

Victor E. Kuz'min · Ludmila N. Ognichenko
Anatoly G. Artemenko

Modeling of the informational field of molecules

Received: 4 January 2001 / Accepted: 21 May 2001 / Published online: 25 July 2001
© Springer-Verlag 2001

Abstract The possibility of model operation of an informational field of an entity (molecule), which can be represented as a discrete parameter system of elements (atoms), is shown. The idea of the approach is based on the application of the Shannon's method (quantitative estimation of the information) not only to the molecular structure, but also to the surrounding space. The performance of an information field of a molecule can be utilized for the solution of the "structure–property" tasks.

Keywords Information of molecular structure · Molecular informational field · Lattice models · QSAR

Introduction

As is known from Latin, the term "information" means "explanation", "statement". According to Claude Shannon the more formal definition of information and its quantitative estimation is the following: "The degree of uncertainty of a situation can be defined by the number of possible variants of development of this situation". [1] Hence, an objective basis of the information is "distinction" or "variety", which agrees with the information concepts of Ashby and Glushkov. [2] The logarithm of the possible number of outcomes (ways out) is used to estimate the quantity of the information according to Hartley's hypothesis. [3] The minimal uncertainty of a situation corresponds to two outcomes (for example, throwing of a coin – "heads" or "tails"). If the binary logarithm is used, the elimination of uncertainty of this situation equals 1 ($I = \log_2 2 = 1$), which corresponds to the

informational quantity in one "bit". Different probabilities (P_i) occur in the case of various outcomes (i) of development of a situation. Hence, it is more correct to estimate the quantity of the information by Shannon's formula. [1]

Shannon's approach is based on discrete representation of the "object" (situation, system), consisting of a set M of "elements" (outcomes). Let the power of this set equal n ($|M|=n$). If the equivalence relation is introduced, all elements of this set form subsets (m_j). Thus, each of subset contains only the following equivalent elements:

$$\begin{aligned}m_1 \cup m_2 \cup \dots \cup m_i \dots \cup m_k &= M, k \leq n \\m_j \cap m_i &= 0 \\|m_i| &= n_i \quad \sum_i n_i = n\end{aligned}$$

The probability of a choice of any element from the i th subset is $p_i = n_i/n$.

According to Shannon, the quantity of the information of one element of the system (object) is defined as:

$$I = - \sum_i p_i \lg p_i$$

where $\lg \equiv \log_2$. For the above-mentioned situation of throwing of a coin 1 bit of the information is:

$$I = -1/2 \lg(1/2) - 1/2 \lg(1/2) = 1$$

This agrees with Hartley's approach.

If uncertainty is absent, i.e. the situation has only one outcome, the system consists of one kind of element only, $I=0$, because $M=m_1; n=n_1$ and $p_1=1$. A maximum quantity of the information is found in the case when the situation has the maximal possible outcomes: the system consists of different elements only: $n_i = 1; p_i = 1/n; I = \lg(n)$.

Thus, the information content of a system is defined by both the quantity of elements (n) and elements method distribution on subsets ($|m_j|=n_j$) only. Such an approach ignores the qualitative content of the information completely. The positive feature of this approach is analysis of any informational processes and systems based on general principles.

V.E. Kuz'min (✉)
O.V. Bogatsky Physico-Chemical Institute
of the National Academy of Sciences of Ukraine,
86 Lustdorfskaya doroga, Odessa, 65080, Ukraine
e-mail: victor@farlep.net
Tel.: +380 482 225127, Fax: +380 482 652012

L.N. Ognichenko · A.G. Artemenko
I.I. Mechnikov Odessa National University,
2 Dvoryanskaya str., Odessa, 65080, Ukraine

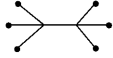
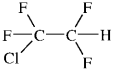
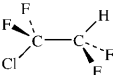
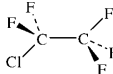
Model	Groups of equivalent elements of structure	Amount of the information (bit)
a) 	(2,6)	0.81
b) <chem>C2F3ClH</chem>	(4, 2, 1, 1)	1.75
c) 	(2, 2, 1, 1, 1, 1)	2.50
d) 	(2, 2, 1, 1, 1, 1)	2.50
e) 	(1, 1, 1, 1, 1, 1, 1, 1, 1)	3.00

Fig. 1 Information content of a molecular structure

The molecular structures are easily formalized as various discrete models. Informational content is easily calculated for each model. [4, 5] For example, we consider halogen-substituted ethane (see Fig. 1). As is evident from Fig. 1, the simplest models of molecules – molecular graphs (a) contain the minimum information. Taking into account the nature of the atoms in a molecular structure, a large quantity of information results (b, c). Differentiation of atoms of one nature depending on their neighbors also leads to an increase of the information quantity (c). Spatial models of the molecular structure are more adequate and the informational content is a maximum (d, e). Thus, the information content in a molecular structure can easily be calculated by the Shannon formula. The informational characteristics of molecules are now used widely in computer synthesis, [6] as well as for the solution of “structure–property” tasks. [5, 6]

Model of the informational field

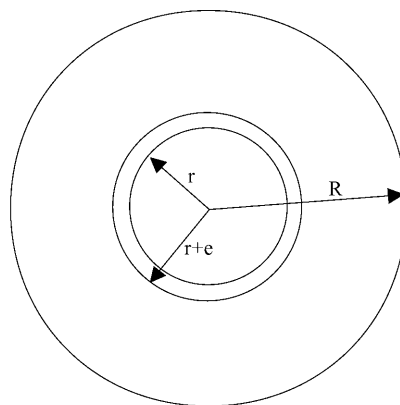
In the literature cited we have tried to investigate a rather fundamental problem “How does the information of the molecular structure influence the surrounding space?”, i.e. to simulate an information field (IF) of a molecule. It is more correct to consider a field of the potential information (FPI not IF), since the information arises only when there is a “receiver” (“observer”). The FPI formal model was constructed previously. [7] According to the formal FPI model, it is necessary to apply Shannon’s approach consistently to research the “object” and surrounding space. In other words, this space, as well as the “object”, have to be discretized and considered as a system of elements.

Let us consider the idealized situation: the “empty” infinite space is divided into identical areas (cells). We

D	D	D	D	D	D	D
D	C	C	C	C	C	D
D	C	B	B	B	C	D
D	C	B	A	B	C	D
D	C	B	B	B	C	D
D	C	C	C	C	C	D
D	D	D	D	D	D	D

$I(A)=1b(N/1)$ $I(B)=1b(N/8)$ $I(C)=1b(N/16)$ $I(D)=1b(N/24)$
 N-number of cells of research area of space

Fig. 2 Fragment of an informational field of a one-element object



2-D – space:

$$I(r) = 1b[(\pi R^2)/(\pi((r+\epsilon)^2 - r^2))] = 21b(R) - 1b(\epsilon) - 1b(2r+\epsilon)$$

$$I(0) = 21b(R) - 21b(\epsilon) = 21b(R/\epsilon)$$

$$I(R) = 21b(R) - 1b(\epsilon) - 1b(2R+\epsilon) \sim 1b(R) - 1b(\epsilon) - 1b(2)$$

$$\Delta I = I(0) - I(R) \sim 1b(R) - 1b(\epsilon) + 1b(2)$$

3-D – space:

$$I(r) = 1b[(4/3)\pi R^3/((4/3)\pi((r+\epsilon)^3 - r^3))] = 31b(R) - 1b(\epsilon) - 1b(3r^2+3r\epsilon+\epsilon^2)$$

$$I(0) = 31b(R) - 31b(\epsilon) = 31b(R/\epsilon)$$

$$I(R) = 31b(R) - 1b(\epsilon) - 1b(3R^2 + 3R\epsilon + \epsilon^2) \sim 1b(R) - 1b(\epsilon) - 1b(3)$$

$$\Delta I = I(0) - I(R) \sim 21b(R) - 21b(\epsilon) + 1b(3)$$

N-D – space:

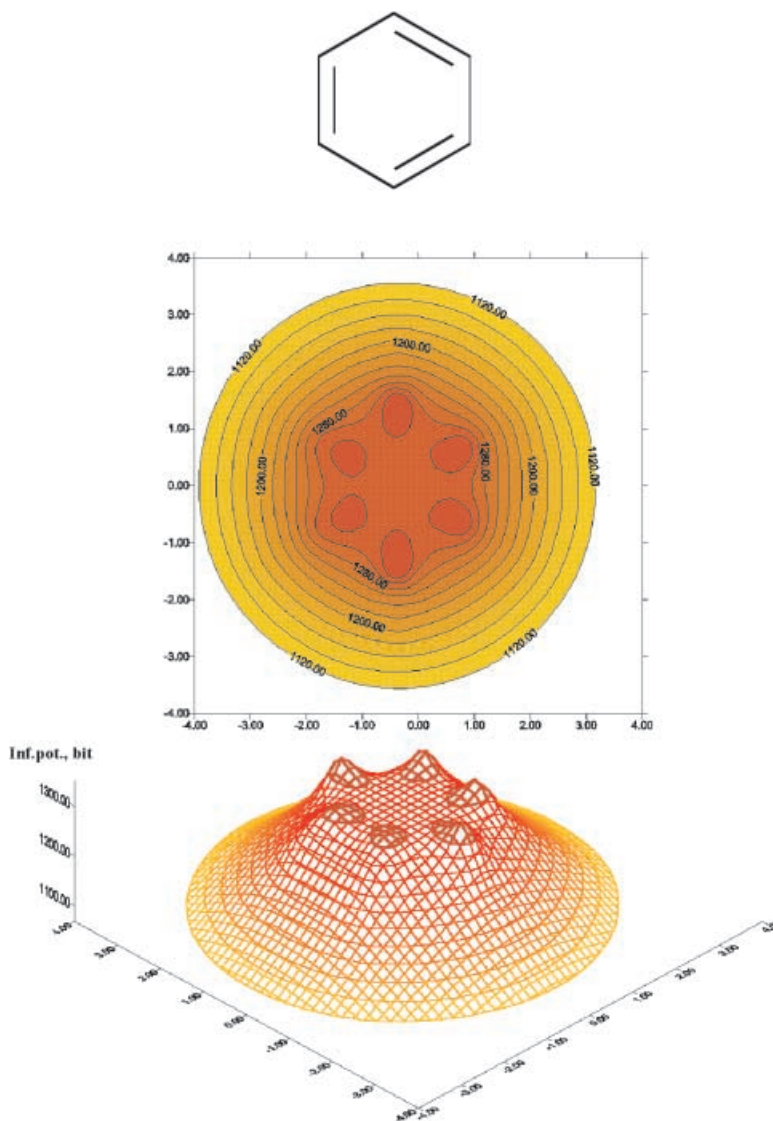
$$I(0) = (N)1b(R/\epsilon)$$

$$(1)R \sim 1b(R) - 1b(\epsilon) - 1b(N)$$

$$\Delta I \sim (N - 1)1b(R) - (N-1)1b(\epsilon) - 1b(N)$$

Fig. 3 Isotropic model of the informational field

Fig. 4 The informational field of benzene



have no bits of information, since all cells are absolutely identical. There is no variety; hence there is no information. However, as soon as we place any object in this space, then the equivalence of many cells is eliminated. Cells become different and their difference depends on the observation place of the object. We (receiver–observer) can receive different quantities of information about this object at various places in the surrounding space. Thus, for modeling the object FPI in a definite area, we should be able to differentiate cells of space surrounding the object. When all cells in the definite area of space are divided into groups, we will need only to apply Shannon’s formula and to calculate the potential information in each cell, in a group of cells and in the whole area.

Generalizing the above, the following conclusion is possible: from a formal position the model of the potential information field for any object describes a situation where the surrounding space is structured by the ob-

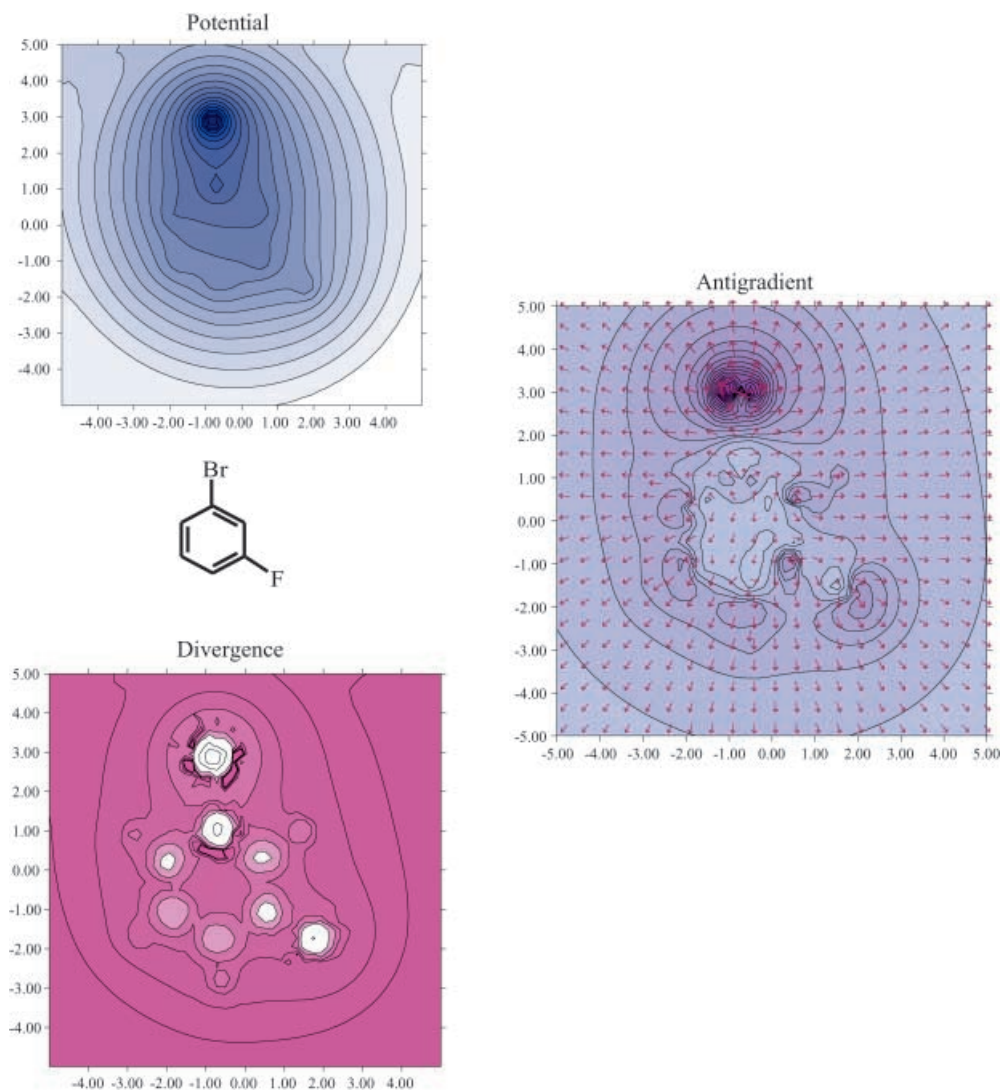
ject, i.e. the object generates the potential information in this area.

What is the reason? A complex of real physical fields of the research object can itself cause structuring of the surrounding space. However, in the given context we do not consider how the structuring of space occurs. We analyze only results. We do not negate physical mechanisms of information transfer from the object in surrounding space; we do not concretize them.

Let us consider the elementary example. The object is one cell; then the surrounding space is divided into concentric layers of cell-groups (see Fig. 2). According to Shannon’s formula the information potential in the given cell depends only on the total amount of cells considered (n) and on the quantity of cells of the given group (n_i)¹. It is evident that in each subsequent layer the informa-

¹ In the given example the model for the presentation was simplified. Angular and non-angular cells were not distinguished.

Fig. 5 Parameters of the information field of bromofluorobenzene



tion potential is less than in the previous one. Thus, the considered model FPI correctly reflects the natural decrease of the potential information when the “receiver” (observer) moves away from the object.

Developing the given approach, it is easy to construct the isotropic model of an information field for spaces of any dimension. Two-dimensional space is given as an illustration (see Fig. 3). Let the discretization step of space be ε , where $\varepsilon \ll R$ (ε is a tuning parameter of the model and R the radius of the area of space considered). It is necessary to calculate the binary logarithm of the relation of the area (volume) of all area of research space (total number of points) to the area (volume) of a layer of space with thickness $(r+\varepsilon)$ (amount of points of the given group) for an estimation of the potential of an information field for a dot object at a distance r .

The procedure for spaces of different dimension is given in Fig. 3, where R determines the research area of space (r changes from 0 up to R); $I(r)$ – informational potential at distance r from the “object”; $I(0)$ – informa-

tional potential at a point corresponding to the object; $I(R)$ – informational potential in the borderline research area.

Only the isotropic model of FPI, where $R=20 \text{ \AA}$; $\varepsilon=0.05 \text{ \AA}$ was used in this work.

It should be considered that the information, as entropy, has additive character for application of the given model FPI to complex objects consisting of several elements. [3] This means that the FPI of a complex object can be modeled as a superposition of the FPI of its elements. Moreover, when dimensionless weight parameters

$$v_i = p_i / \sum_{i=1}^n p_i$$

describing any property (p) of the elements of an object are introduced, a weighted FPI can be constructed. In fact, such a field reflects information about the distribution of the considered property in space. The FPI of benzene weighted by mass is given, for example, in Fig. 4.

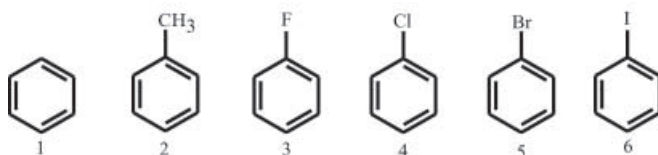
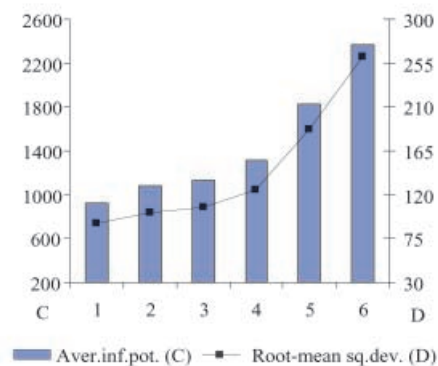
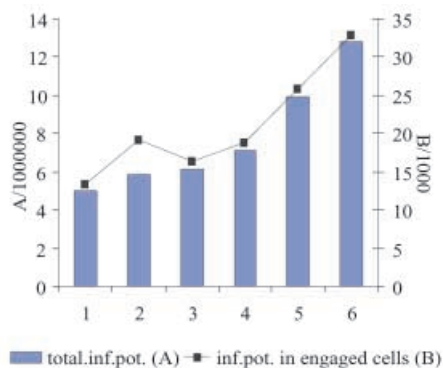


Fig. 6 Parameters of the information field of monosubstituted benzenes (where total inf.pot. (A) is the total informational potential, inf.pot. in engaged cells (B) is the informational potential in engaged cells, aver.inf.pot. (C) is the average informational potential and root-mean sq.dev. (D) is the mean square deviation of potential of an informational field)

In this model the potential (in bit) of the informational field, as well as the antigradient² (in bit Å⁻¹) and the divergence (in bit Å⁻²) of the FPI (see Fig. 5) can be calculated based on the standard procedure. [8]

Molecular informational fields

Different aromatic compounds³ were considered in order to illustrate the opportunities of the given model of an information field. The information fields were weighted by atomic mass for all systems.

The monosubstituted benzenes were considered in the first set of molecules investigated. As can be seen from

² The orientation of vectors of an antigradient of an information field indicates a direction of the information distribution.

³ These molecules are flat and convenient for visual analysis. However, the model of an information field is applied for any molecular structures.

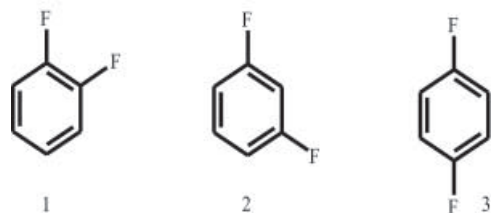
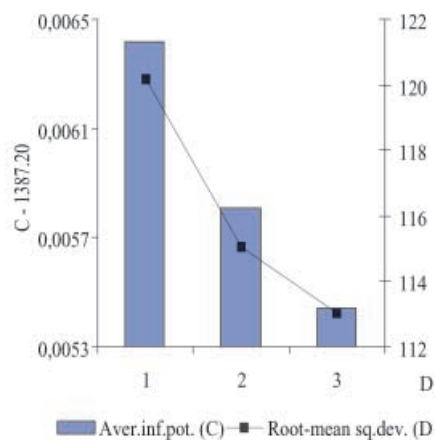
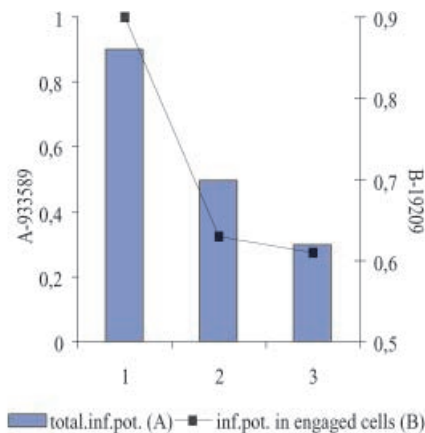


Fig. 7 Parameters of the informational field of isomers of difluorobenzene (where total inf.pot. (A) is the total informational potential, inf.pot. in engaged cells (B) is the informational potential in engaged cells, aver.inf.pot. (C) is the average informational potential and root-mean sq.dev. (D) is the mean square deviation of potential of an informational field)

Fig. 6, the informational potential and the average characteristics of informational fields increase with increase of the substituent mass for all these molecules. The mean square deviation of an informational field potential varies analogously. These results are expected and confirm that the developed concept does not contradict common sense.

All isomers of difluorobenzene were considered in the second set of molecules investigated (see Fig. 7). The characteristics of an informational field for these compounds decrease from the *ortho* to the *para* isomer. It is evident that this is caused by the mutual influence of the informational fields of the fluorine atoms.

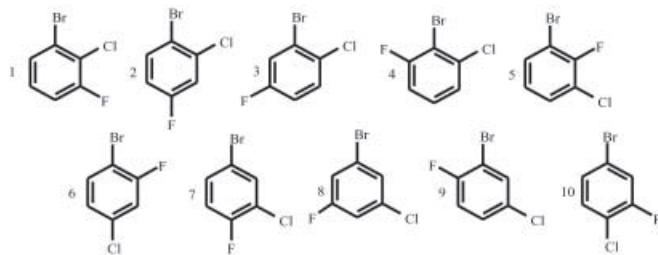
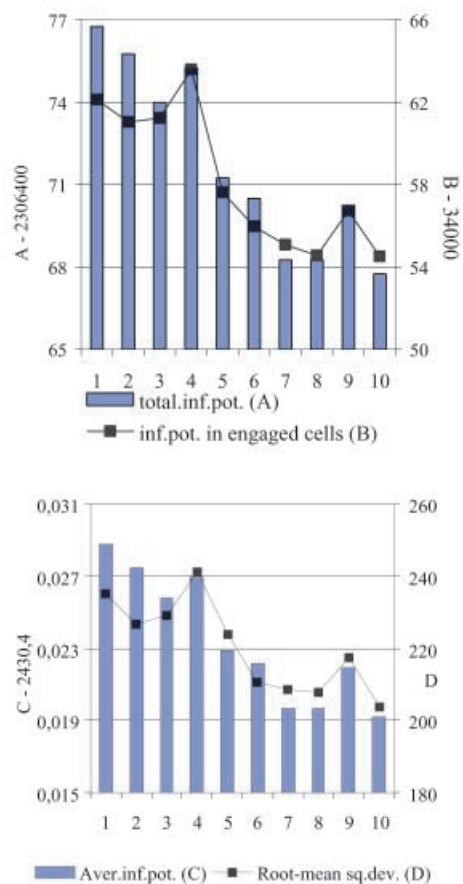


Fig. 8 Parameters of the informational field of trisubstituted benzenes.(where total inf.pot. (A) is the total informational potential, inf.pot. in engaged cells (B) is the informational potential in engaged cells, aver.inf.pot. (C) is the average informational potential and root-mean sq.dev. (D) is the mean square deviation of potential of an informational field)

The isomers of fluorochlorobromobenzene are the most interesting. If the information content of the molecular graph is analyzed for all of them, it is easy to see that it is the same for all compounds. In these cases the information content of the molecular structures equals $lb(9)$ (all atoms⁴ vary).

On the other hand, the characteristics of an informational field vary for all these molecules and allow them to be discriminated (see Fig. 8).

The synergism of the mutual influence of information fields of the substituents is maximal for isomers 1, 2, 3

⁴ The group C–H was considered as a joint atom.

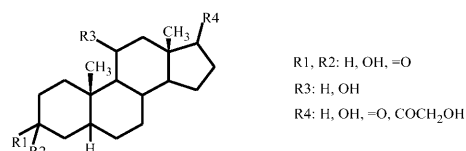


Fig. 9 Chart of the steroids tested

Table 1 Parameters of the correlation equations $A=ax+by+cz+d$ describing the relationship between biological activity and characteristics of an informational field for steroids

Activity (A)	A_{CoBG}	A_{TeBG}
a	$-(0.065 \pm 0.003)$	(0.035 ± 0.002)
t_a^a	10.1	8.0
x	$D_L(56)^b$	$AG_Q(24)^c$
b	$-(0.833 \pm 0.055)$	(0.389 ± 0.029)
t_b^a	7.0	6.1
y	$D_{Rf}(13)^d$	$P_Q(102)^e$
c	(1.531 ± 0.215)	$-(0.562 \pm 0.077)$
t_c^a	3.3	3.4
z	$D_L(123)^b$	$D_Q(39)^f$
d	$-(0.177 \pm 0.45)$	$-(6.875 \pm 0.052)$
t_d^a	0.2	60.7
R^{2g}	0.933	0.912
CVR^{2h}	0.898	0.862
S^i	0.33	0.39
F^j	78	59

^a t is the Student's coefficient (critical value of $t=2.11$ for $\alpha=0.95$)

^b $D_L(56)$ and $D_L(123)$ is the divergence of the informational field weighed by lipophilicity (in cells 56 and 123)

^c $AG_Q(24)$ is the antigradient of the informational field weighed by charge (in cell 24)

^d $D_{Rf}(13)$ is the divergence of the informational field weighed by refraction (in cell 13)

^e $P_Q(102)$ is the potential of the informational field weighed by charge (in cell 102)

^f $D_Q(39)$ is the divergence of the informational field weighed by charge (in cell 39)

^g R is the coefficient of correlation

^h CVR is the coefficient of correlation in requirements for "leave-one-out cross-validation"

ⁱ S is the standard error of estimate

^j F is the value of Fisher's criterion (critical value of $F=3.2$ for $\alpha=0.95$)

and 4. The information interactions of the substituents are minimal for isomers 7, 8 and 10, since these substituents are located at the largest distance from each other.

In the framework of the informational field model we have tried to solve some "structure–property" tasks.

Solution of QSAR tasks by means of structural parameters of the molecular informational fields

The series of the steroid molecules (see chart in Fig. 9) were investigated to assess the application of parameters of a molecular informational field for QSAR tasks. [9, 10] The CoMFA approach, as one of the most popular methods for QSAR tasks solution, was tested on this series of compounds. [9, 10, 11] The affinity of given compounds for Testosterone-Binding Globulin (A_{TeBG}) and

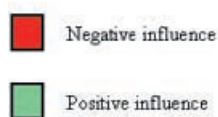
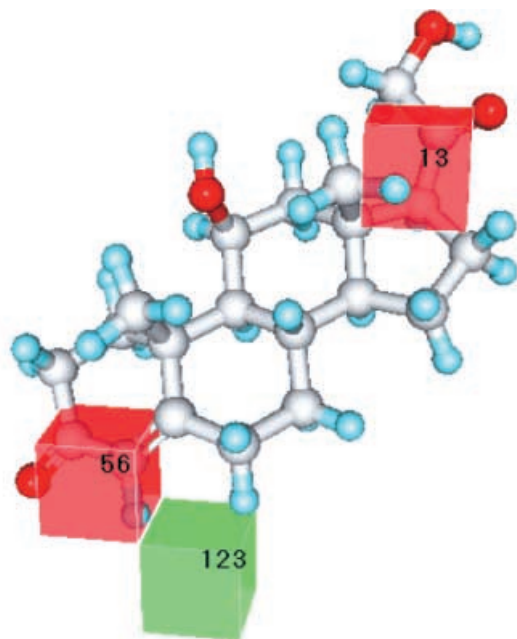


Fig. 10 Affinity of steroids for Corticosteroid-Binding Globulin (CoBG)

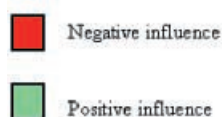
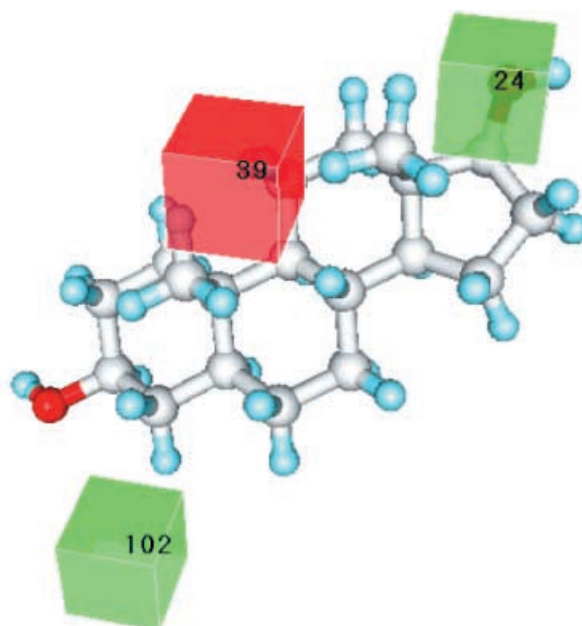


Fig. 11 Affinity of steroids for Testosterone-Binding Globulin (TeBG)

for Corticosteroid-Binding Globulin (A_{CoBG}) were considered as target properties.

The steroids investigated were considered within formalism frameworks of the molecular lattice model. [12] All these molecules were placed in a cubic lattice with the cell size of 2 Å. Structures were aligned on carbon atoms of the steroid skeleton. The investigated area of space is a parallelepiped consisting from 252 cells ($6 \times 6 \times 7$). The steroids are located in the center of this parallelepiped. The potential, antigradient and divergence of an informational field were estimated in each cell. These parameters were weighted by charges, lipophilicity, refraction, mass, polarizability and electronegativity. A set of 4,536 structural parameters was obtained. It is clear that regression analysis cannot be applied for this set of parameters. Therefore, one parameter was selected from each group of cross-correlated parameters (at the level $R=0.9$) at the beginning of the analysis. This procedure allows the total number of parameters be decreased to 590. The equations connecting parameters of an informational field of steroids with their biological activity were calculated by stepwise regression [13] (see Table 1). The corresponding cells are shown in Figs. 10 and 11.

It is noteworthy that the statistical characteristics of these equations (see Table 1) are a rather better than

those given by CoMFA ($R^2=0.897$ for CoBG and $R^2=0.873$ for TeBG). [11]

As seen from Figs. 10 and 11, the biological activity is defined by the characteristics of the information fields in the regions of space near the functional groups of the steroids in both cases. It is quite logical that the equations include the characteristics of information fields weighted by charges, lipophilicity and refraction. These properties of molecular fragments can define the character of the appropriate intermolecular interactions of steroids with globulins.

Thus, the concept of an information field of molecules can be promising for the analysis of the peculiarities of molecular structure and for the solution of various applied "structure–property" tasks.

Acknowledgment This work was partially supported by the INTAS foundation (grant INTAS 97-31528 and grant INTAS 97-1730)

References

1. Shannon K (1963) Works on the theory of the information and cybernetics. In Lit, Moscow
2. Ursul AD (1973) Reflection and information. Nauka, Moscow
3. Brillouin L (1956) Science and information theory. Academic Press, New York

4. Zhdanov UA (1979) Entropy of the information in organic chemistry. RGU, Rostov-na-Donu
5. King RB (ed) (1983) Chemical applications of topology and graph theory. Elsevier, Amsterdam
6. Pierce TH, Hohne BA (eds) (1985) Artificial intelligence: application in chemistry. American Chemical Society, Washington, D.C.
7. Kuz'min VE (2000) Modeling of the informational field of molecules. Rep NAS Ukraine 3:159–163
8. Mc Connell AJ (1957) Application of tensor analysis. Dover Publications, New York
9. Good AC, Sung-Sau S, Richards WG (1993) J Med Chem 36:433–438
10. Good AC, Peterson JS, Richards WG (1993) J Med Chem 36:2929–2937
11. Cramer RD, Patterson DE, Bunce JD (1988) J Am Chem Soc 110:5959–5967
12. Kuz'min VE, Artemenko AG, Kovdienko NA, Tetko IV, Livingstone DJ (2000) J Mol Mod 6:517–526
13. Draper NR, Smith H (1981) Applied regression Analysis. Wiley, New York