Modelling of the 'structure–biological activity' relationship for conformationally mobile molecules

Gennady M. Makeev* and Mikhail I. Kumskov

N. D. Zelinsky Institute of Organic Chemistry, Russian Academy of Sciences, 117913 Moscow, Russian Federation.
Fax: +7 095 135 5328; e-mail: makeev@cacr.ioc.ac.ru

A procedure for the QSAR study of conformationally mobile molecules is proposed.

Progress in methods for the quick construction of three-dimensional structures for small molecules[1] has stimulated studies dealing with the construction of quantitative 'structure–activity' relationships (QSAR) by virtue of descriptors characterising the three-dimensional structures of molecules (3D-QSAR).

The method of comparative field analysis (CoMFA) is widely used within the framework of 3D-QSAR procedures.[2]

The necessity to carry out 'spatial normalisation' of the molecules of the training database, *i.e.* to choose their arrangement (after mutual spatial alignment) with respect to a system of coordinates is the most delicate point in the use of CoMFA. It is known that a change in the position of a coordinate system (for example, its simple rotation) can decrease by half the predicting capacity of a CoMFA model.[3] A description of the characteristic features of spatial structure based on coding of the mutual arrangement of structural fragments in a molecule[4,5] would be free from the problems associated with the choice of the mutual orientation of molecules.

In this study, we present an attempt to describe the spatial structures of molecules of conformationally mobile molecules. We compare the predicting qualities of QSAR models built using two different classes of 3D descriptors: descriptors based on a set of conformations and those based on the minimum molecular energy[6,7] conformation. The form of the descriptors used does not depend on the choice of the mutual orientation of molecules.

Procedure for the construction of 3D descriptors. Let there be a structural database (SDB) of molecules, each of them being matched by a value of biological activity. It is necessary to construct a set of descriptors that would adequately describe variations of the specified activities of compounds included in the SDB, *i.e.* we must find the QSAR models which possess prognostic stability.

We studied descriptors of two types:[7]

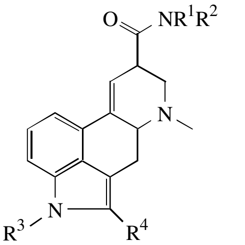
1. new descriptors based on a set of conformations;
2. old descriptors based on the conformations with the lowest energy.[6,7]

The descriptors based on a set of conformations have the following form:

$$(CF_i, CF_j, D_k), X \qquad (1)$$

where (CF_i, CF_j, D_k) is the code (symbolic name) of the descriptor composed of the codes of fragments CF_i, CF_j and D_k is the kth range of distances. $X = 1$ if the molecule contains a pair of CF_i and CF_j, and the distance between them falls into the D_k range at least for one of the conformations calculated for this compound; otherwise $X = 0$.

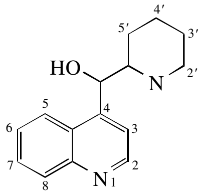
The descriptors based on the conformations with the



Structure	NR^{1}R^2	R^3	R^4	Antiserotonin activities	Hallucinogenic activities
1	NEt$_2$	H	H	100.0	100.0
2	NEt$_2$	H	Br	150.0	7.2
3	NH$_2$	H	H	4.3	0.0
4	NH[CH(C$_2$H$_5$)CH$_2$OH]	Me	H	400.0	0.6
5	NHMe	H	H	6.3	—
6	NMe$_2$	H	H	23.2	10.0
7	NHEt	H	H	11.9	5.0
8	NHEt	Me	H	835.0	4.0
9	NHEt	COMe	H	39.0	7.0
10	NHPr^i	H	H	22.2	—
11	NEt$_2$	Me	H	368.0	36.0
12[a]	NEt$_2$	OMe	H	58.9	66.0
13	NEt$_2$	COMe	H	210.0	100.0
14	NEt$_2$	H	I	57.4	—
15	NEt$_2$	Me	Br	533.0	< 1
16	N(–C$_4$H$_8$–)	Me	H	130.0	< 5
17	N(–C$_4$H$_8$–)	H	H	4.7	5.3
18	N(–CH$_2$CH=CHCH$_2$–)	H	H	4.1	10.0
19	N(–C$_5$H$_{10}$–)	H	H	8.5	—
20	N(–C$_2$H$_4$OC$_2$H$_4$–)	H	H	8.0	11.0

[a] Structure 12 failed to be calculated by the Macromodel program and was removed from the SDB.

Figure 1 A SDB of structural analogues of LSD exhibiting antiserotonin activities.



Structure	Substituents	Antimalarial activities
1	8-CF$_3$, 2-CF$_3$	100
2	6-OMe, 8-CF$_3$, 2-CF$_3$	25
3	7-CF$_3$, 2-CF$_3$	38
4	6-CF$_3$, 2-CF$_3$	12
5	6-OMe, 2-CF$_3$	6
6	6-Me, 2-CF$_3$	2
7	8-Me, 2-CF$_3$	2
8	6-Me, 8-Me, 2-CF$_3$	3
9	6-Me, 8-Me, 4′-Me	3
10	6-Me, 8-Me, 4′-OMe	12
11	6-Me, 8-Me, 4′-Cl	38
12	6-Me, 8-Me, 4′-F	12
13	8-CF$_3$, 4′-H	50
14	8-CF$_3$, 4′-Me	25
15	8-CF$_3$, 4′-OMe	37
16	8-CF$_3$, 4′-Cl	125
17	6-Me, 4′-OMe	3
18	8-Me, 4′-H	6
19	8-Me, 4′-Me	6
20	8-Me, 4′-Cl	50
21	8-Me, 4′-F	25

Figure 2 A SDB of compounds with antimalarial activities.

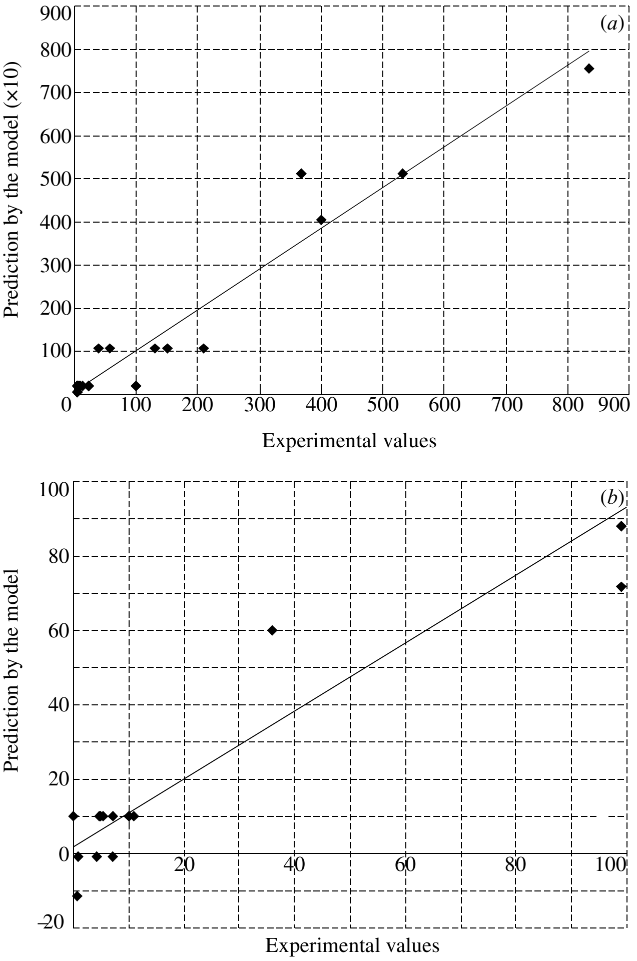


Figure 3 ‘Experiment–prediction’ plots for antiserotonin and hallucinogenic activities constructed for QSAR models on BB descriptors based on the conformations with the lowest energy: (*a*) antiserotonin activity BB ‘bond–bond’ descriptors (R^2 = 0.94, K = 4); (*b*) hallucinogenic activity BB ‘bond–bond’ descriptors (R^2 = 0.91, K = 3).

minimum energy have the same form:

$$(CF_i, CF_j, D_k), Y \qquad (2)$$

but the Y value is equal to the number of pairs of CF_i and CF_j fragments separated by a distance falling in the D_k range for the conformation of the molecule with the lowest energy.

As structural fragments CF_i, we used the simplest substructures consisting of one atom (A) or two atoms linked by a covalent bond (B). For each class of descriptors, the QSAR models were constructed separately using descriptors of three types: ‘atom–atom’ (AA), ‘atom–bond’ (AB) and ‘bond–bond’ (BB) descriptors. During the construction of the CF_i structural fragments, the atoms in the molecule were additionally labelled. The label of an atom takes into account the number and the multiplicity of bonds formed by this atom with other atoms and whether or not this atom is incorporated in a ring.[8] For example if a carbon atom has the label ‘C 2dr’, then marker ‘2’ means that atom has two bonds, ‘d’, the atom has one double bond adjusted, and ‘r’ means that the atom is in the ring. Once the atoms were labelled, a complete list of structural fragments presented in the molecule was enumerated. Two structural fragments were considered to be equivalent if their codes consisting of the symbol labels of the atoms were identical.

Thus, to describe the molecules of the training SDB, we used two types of 3D descriptors. The descriptors of the same type calculated for all compounds included in the training SDB were combined into a ‘molecule–descriptor’ matrix [X], the x_{ij} element of which was equal to the X value in descriptors (1) or Y value in descriptors (2). Component x_{ij} is the magnitude of a

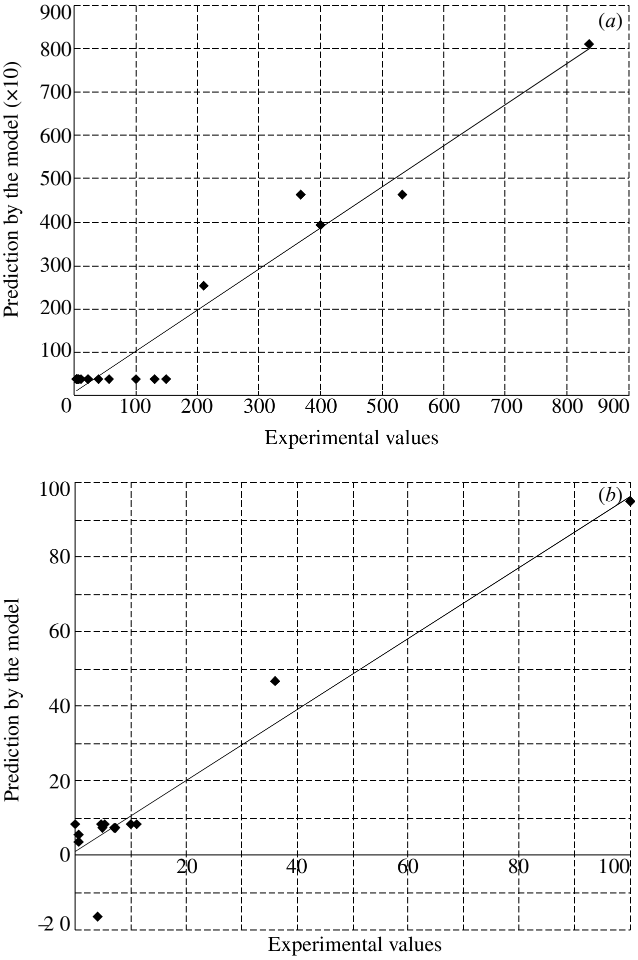


Figure 4 ‘Experiment–prediction’ plots for antiserotonin and hallucinogenic activities constructed for QSAR models on BB descriptors based on a set of conformations: (*a*) antiserotonin activity BB ‘bond–bond’ descriptors (R^2 = 0.95, K = 4); (*b*) hallucinogenic activity BB ‘bond–bond’ descriptors (R^2 = 0.95, K = 4).

jth descriptor in an ith compound: an ith line of the matrix [X] describes an ith compound as a feature’s vector, while a jth column is the vector of the contents of a jth descriptor for all the compounds in the SDB.

Linear QSAR models of the following form were constructed:

$$A_i = b_0 + \sum_{j=1}^{K} b_j x_{ij} + E_i; \; i = 1, \ldots, N; \qquad (3)$$

$$\sum_{i=1}^{N} E_i^2 \rightarrow \min \qquad (4)$$

where A_i is the value for the biological activity of an ith molecule in the training SDB, x_{ij} is the content of a jth descriptor for an ith molecule; K is the number of parameters in the equation; N is the size of the training SDB; b_j are weight coefficients to minimize (4) and E_i is the error for the approximation of the activity of an ith compound.

The search for a minimum for formula (4) involves problems caused by the fact that the resulting QSAR matrix [X] contains a very great number of columns M ($M >> N$, M = 300–1 000), which are strongly correlated with one another. The BIBIGON program, which we used in our study, constructs a family of QSAR models of the form (3)–(4) based on the self-organisation of linear models by virtue of a group method of data handling (GMDH).[9,10] To avoid chance effects[11] we use a cross-validation[12] test (leave-one-out) because the CoMFA method uses it in the same way.[2,3] In the prediction of biological activities, a QSAR model is considered to be significant if the squared cross-validation correlation coefficient[12] is greater than 0.4.[11]

To analyse QSAR models, two training SDBs were used:

Table 1 Parameters of QSAR models for the SDB consisting of mephloquine derivatives ($N = 21$): descriptors based on the lowest-energy conformations.

Activity	AA descriptors	AB descriptors	BB descriptors
A1 is animalarial activity	$K = 3$; $R^2 = 0.85$; Cr. $R^2 = 0.75$	$K = 3$; $R^2 = 0.85$; Cr. $R^2 = 0.75$	$K = 4$; $R^2 = 0.85$; Cr. $R^2 = 0.72$
A2 is the logarithm of A1, $K = 4$	$R^2 = 0.86$; Cr. $R^2 = 0.78$	$R^2 = 0.91$; Cr. $R^2 = 0.88$	$R^2 = 0.90$; Cr. $R^2 = 0.84$

Table 2 Parameters of QSAR models for the SDB consisting of mephloquine derivatives ($N = 21$): descriptors based on a set of conformations.

Activity	AA descriptors	AB descriptors	BB descriptors
A1 is antimalarial activity, $K = 4$	$R^2 = 0.90$; Cr. $R^2 = 0.82$	$R^2 = 0.91$; Cr. $R^2 = 0.81$	$R^2 = 0.91$; Cr. $R^2 = 0.87$
A2 is the logarithm of A1, $K = 4$	$R^2 = 0.85$; Cr. $R^2 = 0.81$	$R^2 = 0.90$; Cr. $R^2 = 0.84$	$R^2 = 0.87$; Cr. $R^2 = 0.80$

Table 3 Parameters of the QSAR models for the SDB consisting of LSD analogues: descriptors based on the lowest-energy conformations.

Activity	AA descriptors	AB descriptors	BB descriptors
A1 is antiserotonin activity, $K = 4$	$R^2 = 0.94$; Cr. $R^2 = 0.79$	$R^2 = 0.95$; Cr. $R^2 = 0.90$	$R^2 = 0.94$; Cr. $R^2 = 0.72$
A2 is the logarithm of A1, $K = 4$	$R^2 = 0.96$; Cr. $R^2 = 0.93$	$R^2 = 0.96$; Cr. $R^2 = 0.93$	$R^2 = 0.98$; Cr. $R^2 = 0.97$
A3 is hallucinogen activity, $K = 3$	$R^2 = 0.872$; Cr. $R^2 = 0.63$	$R^2 = 0.88$; Cr. $R^2 = 0.79$	$R^2 = 0.91$; Cr. $R^2 = 0.81$

Table 4 Parameters of the QSAR models for the SDB consisting of LSD analogues: descriptors based on a set of conformations.

Activity	AA descriptors	AB descriptors	BB descriptors
A1 is antiserotonin activity. $K = 4$	$R^2 = 0.94$; Cr. $R^2 = 0.90$	$R^2 = 0.86$; Cr. $R^2 = 0.80$	$R^2 = 0.96$; Cr. $R^2 = 0.89$
A2 is the logarithm of A1, $K = 4$	$R^2 = 0.93$; Cr. $R^2 = 0.91$	$R^2 = 0.91$; Cr. $R^2 = 0.83$	$R^2 = 0.97$; Cr. $R^2 = 0.95$
A3 is hallucinogen activity	$K = 3$; $R^2 = 0.82$; Cr. $R^2 = 0.58$	$K = 2$; $R^2 = 0.95$; Cr. $R^2 = 0.91$	$K = 2$; $R^2 = 0.94$; Cr. $R^2 = 0.90$

1. A SDB of structural analogues of LSD[13] (Figure 1) exhibiting antiserotonin activity (19 structures). Fifteen of them possessed hallucinogenic activities.

2. A SDB of compounds (Figure 2) with antimalarial activities (21 structures).[14]

The choice of conformation for the 3D-QSAR analysis. Sets of conformations for each molecule in the training SDB were formed in the following way. First, the spatial structure of the molecule was calculated with the aid of the 'Macromodel 4.5' program,[15] for which the energy minimum was found. Then a family of conformations of the molecule was generated using the 'MultiConformer' operation included in the 'Macromodel' program; for this purpose, systematic rotation around each dihedral angle with a step of 60° was carried out. Each structure of the set of structures thus obtained was optimised using an MM3 1990 force field[16] and the diagonal Newton–Raphson optimisation method. Those structures were selected for which the calculated conformation energy exceeded the global minimum energy by not more than 100 kJ mol^{-1}.

Characteristics of the constructed models. The following parameters of the constructed QSAR models are presented below: K is the number of descriptors in model (3), R^2 is the squared multiple correlation coefficient and Cr. R^2 is the squared cross-validation correlation coefficient.[12] In all cases, models with the largest Cr. R^2 values were chosen.

The results of analysis of the SDB consisting of LSD analogues.[13] Antiserotonin ($N = 19$) and hallucinogen ($N = 15$) activities expressed as a percentage of the activity of the LSD molecule were chosen. The activity of LSD was taken to be 100 in both cases.

Figure 3 shows the 'experiment–p rediction' plots for antiserotonin and hallucinogenic activities constructed for the QSAR models with BB descriptors based on the lowest-energy conformations. Figure 4 shows the 'experiment–predicti on' plots for the same activities constructed for the QSAR models with BB descriptors based on a set of conformations. The abscissa axis shows the experimental value for the activity (for logarithms of activity, the initial values are restored), while the ordinate axis shows the values predicted by the model.

The results of analysis of the SDB with antimalarial activity.[14] The antimalarial activities of mephloquine derivatives were estimated with respect to that of mephloquine (activity 100%).

Analysis of the plots shown in Figures 3 and 4 makes it possible to conclude that the main prediction errors refer to compounds with low activities. It can be seen from Tables 1–4 that the cross validation coefficients obtained (Cr. $R^2 = 0.7–0 .97$) imply that the predicted stability of the resulting QSAR models is fairly high. R^2 and Cr. R^2 coefficients increase with an increase in the complexity of the structural fragments under consideration (AA → AB → BB). It can be noted that for the second SDB (antimalarial activity), the descriptors based on sets of conformations provide greater values for the Cr. R^2 coefficient; this is apparently due to the higher flexibility of these compounds.

The models obtained are characterised by fairly good correlation ($R^2 = 0.85–0. 97$) and cross-validation (Cr. $R^2 =$ $= 0.58–0. 96$) coefficients (among the values normally obtained for biological activity); this points to a high stability and high predicting capacity of the models. Thus, the method described above makes it possible to carry out QSAR studies for conformationally mobile molecules.

The implicit form of the descriptors used in the QSAR model (3) makes it possible to distinguish pharmacophore fragments; they will be described in subsequent communications. Since the predicting capacity of the models increases following an increase in the complexity of the structural fragments, it would be expedient to include more complex structural fragments in the construction of descriptors (1)–(2). The BIBIGON program for DOS is freely available from the authors.

This work was partly supported by the Russian Foundation for Basic Research (grant nos. 96-01-01598, 97-07-90307 and 98-01-00324).

References

1 P. A. Willet, *J. Chemom.*, 1992, **6**, 289.

2 R. D. Cramer III, D. E. Patterson and J. D. Bunce, *J. Am. Chem. Soc.*, 1988, **110**, 5959.

3 S. J. Cho and A. Tropsha, *J. Med. Chem.*, 1995, **38**, 1060.

4 P. A. Bath, A. R. Poirrette and P. Willet, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 141.

5 R. Nilakatan, N. Bauman and R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.*, 1993, **33**, 79.

6 G. M. Makeev and M. I. Kumskov, *Mendeleev Commun.*, 1996, 27.

7 G. M. Makeev, M. I. Kumskov, I. V. Svitanko and I. L. Zyryanov, *Pattern Recogn. Image Anal.*, 1996, **6**, 795.

8 L. A. Ponomareva, E. N. Olsuf'eva, M. N. Preobrazhenskaya, M. I. Kumskov and N. S. Zefirov, *Khim.-Farm. Zh.*, 1993, 36 (in Russian).

9 M. I. Kumskov and D. F. Mityushev, *Pattern Recogn. Image Anal.*, 1996, **6**, 497.

10 *Self-organizing methods of modeling: GMDH-type algorithms*, ed. S. Farlow, Marcel Dekker Inc, New York, 1986, p. 320.

11 B. Topliss, *J. Med. Chem.*, 1979, **22**, 1238.

12 M. A. Sharaf, D. L. Yllman and B. P. Kowalsky, *Khemometrika* (*Chemometrics*), Khimiya, Leningrad, 1984 (in Russian).

13 S. P. Gupta, *Chem. Rev.*, 1989, **89**, 1791.

14 S. Polman, S. Kokpol, S. Hannongbua and B. M. Robe, *Analyt. Sci.*, 1989, **5**, 641.

15 *Macromodel User Manual. Version 4.5*, Columbia University, New York, 1994.

16 N. L. Allinger, Y. H. Yuh and J.-H. Lii, *J. Am. Chem. Soc.*, 1989, **111**, 8551.

Received: Moscow, 3rd October 1997
Cambridge, 4th December 1997; Com. 7/07595B