Available online at www.sciencedirect.com

SCIENCE DIRECT®

Mendeleev Commun., 2007, **17**, 90–91

Mendeleev Communications

QSAR modeling on the basis of 3D descriptors representing the electrostatic molecular surface (ambergris fragrances)

Igor V. Svitanko,*[a,b,†] Dmitry A. Devetyarov,[c] Dmitry E. Tcheboukov,[b] Maksim S. Dolmat,[a,‡] Alexey M. Zakharov,[c] Svetlana S. Grigor'eva,[c] Viktoriya T. Chichua,[c] Lyudmila A. Ponomareva[b] and Mikhail I. Kumskov[b]

[a] *Higher Chemical College, Russian Academy of Sciences, 125047 Moscow, Russian Federation. E-mail: svitanko@mail.ru*

[b] *N. D. Zelinsky Institute of Organic Chemistry, Russian Academy of Sciences, 119991 Moscow, Russian Federation. Fax: +7 495 135 5328*

[c] *Department of Mechanics and Mathematics, M. V. Lomonosov Moscow State University, 119992 Moscow, Russian Federation*

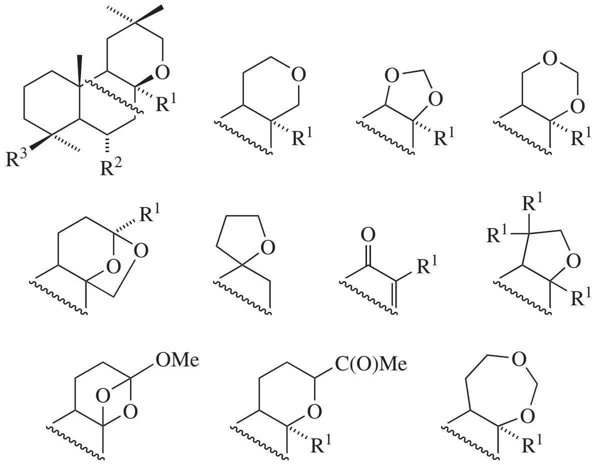
DOI: 10.1016/j.mencom.2007.03.012

A 3D-QSAR approach based on the electrostatic surface of molecules was used for the ambergris odour, and it showed a cross validation coefficient of 0.8.

Previously, we have constructed the models of a musk odour[1] and a model and new structures of the bicycloureas of psychotropic activity[2] (which were successfully synthesised). The method proposed allows us to take into account the spatial electrostatic complementary character of two or several molecules or a molecule and a receptor (long-range interaction). For this purpose, it is necessary to calculate the electrostatic field created by the molecule and then supplement the model thus obtained by structural complementarity (short-range interaction).

The sample consisted of 50 compounds (37 active and 13 inactive, Figure 1) represented by 3D molecular graphs.[3] For every molecule, we considered that the 3D coordinates of nodes (atoms) and their quantitative characteristics (partial atomic charges) are known.

Geometry optimization and charge calculations were performed by Gaussian03. The molecule was represented by a molecular surface (the distances from atoms were equal to a van der Waals radius), which was 'colourised' by a local physical property (LPP), *e.g.*, charge, lipophilicity and the ability to release or accept an electron (donor–acceptor factors) (Figure 2).



R1 = H, Me; R2 = H, OH; R3 = Me, OMe

Figure 1 Examples of compounds in the test sample.

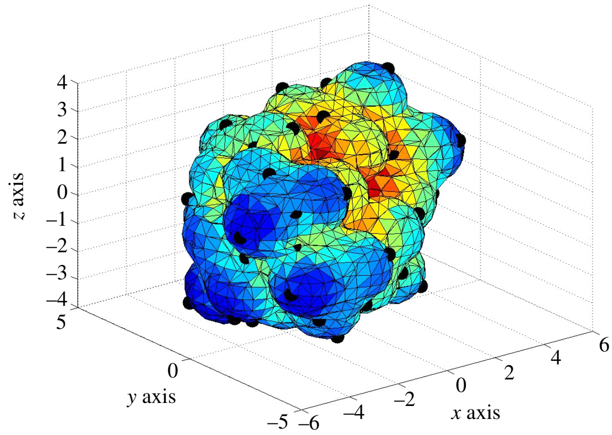


Figure 2 Molecular surface and critical points (see colour version on the cover of this issue).

For every molecule in the test set, we constructed its triangulated molecular surface with excluded solvent using MSMS.[4] We picked critical points on molecular surfaces using Connolly's methodology[5] to describe local knobs and holes. Finally, we calculated simple electrostatic (Coulombic) potentials of every critical point by adding up the effects of electrostatic fields created by individual atoms.

We can now reformulate the QSAR problem as follows: every object (molecule) in the test set is represented by a set of N critical points $(x_i, y_i, z_i, S_i, Q_i)$, $i = 1, ..., N$, where (x_i, y_i, z_i) are the 3D coordinates of a critical point, S_i describes the shape (0 for a hole and 1 for a knob), and Q_i is the electrostatic potential. Our aim was to construct 3D structural descriptors representing the critical points, their pairs and triplets and to check their value for the QSAR modeling.

We first performed a cluster analysis of the electrostatic potentials of all critical points in all molecules to segment out

[†] Vice-chairman of the Higher Chemical College (HCC) of the RAS.

[‡] A former student of the HCC RAS (1995–2000).

three clusters: A, B, and C. Then, we assigned labels to every critical point according to its geometric and electrostatic properties. The label of a critical point is made up of two symbols, where the first one represents electrostatic properties and the second one, the geometric properties (Q_i). Thus, there are the labels A0, A1, B0, B1, C0 and C1.

Further, we calculated the maximum Euclidian distance $d_{\max}$ between all critical points of all molecules in our sample. We divided the segment $[0, d_{\max}]$ into three equal parts:

$$D = \left[0, \frac{d_{\max}}{3}\right],\ E = \left[\frac{d_{\max}}{3}, 2\frac{d_{\max}}{3}\right],\ F = \left[2\frac{d_{\max}}{3}, d_{\max}\right].$$

We constructed a three-level ‘descriptor alphabet‘. The first level of our descriptor alphabet consists of the character labels of all critical points. The second level consists of the character descriptors in the form $F^2 = (P_1, P_2, d) = P_1 + P_2 + d$, where P_1, P_2 are 3-character labels of critical points and d denotes the distance segment (D, E or F) between P_1 and P_2, ‘+’ refers to a concatenation operation. The value of descriptor (P_1, P_2, d) is a number of occurrences of string F^2 on a molecule. Similarly, we constructed the third-level descriptors $F^3 = [(P_1, P_2, d_1), P_3, d_2] = F^2 + P_3 + d_2$, where d_2 denotes the distance between the midpoint of $[P_1, P_2]$ and P_3.

We enumerated all pairs and triplets to calculate F^2 and F^3 descriptors and their values for each molecule. In total, there were 703 descriptors with non-zero values. Thus, we obtained a molecule–descriptor matrix with 50 rows and 703 columns. Finally, we applied a liner version of the group method of data handling (GDMH)[6] to evolve a set of linear models. With GDMH, we found a descriptor set to define a distance on molecules to form two clusters with 28 and 10 points (molecules). On the first cluster, we obtained the linear model with six descriptors and a cross-validation of 0.79 and on the second cluster, with two descriptors and a cross-validation of 0.8. Thus, we can try to predict ambergris fragrances for new molecules located in the neighbourhood of first or second cluster. We refused to do the prediction if a new molecule was located far from cluster centres.

An advantage of the method is that the 3D molecule description is invariant to space rotations of the molecule, which eliminates the need to align molecules in 3D space (as in the CoMFA method).

References

1 I. V. Svitan’ko, M. I. Kumskov, I. L. Zyryanov and I. A. Suslov, *Mendeleev Commun.*, 1994, 161.

2 I. V. Svitan’ko, I. L. Zyryanov, M. I. Kumskov, L. I. Khmel’nitskii, L. I. Suvorova, A. N. Kravchenko, T. B. Markova, O. V. Lebedev, G. A. Orekhova and S. V. Belova, *Mendeleev Commun.*, 1995, 49.

3 *Computational Chemical Graph Theory*, ed. D. H. Rouvray, Nova Publishers, New York, 1989.

4 M. F. Sanner, A. J. Olson and J.-C. Spehner, in *Proc., 11th Annual ACM Symposium on Computational Geometry*, 1995, pp. C6–C7.

5 M. L. Connolly, *Biopolymers*, 1986, **25**, 1229.

6 M. I. Kumskov and D. F. Mityushev, *Pattern Recognition and Image Analysis*, 1996, **6**, 497.

Received: 10th November 2006; Com. 06/2817