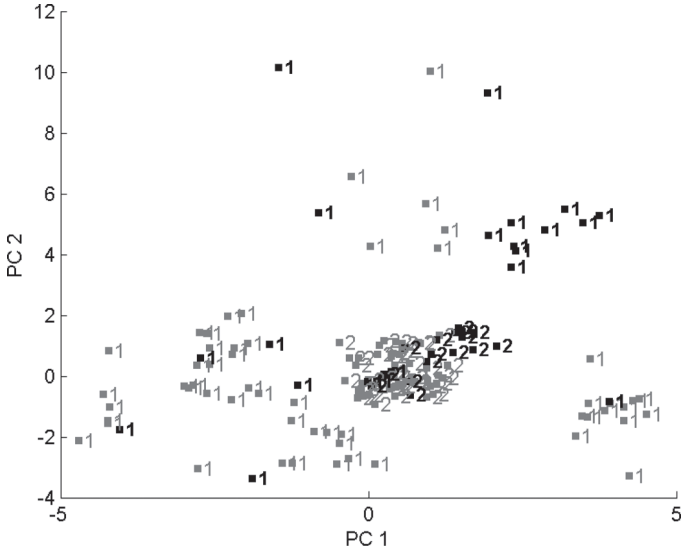

Figure 6 Projection of validation set compounds (black) onto the PC map developed from the training set data (gray). About 34 of the 37 projected compounds lie in a region of the map with compounds that have the same class label. 1, nonmusks and 2, musks

control monitoring where the sample points representing the process under control cluster in a well-defined region of the measurement space whereas the points for the process when it is out of control are equally likely to move away from the cluster in any one of the directions in the high-dimensional measurement space.

Figure 5 shows the corresponding loading plot for the PC score plot in Figure 4. From Figure 5, it is evident that loadings of these three descriptors have values near zero for both PCs. These three descriptors were removed as their contribution to clustering in the PC plot was negligible.

A validation set of 37 compounds (see Table 2) was used to assess the predictive ability of the 45 molecular descriptors identified by the pattern recognition GA. We chose to map the 37 musks and nonmusks directly onto the PC map defined by the 110 compounds and 45 molecular descriptors. Figure 6 shows the projection of the validation set compounds onto the PC map developed from the training set data. About 34 of the 37 projected compounds lie in a region of the map with compounds that have the same class label. The three misclassified compounds (see Figure 7) were nonmusks. These three compounds differ structurally in the lipophilic region, that is, cyclohexane ring, from the tetralin and indan musks. Ohloff et al. (1991) envisioned that interactions between this portion of the molecule and the receptor are more selective than interactions involving the polar functional group. However, the TAE-, PEST-, and CODESSA-derived descriptors used in this study were neither not able to characterize local shape in this region of the molecule nor were they able to adequately delineate changes in shape that occur as a result of alkyl substitution for these closely packed spherical structures.

The validation set included nine aldehyde tetralin musks whose SAR is different from that of the corresponding ketones (see Figure 8). Nevertheless, these compounds were correctly classified by the PC map developed from the training set data. This suggests that information about the underlying SAR of the tetralin and indan musks is captured by these 45 descriptors.

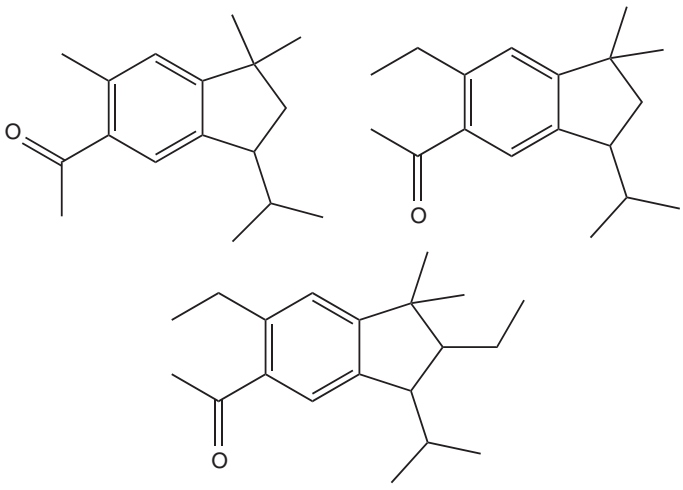

Figure 7 Chemical structures of the three misclassified validation set compounds (nonmusks).

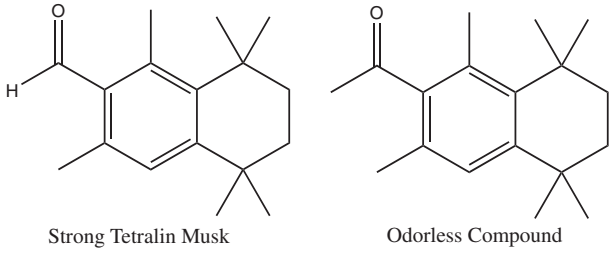

Figure 8 Chemical structures of two validation set compounds correctly classified by the 45 molecular descriptors identified by the pattern recognition GA. For the ketone, the presence of methyl groups near the polar heteroatom of the molecule is correlated to inactivity. It has the opposite effect for the corresponding aldehyde.

A back propagation neural network with topography of 45:2:1 was also used to classify the training set compounds as musks or nonmusks. About 45 descriptors identified by the pattern recognition GA were used as the inputs for a feed-forward neural network. The performance goal of 0.001 was set for the output of the neural network with a momentum constant of 0.99, learning rate of 0.01, ratio to decrease learning rate of 0.7, and ratio to increase learning rate of 1.5. All compounds in the training set were correctly classified. To test the predictive ability of these descriptors and the discriminant associated with them, a validation set of 37 compounds (see Table 2) was employed. Again, 34 of the 37 compounds were correctly classified. The three compounds misclassified by PCA were the same three compounds misclassified by the back propagation neural network.

The 41 TAE-, PEST-, and 4 CODESSA-derived descriptors identified by the pattern recognition GA are listed in Table 3. Most of these descriptors convey information about intermolecular interactions and shape, which suggests their importance in defining musk odor quality for tetralin and indan musks. TAE descriptors selected by the pattern

Table 3 List of TAE, PEST, and CODESSA descriptors

BNP.MIN	BNP.H0	BNP.W16
BNP.B36	BNP.B51	PIP.W3
PIP.W15	PIP.W17	PIP.W31
PIP.W3	FUK.W0	FUK.W4
FUK.W11	FUK.W20	FUK.W28
FUK.B56	EP.H3	EP.W5
EP.W14	EP.W22	EP.W31
EP.B02	G.W10	G.W29
K.W6	K.W21	K.B34
K.B46	G.B12	DKN.STDN
DKN.W27	DGN.AVGN	DGN.W27
DGN.B26	DRN.W0	DRN.W4
LAPL.W10	LAPL.W18	LAPL.W31
ANGLE.B06	ANGLE.B36	WPSA-3
HOMO-LUMO GAP	TOTAL HYBRID.	RCNS

recognition GA include five bare nuclear potential (BNP) descriptors and four Politzer ionization potential (PIP) descriptors. The name "bare nuclear potential" is a reflection of the fact that this quantity is being mapped onto the electron density isosurface. BNP.MIN, BNP.H0, and BNP.W16 capture polar and hydrogen bonding interactions. BNP.B36 and BNP.B51 capture information about shape and polarity. The "3" indicates that rays of average size are important, "6" indicates that relatively high BNP values are important, "5" indicates that long rays are important, and "1" indicates that relatively low BNP values are important. PIP descriptors, which describe several aspects of the surface electrostatic potential distribution of the compounds in the data set, have been shown to be correlated with a number of intermolecular binding modes, not the least of which is induced-dipole interactions. PIP.W3, PIP.W15, PIP.W17, and PIP.W31 are wavelet representations of the local ionization potential of the molecule. PIP.W3 describes the low value range of this property whereas the other three descriptors describe the high value range. These four PIP descriptors are known to convey information about hydrogen bonding and acidity. The seven Fukui radical reactivity indices identified by the pattern recognition GA are similar to the PIP indices, in that both involve a perturbation expression that is meant to describe the spatial distribution of radical reactivity. In the PIP case, the molecular surface is encoded with energy-weighted orbital densities; while in the Fukui case, there is a selectable denominator term that places the reactivity index on a cationic, radical, or anionic scale. Six of the seven Fukui descriptors are wavelet representations with FUK.W0 and FUK.W4 describing the low value range of this property

and FUK.W11, FUK.W20, and FUK.W28 describing the high value range. FUK.B56 is a shape descriptor with near median values of ray length and property.

The six electrostatic potential (EP) descriptors represent a more direct form of scalar EP values on the molecular surface. EP.H3 is a histogram descriptor and EP.W5, EP.W14, EP.W22, and EP.W31 are scale coefficient wavelet descriptors. These five descriptors are correlated to the solvation energy of the molecule with EP.W14, EP.W22, and EP.W31 in the positive part of the EP range. EP.B02 is a shape property descriptor denoting short ray length distances and negative EPs. G.W10, G.W29, K.W6, and K.W21, which are wavelet descriptors derived from the G and K kinetic energy reconstructions normal to and away from the surface of the molecule, describe hydrogen bonding interactions. K.B34, K.B46, and G.B12 are shape descriptors derived from the ray traces of the G and K kinetic energy reconstructions normal to and away from the surface of the molecule. These descriptors capture more of the interior volume, that is, local shape, as opposed to conformational information. DKN.STDN and DKN.W27, which describe the rate of change in the K kinetic energy density normal to and away from the surface of the molecule, are correlated to both hydrophobicity and polarizability. DGN.AVGN and DGN.W27, which characterize the rate of change in the G kinetic energy density normal to and away from the surface of the molecule, describe weak bonding interactions. DGN.B26, which is a PEST descriptor, suggests that molecular shape is important. DRN.W0 and DRN.W4 are wavelet descriptors that represent the rate of fall off of the electron density and are highly correlated to the DGN descriptors. LAPL.W10, LAPL.W18, and LAPL.W31 are also wavelet descriptors. They are derived from the second derivative of the electronic energy distribution and are important in characterizing donor/acceptor relationships.

Other descriptors identified by the pattern recognition GA include ANGLE.B06 and ANGLE.B36, which are pure shape descriptors. Planar, disk shaped molecules are favored by these two descriptors. Four CODESSA descriptors were identified by the pattern recognition GA. These four descriptors contain both complementary and supplementary information to the TAE descriptors about the electronic structure of the molecule and the distribution of charge on the surface.

The SAR of tetralin and indan musks has been documented in the literature (Beets 1978; Narvaez et al. 1986). Geometric considerations have been shown to be important. A particular skeletal arrangement, the position of the polar functional group relative to the bulky substituents in the molecule, has been shown to be strongly correlated with the musk odor modality. Another important variable is the steric environment of the polar functional group. The presence of methyl groups near the polar heteroatom of the molecule has been shown to be correlated to inactivity presumably due to steric hindrance. Although a polar functional group must be present for compounds to elicit musk odor, the introduction of a second functional group will diminish the musk odor modality unless the two functional groups are not too distant from each other in order to allow for some form of cooperation between them. Information about the centers of branching in the molecule, for example, the degree of branching of substituents attached to the ring, is another important variable for identifying musk odorants. Indans appear to require extensive substitution in the five-membered ring if they are to exhibit musk odor. The 147 bicyclic compounds selected for this study exhibit these trends, and the PC map of the descriptors identified by the pattern recognition GA has effectively captured this information by way of a graphically oriented structure–activity correlation. As for Figure 4, the musks and the nonmusks have unequal covariance matrices due to the asymmetric data structure, which is a direct result of the nonmusks being dissimilar with regard to the class defining property of the musks. Since the nonmusks have been chosen to be very similar in structure to the musks, this suggests that TAE- and CODESSA-derived molecular descriptors identified by the pattern recognition GA contain information about the SAR for tetralin and indan musks.

The molecular descriptors identified by the pattern recognition GA contain information about the shape and electronic surface properties of the compounds. No descriptor by itself could correctly classify more than 65% of the compounds in the training set, and no two descriptors had a pair-wise correlation greater than 0.85. From this, one can conclude that musk odor quality does not correlate with any single molecular property. However, musk odor quality of the tetralin and indan musks was correlated to a set of 45 descriptors. The need for such a large number of descriptors to identify indan and tetralin musks can be partly attributed to the nature of the descriptors used. In addition, musk odor is a subtle function of several molecular properties.

The main processes involved in olfaction are volatility, mucous transport, receptor binding, and receptor activation. An examination of the descriptors identified by the pattern recognition GA suggests that volatility and mucous transport were not important factors in this study as molecular descriptors that are correlated to volatility (e.g., gross molecular size and lower order molecular connectivity indices) and transport (e.g., Log P and dipole moment) were not found to be important for differentiating musks from nonmusks. This result, which conflicts with the theory that diffusion rate through the mucous membrane is a major determinant of odor intensity, can probably be attributed to the fact that nonmusks selected for this study are similar in structure to the musks. Receptor binding and activation appear to be the factors for discrimination of musks from nonmusks in the training set. Although the descriptors selected by the pattern recognition GA could be more indicative of structural features lacking in the nonmusks than features involved in the mechanism governing the odor quality of the tetralin and indan musks, the asymmetric data structure encountered for the training set compounds would suggest the opposite.

Conclusion

The results of this study demonstrate the advantages of using TAE-derived descriptors as part of a new methodology to facilitate the design of odorants, for example, musks, which involves using large olfactory databases available through the open scientific literature as input for a new structure–activity correlation methodology. Both the new set of property descriptors and the pattern recognition GA that is able to recognize subtle patterns within descriptor fields are required to develop SARs that span structural manifolds. In this study, a machine learning perspective has been used to investigate OSRs. Misclassifications of validation set compounds that occurred can be directly tied to the molecular descriptors rather than the method used to develop the model.

Acknowledgments

The authors acknowledge Douglas Henry, Javier Narvaez, and Philip S. Magee (deceased) for helpful discussions. C.M.B. and C.M.S. would like to acknowledge the Rensselaer Exploratory Center for Cheminformatics Research (RECCR) and the Rensselaer Center for Biotechnology and Interdisciplinary Studies (CBIS).

Funding

Financial support of the National Science Foundation by the establishment of an REU site at Oklahoma State University (Award #0649162) is gratefully acknowledged by C.W. and B.K.L.

References

Amoore JE. 1970. Molecular basis of odor. Springfield (IL): Charles C. Thomas

Baur A. 1891. Studien uber den Kunstlichen Moschus. Ber. Dtsch. Chem. Ges. 24:2832–2843

Beets MGJ. 1973. Structure-response relationships in chemoreception. In: Cavalitto CJ. editor. Structure activity relationships. New York: Pergamum Press.

Beets MGJ. 1974. Odor, taste, and molecular structure. In: Morton I, Rhodes DN. editors. The contribution of chemistry to food supplies. London: Butterworth and Co. p. 99–152.

Beets MGJ. 1978. Structure activity relationships in human chemoreception. London: Applied Science Publishers.

Bersuker IB, Dimoglo AS, Gorbachov MY, Vlad PF, Pesaro M. 1991. Origin of musk fragrance activity: the electron-topological approach. New J. Chem. 15:307–320.

Breneman CM, Thompson TR, Rhem M, Dung M. 1995. Electron density modeling of large systems using the transferable atom equivalent method. Comput. Chem. 19(3):161–172.

Breneman CM, Sundling CM, Sukumar N, Shen L, Katt W, Embrechts MJ. 2003. New developments in PEST shape/property hybrid descriptors. J. Comp. Aid. Molec. Design. 17(2–4):231–240.

Charastrette M, Zakarya D, Peyraud JF. 1994. Structure-musk odor relationships for tetralins and indans using neural networks (on the contribution of descriptors for classification). Eur. J. Med. Chem. 29:343–348.

Cherqaoui D, Esseffar M, Villemin, Cense JM, Chastrette M, Zakarya D. 1998. Structure-musk odor relationship of tetralin and indan compounds using neural networks. New J Chem. 22:839–843.

Diete C. 1982. Manufacture of perfumery. Philadelphia: Henry Carey & Baird Company.

Dunn WJ, Wold S. 1980. Structure-activity analyzed by pattern recognition: the asymmetric case. J. Med. Chem. 23:595–599.

Dunn WJ, Wold S, 1995. SIMCA pattern recognition and classification. In: van de Waterbeemd H. editor. Chemometric methods in molecular design. New York: VCH. p. 187.

Eiceman GA, Wang M, Pradad S, Schmidt H, Tadjimukhamedov FK, Lavine BK, Mirjankar N. 2006. Pattern recognition analysis of differential mobility spectra with classification by chemical family. Anal. Chim. Acta. 579(1):1–10.

Fehr C, Galindo J, Haubrichs R, Perret R. 1989. New aromatic musk odorants: design and synthesis. Hel. Chim. Acta. 72:1537–1553.

Frater G, Bajgrowicz JA, Kraft P. 1998. Fragrance chemistry. Tetrahedron. 54:7633–7703.

Fuchs, K. US 2759022 19560814; August 14, 1956.

Jennings-White C. 1985. Human primary odors. Perfum. Flavor. 9:46–58.

Karasinski J, Andreescu S, Sadik OA, Lavine BK, Vora MN. 2005. Multiarray sensors with pattern recognition for the detection, classification, and differentiation of bacteria at subspecies and strain levels. Anal. Chem. 77(24):7941–7949.

Klopman G, Ptscelintsev D. 1992. Application of the computer automated structure evaluation methodology to a QSAR study of chemoreception. Aromatic musky odorants. J. Agric. Food Chem. 40:2244–2251.

Kraft P, Bajfrowicz JA. Denis C. Frater G. 2000. Odds and trends: Recent developments in the chemistry of odorants. Angewandte Chemie Int. Ed. 39:2980–3010.

Kraft P. 2004. Brain aided musk design. Chem. Biodiver. 1(12):1957–1974.

Lavine BK, Davidson CE, Moores AJ. 2002. Innovative genetic algorithms for chemoinformatics. Chem. Intell. Lab. Instrumen. 60(1):161–171.

Lavine BK, Davidson CE, Breneman C, Katt W. 2003. Electronic van der Waals surface property descriptors and genetic algorithms for developing structure-activity correlations in olfactory databases. J. Chem. Inf. Comput. Sci. 43:1890–1905.

Lavine BK, Davidson CE, Breneman CM, Katt W. 2004a. Genetic algorithms for clustering and classification of olfactory stimulants. In: Bajorath J. editor. Chemoinformatics: methods and protocols. Methods Mol Biol. Humana Press. Totowa, NJ. 275:399–426.

Lavine BK, Davidson CE, Rayens WT. 2004b. Machine learning based pattern recognition applied to microarray data. Combinatorial Chemistry & High Throughput Screening. 7:115–131.

Lavine BK, Nuguru K, Mirjankar N. 2011. One stop shopping—Feature selection, classification, and prediction in a single step. J. Chem. 25:116–129.

Narvaez JN, Lavine BK, Jurs PC. 1986. Structure-activity studies of musk odorants using pattern recognition: bicyclo and tricyclo-benzenoids. Chem Senses. 11(1):145–156.

Ohloff G, Winter B, Fehr C. 1991. Chemical classification and structure-odor relationships. In: Muller PM, Lamparsky D. editors. Perfumes:

art, science, and technology. Amsterdam: Elsevier Applied Sciences. p. 310–325.

Ohloff G. 1994. Scent and fragrances: the fascination of odors and their chemical perspectives. New York: Springer Verlag.

Pybus DH, Sell CS. editors. 1999. The chemistry of fragrances. Cambridge: Royal Society of Chemistry.

Rose VS, Wood J, MacFie HJH. 1992. Generalized single class discrimination (GSCD). A new method for the analysis of embedded structure-activity relationships. QSAR. 11:492–504.

Rossiter KJ. 1996. Structure-odor relationships. Chem. Rev. 96: 3201–3240.

Sell CE. 2006. The chemistry of fragrances. Cambridge: RSC.

Song M, Breneman CM, Bi J, Sukumar N, Bennett KP, Cramer S, Tugcu N. 2002. Prediction of protein retention times in anion-exchange chromatography systems using support vector regression. J. Chem. Inf. Comput. Sci. 42(6):1347–1357.

Tenahsi R. 1971. Odor and molecular structure. In: Ohloff G, Thomas AF. editors. Gustation and olfaction. London: Academic Press.

Theimer ET. editor. 1982. Fragrance chemistry. The science of the sense of smell. New York: Academic Press.

Whitehead CE, Breneman CM, Sukumar N, Ryan MD. 2003. Transferable atom equivalent multi-centered expansion method. J. Comput. Chem. 24(4):512–529.

Wood TF. 1970. Chemistry of the aromatic musks. Clifton: Givaudanian. p. 1–37.