

SuperScent—a database of flavors and scents

Mathias Dunkel¹, Ulrike Schmidt¹, Swantje Struck¹, Lena Berger¹, Bjoern Gruening¹, Julia Hossbach¹, Ines S. Jaeger^{1,2}, Uta Effmert³, Birgit Piechulla³, Roger Eriksson⁴, Jette Knudsen⁵ and Robert Preissner^{1,*}

¹Structural Bioinformatics Group, Institute of Molecular Biology and Bioinformatics, Charité – University Medicine Berlin, Arnimallee 22, 14195 Berlin, ²Department of Cardiology and Angiology, Charité – University Medicine Berlin, Schumannstr. 20/21, 10117 Berlin, ³Institute of Biological Sciences, Biochemistry, University of Rostock, 18059 Rostock, Germany ⁴Department of Plant and Environmental Sciences, University of Gothenburg, SE 413 19 Göteborg and ⁵Department of Ecology, Lund University, SE 223 62 Lund, Sweden

Received August 15, 2008; Revised September 24, 2008; Accepted September 25, 2008

ABSTRACT

Volatiles are efficient mediators of chemical communication acting universally as attractant, repellent or warning signal in all kingdoms of life. Beside this broad impact volatiles have in nature, scents are also widely used in pharmaceutical, food and cosmetic industries, so the identification of new scents is of great industrial interest. Despite this importance as well as the vast number and diversity of volatile compounds, there is currently no comprehensive public database providing information on structure and chemical classification of volatiles. Therefore, the database SuperScent was established to supply users with detailed information on the variety of odor components. The version of the database presented here comprises the 2D/3D structures of approximately 2100 volatiles and around 9200 synonyms as well as physicochemical properties, commercial availability and references. The volatiles are classified according to their origin, functionality and odorant groups. The information was extracted from the literature and web resources. SuperScent offers several search options, e.g. name, Pubchem ID number, species, functional groups, or molecular weight. SuperScent is available online at: <http://bioinformatics.charite.de/superscent>.

INTRODUCTION

In scientific terms, scents are mixtures of volatile compounds with a high vapor pressure and a molecular weight, which is usually $<300 \text{ g mol}^{-1}$ (1). Human beings often associate scents with volatiles that can be perceived by the human nose and have a pleasant smell. But the

entire group of volatile compounds comprises thousands of inorganic and organic compounds stemming from major pathways of secondary metabolisms of many organisms (1). These volatiles may affect living organisms in one way or the other. Over the past decades, scientific investigations have revealed that volatiles play a key role in life by acting as semiochemicals, mediating inter- and intraspecies interactions of living organisms.

Volatiles allow animals to recognize or detect individuals. The volatiles called pheromones are indispensable for mating choices, sexual behavior and fertilization, and also for nursing. They are important for the maintenance of social relationships, especially in animal communities such as hives, ant colonies, prairie dog towns or even packs/prides/herds of larger animals. Volatiles also support foraging and the detection of prey. They can also serve as signals to warn kin in situations of danger (alarm pheromones) or they are even used to defend against predators. The relevance of volatiles is not restricted to the animal kingdom. Volatile semiochemicals are involved in plant–plant interactions (2) and many play a crucial role in plant–animal interactions, e.g. pollination (3), herbivory (4) and the plants' response to defend against herbivores (5). Recently, it was shown that bacteria also emit a wealth of volatiles with an impact on plants, fungi, animals and bacteria (M. Kai *et al.*, Bacterial volatiles and their action potential, Applied Microbiology Biotechnology, submitted for publication).

Last but not least, scent components are of tremendous commercial interest, resulting in many applications of volatiles in science and industry. Certain pleasant odors have a positive effect on customer and are therefore used in shopping malls, for wellness applications, and for the production of perfumes, cosmetics and household cleaning agents, to name a few.

For the recognition of odorant molecules, a large variety of olfactory receptors is known in humans and animals. To discriminate between scent components,

*To whom correspondence should be addressed. Tel: +49 30 8445 1649; Fax: +49 30 8445 1551; Email: robert.preissner@charite.de

each receptor has affinities for a range of molecules and, by combination of activated receptors, many different smells can be perceived (6). All olfactory receptors are members of class A rhodopsin-like family of G-protein-coupled receptors (GPCRs) (7), a group of seven transmembrane domain receptors activating signal transductions in cells. Crystal structures are currently available for rhodopsin (8) and β 2-adrenoreceptor (9). An excellent source for information about olfactory receptors is the Olfactory Receptor database (10). This database provides information about the perceiving proteins and complexes, and a database gathering comprehensive knowledge on volatiles should give information about their ligands. The work of Schmuker *et al.* (11,12) bridges the gap between the structure of volatile compounds and the receptor response by prediction via machine learning techniques using experimental data from *in vivo* receptor recordings.

Hitherto, a comprehensive compilation of volatiles was not publicly available for scientific use. It was therefore our goal to establish the database 'SuperScent'. A variety of database resources of scent components is already known, but they are limited in scope and focus on certain subgroups of scents, e.g. the OdorDB, part of the neuroscience Senselab project (13), centers on odorant molecules experimentally shown to bind olfactory receptor proteins, whereas the Pherobase database (<http://www.pherobase.com/>) focuses on pheromones but meanwhile also covers a broad variety of semiochemicals. A compilation of floral scent components is found in the ScentBase (<http://www2.dpes.gu.se/SCENTbase.html>) and the Flavornet (<http://www.flavornet.org/flavornet.html>) summarizes volatile compounds found in the human olfactory perception space. Consequently, these databases are useful for special purposes, but there is still a need for a comprehensive listing of volatiles regarding their properties and their uses in science and industry.

The SuperScent database, which is presented here, comprises more than 2100 compounds, together with several classification criteria. These volatile compounds were collected from a variety of sources, e.g. the literature or other databases, leading to a comprehensive dataset of scent components, together with information about their chemical properties and commercial availability. These features provide the user with an extensive database, together with substantial options, such as a search within particular group odor subgroups.

THE DATABASE

With 2147 compounds, 9214 synonyms and references to more than 20 different suppliers, SuperScent provides the largest diversity of volatile compounds and corresponding information available online. The data are accessible with two different search options: 'Scent Search' and 'Structure Search'.

The 'Scent Search' enables the user to look for compounds by means of the PubChem-ID or the name. Additionally, by choosing certain functional groups,

species, or range of molecular weights, all database entries meeting the search criteria can be accessed.

The second way to screen the database requires a molecular structure ('Structure Search'). Here, a SMILES code (Simplified Molecular Input Line Entry System) or a MOL-file of the search compound can be uploaded. With the help of MarvinView (<http://www.chemaxon.com>), the user can draw either the whole structure of the search molecule, or parts of it (I-A in Figure 1). Thus, it is possible to screen the database with self-edited molecule structures. In order to find resembling structures in the database, a similarity search is performed. The 10 most similar database entries are listed in the order of similarity (II in Figure 1). For each compound, the name, the PubChem-ID, the 2D structure and the Tanimoto coefficient are presented. Furthermore, a similarity search to find the 10, 20 or 30 most similar compounds is provided.

Additional information is available in a separate 'Properties' window (II-A/B in Figure 1). The user can find synonyms, functional groups, the molecular weight and all species in which the compound has been found, together with the corresponding reference. Furthermore, the compounds are classified according to their structure (e.g. benzenoids), chemical features (Figure 2), quality of scent (e.g. fruity-peach) and ordering information (supplier; ID).

Another useful feature of the web interface is the 'Scent Tree' (I-B in Figure 1). Here, the database entries have been clustered according to the quality of their aroma. There are 29 classes, such as balsamic, floral or spicy, which are again divided into several subclasses. For instance, the fruity scents include subclasses such as apple, banana or coconut, leading to an overall number of 121 different groups.

A manually verified upload option allows the scientific community to contribute to the database. Here, the user can import a MOL-file together with corresponding information of the compound. The SuperScent database will be updated twice a year.

METHODS

Data were collected from the literature and various web resources, such as a collection of floral scents (1), a review outlining bacterial volatiles (14) and PubChem (<http://pubchem.ncbi.nlm.nih.gov/>). Abstracts of the literature database PubMed were filtered for relevant articles using specific keywords. The abstracts were screened against names and synonyms of chemical compounds, as well as a distinct set of substrings of IUPAC names. The text passages containing matches were manually curated by a scientific team of biologists that confirmed the matching compounds and verified them. Several web resources, e.g. the Riechstoff-Lexikon (<http://omikron-online.de>) and the Flavors and Fragrances catalog (<http://safcsupplysolutions.com>) supplied by Sigma Aldrich were checked.

SuperScent is designed as a relational database, which is implemented on a MySQL server. For chemical functionality, the MyChem package, which aims to provide a

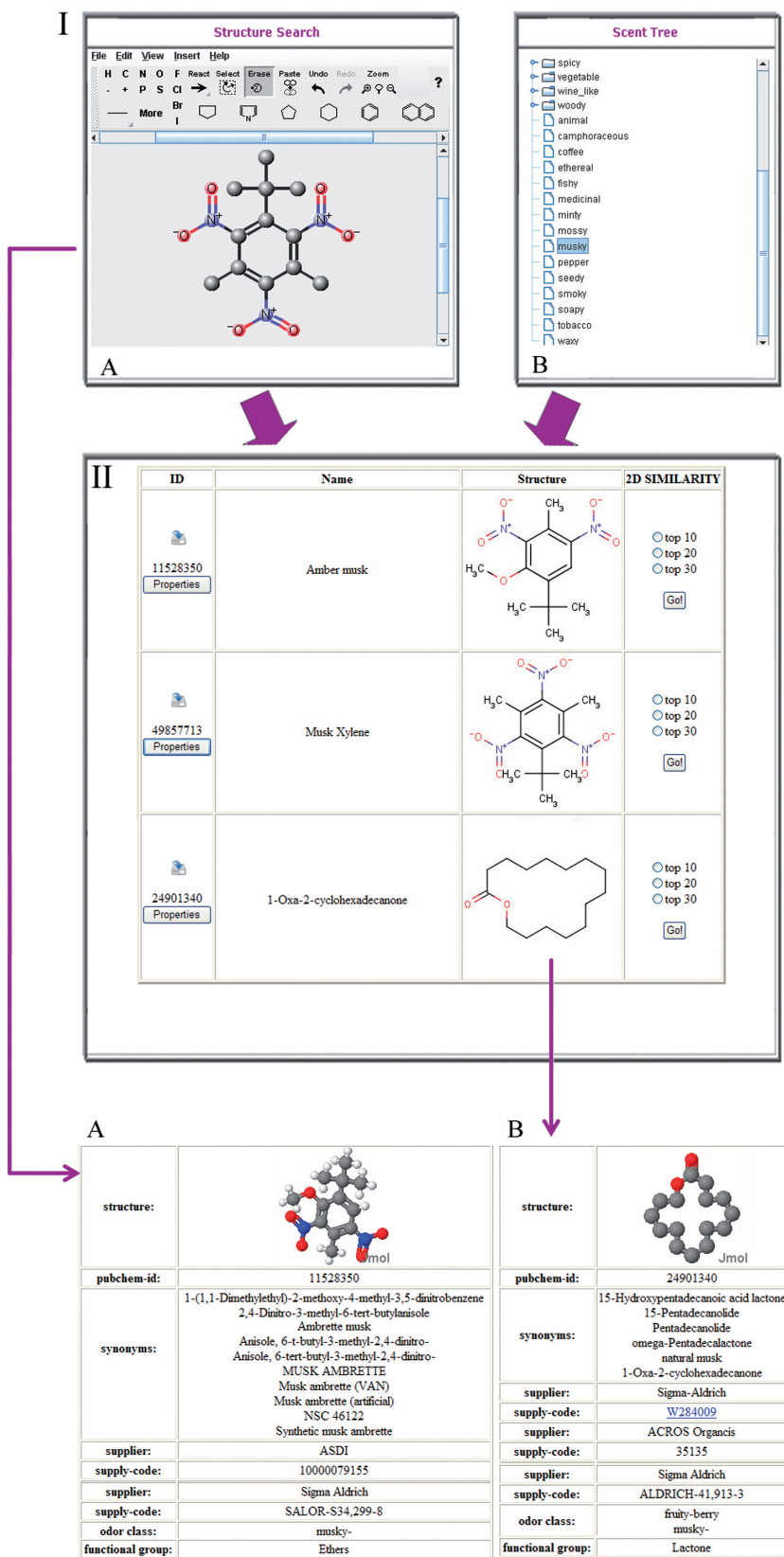


Figure 1. Flow chart of a scent search in the SuperScent database. (I) Different search options are provided: (A) Structure search: upload an MOL-file or SMILES code, but it is also possible to draw one's own structure. (B) Scent tree: structures are clustered into scent classes. The branch can be expanded by clicking on the nodes and a click on one subclass shows a list of compounds. (II) Result table of a scent search. The detailed view for two molecules is depicted: (A) synthetic musk. (B) natural musk.

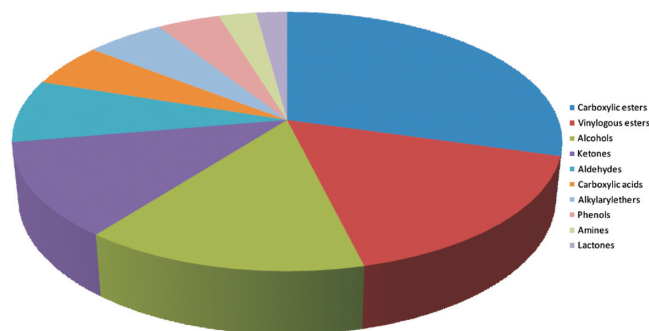


Figure 2. Pie chart of chemical classes found in the SuperScent database.

complete set of functions for handling chemical data within MySQL, is added. Most of the functions used by MyChem depend upon Open Babel (15). The fingerprint algorithm implemented in Open Babel follows the Daylight approach (<http://www.daylight.com/dayhtml/doc/theory/>). As similarity index, the Tanimoto coefficient is used, calculating the number of bit positions set to 1 in both fingerprints, divided by the number of bit positions set to 1, in at least one of the fingerprints. If a set bit is considered as a feature present in the molecule, the Tanimoto coefficient is a measure of the number of common features in both molecules (16). A Tanimoto coefficient of >0.85 indicates that two molecules may have similar activities (17). For displaying 3D structures, Jmol, an open-source Java viewer for chemical structures in 3D (<http://www.jmol.org/>) is used. Marvin ChemSketch was applied for the built-in molecule editor, which allows structural screening with self-edited molecules. The website is built with php and javascript, and web access is enabled via Apache HTTP Server 2.2.

CONCLUSION AND FUTURE DIRECTION

The SuperScent database has become a useful tool to retrieve information about scents or to get an overview of the known volatile organic compounds. The included data on purchasability of scents will enable systematic experimental approaches on the relation between structural similarities and scent classes. Furthermore, structure comparisons of self-edited molecules with the annotated scents may allow a first rough estimation of the potential aroma of new chemicals. The SuperScent database is a free resource with embedded screening functions for chemical compounds. The extension of the database allows the scientific community simple access to a growing number of available scents. We plan to extend the database by including olfactory receptors in the near future.

FUNDING

Deutsche Forschungsgemeinschaft (SFB 449); Investitionsbank Berlin (IBB); International Research Training Group (IRTG); Berlin-Boston-Kyoto and Deutsche Krebshilfe. Funding for open access charge: SFB 449.

Conflict of interest statement. None declared.

REFERENCES

- Knudsen, J.T., Eriksson, R., Gershenzon, J. and Ståhl, B. (2006) Diversity and distribution of floral scent. *Bot. Rev.*, **72**, 1–120.
- Baldwin, I.T., Halitschke, R., Paschold, A., von Dahl, C.C. and Preston, C.A. (2006) Volatile signaling in plant-plant interactions: “talking trees” in the genomics era. *Science*, **311**, 812–815.
- Dobson, H.E.M. (2006) Relationship between floral fragrance composition and type of pollinator. In Dudareva, N. and Pichersky, E. (eds), *Biology of Floral Scent*. Taylor and Francis, Boca Raton, FL, pp. 147–198.
- Vet, L.E.M. and Dicke, M. (1992) Ecology of infochemical use by natural enemies in a tritrophic context. *Annu. Rev. Entomol.*, **37**, 141–172.
- Heil, M. and Silva Bueno, J.C. (2007) Within-plant signaling by volatiles leads to induction and priming of an indirect plant defense in nature. *Proc. Natl Acad. Sci. USA*, **104**, 5467–5472.
- Firestein, S. (2001) How the olfactory system makes sense of scents. *Nature*, **413**, 211–218.
- Krautwurst, D. (2008) Human olfactory receptor families and their odorants. *Chem. Biodivers.*, **5**, 842–852.
- Palczewski, K., Kumasaka, T., Hori, T., Behnke, C.A., Motoshima, H., Fox, B.A., Le Trong, I., Teller, D.C., Okada, T., Stenkamp, R.E. *et al.* (2000) Crystal structure of rhodopsin: a G protein-coupled receptor. *Science*, **289**, 739–745.
- Cherezov, V., Rosenbaum, D.M., Hanson, M.A., Rasmussen, S.G., Thian, F.S., Kobilka, T.S., Choi, H.J., Kuhn, P., Weis, W.I., Kobilka, B.K. *et al.* (2007) High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science*, **318**, 1258–1265.
- Crasto, C., Marengo, L., Miller, P. and Shepherd, G. (2002) Olfactory Receptor Database: a metadata-driven automated population from sources of gene and protein sequences. *Nucleic Acids Res.*, **30**, 354–360.
- Schmucker, M., de Bruyne, M., Hahnel, M. and Schneider, G. (2007) Predicting olfactory receptor neuron responses from odorant structure. *Chem. Cent. J.*, **1**, 11.
- Schmucker, M. and Schneider, G. (2007) Processing and classification of chemical data inspired by insect olfaction. *Proc. Natl Acad. Sci. USA*, **104**, 20285–20289.
- Crasto, C.J., Marengo, L.N., Liu, N., Morse, T.M., Cheung, K.H., Lai, P.C., Bahl, G., Masiar, P., Lam, H.Y., Lim, E. *et al.* (2007) SenseLab: new developments in disseminating neuroscience information. *Brief. Bioinform.*, **8**, 150–162.
- Schulz, S. and Dickschat, J.S. (2007) Bacterial volatiles: the smell of small organisms. *Nat. Prod. Rep.*, **24**, 814–842.
- Guha, R., Howard, M.T., Hutchison, G.R., Murray-Rust, P., Rzepa, H., Steinbeck, C., Wegner, J. and Willighagen, E.L. (2006) The Blue Obelisk-interoperability in chemical informatics. *J. Chem. Inf. Model.*, **46**, 991–998.
- Delaney, J.S. (1996) Assessing the ability of chemical similarity measures to discriminate between active and inactive compounds. *Mol. Divers.*, **1**, 217–222.
- Patterson, D.E., Cramer, R.D., Ferguson, A.M., Clark, R.D. and Weinberger, L.E. (1996) Neighborhood behavior: a useful concept for validation of “molecular diversity” descriptors. *J. Med. Chem.*, **39**, 3049–3059.