ISSN 0027-1314, Moscow University Chemistry Bulletin, 2007, Vol. 62, No. 5, pp. 249–251. © Allerton Press, Inc., 2007.
Original Russian Text © S.S. Grigor'eva, V.T. Chichua, D.A. Devet'yarov, M.I. Kumskov, 2007, published in Vestnik Moskovskogo Universiteta. Khimiya, 2007, No. 5, pp. 305–307.

Selection of the Optimal Description of the Structure of a Molecule with Specified Biological Activity in the QSAR Problem

S. S. Grigor'eva, V. T. Chichua, D. A. Devet'yarov, and M. I. Kumskov

Department of Computational Mathematics, Faculty of Mechanics and Mathematics

E-mail: ss_grigoreva@mail.ru

Received April 9, 2007

Abstract—A special algorithm for solving the QSAR problem for amber odorants has been considered.

DOI: 10.3103/S0027131407050057

The problem of the search for relationships between the structures of chemical compounds and their properties is still important. In this work, we described a special algorithm applicable to solving this problem for a set of molecules of amber odorants (small molecules possessing an amber odor). A specific feature of the approach to the QSAR problem used in this work consists in that prediction of some (specified) properties of a molecule is based not on initial molecular graphs but on three-dimensional graphs derived from the initial ones, and the vertices of these graphs are formed by stationary points rather than by the atoms of molecules [1, 2]. As a result, the object to be analyzed is a spatial graph whose vertices are located at the stationary points of the molecular surface. Only after then, based on the resulting three-dimensional labeled graph, the values of structural 3D descriptors are calculated, which can be used for making QSAR predictions.

The above molecular surface is constructed based on the van der Waas radii of atoms. Around each atom of a molecule, a sphere of a given radius is constructed, and the aggregate of these spheres is considered. The resulting region is the base of the molecular surface at which, after "smoothing," stationary points are distinguished. They are defined as geometric local extrema of the surface (the points at the surface nearest to or farthest from definite groups of atoms) or as physicochemical extrema. A set of stationary points is calculated for each molecule; these points are specified by their coordinates, geometric type, and potential (Fig. 1).

Each stationary point is assigned a symbolic label (marker). Marking is determined by the description parameters optimized in the problem and can be calculated based on both the geometric type and the electrostatic potential at the surface. The electrostatic potential is calculated by the formula

$$\phi = \sum (Q_i/4\pi\varepsilon\varepsilon' R_i),$$

where R_i is the distance from the ith atom to the SP, and Q_i is the charge on this atom. A set of all electrostatic potential values thus obtained is considered. The range containing this set is divided into several partition intervals. Three intervals are distinguished: the interval of values close to zero (the magnitude of the potential is smaller than the specified threshold value) and the intervals of positive and negative values (the magnitude is larger than the threshold value).

As a result, each stationary point has two characteristics: (1) the local maximum or local minimum (two variants) and (2) and the interval of the electrostatic potential at the point (three variants). Combinations of these characteristics gave six types of stationary points. According to these types, the stationary points are assigned symbolic labels. In this work, the numbers from 1 to 6 are used as the labels.

At the next stage, we need to construct a family of descriptors adapted to a given property (activity) and then formulate a structure–descriptor matrix. A known model used for studying structure–biological activity relationships is a "spatial triangle" in which the vertices have specified local physicochemical properties and the sides are specified by intervals of distances. If a 3D conformation of a molecule exists that "contains" such a triangle, this molecule is believed to have the specified biological property. A more complicated variant of such a model is a spatial pyramid with specified properties of the "vertices" and "edges." Based on such a model, we attempted to construct the descriptor alphabet so that to describe the mutual arrangement of pairs, triads, and tetrads of stationary points of the molecular surface.

Then, we formulated a molecule–descriptor matrix in which the rows correspond to the molecules of the training set and the columns, to the descriptors. It is worth noting that, in addition to structural 3D descriptors, scalar descriptors can be used: general physico-

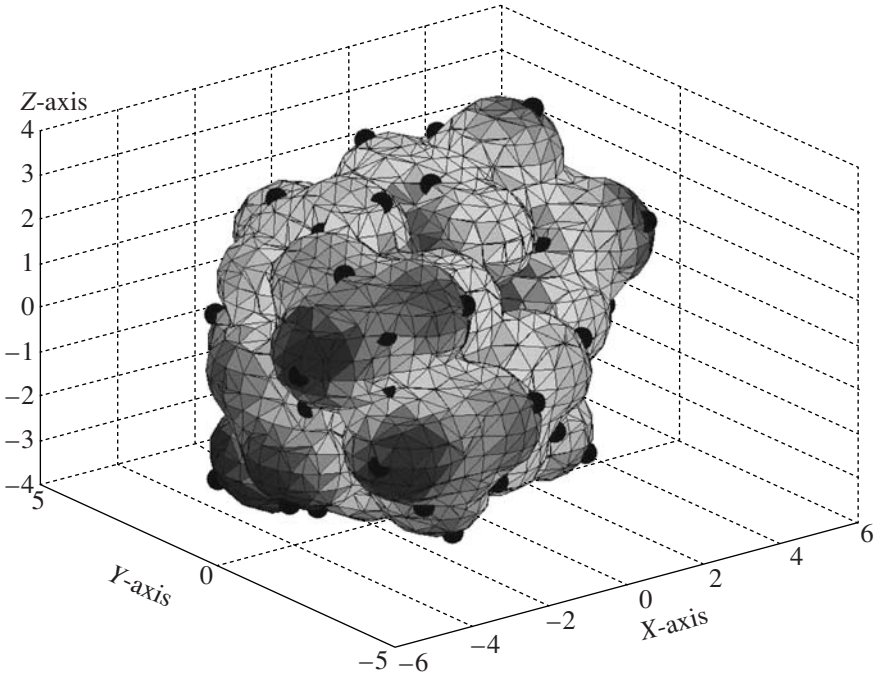
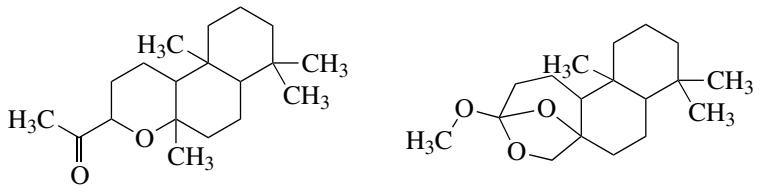


Fig. 1. Stationary points at the molecular surface.



Scheme 1.

chemical characteristics of molecules (e.g., molecular weight, volume, refraction, surface tension, density, dielectric constant, polarizability) and classical topological indices.

At the fourth stage, based on the formulated molecule–attribute matrix X of the size $N \times M$ (N is the number of objects of the training set, M is the number of revealed attributes of the objects) and on the column of properties (depending on whether or not the ith molecule has a given property, the column has in the ith row, respectively, 1 or –1), a prediction algorithm was "launched" [3]. As such an algorithm, we used the group method of data handling (GMDH), which makes it possible to select significant properties of the molecules of the matrix column (since the constructed matrix is often very "wide," i.e., $M \gg N$) [4].

Inasmuch as the spatial forms of molecules of the processed set are "inhomogeneous", the structure-activity relationship in the framework of the GMDH method is sought as a tree of solutions in such a way that the initial set of molecules is divided, upon training, into groups (clusters), and, then, a separate classifier is found for each cluster (Fig. 2).

The hierarchical cluster analysis method is used for identification of clusters [3]. An important advantage of this method is the possibility to reject the prediction if the compound studied is not similar to the molecules of the training set. To assess the predictive stability of the model, a cross-validation procedure is used [3] and the multiple correlation coefficient is calculated. This coefficient allows us to evaluate the quality of the description of the set in the given classification model constructed on selected parameters. The best model can be chosen by varying the parameters of calculation of descriptors, and the corresponding set of descriptors is referred to as optimal for a specified property.

As mentioned above, this algorithm was tested for the set of 50 molecules, 37 of which were amber odor-

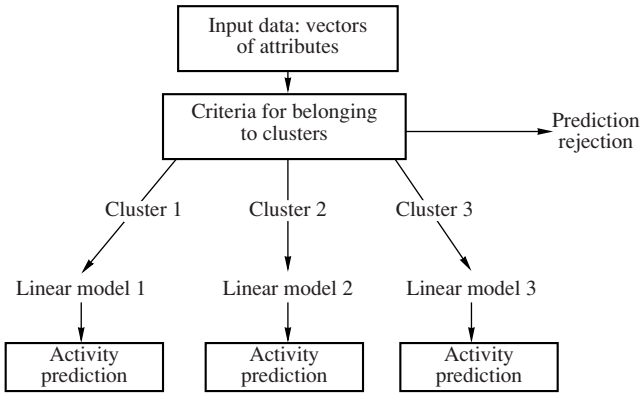


Fig. 2. Scheme of use of the solution tree in prediction of biological activity.

ants (have an odor) (Scheme 1), i.e., were biologically active, and the other 13 molecules with similar structures did not exhibit biological activity. For each compound of the set, the 3D description of the corresponding molecular graph was obtained: the graph vertices (atoms) were enumerated with additional attributes, such as chemical symbol, three-dimensional coordinates in angstroms, and electric charge. One of the molecular surfaces with stationary points is shown in Fig. 1.

After pairs and triads for each compound were found, descriptors of length 703 were obtained. Based on the formulated molecule–attribute matrix, the GMDH algorithm was implemented. As a result, two large clusters containing 28 and 10 elements were distinguished with the cross-validation estimate 78.6 and 80%, respectively [5]. This means that, by means of the suggested algorithm, the labeled stationary points describe rather accurately the activity in a series of amber odorants.

In further description of the molecular surface in the framework of the suggested approach, other properties of molecules can be considered, for example, their lipophilicity or reactivity (the propensity to donate or accept an electron or proton), thereby improving the predictive quality of the model.

REFERENCES

1. Lee, B. and Richards, F.M., *J. Mol. Biol.*, 1971, vol. 3, no. 55, p. 379.
2. Rouvray, D.H., *Computational Chemical Graph Theory*, New York, 1989.
3. Aivazyan, S.A., Bukhshtaber, V.M., Enyukov, I.S., and Meshalkin, L.D., *Prikladnaya statistika. Klassifikatsiya i snizhenie razmernosti* (Applied Statistics. Classification and Dimension Reduction), Moscow, 1989.
4. Kumskov, M.I. and Mityushev, D.F., *Pattern Recognit. Image Anal.*, 1996, vol. 6, p. 497.
5. Svitanko, I.V., Kumskov, M.I., Tcheboukov, D.E., Dolmat, M.S., Zakharov, A.M., Ponomareva, L.A., Grigor'eva, S.S., and Chichua, V.T., *Proc. of the 16th European Symposium on Quantitative Structure–Activity Relationships & Molecular Modelling*, 2006.