MATCH
Communications in Mathematical
and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **55** (2006) 315-329

ISSN 0340 - 6253

QSAR Modeling of Steroid Hormones

Adina Costescu, Cristina Moldovan and Mircea V. Diudea

Faculty of Chemistry and Chemical Engineering
400028 Cluj, Romania

(Received September 9, 2005)

Abstract

A quantitative structure-activity relationship (QSAR) study on steroid hormones was performed in order to find a relation between some topochemical and electronic parameters and two measured activities: hormones binding affinity to the gestagenic receptor and corticosteroid binding globulin (CBG). Principal component analysis (PCA) was used for data reduction. The loading plot values gave information about descriptors showing significant correlation with the biological property. The models revealed the main feature describing the binding activity of these compounds is a partial charge-based descriptor, a measure of the molecular electronic properties, used within an auto-correlation weighting scheme of substituted positions descriptors. The linear regression equations allowed a pertinent prediction of the biological activities herein studied. Comparison with literature data shows the quality of the proposed models being comparable to the models provided by some more sophisticated 3D-based methods.

INTRODUCTION

Quantitative structure-activity relationship (QSAR) techniques have become indispensable in all aspects of research regarding the molecular interpretation of biological properties.[1] It is obvious that physical, chemical, or biological properties of a compound depend on the three-dimensional (3D) arrangement of atoms in the molecule. The ability to produce quantitative correlation between 3D structure of molecules and their biological activity is important in deciding upon the synthetic ways of bioactive chemicals.[2]

Biological activity of steroids varies considerably with seemingly small structural changes. This important molecular family presents very challenging features for any prediction method, particularly due to the relatively low flexibility of the cyclopentanoperhydrophenanthrene skeleton. Due to this reason, many excellent QSAR models based on 2D properties, such as topological descriptors, have a quality comparable to the models provided by some more complex 3D-based methods.[3,4]

In this paper, we try to identify those aspects of molecular structure that can be relevant to a particular biological activity of some steroid derivatives: the receptor-binding affinity.

DATA SET

The data set of 31 (androstan) steroids ASs, showing corticosteroid binding globulin (CBG) affinity, was taken from the publications by Dunn *et al*.[5] and Mickelson *et al*.[6] This set has repeatedly served as a benchmark in evaluating the performance of new QSAR methods. The structures used in this work have been carefully checked in order to avoid any further propagation of errors. Qualitatively, light substituents, such as oxygen and hydroxyl, at position 17 of steroid skeleton, seem to increase the CBG activity[7,8] whereas the presence of the bulky chain, such as $COCH_2OH$, enhances the activity.[6]

A set of testosterone steroids [TSs] comprising of 39 molecules[9], tested for the binding affinity to gestagenic receptor was also considered. Testosterone derivatives are relatively rigid systems with the exception of the side chains in position 17.

General scaffold of the ASs herein investigated is shown in Figure 1 while the description of each molecule with its activity is given in Table 1.

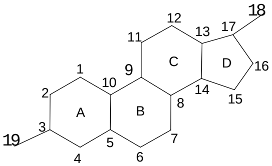
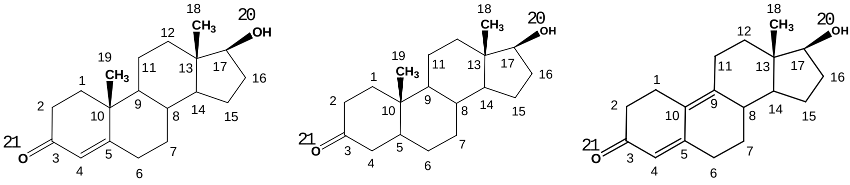


Figure 1. Androstan scaffold

Table 1. Androstan derivatives ASs:

	Compound	$\log P_i$ obs		Compound	$\log P_i$ obs
1	aldosterone	6.279	17	pregnenolone	5.255
2	androstanediol	5.000	18	17-hydroxypregnenolone	5.000
3	androstenediol	5.000	19	progesterone	7.380
4	androstenedione	5.763	20	17-hydroxyprogesterone	7.740
5	androsterone	5.613	21	testosterone	6.724
6	corticosterone	7.881	22	prednisolone	7.512
7	cortisol	7.881	23	cortisol 21-acetate	7.553
8	cortisone	6.892	24	4-pregnene-3,11,20-trione	6.779
9	dehydroepiandrosterone	5.000	25	epicorticosterone	7.200
10	deoxycorticosterone	7.653	26	19-nortestosterone	6.144
11	deoxycortisol	7.881	27	16R,17-dihydroxy-4-pregnene-3,20-dione	6.247
12	dihydrotestosterone	5.919	28	16-methyl-4-pregnene-3,20-dione	7.120
13	estradiol	5.000	29	19-norprogesterone	6.817
14	estriol	5.000	30	11β,17,21-trihydroxy-2R-methyl-4-pregnene-3,20-dione	7.688
15	estrone	5.000	31	11β,17,21-trihydroxy-2R-methyl-9R-fluoro-4-pregnene-3,20-dione	5.797
16	etiocholanolone	5.255			

General structures of the TS set are presented in Figure 2. Among the 39 molecules (Table 1), 27 are 4-androsten-3-one derivatives (Figure 2a), 4 are 5αH-androstan-3-one derivatives (Figure 2b) and 8 are 4,9-androstadien-3-one derivatives (Figure 2c).



(a) 17β-OH-4-androstene-3-one(testosterone)

(b) 17β-OH-5αH-androstene-3-one (5αH-testosterone)

(c) 17β-OH-4,9-androstadiene-3-one (9Δ-testosterone)

Figure 2. General structures of the testosterone derivatives TS set

The description of each TS molecule, along with the corresponding activity, is given in Table 2.

Table 2. Testosterone derivatives TSs:

	Compound	logA_i obs		Compound	logA_i obs
1	18-Me-NE[1]	2.55	20	18-Me-NT[1]	1.53
2	11=CH_2-18-Me-NE[1]	2.54	21	18-Me[3]	1.42
3	11=CH_2,17α-CCH-NT[1]	2.48	22	7α-Me-NT[1]	1.34
4	7α,17α-diMe-NT[1]	2.33	23	3-deoxi-17α-Me-NT[1]	1.3
5	11β-Cl-NE[1]	2.32	24	E[1]	1.28
6	7α,17α-diMe[3]	2.3	25	17α-CCH-19-nor[2]	1.23
7	NE[1]	2.19	26	9Δ-testosterone[3]	1.23
8	11β-Me,17α-CCH-18-Me-NT[1]	2.18	27	3-deoxi-11=CH_2-18-Me-NE[1]	1.20
9	17α-Me-NT[1]	2.00	28	17α-CCH18-Et[3]	1.04
10	11=CH_2,17α-CCH-18Me-19-nor[2]	1.99	29	3-deoxi-7-Meα-NE[1]	0.95
11	17α-CH_2Cl-NT[1]	1.93	30	3-deoxi-NE[1]	0.85
12	17α-Et-NT[1]	1.89	31	NT[1]	0.78
13	18-Et-NE[1]	1.86	32	17α-CCH[2]	0.70
14	17α-Me[3]	1.85	33	18-Et[3]	0.66
15	17α-Pr-18-norT[1]	1.83	34	18-Et-NT[1]	0.65
16	17α-CCH-18-Me[3]	1.83	35	4-Cl-17α-Me-T[1]	0.11
17	17α-CH=CH_2-NT[1]	1.81	36	Testosterone[1]	-0.22
18	17αCH_2CMeCHMe=CH_2-NT[1]	1.68	37	5Hα-Testosterone[2]	-0.30
19	17α-CCH[3]	1.62	38	17α-CH_2CN-T[1]	-0.52
			39	4-Cl-11β-OH-17α-Me-T[1]	-1.00

[1] testosteone derivatives (Figure 2a)

[2] 5Hα-testosterone derivatives (Figure 2b)

[3] 9Δ-testosterone derivatives (Figure 2c)

NE – norethysterone (ethysterone outof C19), E – ethysterone (17α-ethynil-17β-hydroxy-4androsten-3-one), NT – nortestosterone (testosterone outof C19)

CALCULATION OF MOLECULAR DESCRIPTORS

Most of the applications of molecular descriptors have been dedicated to QSAR studies because of the great importance for biology of the structure-activity relationship.[10] The computation of such descriptors is accessible by using available software products. The complete set of molecular descriptors used in this study (some of them will be defined below), was calculated by TOPOCLUJ[11] (electronic descriptors – partial charges) and by Dragon[12] software packages. The structures were optimized by using the semiempirical PM3 Hamiltonian, available in HyperChem.[13]

The subset of electronic parameters includes molecular descriptors derived on atomic partial charges. Within TOPOCLUJ program, the partial charges Ch_i are calculated as follows:[11]

$$Ch_{i,j} = \log(S_j / S_i)^{1/(d_{i,j})^2} \qquad (1)$$

$$Ch_i = \sum_j ch_{i,j} \qquad (2)$$

In the above, S_i, S_j represent the Sanderson group electronegativities calculated for the hydride groups (*i.e.*, the heavy atoms with their surrounding hydrogen atoms) in the molecule while d_{ij} is the Euclidean distance separating atoms *i* and *j* in a minimal energy optimized chemical structure (HyperChem).[13] $Ch_{i,j}$ is the perturbation of the electronegativity of atom *i* produced by any *j* atom in molecule while Ch_i is the resultant of these perturbations over the atom *i*. For other topological partial charge calculations see refs.[14, 15]

Any steroid compound can be described by these partial charges that characterize both the substituted/unsubstituted positions and the heteroatom (oxygen). On this ground, a flexible global descriptor (*CD*) can be defined as an additive function of autocorrelation weights[16, 17] of the partial charges corresponding to considered atoms *j*:

$$CD = \sum_j c_j \cdot Ch_j \qquad (3)$$

where c_j is the regression coefficient (*i.e.*, the correlation weight) as given by the multivariate regression log($A_i$obs) = f(Ch_j). These “ad-hoc” weightings depend on the set of molecules under consideration and the used local descriptors, as well. Partial charges (Ch_j) were calculated for the following positions of the scaffolds: 3, 10, 11, 13, 17, 18, 19 (Figure 1) and 3, 4, 7, 10, 11, 13, 17, 19, 20, 21 (Figure 2).

Dragon 2.1 software was used to calculate a total of 1600 molecular descriptors, for all the studied compounds. The most relevant of these descriptors, in our studies, were: radial distribution functions (RDF), autocorrelation indices and geometrical descriptors.

Descriptors belonging to the class of radial distribution function[18] are based on the distance distribution in the geometrical representation of the molecule. In addition to interatomic distances in the entire molecule, the RDF provides valuable information about bond distances, ring types, planar and non-planar systems, atom types and other important structural motifs. By using different weighting schemes, which include atom types, electronegativity, atomic mass (*RDF090m*) or van der Waals radii, RDF can be adjusted to give rise to important descriptors in deriving an appropriate QSAR.

The next group of descriptors is based on 2-dimensional autocorrelation functions applied to a molecular graph. Such descriptors express a correlation between numerical values, which can be statistically weighted using various atomic properties, at intervals equal to the given lag value.[19] For example, *MATS1p*-Moran autocorrelation - lag 1 / weighted by atomic polarizabilities; *MATS4e*-Moran autocorrelation - lag 4 / weighted by atomic Sanderson electronegativities.[12] Application of the Sanderson electronegativities as weighting coefficients, takes into account, in some degree, charge distribution inside a molecule.

The geometrical descriptors indicate the size of molecules; they are derived from the three-dimensional coordinates of the atomic nuclei, the atomic masses and/ or the atomic radii in the molecule. A descriptor of this kind, used in our models, is *L/Bw* - length-to-breadth ratio by WHIM.[12]

DATA ANALYSIS

Due to the high complexity of interactions between the receptor molecule and potential inhibitor molecules, it is quite difficult to model TSs and ASs using simple linear regression models.

Principal components analysis *PCA* is a very powerful statistical technique useful to reduce the noise of the data set and to eliminate uncorrelated variables. Loading factors can be used to evaluate the relevant descriptors (i.e., those contributing highly to the data variance). A high loading value indicates that the principal component (PC) is aligned in a direction close to the original descriptor response. Each PC can be examined to determine which descriptors contribute significantly to that PC. Additionally, the relation of the descriptors to each other can be explored. Loading plots or tables can be used to determine which descriptors provide unique information and which ones give similar information.

From 1600 descriptors we obtained 20 (in TSs) and respectively 14 (in ASs) PCs, which account for 98% of the variance. From these selected PCs, we have chosen 3 factors (for each PC), with the greatest loadings, as independent variables. Thus, we drastically reduced the descriptor space, with two orders of magnitude (from 1600 to 60 descriptors for TS set and from 1600 to 42 descriptors for AS set).

The tried models were either simple, monovariate, linear ones, assuming that X and Y are linked in a dependence of the form:

$$Y = b_0 + b_1 \cdot X \tag{4}$$

or multiple linear regression models, as:

$$Y_i = b_0 + \sum_{j=1}^{p} b_j \cdot X_{ij} \tag{5}$$

where Y_i is the dependent variable, X_{ij} are the predictor values ($j = \overline{1, p}$, where $p < n$, n being the number of experiments Y_1, Y_2,...., Y_n), b_j are the regression coefficients and b_0 is a constant. The quality of the models was estimated by: the Pearson correlation coefficient (r), the standard error of estimate (s), the Fischer ratio (F) and the coefficient of variance[20] (CV%). The regression equations were derived by using the STATISTICA 6.0 software package.[21]

The QSAR analysis followed the steps: (1) structure optimization; (2) calculation of molecular descriptors; (2a) multivariate regression, to find the autocorrelation coefficients; (3) Splitting the data set into a training set (for the regression calibration) and a predicting set (for the model validation); (4) principal component analysis (*PCA*); (5) finding a regression function for the model; (6) testing the predictive capability of the model; (7) interpretation of the model.

The step (2a) was described by eq. (3) and the sentence above. In the step (3) selection was performed randomly to ensure a great structural diversity within sets.

The step (5) starts with a monovariate regression, for which the best found descriptor was *CD*, which explains 93.4% (in TS set) and 89.1% (in AS set), see Table 3. It is obvious that the main contribution in explaining the receptor binding affinity is due to the partial charge descriptor (*CD*). Leave-one-out analysis [22, 23] was performed on all subsets in view of finding the outliers (if any). Points which do not fall within any specified error limits are considered outliers (standard residual >2×s).

In both monovariate and bivariate regression (Table 3) compounds 1 (in TSs) and 13 (in ASs) appear to be outliers. These compounds were not included in further analysis. With these outliers removed, we observed a somewhat improved correlation with the same descriptors.

Table 3. Models used and results of cross-validation procedure.

Data sets	Number of observations (n)	Model	Correlation coefficient (r^2) (before LOO)	Outlier structure	Correlation coefficient (r^2) (after LOO)
TS	39	$\log(A_i\text{calc}) = \mathrm{f}(CD)$	0.934	1	0.945
		$\log(A_i\text{calc}) = \mathrm{f}(CD, JGI9)$	0.947	1	0.956
AS	31	$\log(P_i\text{calc}) = \mathrm{f}(CD)$	0.891	13	0.920
		$\log(P_i\text{calc}) = \mathrm{f}(CD, L/Bw)$	0.931	13	0.939

A trivariate regression gave similar results, but no essential improvement of statistics could be noticed.

The results will be further presented on each class of compounds.

1. *TS - SET*

In view of developing the model, we split the TS set into training (n = 26) and prediction (validation) set (n = 12), Table 4.

(A) TRAINING SET (N = 26)

The electronic descriptor *CD,* actually *CDP,* was calculated *de nuovo* on the training set according to eq 3. This is because the correlating weights c_j fit only for a selected property and a given set (in this case, the training set of 26 structures). Table 4 shows the relevant descriptors.

The best models for the set TS are:

Monovariate regression:

$$\log A_i\text{calc} = 1.913 + 0.999 \cdot CDP_i \qquad (6)$$

$n = 26$ $R^2 = 0.903$ s = 0.062 F = 274.46

Bivariate regression:

$$\log A_i\text{calc} = 1.132 + 0.936 \cdot CDP_i + 50.428\ JGI9_i \qquad (7)$$

$n = 26$ $R^2 = 0.934$ s = 0.16 F = 152.02

Multiple regressions:

$$\log A_i\text{calc} = 1.5061 + 106.72 \cdot JGI9_i - 6.211 \cdot MATS4e_i - 16.929 \cdot MATS1p_i + 0.787 \cdot CDP_i \qquad (8)$$

$n = 26$ $R^2 = 0.966$ s = 0.18 F = 148.89

Table 4. Topological descriptors and observed activity for the TS set.

Structure	JGI9	MATS4e	MATS1p	CDP	log A obs
Training set					
1	0.018	-0.099	0.108	-0.090	2.55
2	0.018	-0.092	0.103	0.462	2.54
3	0.015	-0.102	0.107	0.540	2.48
4	0.019	-0.08	0.114	0.340	2.33
5	0.015	-0.086	0.1	0.156	2.32
7	0.015	-0.109	0.113	-0.029	2.19
8	0.018	-0.088	0.103	0.006	2.18
9	0.019	-0.088	0.119	0.079	2
11	0.015	-0.078	0.106	0.058	1.93
12	0.015	-0.125	0.114	0.082	1.89
13	0.015	-0.09	0.103	0.041	1.86
15	0.018	-0.126	0.107	-0.011	1.83
17	0.015	-0.117	0.113	0.025	1.81
18	0.014	-0.08	0.096	-0.095	1.68
20	0.019	-0.076	0.119	-0.808	1.53
22	0.015	-0.074	0.119	-0.461	1.34
23	0.016	-0.135	0.119	-1.051	1.3
24	0.015	-0.109	0.113	-0.225	1.28
27	0.011	-0.123	0.103	-0.535	1.2
9	0.01	-0.144	0.108	-0.834	0.95
30	0.01	-0.159	0.113	-1.119	0.85
31	0.015	-0.083	0.125	-0.852	0.78
34	0.015	-0.068	0.114	-0.865	0.65
35	0.014	-0.088	0.101	-2.185	0.11
36	0.015	-0.087	0.119	-2.636	-0.22
38	0.012	-0.037	0.108	-2.053	-0.52
39	0.014	-0.057	0.128	-2.623	-1
Validation set					
6	0.019	-0.097	0.113	0.423	2.3
10	0.018	-0.089	0.103	0.411	1.99
14	0.019	-0.11	0.119	0.142	1.85
16	0.014	-0.068	0.109	-0.082	1.83
19	0.015	-0.138	0.112	0.076	1.62
21	0.019	-0.1	0.119	-0.721	1.42
25	0.015	-0.108	0.125	-0.069	1.23
26	0.015	-0.105	0.113	-0.620	1.23
28	0.015	-0.115	0.103	-0.067	1.04
32	0.015	-0.105	0.108	-1.792	0.7
33	0.015	-0.089	0.113	-0.784	0.66
37	0.015	-0.084	0.12	-2.438	-0.3

(B) PREDICTION SET ($N = 12$)

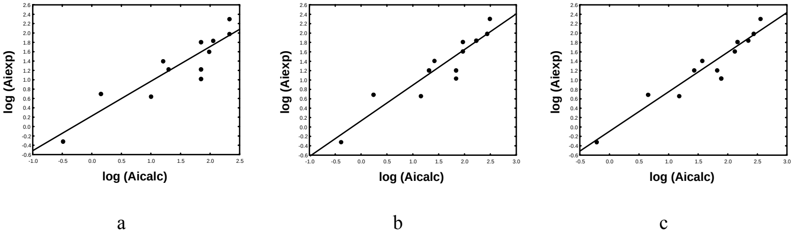
Any QSAR model must be validated on an external predicting set. The calculation of *CDP* in the predicting set used the actual partial charges and the $c_j^{'}$ parameters priory generated for the training set (see Table 5). We assume the biological activity in the predicting set is unknown.

Table 5. TS predicting set (n = 12)

Structure	log $A_i$obs	log $A_i$calc (eq 6)	log $A_i$calc (eq 7)	log $A_i$calc (eq 8)
6	2.300	2.327	2.486	2.538
10	1.990	2.315	2.423	2.418
14	1.850	2.050	2.223	2.340
16	1.830	1.830	1.963	2.154
19	1.620	1.985	1.957	2.103
21	1.420	1.200	1.417	1.556
25	1.230	1.842	1.821	1.415
26	1.230	1.299	1.306	1.805
28	1.040	1.844	1.823	1.875
32	0.700	0.146	0.211	0.651
33	0.660	0.990	1.153	1.169
37	-0.300	-0.490	-0.392	-0.218
R^2		**0.82**	**0.86**	**0.91**
CV%		17.98	17.14	15.057

In the validation set, the results are in agreement with those in training set, the best prediction being obtained from the model by eq.8

Plots of calculated versus observed values are shown in Figure 3 a-c: (a) calculated *cf.* eq.6 (b) calculated *cf.* eq.7 and (c) calculated *cf.* eq.8.



(a) log ($A_i$obs) = 0.22687 + 0.74101 × log ($A_i$calc); R = 0.90769.

(b) log ($A_i$obs) = 0.13087 + 0.75976 × log ($A_i$calc); R = 0.92747.

(c) log ($A_i$obs) = -0.0930 + 0.84245 × log ($A_i$calc); R = 0.95477.

Figure 3. The plots of experimental *vs*. calculated values for the receptor binding affinity of TSs.

For the second set (AS) we used the same algorithm as for the TS set.

2. *AS - SET*

We split the AS set in training ($n = 20$) and predicting set ($n = 11$), as shown in Table 6.

(A) TRAINING SET ($N = 20$)

Table 6. Topological descriptors and observed partition coefficient log P for the AS set.

Structure	L/Bw	RDF090m	CDP	log P obs.
		Training test		
1	7.3	3.17	-1.722	5
2	6.5	3.127	-2.101	5
4	6.5	3.505	-2.062	5
5	6.9	1.755	-2.456	5
7	6.3	2.803	-2.188	5
8	4.3	0.191	-1.950	5.255
9	7.1	1.91	-2.087	5.255
12	7	4.01	-0.880	5.797
13	6.1	3.497	-0.187	5.919
14	7.2	0.61	-1.212	6.144
17	6.9	2.729	-0.336	6.724
18	6	3.288	0.492	6.779
20	7.1	3.976	-0.243	6.892
22	6.6	2.525	-0.046	7.2
23	7.1	0.942	0.085	7.38
25	9.5	1.704	0.205	7.553
26	9.1	0.673	0.006	7.653
28	7.8	1.122	0.317	7.74
29	8.7	2.31	0.040	7.881
30	9	1.077	0.309	7.881
		Validation set		
3	6.7	2.669	-2.350	5
6	6.6	0.811	-2.117	5
10	6.3	1.338	-2.184	5.613
11	6.2	1.272	-0.479	5.763
15	6.9	1.986	0.400	6.247
16	6.4	1.759	-0.716	6.279
19	9.1	0.406	-0.810	6.817
21	6.9	2.014	0.046	7.12
24	5.9	2.777	0.110	7.512
27	6.9	3.512	0.675	7.688
31	8.7	1.423	0.225	7.881

The best models for the training set (AS-set)

Monovariate regression:

$$\log P_i\text{calc} = 7.236 + 1.033\, CDP_i \qquad (9)$$

$n = 20$ $R^2 = 0.903$ s = 0.10 F = 159.99

Bivariate regression:

$$\log P_i\text{calc} = 5.737 + 0.914\, CDP_i + 0.194\, L/B_w i \qquad (10)$$

$n = 20$ $R^2 = 0.931$ s = 0.31 F = 114.12

Multiple regressions:

$$\log P_i\text{calc} = 6.268 + 0.17\ L/B_w i - 0.166\ RDF090m_i + 0.904\ CDP_i \quad (11)$$

$n = 20$ $R^2 = 0.962$ s = 0.24 F = 129.73

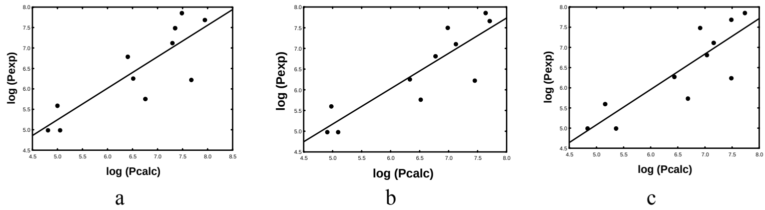
(B) PREDICTION SET (N = 11)

Table 7. AS predicting set (n = 11)

Structure	log $P_i$obs	log $P_i$calc (eq.9)	log $P_i$calc (eq.10)	log $P_i$calc (eq.11)
3	5.000	4.808	4.890	4.836
6	5.000	5.049	5.083	5.339
10	5.613	4.980	4.964	5.140
11	5.763	6.741	6.504	6.676
15	6.247	7.650	7.444	7.472
16	6.279	6.496	6.326	6.415
19	6.817	6.400	6.765	7.015
21	7.120	7.284	7.121	7.147
24	7.512	7.350	6.984	6.907
27	7.688	7.934	7.696	7.466
31	7.881	7.468	7.633	7.713
R^2		**0.716**	**0.766**	**0.728**
CV%		7.089	5.010	6.586

For the TS set, the variance percentage CV%, in prediction, decreases with increasing the number of variables (Table 5, eq 8), for the AS set the predicting ability seems to be better in bivariate regression (Table 7, eq 10).

Figures 4 a-c display the plot of experimental *vs*. calculated values for the receptor binding affinity of ASs: (a) calculated values *cf*. eq.9; (b) calculated values *cf*. eq.10 and (c) calculated values *cf*. eq.11.



(a) log ($P_i$obs) = 1.3932 + 0.77043 × log ($P_i$calc); R = 0.84636;

(b) log ($P_i$obs) = 0.90106 + .85433 × log ($P_i$calc);R = 0.87565;

(c) log ($P_i$obs) = 0.69650 + 0.87704 × log ($P_i$calc); R = 0.85352

Figure 4. The plot of experimental *vs*. calculated values for the receptor binding affinity of ASs.

CONCLUSIONS

In order to explain the contribution of each substituent position to the receptor binding affinity of TSs and ASs we generated a simple electronic descriptor CD, based on atomic partial charges, *ad-hoc* correlated with the studied property.

The QSAR models of this study indicate that this global descriptor is the most significant one in predicting the activities of our compounds. It can indicate the most important substituent positions. Thus, CD calculated for the atoms in positions above-mentioned, without the substituents in position 17 of the steroid skeleton, accounts for 25% of the variance (of CBG activity in AS set) while including position 17 this raised up to 89%. Similarly, CD calculated without the substituent in position 19 (TS set, hormones binding affinity to the gestagenic receptor) explains about 81% of the variance while it raised up to 92%, after including the substituent in that position. These results confirm the previous qualitative conclusions.

Both models have been validated on external prediction sets. The models derived for the molecular activity/property, by CD and the descriptors obtained from the factor loadings of PC, are comparable with those reported in literature,[4,6,8,24-26] with good predictive ability. Noticeable is the fact that, simple 2D models, like those herein developed, are comparable to the models provided by some more complex 3D-based methods (CoMFA, COMSA, GRIND, EEVA, etc), requiring much more computational resources.

REFERENCES

[1] A. R Katritzky, U. Maran, V. S. Lobanov, M. Karelson, Perspective: Structurally diverse quantitative structure-property relationship correlations of technologically relevant physical properties. *J. Chem. Inf. Comput. Sci.* **40**, 1-18, (2000).

[2] D. D. Robinson, P. J. Winn, P. D. Lyne, W. G Richards, Self-Organizing Molecular Field Analysis: A Tool for Structure-Activity Studies. *J. Med. Chem*. **42**, 573-583, (1999).

[3] R. D. III. Cramer, D. E. Patterson, J. D. Bunce, Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc*. **110**, 5959-5967, (1988).

[4] E. Coats, The CoMFA Steroids as a Benchmark Dataset for Development of 3D QSAR Methods. *Perspect. Drug Disco*V. *Design*. **12/13/14**, 199-213, (1998).

[5] J. F. Dunn, B. C. Nisula, D. Rodbard, Transport of Steroid Hormones: Binding of 21 Endogeneous Steroids to Both Testosterone- Binding Globulin and Corticosteroid-Binding Globulin in Human Plasma. *J. Clin. Endocrin. Metab*. **53**, 58-68, (1981).

[6] Kari Tuppurainen, Marja Viisas, Reino Laatikainen, and Mikael Pera¨kyla: Evaluation of a Novel Electronic Eigenvalue (EEVA) Molecular Descriptor for QSAR/QSPR Studies: Validation Using a Benchmark Steroid Data Set. *J. Chem. Inf. Comput. Sci*. **42***,* 607-613,. (2002).

[7] M. Wagener, J. Sadowski, J. Gasteiger, Autocorrelation of Molecular Surface Properties for Modeling *Corticosteroid Binding Globulin* and Cytosolic *Ah* Receptor Activity by Neural Networks. *J. Am. Chem. Soc*. **117**, 7769-7775, (1995).

[8] J. Polanski, B. Walczak, The Comparative Molecular Surface Analysis (COMSA): a novel Tool for Molecular Design. *Comput. Chem*. **24**, 615-625, (2000).

[9] W. L. Duax, J. F. Griffin, D. C. Rohrer, D. C. Swenson and C. M. Weeks, Molecular details of receptor binding and hormonal action of steroids derived from X-ray crystallographic investigations. *J. Steroid Biochem.* **15**, 41-47, (1981).

[10] S. C. Basak, In Practical Applications of Quantitative Structure-Activity Relationships QSAR in Environmental Chemistry and Toxicology, (Karcher W, Devillers J Eds, Kluwer Academic Publishers: Dordrecht, The Netherlands), *83*, (1990).

[11] M. V. Diudea and O. Ursu, *TOPOCLUJ* (Copyright Babes-Bolyai Univ. Cluj), (2002).

[12] *Dragon 2.1* software (http://www.disat.unimib.it/chm/Dragon.htm).

[13] HyperChem [TM], release 4.5 for SGI, © 1991-1995, Hypercube, Inc.

[14] I. Rios-Santamaria, R. Garcia-Domenech, J. Cortijo, P. Santamaria, E. J. Morcillo and J. Galvez, Natural Compounds with Bronchodilator Activity Selected by Molecular Topology. *Internet Electron J. Mol. Des.* **1**, 70-79, (2002).

[15] J. Galvez, R. Garcia-Domenech, M. T. Salabert and R. Soler, Charge Indexes. New Topological Descriptors. *J. Chem*. *Inf*. *Comput*. *Sci*. **34**, 520-525, (1994).

[16] A. A. Toropov and A. P. Toropova, QSAR Modeling of Mutagenicity Based on Graphs of Atomic Orbitals. *Internet Electron J. Mol. Des.* **1**, 108-114, (2002).

[17] D. J. G. Marino, P. J. Peruzzo, E. A. Castro and A. A. Toropov, QSAR Carcinogenic Study of Methylated Polycyclic Aromatic Hydrocarbons Based on Topological

Descriptors Derived from Distance Matrices and Correlation Weights of Local Graph Invariants. *Internet Electron J. Mol. Des.* **1**, 115-133, (2002).

[18] M. C. Hemmer, V. Steinhauer, J. Gasteiger, Deriving the 3D structure of organic molecules from their infrared spectra. *Vib. Spectrosc.* **19**, 151-164, (1999).

[19] R. Todeschini, V. Consonni,. *Handbook of Molecular Descriptors.* Wiley-VCH: Weinheim, Germany, (2000).

[20] M. V. Diudea and O. Ivanciuc, *Topologie Moleculara*, (Ed. COMPREX, Cluj), (1995).

[21] StatSoft, Inc. (2001). STATISTICA (data analysis software system), version 6. www.statsoft.com

[22] M. Randic and S. C. Basak, A New Descriptor for Structure-Property and Structure-Activity Correlations. *J. Chem. Inf. Comput. Sci*. **41**, 650-656, (2001).

[23] V. N. Viswanadhan, G. A. Mueller, S. C. Basak and J. N. Weistein, Comparison of a Neural Net-Based QSAR Algorithm (PCANN) with Hologram- and Multiple Linear Regression-Based QSAR Approaches: Application to 1,4-Dihydropyridine-Based Calcium Channel Antagonists**.** *J Chem Inf Comput Sci*. **41**, 505-511, (2001).

[24] A. N. Jain, K. Koile, D. Chapman, Compass: Predicting Biological Activities from Molecular Surface Properties. Performance Comparisons on a Steroid Benchmark. *J. Med. Chem*. **37**, 2315-2327, (1994).

[25] M. Pastor, G. Cruciani, I. McLay, S. Pickett, S. Clementi, GRid-INdependent Descriptors (GRIND): A Novel Class of Alignment- Independent Three-Dimensional Molecular Descriptors. *J. Med. Chem*. **43**, 3233-3243, (2000).

[26] D. A. Dragos, Contributions to QSAR study: applications to the serotoninergic neororeceptors, PhD thesis, Timisoara, (2005).