Automatic formation method for structural descriptors of organic compounds for quantitative structure—property relationships

M. I. Kumskov, L. A. Ponomareva,* E. A. Smolenskii, D. F. Mitushev, and N. S. Zefirov

N. D. Zelinsky Institute of Organic Chemistry, Russian Academy of Sciences,
47 Leninsky prosp., 117913 Moscow, Russian Federation.
Fax: +7 (095) 135 5328

A graph theory approach for constructing structural descriptors for finding quantitative structure—property relationships (QSPR) has been suggested. A complete enumeration of the linear fragments (chains of atoms) of molecular structures has been performed. Atoms in chains differ both by the type of element and by the marker that reflects their chemical and topological peculiarities. QSPR-models have been developed for various classes of chemical compounds for estimation of lipophilicity, enthalpy of formation, boiling points, and chromatographic retention time. All of these QSPR-models have been simulated by a BIBIGON MATCH PC-computer system (Basic Instrument for Building/Interactive Generation of Optimized Networks of Marked-Atom Chains), which implements the method suggested.

Key words: structural descriptors, markers, chemical graph, quantitative structure—property relationships (QSPR), lipophilicity, enthalpy of formation, boiling point, chromatographic mobility, polarizability.

The establishment of dependences between the structures of chemical compounds and their properties is one of the prime problems of theoretical chemistry. Different methods for describing molecules by various descriptors are used for building computerized "structure—property" correlation dependences.

A large class of descriptors[1,2] can be reduced to the following scheme of calculations: 1) a molecular chemical graph (MC-graph) is divided into subgraphs; 2) a number of repetitions of the j-th fragment in the considered MC-graph is calculated; 3) a weighted coefficient b_j is assigned to each fragment; 4) a descriptor value is calculated. Classical indices of Wiener, Balaban, Randič, Kier—Hall, and others[2,3] have been found by this scheme. For example, for the calculation of the Wiener index, a molecule is divided into atoms, and the b_j value assigned to a vertex of an MC-graph is determined by the sum of the j-th row of the distance matrix.[2] For the calculation of the Randich index, a molecular graph is represented by edges; the vertices of this graph do not differ, but the multiplicity of the bonds is implicitly considered by the coefficients b_j.

The classification of fragments based on families of Kier—Hall connection indices is the most developed. The method of division is determined by the number of atoms (K) of a fragment: only atoms are in the fragment when $K = 1$, only edges are included when $K = 2$, double chains are included when $K = 3$, and triple chains or "clusters" (subgraphs with a central vertex and three adjacent edges) are in the fragment when $K = 4$. The subgraphs implicitly differ in their weighted coefficients b_j.[3] Let us note in this context that for our approach the coefficients b_j are not rigidly assigned to the fragments, but they are established for every new property by the least-squares method.

Despite their differences, these methods have one commom feature: sets of descriptors specified beforehand are used for describing molecular structures. These sets can be amplified, if it is found in the course of the QSPR calculation that they are not sufficiently adequate. As a rule, the list of descriptors used in calculations of one of the properties, for example, boiling point, is not very suitable for calculations of another property, for example, lipophilicity.

The method of descriptor formation

We suggest another approach to the solution of the QSPR-problem based on the automatic formation of structural descriptors and their weighted coefficients for a set of studied substances (a learning set) according to specified rules. Lists of descriptors specified beforehand are not used. It is important that the obtained QSPR are prognostic and are characterized by an informative interpretation of the parameters of the molecules involved in the sample.

Translated from *Izvestiya Akademii Nauk. Seriya Khimicheskaya,* No. 8, pp. 1391—1393, August, 1994.

This method of building the QSPR logically develops a method suggested by one of the authors.[1] A chain, *i.e.*, an open sequence of connected atoms, is taken as the basis for building an index. The use of chains as structural fragments for calculating indices has been previously justified.[1] Markers and symbols are introduced to take into consideration specific features of the structure of organic molecules. Markers allow one to define the atoms included in symbols in the following way: "ATOM'S_SYMBOL" = "ATOM'S_NAME", "MARKER_1",...,"MARKER_p".

There are presently three reference markers *d*, *b*, and *r*, in the BIBIGON MATCH system for the primary classification of atoms. Marker *d* specifies the number of adjacent atoms (except H) and corresponds in the graph theory sense to the degree of the vertex of a MC-graph with a "rubbed" H atom. Marker *b* enumerates the types of chemical bonds for a given atom using the designation *s* for the case when all of the bonds are single, *d* for the case when there is one double bond, *t* for the existence of a triple bond, *a* for the existence of an aromatic bond, and *w* for two double bonds. Marker *r* characterizes atoms in cyclic compounds as *c* for an acyclic atom, *r* for an intracyclic atom, and *s* for an intracyclic atom with an attached substituent. The symbol of an atom is "ATOM'S_SYMBOL" = "ATOM'S_NAME", "*d*", "*b*", "*r*".

The system makes it possible to design new markers and introduce them to the BIBIGON MATCH software complex. Let some numerical function $h(v_i)$ be specified for atoms v_i in a molecule, such that the new marker can accept, for example, three symbolic values: *a*, *b*, and *c*. Then let us divide the domain of existence for the function $h(v_i)$ into three nonoverlapping intervals (H1, H2, H3) and define the H-marker as "*a*", "*b*", and "*c*", taking into account that $h(v_i)$ belongs to the interval H1, H2, and H3, respectively. The division of the region into intervals H1, H2, and H3 must be performed by the chemist-investigator, taking into consideration the information content of the interpretation and the character of the studied property of the chemical compounds.

It is not necessary to include the name of the element in the symbol of the atom: it may be reflected by the marker itself, which is called in this case the "coloring" of the atom. Then the symbols of atoms have the form: "ATOM'S_COLORING" = "MARKER_1",..., "MARKER_p".[5] If it is convenient to present a certain functional group of atoms as a whole for the description of some property, this fragment may be designated by a special symbol and then a preliminary processing of molecular graphs can be performed, converting their corresponding components to one vertex of a new graph. The latter may be called a "superatom", and a special designation may be introduced for it.

After dividing the MC-graphs of organic compounds into chains of marked atoms, a description matrix X with elements X_{ij}, which reflect the number of repetitions of the j-th chain in the i-th molecule, is formed. The j column of the X matrix corresponds to the specified j-th descriptor (a chain of marked atoms). When a new H-marker is included in the formation of symbols, another collection of properties of the molecules of the learning set and another X matrix are created.

The detailed classification of atoms with the simultaneous use of several markers results in the fact that the number of columns of the description matrix is many times the number of its rows (dimension of the learning set). This fact makes it almost impossible to use all of the found descriptors. The method of the self-organization of models[6] in combination with special algorithms and programs included in the BIBIGON MATCH system allows one to find informative descriptors,[7] which are called basic descriptors, and their number usually does not exceed 1/5 of the dimension of the learning set.[8] A physicochemical value is calculated from the linear equation

$$Y = b_0 + \sum_j b_j \cdot F(\bar{x}_j) + \bar{\varepsilon}_i \qquad (1)$$

where b_j is a parameter and $F(x_j)$ is a function chosen out of the function list for each descriptor by the program. This list may be changed or supplemented. The use of functions of descriptors helps in several cases to increase the calculation accuracy, because the nonlinear dependences of the contributions of some of the functional groups in a molecule are taken into account.

The prognostication of the obtained results was checked by the "cross validation" method:[7] 1) the i-th compound was removed from the learning set and new weighted coefficients were found for a given set of chains; 2) a property of the removed i-th structure was predicted in terms of the obtained linear model, and the prediction error was remembered. This procedure was repeated for each compound of the studied set, resulting in the formation of a "cross" error vector necessary for the calculation of the multiple correlation coefficient (R), the standard deviation (S), and the Fischer criterion (F), *i.e.*, "prognostication parameters".

Results and Discussion

Some properties of various classes of organic compounds were calculated by formula (1). A reference sample containing 172 compounds was taken from Ref. 9 for describing the lipophilicity of substituted benzenes. $R = 0.976$, $S = 0.194$, and $F = 393.53$ for 16 descriptors and $R = 0.968$, $S = 0.272$, and $F = 185.95$ for the "cross validation" method were obtained. by the BIBIGON MATCH program. These values were compared[10] in detail with the similar data of the commercial Med-Chem program and the data of Ref. 9.

A series of 201 compounds[11] was used for the calculating enthalpies of formation for different chemical classes. This resulted in the determination of the following values of the parameters: $R = 0.991$, $S = 0.102$, and

$F = 312.60$ for the linear regression method (16 descriptors) and $R = 0.975$, $S = 0.231$, and $F = 224.67$ for the "moving control" method. In Ref. 11 the enthalpy of formation was calculated using 20 structural descriptors, and the series was divided into three classes. The standard deviation (S) of the experimental values from the calculated values did not exceed 0.5.

The applicability of the described method for the analysis of biological activity was checked for a reference sample containing 54 compounds,[12] and the parameters $R = 0.937$, $S = 0.384$, and $F = 47.38$ were obtained for 7 descriptors, while $R = 0.915$, $S = 0.445$, and $F = 33.61$ were obtained for the "moving control" method. According to Ref. 12, $R = 0.86$ corresponds to the estimates of coccidic activity, whereas the authors of Ref. 13 obtained in this case $R = 0.900$, $S = 0.480$, and $F = 28$, using the same series and a prognostic equation with 7 parameters.

Polarizabilities of molecules were also calculated using a reference sample that included 293 compounds[14] and the BIBIGON MATCH program. $R = 0.982$, $S = 6.308$, and $F = 873.05$ and $R = 0.981$, $S = 6.593$, and $F = 796.67$ were found for 8 descriptors and for the "cross validation" method, respectively. The calculations of chromatographic retention times for anthracyclinic antibiotics of the downorubicine series have been performed for a series containing 84 compounds.[15] Equation (1) contained 25 descriptors, which resulted in $R = 0.981$, $S = 0.094$, and $F = 64.12$. In the case of the "cross validation" method $R = 0.980$, $S = 0.047$, and $F = 58.32$.

A series containing 323 compounds was used for the calculation of boiling points of furans, tetrahydrofurans, and thiophenes.[16] $R = 0.921$, $S = 18.126$, and $F = 194.12$, and $R = 0.904$, $S = 19.246$, and $F = 147.49$ were obtained for 16 descriptors and for the "cross validation" method, respectively.

Thus, the estimation of different properties of organic compounds by the method suggested in several cases turned out to be more exact than those published in the literature. As shown, the design of new markers allows the easy approximation and prediction of physicochemical properties that presently require the use of labor consuming methods or can be hardly estimated at all, for example, the biological activity of organic compounds.

The authors are grateful to Prof. V. A. Shlyapochnikov for his helpful discussions and critical remarks.

This work was partially supported by the Russian Foundation for Basic Research (Projects 94-03-09323, 94-01-00041).

References

1. E. A. Smolenskii, *Zh. Fiz. Khim.*, 1964, **38**, 1288 [*J. Phys. Chem. USSR*, 1964, **38** (Engl. Transl.)].
2. A. J. Stuper, W. E. Brugger, and P. C. Jurs, *Computer Assisted Studies of Chemical Structure and Biological Function,* John Wiley & Sons Inc., New York, 1979.
3. D. Rouvray, in *Chemical Applications of Topology and Graph Theory*, Ed. by R. B. King, Elsevier, New York, 1983.
4. L. B. Kier and L. H. Hall, *Molecular Connectivity in Structure Activity Analysis*, Wiley, London, 1986, 300.
5. M. I. Kumskov, L. A. Ponomareva, and N. S. Zefirov, *Materialy 4 Vsesoyuznoi shkoly-seminara "Statisticheskii i diskretnyi analiz dannykh i ekspertnye otsenki"* [*Proceedings of the 4th All-Union School-Seminar "Statistical and Discrete Analysis of Data and Expert Estimations"*], *Odessa*, 1991, 90 (in Russian).
6. N. S. Zefirov, M. I. Kumskov, L. A. Ponomareva, D. F. Mityushev, and E. A. Smolenskii, *Tez.. dokl. IX Vsesoyuz. konf. "Khimicheskaya informatika"* [*Proceedings of the IX All-Union Conf. "Chemical informatics"*, Chernogolovka, 1992, 159 (in Russian).
7. M. A. Sharaf, D. A. Ilamen, and B. R. Koval'skii, *Khemometrika* [*Chemometrics*], Khimiya, Leningrad, 1989, 269 (in Russian).
8. J. G. Topliss and R. P. Edwards, *Med. Chem.*, 1979, **22**, 1238.
9. P. Camilleri, S. A. Watts, and J. A. Boraston, *J. Chem. Soc. Perkin Trans. II*, 1988, 1699.
10. E. A. Smolenskii, L. A. Ponomareva, and N. S. Zefirov, *Dokl. Akad. Nauk SSSR*, 1990, **312**, 155—159 [*Dokl. Chem.*, 1990, **312** (Engl. Transl.].
11. F. DeLos and DeTar, *J. Org. Chem.*, 1991, **56**, 1478—1481.
12. J. W. McFarland, C. B. Cooper, and D. M. Newcomb, *J. Med. Chem.*, 1991, **34**, 1908.
13. N. S. Zefirov, D. E. Petelin, V. A. Palyulin, and J. W. McFarland, *Dokl. Akad. Nauk*, 1992, **327**, 504 [*Dokl. Chem.*, 1992, **327** (Engl. Transl.].
14. K. J. Miller, *J. Am. Chem. Soc.*, 1990, **112**, 8533.
15. L. A. Ponomareva, E. N. Olsuf'eva, M. N. Preobrazhenskaya, M. I. Kumskov, and N. S. Zefirov, *Khim. Farm. Zh.*, 1993, 36—40 [*Chem. Pharm. J.*, 1993 (Engl. Transl.)].
16. D. T. Stanton, P. C. Jurs, and M. G. Hicks, *J. Am. Chem. Soc.*, 1991, **31**, 301.

Received February 21, 1994