The relationship between the structure and anti-AIDS activity of polyhydroxypiperidines and polyhydroxypyrrolidines

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 mathematical model has been constructed for the anti-AIDS activity of polyhydroxypiperidines and polyhydroxypyrrolidines based on a systematic search of such sets of pairs of fragments, whose presence or absence in the molecules studied considerably affects their biological activity.

Various methods for describing the spatial structure of molecules are used for constructing quantitative 'structure–activity' relationships (QSAR) of biologically active compounds.[1–5] Previously, the method of molecular electrostatic potential (MEP) applied within the BIBIGON program package[5–8] was successfully used for QSAR equations for musk odourants[3] and for compounds possessing psychotropic activity.[4] This method uses descriptors [describing the mutual spatial location of so-called 'representation points' (RP)]. The descriptors have the form:

$$(\mathrm{E1},\ \mathrm{E2},\ \mathrm{D}),\ \mathrm{X} \qquad (1)$$

where E1 and E2 are symbolic names of the intervals containing the MEP values of the first and second RP, respectively; D is the name of the interval containing the geometric distance between the RPs; and X is the number of occurrences of the (E1, E2, D) descriptor in a molecule.

The advantage of descriptors (1) is their clear physical interpretation. In addition, the possibility exists for a versatile expert control of the precision of describing the spatial structures of molecules by choosing the intervals that partition the MEP and D values in (1). On the other hand, a drawback of these descriptors is that they require laborious computations: optimisation of the geometry of the training database is performed in two stages. First, the conformation is searched by molecular mechanics methods, then they are refined by quantum-chemical calculations.

In the present work we use a modification of descriptors (1) developed by the authors for performing QSAR analyses of biologically active molecules.[9] Using this method, all advantages of descriptors (1) are maintained, but the required computation time is reduced several times.

An analysis of the meaning of descriptors (1) makes it possible to note that introduction of the E1, E2 and D intervals results in a considerable roughening in the description of the fine geometric configurations of molecules, which were refined at the quantum-chemical optimisation step requiring a lengthy computation time. For this reason, we considered that for forming modified descriptors it is sufficient to use 'approximate' geometric configurations obtained by molecular mechanics methods. In addition, according to ref. 9, the choice of representation points can be based on enumeration of all atoms or bonds in a molecule. The RPs of this type can be formed much more easily than the RPs located on an equidistant surface surrounding the molecule, as proposed in refs. 3–5.

The descriptors used by us are similar to those in (1) and have the form:

$$(\mathrm{f_1},\ \mathrm{f_2},\ \mathrm{D}),\ \mathrm{X} \qquad (2)$$

where $\mathrm{f_1}$ and $\mathrm{f_2}$ are base fragments chosen from some lists; D is the number of the interval of distances between the $\mathrm{f_1}$ and $\mathrm{f_2}$ fragments in the molecule; and X is the number of occurrences of a ($\mathrm{f_1}$, $\mathrm{f_2}$, D) descriptor in the molecule. (The distance between the fragments is determined as the closest distance between atoms belonging to different fragments).

Two descriptors, ($\mathrm{f_1}$, $\mathrm{f_2}$, $\mathrm{D_1}$) and ($\mathrm{g_1}$, $\mathrm{g_2}$, $\mathrm{D_2}$), are considered equal if the corresponding base fragments and distance intervals are pairwise equal:

$$\mathrm{f_1} = \mathrm{g_1};\ \mathrm{f_2} = \mathrm{g_2};\ \mathrm{D_1} = \mathrm{D_2} \qquad (3)$$

The comparison of base fragments in (3) takes into account the additional classification of elements used in the BIBIGON system.[6–8] The geometric differences between the base fragments are not considered.

Thus, unlike (1) and (2) we use instead of RP 'representation fragments', a list of which can be provided by a chemist expert who deals with QSAR modelling. We used two lists of base fragments generated by the BIBIGON system automatically from the molecules of a training data base. List 'A' contained fragments consisting of one atom, and list 'B' contained fragments consisting of two atoms linked in the molecule by a covalent bond. The D distance in the molecule between the base fragments can be determined as a geometric distance (D → G) or as a topological distance (D → T). Hence, six types of descriptors can be formed from a training data base:

three 'topological' types: AAT, ABT, and BBT

three 'geometric' types: AAG, ABG, and BBG,

where the G and T symbols indicate the type of distance used, and symbols A and B indicate that atoms or bonds, respectively, are used as fragments. The QSAR equations formed by the BIBIGON system had the same form as in refs. 3–5:

$$Y_j = b_0 + \sum_{j=1}^{k} b_i x_{ji} + E_j;\ j = 1 \ldots N \qquad (4)$$

$$\sum_{j=1}^{k} E_j^2 \rightarrow \min \qquad (5)$$

where Y_j is the activity of the jth molecule in the training database, x_{ji} is the number of occurrences of the ith descriptor in the jth structure; k is the number of parameters of the equation; N is the size of the training database; b_i are weight coefficients; and E_j are errors in approximation of the activity of the jth compound.

The minimum (5) is found by the BIBIGON system by choosing $\{x_i\}$ sets and their weight coefficients $\{b_i\}$ by self-organizing of the models,[10] which is related to the necessity of processing a very large number of descriptors. For example, when the database size is $N = 50$–100, the number of different descriptors (2) can reach 1000–2000.

To construct QSAR dependencies, we used a database containing 45 compounds corresponding to polyhydroxypiperidines and polyhydroxypyrrolidines (Figure 1), 19 of which possess anti-AIDS activity.[11] The database was prepared for searching a 'structure–property' dependence by

Table 1 Results of QSAR construction on geometric and topological descriptors.[a]

N - number of descriptors	Descriptors of AAG type		Descriptors of ABG type		Descriptors of BBG type		Descriptors of AAT type		Descriptors of ABT type		Descriptors of BBT type	
	R^2	Cr. R^2	R^2	Cr. R^2	R^2	Cr. R^2	R^2	Cr. R^2	R^2	Cr. R^2	R^2	Cr. R^2
1	0.45	0.39	0.47	0.39	0.46	0.42	0.54	0.44	0.56	0.51	0.56	0.51
2	0.55	0.46	0.63	0.57	0.69	0.63	0.66	0.45	0.64	0.50	0.78	0.76
3	0.63	0.50	0.75	0.64	0.81	0.64	0.70	0.47	0.68	0.59*	0.83	0.81
4	0.70	0.64	0.79	0.69	0.83	0.69	0.73	0.52*	0.71	0.51	0.85	0.83*
5	0.75	0.64	0.82	0.61	0.85	0.71	0.76	0.52	0.74	0.57	0.87	0.82
6	0.80	0.66	0.84	0.70	0.87	0.72	0.78	0.55*	0.77	0.62*	0.89	0.83
7	0.84	0.70	0.86	0.82	0.88	0.76*	0.79	0.51	0.78	0.60	0.89	0.83
8	0.87	0.71	0.88	0.84*	0.89	0.74	0.81	0.49	0.80	0.52	0.90	0.83
9	0.88	0.72*	0.89	0.84*	0.91	0.76*	0.82	0.48	0.82	0.36	0.91	0.84

[a] Designations used in Table 1: R^2 and Cr. R^2 – squares of the multiple correlation coefficient and the 'cross validation' coefficient, respectively. Asterisks indicate the 'best' models possessing the best predicting ability.

calculating the geometry of the molecules by molecular mechanics programs[9] using the MM2 force field and taking hydrogen bonds into account. Then, absolute configurations were determined, and carbon atoms with *R* and *S* absolute configurations were assigned marks CR and CS, respectively. Thus, the absolute configurations of optical centres were explicitly taken into account in marking the vertices of molecular graphs. The anti-AIDS activity was approximated to an equal concentration by the procedure reported in ref. 12. The following partitioning of distances into intervals was used for constructing the code of descriptor (2):

Geometric intervals: (Å): 1, [1.0–1.2); 2, [1.2–1.5); 3, [1.5–2.0); 4, [2.0–2.5); 5, [2.5–3.0); 6, [3.0–4.5); 7, [4.5–5.0); 8, [5.0–5.5); 9, [5.5–6.0); 10, [6.0–∞).

Topological intervals: 1, [0–1); 2, [1–2); 3, [2–3); 4, [3–4); 5, [4–5); 6, [5–6); 7, [6–7); 8, [7–8); 9, [8–9); 10, [9–∞).

A separate calculation was carried out for each of the six types of descriptors (Table 1). To estimate the predicting quality of QSAR equations, we used the standard 'cross validation' method.[13] As follows from Table 1, the best results were attained for ABG and BBT type descriptors. Tables 2 and 3 present the weight coefficients b_i and the type of descriptors for the best models. Table 4 shows prediction results for the best ABG descriptor based model from those presented in Tables 1 and 2 ($R^2 = 0.89$; cross validated $R^2 = 0.84$).

The calculations for a subset of 22 piperidine derivatives was performed using the BBG and BBT descriptors, since the above results suggest that QSAR equations based on them have the best predictive force. The R^2 values obtained for the best models ($R^2 = 0.986$–0.992) are much higher than those obtained previously ($R^2 = 0.874$).

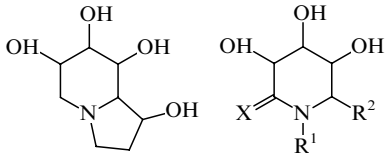
The structures obtained by computer prediction were passed for performing chemical syntheses and testing.

The results obtained in this work allow us to draw the following conclusions:

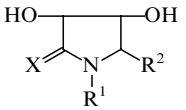
1. The descriptors proposed make it possible to obtain predictive QSAR equations for biological activity with much lower computational requirements than descriptors (1).

2. To find an adequate description of molecules, one should progressively increase the complexity of descriptors used.[1] For example, the 'topological' descriptors allowed us to obtain better QSAR equations than the 'geometric' descriptors.

3. It seems promising to study pairs of molecular fragments as structural descriptors using more complex base fragments.



R^1 = H, Me, Et, Bu, $PhCH_2$, Ac, $MeOC(O)(CH_2)_5$
R^2 = CH_2OH, $CH(OH)CH_2OH$, Me, CO_2H
X = 2H, O



R^1 = H, Me, Bu, $PhCH_2$
R^2 = CH_2OH, $CH(OH)CH_2OH$, CO_2H
X = 2H, O, H + OH

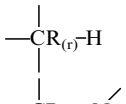
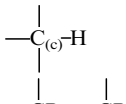
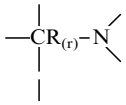
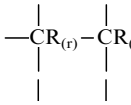
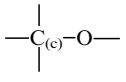
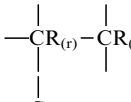
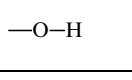
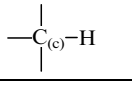
Figure 1 Main types of structures in the training database.

Table 2 View of descriptors of the QSAR equation constructed on BBG geometric descriptors. The training database includes 45 compounds.[a]

Number of descriptor	b_i	The f_1 fragment	The f_2 fragment	The G interval
0	6.166	CONST		
1	49.831	—$C_{(c)}$—	$N_{(r)}$—$C_{(c)}$—	$2.0 \leqslant D_g < 2.5$
2	14.902	—$C_{(c)}$—	—$O_{(c)}$—$C_{(c)}$—	$2.5 \leqslant D_g < 3.0$
3	13.894	—$C_{(r)}$—	—$O_{(c)}$—$CS_{(r)}$—	$1.5 \leqslant D_g < 2.0$
4	−5.777	—$CR_{(c)}$—	$N_{(r)}$—$C_{(r)}$—	$2.0 \leqslant D_g < 2.5$
5	6.126	H—	—$O_{(c)}$—$CR_{(r)}$—	$4.5 \leqslant D_g < 5.0$
6	49.831	$N_{(r)}$—	—$CS_{(r)}$ —$CR_{(r)}$—	$2.5 \leqslant D_g < 3.0$
7	−9.526	—O—	—$CR_{(r)}$—$CR_{(c)}$—	$3.0 \leqslant D_g < 4.5$
8	−7.835	—O—	—$CS_{(r)}$—$CS_{(r)}$ —	$3.0 \leqslant D_g < 4.5$
9	−3.368	—O—	H—$C_{(c)}$—	$2.5 \leqslant D_g < 3.0$

[a] Designations used in Table 2: CR – carbon atoms with R absolute configuration; subscripts: (r) – atom in a ring; (c) – atom in a chain.

Table 3 View of descriptors of the QSAR equation constructed on BBT topological descriptors. The training database includes 45 compounds.[a]

Number of descriptor	b_i	The f_1 fragment	The f_2 fragment	The T interval
0	4.135	CONST		
1	7.031			$1 \leqslant D_t < 2$
2	70.958			$2 \leqslant D_t < 3$
3	11.856			$2 \leqslant D_t < 3$
4	7.638			$7 \leqslant D_t < 8$

[a] For designations, see Table 2.

The authors are grateful to Professor G. V. Grishina and to leading scientist E. A. Smolenskii for a fruitful discussion of the results. This study was financially supported in part by the Russian Foundation for Basic Research (grant nos. 94-01-00041 and 94-03-09323).

References

1 H. V. Waterbeemd and B. Testa, in *Advances in Drug Design*, ed. B. Tesfa, Academic Press, London, 1987, pp. 85–225.
2 *Chemical Applications of Topology and Graph Theory*, ed. R. B. King, Elsevier, New York, 1983, vol. 28.
3 I. V. Svitan'ko, M. I. Kumskov, I. L. Zyryanov and I. A. Suslov, *Mendeleev Commun.*, 1994, 161.
4 I. V. Svitan'ko, M. I. Kumskov, I. L. Zyryanov, L. I. Khmel'nitskii, L. I. Suvorova, A. N. Kravchenko, T. B. Markova, O. V. Lebedev, G. A. Orekhova and S. V. Belova, *Mendeleev Commun.*, 1995, 49.
5 M. I. Kumskov, I. L. Zyryanov and I. V. Svitan'ko, *Pattern Recognition and Image Analysis* (in press).
6 M. I. Kumskov, E. A. Smolenskii, L. A. Ponomareva, D. F. Mityushev and N. S. Zefirov, *Dokl. Akad. Nauk*, 1994, **336**, 64 [*Dokl. Chem. (Engl. Transl.)*, 1994, **336**, 83].
7 M. I. Kumskov, L. A. Ponomareva, E. A. Smolenskii, D. F. Mityushev and N. S. Zefirov, *Izv. Akad. Nauk, Ser. Khim.*, 1994, 1391 (*Russ. Chem. Bull.*, 1994, **43**, 1317).
8 D. F. Mityushev and M. I. Kumskov, *Organizatsiya pol'zovatel'skogo interfeisa v programmakh poiska zavisimostei 'struktura-svoistvo' khimicheskikh soedinenii (Organising the user interface in programs for searching 'structure-property' relationships for chemical compounds)*, N. D. Zelinsky Institute of Organic Chemistry, RAS, Moscow, 1995, deposited in VINITI 01.02.95, No. 288-V95 (in Russian).
9 G. M. Makeev and M. I. Kumskov, *Gibkaya metodika postroeniya deskriptorov trekhmernykh struktur (A versatile procedure for constructing descriptors of three-dimensional structures)*, N. D. Zelinsky Institute of Organic Chemistry, RAS, Moscow, 1995, deposited in VINITI 01.02.95, No. 287-V95 (in Russian).
10 M. I. Kumskov and D. F. Mityushev, *Pattern Recognition and Image Analysis* (submitted for publication).
11 G. W. J. Fleet, A. Karpas, R. A. Dwek, L. E. Fellows, A. S. Tyms, S. Petursson, S. K. Namgoong, N. G. Ramsden, P. W. Smith, J. C. Son, F. Wilson, D. R. Witty, G. S. Jacob and T. W. Rademacher, *FEBS*, 1988, **237**, 128.

Table 4 Approximated and predicted values for the best ABG based model given in Table 2.

Number of compound according to ref. 11	Appoximated value of anti-HIV activity	Predicted value of anti-HIV activity	Error
16	75.0	70.8	4.1
63	0.0	−0.3	0.3
62	0.0	9.7	−9.7
55	10.0	20.8	−10.8
36	100.0	84.8	15.1
56	100.0	111.0	−11.0
57	100.0	94.2	5.7
17	8.0	6.2	1.7
35	25.0	14.8	10.1
7	15.0	9.8	5.1
33	0.0	6.2	−6.2
38	50.0	17.9	32.0
34	25.0	22.9	2.0
11	0.0	−0.2	0.2
51	0.0	1.5	−1.5
32	3.0	−3.6	6.6
31	0.0	−7.1	7.1
42	0.0	−6.5	6.5
12	16.0	14.4	1.5
53	0.0	−1.8	1.8
5	0.0	9.1	−9.1
37	8.0	14.1	−6.1
2	50.0	51.3	−1.3
1	0.0	2.1	−2.1
45	25.0	29.1	−4.1
46	0.0	5.0	−5.0
47	75.0	76.9	−1.9
15	38.0	14.4	23.5
39	0.0	17.9	−17.9
48	25.0	29.9	−4.9
14	19.0	3.4	15.5
40	25.0	18.3	6.6
58	0.0	−0.2	0.2
41	0.0	−3.6	3.6
13	0.0	7.9	−7.9
52	0.0	10.8	−10.8
10	0.0	3.4	−3.4
8	0.0	3.4	−3.4
9	0.0	−4.7	4.7
60	0.0	1.5	−1.5
61	0.0	−0.4	0.4
59	0.0	20.8	−20.8
6	0.0	3.4	−3.4
4	0.0	9.1	−9.1
3	0.0	2.7	−2.7

12 E. A. Smolenskii, G. V. Grishina, G. M. Makeev and N. S. Zefirov, *Dokl. Akad. Nauk*, 1993, **332**, 603 (in Russian).
13 M. A. Sharaf, D. A. Ilamen and B. R. Koval'skii, *Khemometrika (Chemometrics)*, Khimiya, Leningrad, 1989 (in Russian).

Received: Moscow, 22nd May 1995
Cambridge, 7th July 1995; Com. 5/03639I