ISSN 1070-3632, Russian Journal of General Chemistry, 2011, Vol. 81, No. 3, pp. 449–455. © Pleiades Publishing, Ltd., 2011.
Original Russian Text © V.A. Grigor'ev, O.A. Raevskii, 2011, published in Zhurnal Obshchei Khimii, 2011, Vol. 81, No. 3, pp. 353–359.

Fractal Dimension of the Interatomic Distance Histogram: New 3D Descriptor of Molecular Structure

V. A. Grigor'ev and O. A. Raevskii

Institute of Physiologically Active Substances, Russian Academy of Sciences, Severnyi pr. 1, Chernogolovka, Moscow Oblast, 142432, Russia
e-mail: beng@ipac.ac.ru

Received January 14, 2010

Abstract—A method is developed for the calculation of a new 3D descriptor of molecular structure, which is the fractal dimension of the histogram of interatomic distances serving as a measure of the structure complexity in the geometric aspect. The new descriptor was found to depend mainly on three factors: the number of atoms in a molecule, the number of bars in the histogram, and the point symmetry group. The fractal dimension of the histogram of interatomic distances was found to be of a very specific character.

DOI: 10.1134/S1070363211030017

In recent decades the fractal geometry became widespread in the study of various objects and processes [1, 2]. The concept of fractals and the related concept of fractal dimension are now successfully used in various fields of science [3–7]. In particular, when studying the properties of nano-objects, we study the relationship between the fractal dimension and the material properties [8, 9]. The existence of such links is obviously due to the presence of fractal properties in the nano-sized molecular clusters. However, it is necessary to note the lack of similar work in the field of relationships *structure–property* and *structure–activity*, at least with respect to usual small molecules. This is probably due to the fact that such molecules do not possess fractal properties. The same was noted in one of the few publications [10] concerning the use of the fractal ideas for exploring the *structure–activity* relationship. It is noteworthy that the lack of invariance at varying the scale (self-similarity) of geometric structures of molecules expressed, for example, as the atomic Cartesian coordinates, does not exclude the presence of self-similarity at other ways of representation of a spatial structure.

The purpose of this study is estimating the fractal dimension of the geometric structure of molecules expressed through a histogram of interatomic distances.

Molecular structures were calculated using the semi-empirical quantum-chemical method AM1 implemented in the HyperChem software package [11]. As the starting molecules for the full optimization were taken the three-dimensional structures generated by the program derived from the corresponding two-dimensional structures. Point symmetry group of molecules were determines using the information systems KnowItAll [12] and HyperChem [11]. For each compound using a computer software DRAGON 3224 descriptors were calculated [13]. The values of the alkanes density were taken from [14].

Histograms (the discrete distribution function) of the interatomic distances were represented as planar graphs consisting of a set of shaded rectangles (bars). The abscissa axis was divided with a certain step (resolution) ΔR into intervals, whose number depended on the values of the interatomic distances R varying from a minimum value R_{min} to a maximum value R_{max}. The ordinate (vertical) axis corresponded to the frequency f, that is, to the number of interatomic distances that fall within the specified interval. This is an integer number varying from 0 to f_{max}. This axis was also divided into intervals with a certain step Δf. The histograms obtained were digitized through representing them as a square binary matrix with 1 corresponding to the shaded parts of a histogram, and 0 to the rest.

The fractal dimension was determined by the *cell* method using a computer program whose algorithm is given in [15], by covering the digitized image with the cells of variable size and counting the number of cells

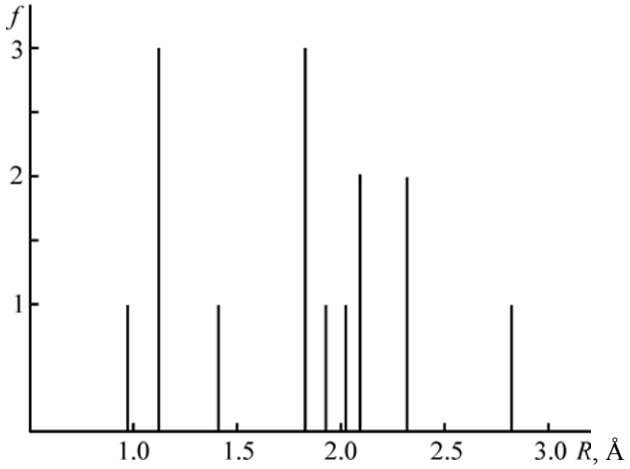

Fig. 1. Histogram of distances in a molecule of methanol calculated with 0.01 Å resolution.

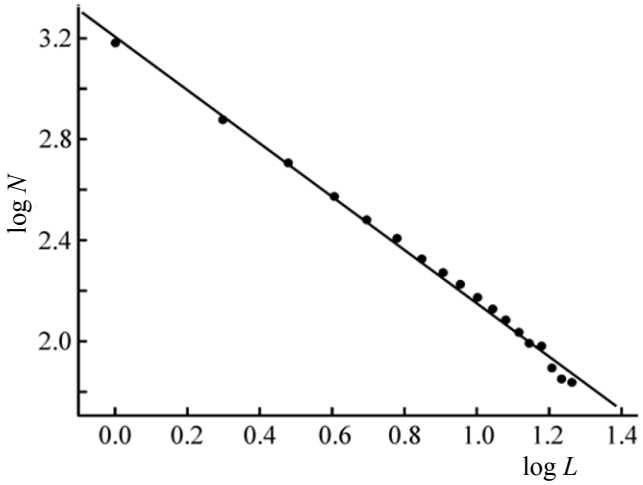

Fig. 2. Dependence of log N vs. log L [Eq. (1)] for the methanol molecule.

that intersected the shaded areas of the histograms. The local fractal dimension was calculated using the equation:

$$\log N(L) = \log C - D \log L, \quad (1)$$

where $N(L)$ is the minimum number of cells with the side L, necessary to cover a fractal, D is the *cellular* fractal dimension (with a minus sign), C is a constant.

To determine the value of D with Eq. (1), it is necessary to carry out two measurements at two different values of L. However, to estimate the fractal dimension more reliably, the number of measurements should be large. We used a set of cells with dimensions of sides L = 1, 2, ..., L_{max}. For each molecule the value of L_{max} was determined by dividing the less of two values, $(R_{max} - R_{min})/\Delta R$ and $f_{max}/\Delta f$, by 10.

When calculating the D the fact should be taken into account that the histograms have different coefficients of similarity along different axes, since the dimensions of the abscissa and ordinate have different physical meaning [2]. Therefore, for comparative fractal analysis using the molecular *cell* method, it is necessary to fix the standard model with an initial minimum cell, whose dimensions must be small compared to the scale of the histogram on both axes. Such a minimal cell corresponds to a value of L = 1 in Eq. (1). Since the minimum value of the interatomic distance in the molecule is about 1 Å, and the accuracy of the interatomic distances obtained by semiempirical quantum-chemical methods is 0.01 Å [16], this quantity was chosen as the minimum possible cell size ΔR along abscissa. Due to the fact that the minimum non-zero value of a bar on the ordinate is 1, then as a reasonable cell size Δf along the ordinate was taken 0.01. The size of digitized square fractal matrix is defined based on a maximum of two values: $R_{max}/0.01$ and $f_{max}/0.01$.

Regression analysis was performed using a computer program SVD [17] modified by the authors. As the statistical characteristics of the equations were taken the following: n, the number of points; r^2, the squared linear correlation coefficient; s, the standard deviation; F, the Fisher criterion, q^2, the squared linear correlation coefficient at the sliding mode control one by one.

As an example, Figs. 1 and 2 show the results of calculating the fractal dimension of the histogram of interatomic distances for the molecule of methanol. The molecule contains 6 atoms, so the total number of interatomic distances is equal to 6×(6–1)/2 = 15, which corresponds to the sum of all bars in Fig. 1. To digitize the histogram a fractal matrix is required of 300%× 300 pixels (3/0.01 = 300). As a result of operation of the *cellular* algorithm the data were obtained shown in Fig. 2. They are described by a straight line with good statistical indicators. The obtained value of fractal dimension D = 1.06.

Table 1 shows the calculation results for 73 organic molecules belonging to 6 different classes: alkanes, alcohols, esters, ketones, amines, aromatics. The number of atoms N is varied from 5 to 50, and the maximum length of the interatomic distances R_{max} varies from 1.82 Å to 20.57 Å. The minimum value of the fractal dimension D is 1.00 and the maximum is

Table 1. Point symmetry group (S), number of atoms (N_a), number of bars in the histogram of interatomic distances (N_b), fractal dimension (D) of molecules, and the statistical characteristics of the regression equations

Molecule	S	N_a	N_b	D	n	r^2	s	F	q^2
Methane	T_d	5	2	1.00(±0.01)	7	0.999	0.01	358483	0.999
Ethane	D_{3d}	8	6	1.00(±0.01)	19	0.999	0.01	307988	0.999
Propane	C_{2v}	11	16	1.08(±0.01)	32	0.998	0.02	17564	0.998
Butane	C_{2h}	14	23	1.13(±0.01)	44	0.997	0.02	16044	0.997
Pentane	C_{2v}	17	32	1.20(±0.01)	57	0.992	0.04	6951	0.991
Hexane	C_{2h}	20	43	1.23(±0.01)	69	0.992	0.04	7938	0.991
Heptane	C_{2v}	23	54	1.30(±0.01)	82	0.991	0.05	9326	0.990
Octane	C_{2h}	26	64	1.31(±0.01)	94	0.993	0.05	12260	0.992
Ninane	C_{2v}	29	76	1.33(±0.01)	107	0.991	0.05	11748	0.990
Decane	C_{2h}	32	93	1.36(±0.01)	119	0.992	0.05	15083	0.992
Undecane	C_{2v}	35	96	1.36(±0.01)	132	0.992	0.05	16576	0.991
Dodecane	C_{2h}	38	106	1.37(±0.01)	144	0.993	0.05	19326	0.992
Tridecane	C_{2v}	41	116	1.37(±0.01)	157	0.993	0.05	22858	0.993
Tetradecane	C_{2h}	44	132	1.37(±0.01)	169	0.993	0.05	23095	0.992
Pentadecane	C_{2v}	47	139	1.38(±0.01)	182	0.992	0.05	22773	0.992
Hexadecane	C_{2h}	50	153	1.39(±0.01)	194	0.992	0.05	23793	0.991
Methanol	C_s	6	10	1.06(±0.02)	18	0.995	0.03	3269	0.993
Ethanol	C_s	9	23	1.22(±0.02)	30	0.993	0.04	3777	0.991
1-Propanol	C_s	12	33	1.23(±0.01)	43	0.996	0.03	10460	0.995
2-Propanol	C_1	12	39	1.21(±0.01)	33	0.997	0.02	12298	0.994
1-Butanol	C_s	15	49	1.25(±0.01)	55	0.995	0.03	10609	0.994
2-Methyl-1-propanol	C_1	15	68	1.33(±0.01)	43	0.998	0.02	20631	0.998
2-Butanol	C_1	15	71	1.38(±0.01)	46	0.996	0.03	10823	0.995
2-Methyl-2-propanol	C_s	15	36	1.20(±0.01)	33	0.999	0.01	32528	0.999
1-Pentanol	C_s	18	60	1.27(±0.01)	68	0.995	0.04	12865	0.994
1-Hexanol	C_s	21	80	1.32(±0.01)	80	0.993	0.05	10518	0.992
1-Heptanol	C_s	24	98	1.36(±0.01)	93	0.992	0.05	11146	0.991
1-Octanol	C_s	27	112	1.38(±0.01)	105	0.993	0.05	13949	0.992
Dimethyl ether	C_{2v}	9	13	1.03(±0.01)	29	0.998	0.01	17938	0.998
Diethyl ether	C_{2v}	15	34	1.26(±0.02)	55	0.993	0.04	7545	0.992
Butyl methyl ether	C_s	18	70	1.38(±0.01)	67	0.994	0.04	10353	0.993
sec-Butyl methyl ether	C_1	18	105	1.42(±0.01)	48	0.998	0.03	21181	0.997
tert-Butyl methyl ether	C_s	18	65	1.36(±0.01)	42	0.998	0.03	14145	0.997
Dipropyl ether	C_{2v}	21	59	1.36(±0.01)	79	0.992	0.05	9060	0.990
Diisopropyl ether	C_2	21	83	1.41(±0.01)	55	0.995	0.04	11518	0.994
tert-Pentyl ethyl ether	C_1	21	138	1.44(±0.01)	47	0.998	0.02	25553	0.998
Butyl ethyl ether	C_s	21	86	1.40(±0.01)	80	0.995	0.04	16483	0.995
Dibutyl ether	C_{2v}	27	86	1.41(±0.01)	104	0.994	0.05	16003	0.993
Dipentyl ether	C_{2v}	33	125	1.45(±0.01)	129	0.992	0.05	15857	0.991

Table 1. (Contd.)

Molecule	S	N_a	N_b	D	n	r^2	s	F	q^2
Diisopentyl ether	C_2	33	202	1.52(±0.01)	104	0.997	0.03	33317	0.996
Acetone	C_s	10	20	1.12(±0.01)	30	0.997	0.02	10729	0.997
2-Butanone	C_s	13	40	1.24(±0.01)	42	0.998	0.02	18576	0.998
3-Methyl-2-butanone	C_1	16	78	1.38(±0.01)	43	0.997	0.03	14187	0.996
2-Pentanone	C_s	16	58	1.28(±0.01)	57	0.995	0.04	11109	0.995
3-Pentanone	C_s	16	35	1.25(±0.01)	57	0.993	0.04	7646	0.992
2-Hexanone	C_s	19	67	1.31(±0.01)	69	0.995	0.04	14486	0.995
3-Hexanone	C_s	19	67	1.31(±0.01)	69	0.994	0.04	10935	0.993
2-Methyl-3-pentanone	C_1	19	110	1.38(±0.01)	57	0.996	0.03	15628	0.996
4-Methyl-2-pentanone	C_1	19	116	1.41(±0.01)	57	0.998	0.03	23218	0.997
Methylamine	C_s	7	12	1.14(±0.03)	19	0.990	0.04	1648	0.986
Dimethylamine	C_s	10	22	1.15(±0.01)	31	0.996	0.03	6742	0.994
Ethylamine	C_1	10	35	1.30(±0.01)	31	0.997	0.03	9338	0.995
Ethylmethylamine	C_1	13	56	1.37(±0.01)	44	0.996	0.03	10269	0.995
Isopropylamine	C_1	13	53	1.32(±0.01)	33	0.997	0.03	10903	0.996
Propylamine	C_1	13	55	1.35(±0.01)	40	0.996	0.03	9001	0.995
Trimethylamine	C_{3v}	13	15	1.08(±0.01)	30	0.996	0.03	6699	0.995
Butylamine	C_1	16	76	1.39(±0.01)	56	0.994	0.04	9355	0.993
sec-Butylamine	C_1	16	80	1.42(±0.01)	46	0.998	0.03	20548	0.997
tert-Butylamine	C_s	16	43	1.27±0.01)	33	0.998	0.02	14740	0.998
Diethylamine	C_s	16	54	1.39(±0.02)	57	0.993	0.05	7333	0.991
Dimethylethylamine	C_1	16	81	1.36(±0.01)	43	0.997	0.03	15907	0.997
Isobutylamine	C_1	16	80	1.37(±0.01)	44	0.998	0.02	21732	0.998
Methylisopropylamine	C_1	16	87	1.38(±0.01)	44	0.998	0.02	26579	0.998
N-Methylpropylamine	C_1	16	78	1.36(±0.01)	57	0.996	0.03	14017	0.996
Benzene	D_{6h}	12	10	1.07(±0.01)	38	0.995	0.03	6572	0.994
Naphthalene	D_{2h}	18	39	1.31(±0.01)	61	0.993	0.04	8584	0.992
Anthracene	D_{2h}	24	66	1.41(±0.01)	84	0.992	0.05	10382	0.991
Phenanthrene	C_{2v}	24	113	1.48(±0.01)	81	0.996	0.04	18046	0.995
2,3-Benzanthracene	D_{2h}	30	99	1.48(±0.01)	108	0.993	0.05	14368	0.992
1,2-Benzanthracene	C_s	30	252	1.52(±0.01)	104	0.997	0.03	39943	0.997
Pentacene	D_{2h}	36	133	1.50(±0.01)	132	0.994	0.05	22053	0.993
1,2:3,4-Dibenzanthracene	C_{2v}	36	199	1.51(±0.01)	103	0.997	0.04	29468	0.996
1,2:5,6-Dibenzanthracene	C_{2h}	36	235	1.58(±0.01)	124	0.996	0.04	30875	0.996

1.58. For comparison: the measurement of fractal dimension carried out as described above gives for a segment of straight line the value $D = 1$, and for the square $D = 2$, corresponding to their topological dimension.

The data presented in Fig. 3 indicate the existence of trend in the growth of the fractal dimension with the number of atoms in the molecule. However, there are significant deviations from this trend, indicating that

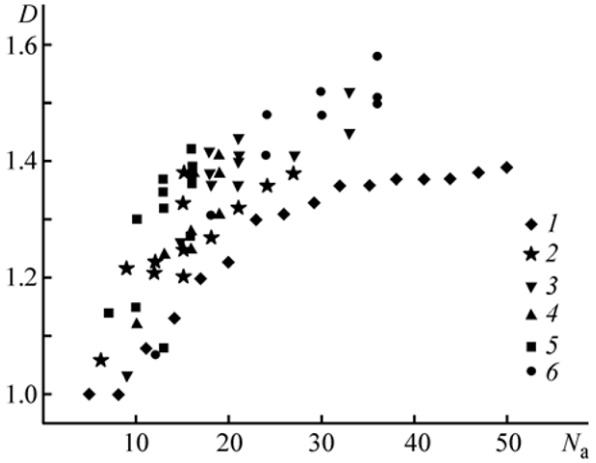

Fig. 3. Dependence of the fractal dimension on the number of atoms in different molecules: (*1*) alkanes, (*2*) alcohols, (*3*) ethers, (*4*) ketones, (*5*) amines, and (*6*) aromatic compounds.

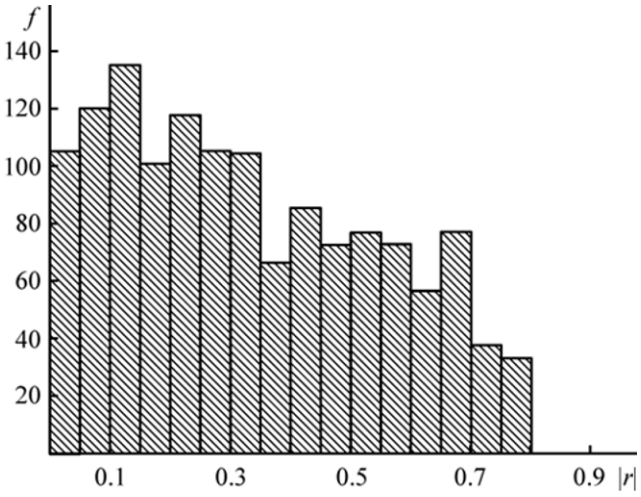

Fig. 4. Histogram of distribution of the correlation coefficients of the fractal dimension with 1363 descriptors.

the number of atoms is not the unique factor. For example, each of the four isomeric alcohols with the general molecular formula $C_4H_{10}O$, 1-butanol, 2-methyl-1-propanol, 2-butanol, and 2-methyl-2-propanol have 15 atoms, but they differ in the fractal dimension: the D value varies from 1.20 to 1.38.

The observed phenomenon could be understood if we take into account the point group symmetry of molecules. Indeed, compounds 1-butanol and 2-methyl-1-propanol belong to the group C_s and have D = 1.25 and 1.20, respectively. In contrast, compounds 2-methyl-1-propanol and 2-butanol belong to the group C_1, and their fractal dimensions are 1.33 and 1.38, respectively. The quantitative measure corresponding to the number of atoms in a molecule and its belonging to a particular point group symmetry, to a first approximation, can be the number of bars in the histograms of interatomic distances N_b. For 1-butanol N_b = 49, for 2-methyl-2-propanol 36, for 2-methyl-1-propanol 68, and for 2-butanol 71. Obviously, the larger the number of symmetry elements of a molecule, the greater the number of coincident interatomic distances and the less different bars contains the histograms. However, further study showed that a better quantitative measure for the estimation of the influence of the number of atoms and point group symmetry in the molecule is the normalized value obtained by dividing the number of bars on the number of atoms: N_b/N_a.

The regression analysis of the dependencies for individual point symmetry groups yielded the following results:

$$C_{2v}:\ D = 0.97(\pm0.03) + 0.81(\pm0.06)\log (N_b/N_a), \quad (2)$$
$$n\ 14,\ r^2\ 0.935,\ s\ 0.04,\ F\ 172.4,\ q^2\ 0.889,$$

$$C_s:\ D = 0.91(\pm0.03) + 0.70(\pm0.06)\log (N_b/N_a), \quad (3)$$
$$n\ 23,\ r^2\ 0.851,\ s\ 0.04,\ F\ 119.8,\ q^2\ 0.810,$$

$$C_1:\ D = 0.91(\pm0.06) + 0.63(\pm0.08)\log (N_b/N_a),$$
$$(4) n\ 18,\ r^2\ 0.785,\ s\ 0.03,\ F\ 58.5,\ q^2\ 0.682.$$

Equations (2)–(4) indicate the existence of a correlation between the fractal dimension and the normalized number of bars. This relationship becomes clear if we consider the physical meaning of N_b/N_a. In essence, it represents the number of bars per a single atom, i.e., characterizes the density of filling the histogram space. Therefore, the fractal dimension as a quantitative measure of the degree of irregularity or complexity of the spatial structure of molecule (in the limit of the extreme irregularity it is a completely filled space), is in a statistical relationship with the N_b/N_a value.

To determine the significance of the fractal dimension histograms of the interatomic distances among other descriptors, we analyzed the absolute values of paired correlation coefficients of the 3224 linear descriptors with the value of D for 73 compounds studied. From the set of descriptors the

Table 2. The fractal dimension (D) and density (d_4^{20}) in the homologous series of alkanes

Alkane	D	d_4^{20}, g cm^{-3}
Pentane	1.20	0.6262
Hexane	1.23	0.6603
Heptane	1.30	0.6838
Octane	1.31	0.7025
Nonane	1.33	0.7176
Decane	1.36	0.7300
Undecane	1.36	0.7402
Dodecane	1.37	0.7487
Tridecane	1.37	0.7564
Tetradecane	1.37	0.7628
Pentadecane	1.38	0.7685
Hexadecane	1.39	0.7730
Heptadecane	1.39	0.7780
Octadecane	1.41	0.7820
Nonadecane	1.42	0.7855
Eicosane	1.42	0.7886
Pentacosane	1.46	0.8012
Triacontane	1.47	0.8097

uninformative variables, which contained over 95% of constant values, were removed. Such screening left 1363 descriptors for further analysis. Figure 4 shows the histogram of correlation coefficients calculated with the increment 0.05. In this case, the minimum value of $|r|$ is equal to 0.002, the maximum is 0.859, and the average value is 0.332.

In [18] a classification was suggested of molecular descriptors depending on their correlation coefficients $|r|$ with respect to the gas chromatographic retention time. In accordance with these data, the descriptors with $|r| \geq 0.99$ are fundamental, $0.99 > |r| \geq 0.80$ are important, $0.80 > |r| \geq 0.50$, are probable, and $0.50 > |r|$ are specific. This classification can be used not only for estimation of the descriptor performance in relation to a property, but also of a given property, in particular, the fractal dimension, in the relation to the descriptors. Indeed, from the data in Fig. 4 follows that 1010 of 1363 descriptors have a value of $|r| < 0.5$, 351 are in the range $0.80 > |r| \geq 0.50$ and 2 descriptors are in the range $0.99 > |r| \geq 0.80$. It is interesting to note that the last two descriptors are topological indices: the structural information content of 1st and 2nd orders. The fraction of *specific* in the fractal dimensions with respect to the 1363 dscriptors is 100%×1010/1363 = 74.1%, the fraction of *probable* is 100%×351/1363 = 25.8%, while the remaining fraction of *important* is 0.1%. That is, the fractal dimension can be regarded as a very *specific* descriptor.

The usefulness of the new proposed descriptor can be demonstrated by establishing the quantitative relations between the fractal dimension and the density in a homologous series of alkanes (pentane, hexane, eicosane ..., pentacosane, triacontane), which under normal conditions are liquids or solids (Table 2).

$$d_4^{20}\ (\text{g cm}^{-3}) = -0.21(\pm 0.05) + 0.70(\pm 0.03)D, \qquad (5)$$
$$n\ 18,\ r^2\ 0.965,\ s\ 0.01,\ F\ 441.6,\ q^2\ 0.956.$$

The resulting regression equation indicates the existence of a close correlation between the micro- and macromolecular characteristics. In the studied series of compounds an increase in the complexity of the molecular structure is accompanied by an increase in density. Thus, the fractal dimension can be recommended for the use as a descriptor for the study of the *structure–property* and *structure–activity* relations.

In conclusion, it should be noted that the proposed descriptor is fully consistent with the basic requirements to the new descriptors [19]. It is invariant with respect to the marking and numbering of atoms in a molecule. The value of D does not depend on the molecular rotation and translation, as in the basis of its calculation lies an invariant: the interatomic distance. The above calculation of the fractal dimension of the histogram of interatomic distances is not ambiguous. And finally, the range of D from 1 to 2 is convenient from the practical point of view.

REFERENCES

1. Mandelbrot, B., *The Fractal Geometry of Nature*, Moscow: Institute of Computer Sciences, 2002.
2. Feder, E., *Fractals*, Moscow: Mir, 1991.
3. Risovic, D., Mahovic Poljacek, S., and Gojo, M., *Applied Surface Science*, 2009, vol. 255, no. 7, pp. 4283–4288.
4. Eftekhari, A., *Physica, B: Condensed Matter*, 2007, vol. 387, nos. 1–2, pp. 92–97.
5. Ghanbari, Kh., Mousavi, M.F., Shamsipur, M., Rahmanifar, M.S., and Heli, H., *Synthetic Metals*, 2006, vol. 156, nos. 14–15, pp. 911–916.
6. Koňas, P., Buchar, J., and Severa, L., *European Journal of Mechanics, A, Solids*, 2009, vol. 28, no. 3, pp. 545–550.
7. Chang-Yong, Lee, *Physical Review, E,* 2006, vol. 73, p. 042901.
8. Mihranyan, A. and Strømme, M., *Surface Science*, 2007, vol. 601, pp. 315–319.

9. Ono, Y., Mayama, H., Furó, I., Sagidullin, A.I., Matsushima, K., Ura, H., Uchiyama, T., and Tsujii, K., *J. Colloid and Interface Science*, 2009, vol. 336, no. 1, pp. 215–225.
10. Artemenko, A.G., Kovdienko, N.A., Kuz'min, V.E., Kamalov, G.L., Lozitskaya, R.N., Fedchuk, A.S., Lozitsky, V.P., Dyachenko, N.S., and Nosach, L.N., *Experimental Oncology*, 2002, vol. 24, pp. 123–127.
11. *http://www.hyper.com/*.
12. *http://www.knowitall.com/*.
13. *http://www.talete.mi.it/*.
14. *Svoistva organicheskikh soedinenii. Spravochnik* (Properties of Organic Compounds. Handbook), Potekhin, A.A., Ed., Leningrad: Khimiya, 1984.
15. Kronover, R.M., *Fraktaly i khaos v dinamicheskikh sistemakh. Osnovy teorii* (Fractals and Chaos in Dynamic Systems. Theoretical Fundamentals), Moscow: Postmarket, 2000, p. 137.
16. Blatov, V.A., Shevchenko, A.P., and Peresypkina, E.V., *Poluempiricheskie raschetnye metody kvantovoi khimii: Uchebnoe posobie* (Semiempirocal Calculation Methods of Quantum Chemistry. A Tutorial), Samara: Univers-Grupp, 2005.
17. Forsythe, G.E., Malcolm, M.A., and Moler, C.B., *Computer Methods for Mathematical Computations*, Moscow: Mir, 1980.
18. Dimov, N., Osman, A., Mekenyan, O., and Papazova, D., *Analytica Chimica Acta*, 1994, vol. 298, no. 3, pp. 303–317.
19. *http://www.moleculardescriptors.eu/*.