Available online at www.sciencedirect.com

SCIENCE DIRECT®

Bioorganic & Medicinal Chemistry 13 (2005) 6830–6835

Bioorganic & Medicinal Chemistry

QSAR of the testosterone binding globulin affinity by means of correlation weighting of local invariants of the graph of atomic orbitals

Ivan Raska, Jr.[a,b] and Andrey Toropov[c,*]

[a]3rd Medical Department, 1st Faculty of Medicine, Charles University in Prague, Unemocnice 1, 128 08 Prague 2, Czech Republic
[b]Laboratory of Molecular Biology, Centro Catullo e Daniela Borgomainerio, Istituto di RicercheFarmacomogiche "Mario Negri," via Eritrea 62, 20157 Milano, Italy
[c]Uzbek Academy of Sciences Institute of Geology & Geophysics, Khodzhibaev street 49, Tashkent 700041, Uzbekistan

Received 20 June 2005; revised 23 July 2005; accepted 25 July 2005
Available online 22 September 2005

Abstract—Numerical values of the testosterone binding globulin affinity have been modeled as a mathematical function of molecular structure in two versions of molecular structure elucidation: first, by hydrogen-filled molecular graphs (HFG); second, by the so-called graphs of atomic orbitals (GAO). Increased orders of Morgan extended connectivity in the HFG and GAO have been examined as local invariants. Using optimisation of the correlation weights of the above-mentioned invariants, quantitative structure–activity relationships (QSAR) have been obtained. Best statistical characteristics of these QSARs are derived in the case of the Morgan extended connectivity of first order in the GAO. They are as follows: $n = 11$, $r^2 = 0.6540$, $s = 0.824$, $F = 17$ (training set); $n = 9$, $r^2 = 0.8791$, $s = 0.388$, $F = 51$ (test set).

1. Introduction

Testosterone binding globulin (TeBG) affinity is an important biological parameter from the point of view of endocrinology.[1–3] Singh and co-workers have reported results of QSAR (quantitative structure–activity relationships) analysis based on quantum chemical data (absolute hardness and electronegativity[3]).

Recently, the so-called optimization of correlation weights of local graph invariants (OCWLI) has been tested as a tool of the QSAR analysis.[4,5] The OCWLI may be based on the hydrogen-filled molecular graphs (HFG), as well as on the graph of atomic orbitals (GAO).[6,7]

The aim of the present study was to carry out a comparative QSAR analysis of the data taken from Ref. 3, based on the OCWLI of the HFG and the GAO.

2. Methods

Descriptors used in this QSAR study are defined as

$$\mathrm{DCW}(G,x) = \sum_{k=1}^{n} \mathrm{CW}(V_k) \cdot \mathrm{CW}(^{x}\mathrm{EC}_k), \qquad (1)$$

where G is the type of molecular graph, that is, the HFG or the GAO; x is the order of Morgan extended connectivity (EC); V_k are atoms in the case of HFG (H,C,O) and atomic orbitals in the case of the GAO ($1s^1$, $2s^2$, $2p^2$, etc.); and $\mathrm{CW}(V_k)$ and $\mathrm{CW}(^{x}\mathrm{EC}_k)$ are correlation weights of V_k and $^{x}\mathrm{EC}_k$, respectively.

The GAO can be obtained from the HFG by the following scheme:

First, each atom in the HFG should be replaced by a group of atomic orbitals (AO), according to the description given in Table 1.

Second, the (0,1)-adjacency matrix should be constructed according to the following rules: (a) size of the matrix

Keywords: QSAR; Testosterone binding globulin affinity; Graph of atomic orbitals.

* Corresponding author. E-mail: aatoropov@yahoo.com

0968-0896/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2005.07.059

is defined as the number of AO present in the given structure; the (*i*,*j*) element of the matrix is defined as 1 if (and only if) *i*th and *j*th AO are from different atoms in the HFG and if (and only if) these atoms are connected each to other in the HFG; otherwise, the (*i*,*j*) element is defined as 0.

Table 1. AO groups on the chemical elements that take place in compounds under consideration

Chemical element	AO group of the chemical element
H	$1s^1$
C	$1s^2,2s^2,2p^2$
O	$1s^2,2s^2,2p^4$

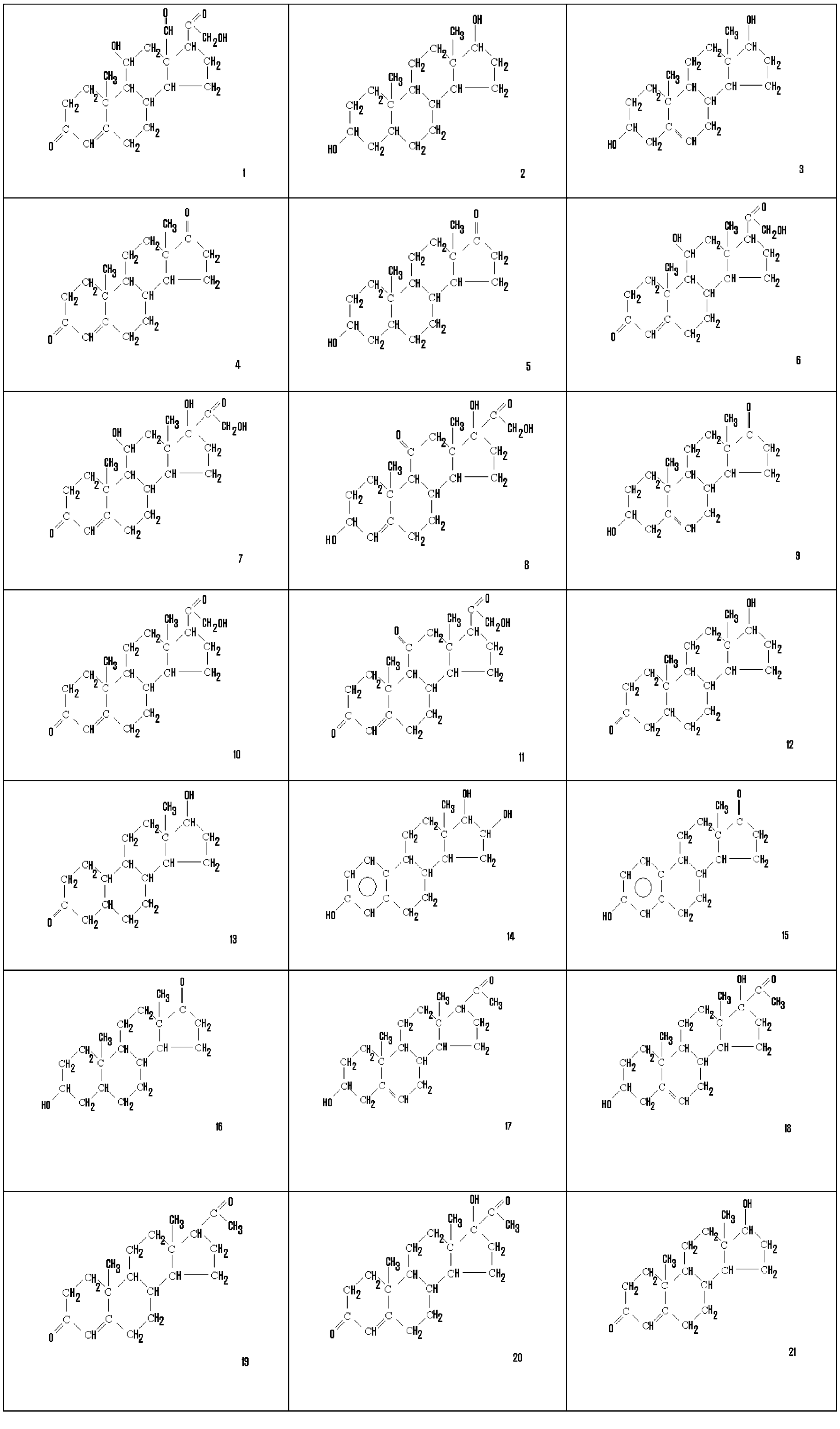


Figure 1. Molecular structure of 21 compounds under consideration.

A detailed description of the conversion HFG into GAO has been reported in Refs. 6 and 7.

3. Results and discussion

In analyzing biological interactions, stereochemistry is a very important component. However, this investigation is based on molecular topology together with the information on chemical elements that are involved in molecular systems. In other words, molecular geometry (3D) has not been taken into account. Molecular structures of the derivatives of testosterone used in the present QSAR analysis are shown in Figure 1. Probably, the above-mentioned geometrical features of the etiocholanolone (ID 16 in Fig. 1) define the uniqueness of this substance, because there is no other model that is able to predict reasonably well the TeBG value on this compound. It should be noted that androsterone (ID 5), being identical to etiocholanolone from a topology point of view, is not an outlier in the models under consideration. Results discussed below are obtained for 20 testosterone derivatives without an etiocholanolone. Six compounds (ID 4, 5, 13, 15, 16, 18, and 20) were outliers in a two-variable modeling, based on the geometry and quantum chemical features of molecules. This model was described in Ref. 3.

Table 2. Statistical characteristics of QSARs based on different descriptors

Descriptor	Probe	C_1	C_0	Training set, $n = 11$			Test set, $n = 9$		
				r^2	s	F	r^2	s	F
DCW (HFG,^{0}EC)	1	0.206	−15.840	0.6137	0.870	14	0.8458	0.471	38
	2	0.173	−15.889	0.6137	0.870	14	0.8454	0.472	38
	3	0.211	−15.816	0.6137	0.870	14	0.8455	0.471	38
DCW (HFG,^{1}EC)	1	0.357	−37.609	0.9813	0.191	474	0.1716	1.967	1
	2	0.322	−36.723	0.9765	0.214	375	0.2179	1.905	2
	3	0.261	−36.067	0.9682	0.250	274	0.2892	1.879	3
DCW (GAO,^{0}EC)	1	0.039	−14.807	0.6540	0.824	17	0.8791	0.389	51
	2	0.016	−14.784	0.6539	0.824	17	0.8810	0.385	52
	3	0.003	−14.552	0.6538	0.824	17	0.8854	0.373	54
DCW (GAO,^{1}EC)	1	0.142	−27.483	1.0000	0.006	528409	0.8431	1.029	38
	2	0.103	−21.975	0.9999	0.012	113353	0.7870	1.077	26
	3	0.152	−24.923	1.0000	0.009	217044	0.7606	1.296	22

n is the number of compounds in the set; r, s, and F are correlation coefficient, standard error estimation, and Fischer F-ratio, respectively.

Table 3. Training and test sets of testosterone derivatives, numerical values of DCW (GA0,0), and experimental and calculated values of the TeBG

No	ID[a]	Structures	DCW (GAO,^{0}EC)	$TeBG_{expr}$	$TeBG_{calc}$	expr. − calc.
Training set						
1	1	Aldosterone	245.096	−5.320	−5.385	0.065
2	3	Androstenediol	168.996	−9.170	−8.330	−0.840
3	4	Androstenedione	174.338	−7.460	−8.123	0.663
4	5	Androsterone	158.300	−7.140	−8.744	1.604
5	11	Deoxycortisol	211.046	−7.200	−6.703	−0.497
6	12	Dihydrotestosterone	158.300	−9.740	−8.744	−0.996
7	13	Estradiol	146.695	−8.830	−9.193	0.363
8	14	Estriol	201.317	−6.630	−7.079	0.449
9	18	17-Hydroxypregnenolone	200.290	−6.360	−7.119	0.759
10	19	Progesterone	203.543	−6.940	−6.993	0.053
11	21	Testosterone	171.667	−9.200	−8.226	−0.974
Test set						
1	2	Androstanediol	155.629	−9.110	−8.787	−0.323
2	6	Corticosterone	208.376	−6.340	−6.746	0.406
3	7	Cortisol	207.794	−6.200	−6.768	0.568
4	8	Cortisone	207.794	−6.410	−6.768	0.358
5	9	Dehydroepiandrosterone	171.667	−7.810	−8.166	0.356
6	10	Dehydroxycorticosterone	191.185	−7.380	−7.411	0.031
7	15	Estrone	186.798	−8.170	−7.581	−0.589
8	17	Pregnenolone	200.872	−7.140	−7.036	−0.104
9	20	17-Hydroxyprogesterone	202.961	−6.990	−6.955	−0.035

[a] Identifying number of the compound in Figure 1.

Separation into the training and test sets was done in a random manner. The only limitation is as follows: all invariants must take part in the training set. Variations of this separation produce different statistical characteristics of the models. However, in fact, statistical quality is stable if all invariants take part in training procedures.

Correlations of biological parameters with different descriptors, as a rule, are not high. Mainly, they are caused by the existence of a large number of conditions that are able to influence the biochemical process. The TeBG is not an exception. Under such circumstances, statistical characteristics of models based on the DCW(HFG,0) and DCW(GAO, 0), which are listed in Table 2, should be estimated as reasonably good ones, whereas models based on the DCW(HFG,1) and DCW(GAO,1) are examples of overtraining (overfitting); i.e., it is a situation when good enough statistical characteristics of models on training set are accompanied by low correlations on the test set. In other words, overtraining is a situation when too many parameters are involved in optimization, and, as a result, the model is too sensitive to all fragments of the architecture of molecules from the training set. In fact, some of these fragments are not images of significant agents of the TeBG.

The best model of the TeBG is as follows:

$$\mathrm{TeBG} = -14.807 + 0.0387 \cdot \mathrm{DCW(GA0, 0)} \qquad (2)$$

$$n = 11, r^2 = 0.6540, s = 0.824, F = 17(\text{training set})$$

$$n = 9, r^2 = 0.8791, s = 0.388, F = 51(\text{test set})$$

Calculation of the TeBG with Eq. 2 is given in Table 3. Correlation weights of the first OCWLI probe have been used here. A graphical presentation of this model is shown in Figure 2 (training set) and Figure 3 (test set).

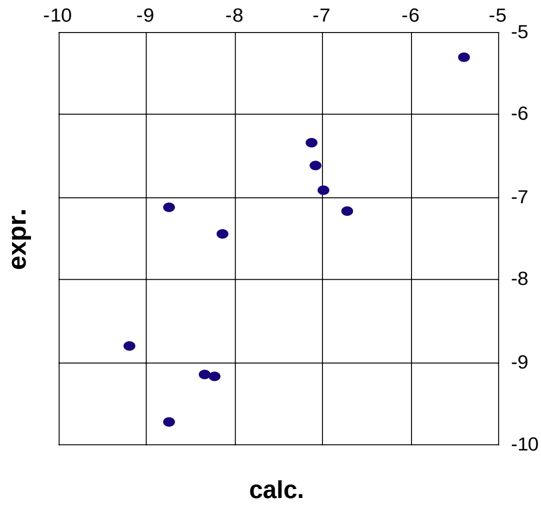


Figure 2. Plot of experimental versus calculated values of the TeBG on the training set.

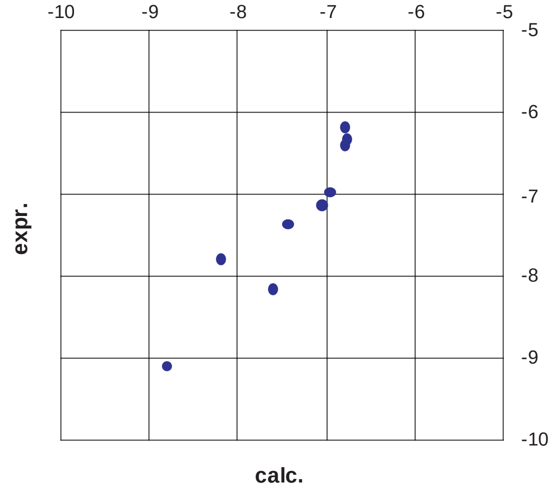


Figure 3. Plot of experimental versus calculated values of the TeBG on the test set.

In Table 4, numerical values of correlation weights, over all three probes of the OCWLI for the DCW(GAO,0), obtained by the Monte Carlo method optimization procedure are listed. An example of the calculation of the DCW(GAO,0) for testosterone (ID 21) is shown in Table 5.

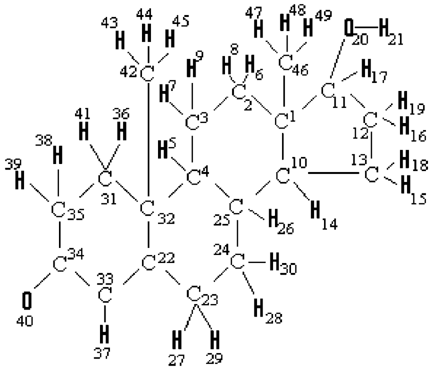
4. Conclusions

Optimization of the correlation weights of local graph invariants gives a reasonably good prediction for the testosterone binding globulin affinity, in the case of Morgan extended connectivity of zero order (in other words, it is the 'classical' vertex degree in molecular graph). Even in the case of an extended connectivity of first order, overtraining takes place. The best accuracy of prediction takes place for optimization of correlation weights in the graph of atomic orbitals. Thus, information on the structure of chemical elements (atoms) is use-

Table 4. Correlation weights on three probes of the OCWLI based on the DCW (GAO,0)

GAO invariant	Optimization of correlation weights of local graph invariants in the GAO		
	Probe 1	Probe 2	Probe 3
AO_j			
$1s^1$	−0.00773	−0.00731	−0.00547
$1s^2$	4.64259	17.13569	87.34067
$2s^2$	5.82442	11.06097	43.24330
$2p^2$	4.39740	5.71119	9.05563
$2p^4$	0.54829	0.42205	0.13249
^0EC_j			
3	1.77528	1.70519	1.75346
4	0.03433	0.05714	0.14885
6	0.84709	0.93121	1.07359
7	2.00007	2.16696	2.35576
8	−0.00968	−0.00507	−0.00886
9	0.00904	0.00574	−0.00385
10	1.12133	1.19979	1.25657
12	1.05675	1.12095	1.12907

Table 5. Calculation of the DCW(GAO,0) on testosterone (ID 21) with correlation weights of the first OCWLI probe (Table 4)



Atom (A_k)	k	^0EC_k	AO_j	j	^0EC_j	$CW(AO_j)$	$CW(^0EC_j)$
C	1	4	$1s^2$	1	12	4.64259	1.05675
			$2s^2$	2	12	5.82442	1.05675
			$2p^2$	3	12	4.39740	1.05675
C	2	4	$1s^2$	4	8	4.64259	−0.00968
			$2s^2$	5	8	5.82442	−0.00968
			$2p^2$	6	8	4.39740	−0.00968
C	3	4	$1s^2$	7	8	4.64259	−0.00968
			$2s^2$	8	8	5.82442	−0.00968
			$2p^2$	9	8	4.39740	−0.00968
C	4	4	$1s^2$	10	10	4.64259	1.12133
			$2s^2$	11	10	5.82442	1.12133
			$2p^2$	12	10	4.39740	1.12133
H	5	1	$1s^1$	13	3	−0.00773	1.77528
H	6	1	$1s^1$	14	3	−0.00773	1.77528
H	7	1	$1s^1$	15	3	−0.00773	1.77528
H	8	1	$1s^1$	16	3	−0.00773	1.77528
H	9	1	$1s^1$	17	3	−0.00773	1.77528
C	10	4	$1s^2$	18	10	4.64259	1.12133
			$2s^2$	19	10	5.82442	1.12133
			$2p^2$	20	10	4.39740	1.12133
C	11	4	$1s^2$	21	10	4.64259	1.12133
			$2s^2$	22	10	5.82442	1.12133
			$2p^2$	23	10	4.39740	1.12133
C	12	4	$1s^2$	24	8	4.64259	−0.00968
			$2s^2$	25	8	5.82442	−0.00968
			$2p^2$	26	8	4.39740	−0.00968
C	13	4	$1s^2$	27	8	4.64259	−0.00968
			$2s^2$	28	8	5.82442	−0.00968
			$2p^2$	29	8	4.39740	−0.00968
H	14	1	$1s^1$	30	3	−0.00773	1.77528
H	15	1	$1s^1$	31	3	−0.00773	1.77528
H	16	1	$1s^1$	32	3	−0.00773	1.77528
H	17	1	$1s^1$	33	3	−0.00773	1.77528
H	18	1	$1s^1$	34	3	−0.00773	1.77528
H	19	1	$1s^1$	35	3	−0.00773	1.77528
O	20	2	$1s^2$	36	4	4.64259	0.03433
			$2s^2$	37	4	5.82442	0.03433
			$2p^4$	38	4	0.54829	0.03433
H	21	1	$1s^1$	39	3	−0.00773	1.77528
C	22	3	$1s^2$	40	9	4.64259	0.00904
			$2s^2$	41	9	5.82442	0.00904
			$2p^2$	42	9	4.39740	0.00904
C	23	4	$1s^2$	43	8	4.64259	−0.00968
			$2s^2$	44	8	5.82442	−0.00968
			$2p^2$	45	8	4.39740	−0.00968
C	24	4	$1s^2$	46	8	4.64259	−0.00968
			$2s^2$	47	8	5.82442	−0.00968
			$2p^2$	48	8	4.39740	−0.00968
C	25	4	$1s^2$	49	10	4.64259	1.12133
			$2s^2$	50	10	5.82442	1.12133
			$2p^2$	51	10	4.39740	1.12133

Table 5 (*continued*)

Atom (A_k)	k	^0EC_k	AO_j	j	^0EC_j	CW(AO_j)	CW(^0EC_j)
H	26	1	$1s^1$	52	3	−0.00773	1.77528
H	27	1	$1s^1$	53	3	−0.00773	1.77528
H	28	1	$1s^1$	54	3	−0.00773	1.77528
H	29	1	$1s^1$	55	3	−0.00773	1.77528
H	30	1	$1s^1$	56	3	−0.00773	1.77528
C	31	4	$1s^2$	57	8	4.64259	−0.00968
			$2s^2$	58	8	5.82442	−0.00968
			$2p^2$	59	8	4.39740	−0.00968
C	32	4	$1s^2$	60	12	4.64259	1.05675
			$2s^2$	61	12	5.82442	1.05675
			$2p^2$	62	12	4.39740	1.05675
C	33	3	$1s^2$	63	7	4.64259	2.00007
			$2s^2$	64	7	5.82442	2.00007
			$2p^2$	65	7	4.39740	2.00007
C	34	3	$1s^2$	66	9	4.64259	0.00904
			$2s^2$	67	9	5.82442	0.00904
			$2p^2$	68	9	4.39740	0.00904
C	35	4	$1s^2$	69	8	4.64259	−0.00968
			$2s^2$	70	8	5.82442	−0.00968
			$2p^2$	71	8	4.39740	−0.00968
H	36	1	$1s^1$	72	3	−0.00773	1.77528
H	37	1	$1s^1$	73	3	−0.00773	1.77528
H	38	1	$1s^1$	74	3	−0.00773	1.77528
H	39	1	$1s^1$	75	3	−0.00773	1.77528
O	40	1	$1s^2$	76	3	4.64259	1.77528
			$2s^2$	77	3	5.82442	1.77528
			$2p^4$	78	3	0.54829	1.77528
H	41	1	$1s^1$	79	3	−0.00773	1.77528
C	42	4	$1s^2$	80	6	4.64259	0.84709
			$2s^2$	81	6	5.82442	0.84709
			$2p^2$	82	6	4.39740	0.84709
H	43	1	$1s^1$	83	3	−0.00773	1.77528
H	44	1	$1s^1$	84	3	−0.00773	1.77528
H	45	1	$1s^1$	85	3	−0.00773	1.77528
C	46	4	$1s^2$	86	6	4.64259	0.84709
			$2s^2$	87	6	5.82442	0.84709
			$2p^2$	88	6	4.39740	0.84709
H	47	1	$1s^1$	89	3	−0.00773	1.77528
H	48	1	$1s^1$	90	3	−0.00773	1.77528
H	49	1	$1s^1$	91	3	−0.00773	1.77528

DCW(GAO,0) = 171.66722.

ful for the prediction of the TeBG activity under consideration.

References and notes

1. Hu, J.-Y.; Aizawe, T. *Water Res.* **2003**, *37*, 1213.
2. Delisle, R. K.; Yu, S. J.; Nair, A. V.; Welsh, W. J. *J. Mol. Graph. Model.* **2001**, *20*, 155.
3. Singh, P. P.; Srivastava, H. K.; Pasha, F. A. *Bioorg. Med. Chem.* **2004**, *12*, 171.
4. Toropov, A. A.; Schultz, T. W. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 560.
5. Toropov, A. A.; Roy, K. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 179.
6. Toropov, A. A.; Toropova, A. P. *J. Mol. Struct. (Theochem)* **2001**, *538*, 287.
7. Toropov, A. A.; Toropova, A. P. *J. Mol. Struct. (Theochem)* **2003**, *637*, 1.