STUDIA UBB CHEMIA, LXI, 1, 2016 (p. 71-78)
(RECOMMENDED CITATION)

Dedicated to Professor Mircea Diudea on the Occasion of His 65th Anniversary

A QSPR MODEL FOR STEROIDS

LAVINIA L. PRUTEANU[a,*], SARA ERSALI[a], SORANA D. BOLBOACA[b]

ABSTRACT. A QSPR (Quantitative Structure-Property Relationship) model was derived for a set of forty 7 beta-hydroxysteroid compounds selected from PubChem database in order to assess the link between structural features and lipophilicity expressed as logP. After optimization and topological indices data collecting, the cluster of molecules was superposed onto a representative hypermolecule. Based on each molecule atoms positions, a binary vector and its weighted by mass fragments was computed for each molecule in the set. A model relating the structure with logP was identified based on the contributions of statistically significant positions of each molecule superposed on the hypermolecule and based on structural descriptors. The obtained model was validated in leave-one-out analysis as well as on training versus test analysis.

Keywords: *7 beta-Hydroxysteroid, QSPR (Quantitative Structure-Property Relationships), logP, hypermolecule*

INTRODUCTION

The interest for production of steroidal drugs began in 1952 when Murray and Peterson used *Rhizopus* species and patented the process of 11 alfa-hydroxylation of progesterone [1]. Since then, numerous studies based on transformation of steroids have been developed in order to find new drugs and also hormones derived from steroids, hydroxylation being one of the most widely applied transformations [2-4]. The 7 beta-hydroxysteroid derivatives are steroid compounds having one hydrogen atom replaced by a hydroxyl group at the carbon atom in position 7. For this study, forty molecules of this group have

[a] *Babeş-Bolyai University, Faculty of Chemistry and Chemical Engineering, 11 Arany Janos str., RO-400028, Cluj-Napoca, Romania*
[b] *Iuliu Haţieganu University of Medicine and Pharmacy, Department of Medical Informatics and Biostatistics, 6 Louis Pasteur str., RO-400349, Cluj-Napoca, Romania*
** Corresponding Author: pruteanulavinia@gmail.com*

been downloaded from the PubChem database, namely those molecules with a high structural similarity. One of the predictive methods used for modeling different properties is represented by Quantitative Structure Property Relationships (QSPR).

There are some more elaborated methods for prediction of molecular properties, such as Comparative Molecular Field Analysis (CoMFA) [5, 6], CODESSA (Comprehensive Descriptors for Structural and Statistical Analysis), Hologram QSAR [7], PDTA (Photodynamic Therapy Activity) [7], along with the traditional QSPR procedures using simple or multiple linear regression analysis (MLRA) [8-11].

In this study, the concept of reunion, of molecular structural features of the studied set, as a hypermolecule [12], was used to investigate the relation between structural features of a sample 7 beta-Hydroxysteroid and its lipophilicity.

RESULTS AND DISCUSSION

A significant regression model with estimation abilities was obtained with seven variables identified as significant positions (Eq. 1 is represented in Table 1).

$$\log P = 36.2431 + 0.0180 \cdot CjDi - 1.6780 \cdot AD - 0.0353 \cdot DI + 0.0228 \cdot CjDe - 0.0605 \cdot p18 - 0.0542 \cdot p33 - 0.0495 \cdot p35 \quad \textbf{(1)}$$

$R^2 = 0.9610$, $R^2_{adj} = 0.9525$, $Q^2 = 0.9413$; s = 0.4808, n = 40
F-statistics (p-value) = 113 ($1.02 \cdot 10^{-20}$)

where R^2 = determination coefficient, R^2_{adj} = adjusted determination coefficient; Q^2 = determination coefficient in leave-one-out analysis; s = standard error of estimate; n = sample size; F-statistics = Fisher statistics, p-value = probability to obtain the model by chance.

Table 1. Significant positions and their regression coefficients

Variable	Coefficients	Standard Error	t Stat (p-value)
Intercept	36.2431	4.4755	8.10 ($3.01 \cdot 10^{-9}$)
CjDi	0.0180	0.0020	8.82 ($4.41 \cdot 10^{-10}$)
AD	-1.6780	0.2383	-7.04 ($5.53 \cdot 10^{-8}$)
DI	-0.0353	0.0060	-5.85 ($1.69 \cdot 10^{-6}$)
CjDe	0.0228	0.0043	5.35 ($7.12 \cdot 10^{-6}$)
P18	-0.0605	0.0197	-3.07 ($4.37 \cdot 10^{-3}$)
P33	-0.0542	0.0157	-3.44 ($1.63 \cdot 10^{-3}$)
P35	-0.0495	0.0137	-3.61 ($1.04 \cdot 10^{-3}$)

CjDi = Cluj distance; AD = Adjacency;
DI = Distance; CjDe = Cluj detour;
P18 = Position 18; P33 = Position 33; P35 = Position 35

The model was selected as the best alternative, being the one with high explanatory power at the smallest number of predictors. The model was obtained by applying successively the forward stepwise method for the set of descriptors given in Tables 3 and 4. Having a insignificant contribution to the model able to explain the logP values, the following descriptors were excluded from the model (in this order: CON, p28, p36, p37 at 5% risk being in error; p43, p40 at 1% risk being in error followed by a procedure of backward stepwise which removed the rest of the non-explanatory variables: CFDi, CFDe, p17, p26, p34, p50, D3D and DE).

The obtained model seems to have the errors between the predicted and the observed values homogenously distributed between observations, as the goodness-of fit plot reveals (Figure 1).

The explanatory power of the model was analyzed with leave-one-out strategy, when the obtained explanatory power was 0.9413 (see Eq.1).

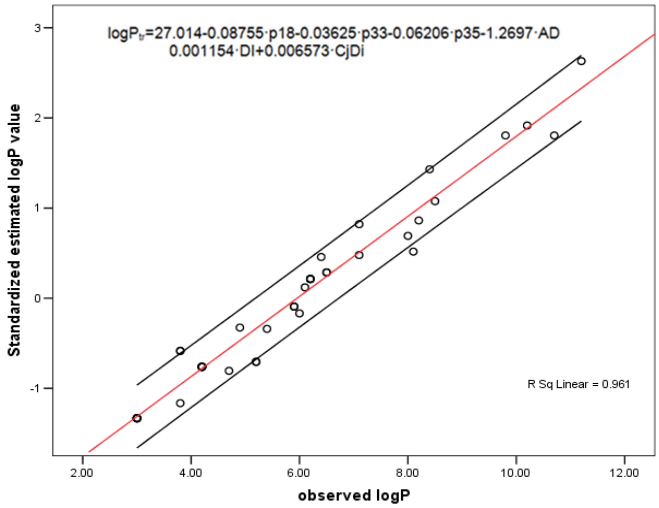

Figure 1. Goodness-of-fit of estimation model (the red line is the model fit and the black line is associated 95% confidence interval)

A training vs. test analysis was conducted on the selected pool of descriptors, when the set of 40 molecules were split in 24 molecules as the training set and 16 of them were used for the test of the model obtained. Following molecules were randomly selected to belong to the training set: 57396177, 49823443, 16082386, 16758147, 22216291, 9922115, 70688976, 70682680, 22213946, 11647965, 57390981, 12358742, 313039, 76325907, 57401396, 76327928, 76322252, 16759984, 24982302, 52947587, 12760132, 76336739, 76310266, and 70697302. The regression equation obtained with these molecules was used to predict the logP values for the rest of the molecules

(test set). The equation is:

$$\log P_{tr}=27.014-0.08755\cdot p18-0.03625\cdot p33-0.06206\cdot p35-1.2697\cdot AD\ 0.001154\cdot DI+0.006573\cdot CjDi \quad \textbf{(2)}$$
$$r^2_{tr}=0.9337,\ F_{tr}=40 \text{ (probability of wrong model } p_F<5\cdot 10^{-9})$$
$$r^2_{ts}=0.8730,\ F_{ts}=9 \text{ (probability of wrong model } p_F<2.6\cdot 10^{-3})$$

where tr = training set; ts = test set

As the training vs. test analysis shown, it is a little drop in the explanatory power when the model is not fed with the whole pool of molecules, and this fact can be explained by the large number of descriptors used to construct de structure-property relationship (in this case, a number of 7 variables were used, with an average of 5.7 molecules per descriptor for the whole pool of molecules and 5.0 molecules if we count the intercept too and a number of 3.42 molecules per descriptor for the training set and 3.0 molecules if we count the intercept too). Therefore, it is expected for a model having a small ratio between the number of molecules and the number of descriptors to produce such drop in the explanatory power when the input data are reduced in size.

CONCLUSIONS

The analysis drawn with the hypermolecule constructed from superposition of the molecules from the dataset shows a series of advantages, such as the natural reconstruction of the expected profile of action, as well as a series of disadvantages, such as the dropping of the explanatory power for the analysis conducted with a test set. Based on the selected model, which includes a series of positions in the hypermolecule, one can say that the positions 18, 33 and 35 are the ones which decreases the most (all these positions have a negative effect on the logP value, coefficients of it being negative) the value of logP.

MATERIALS AND METHODS

The set of forty 7β-hydroxysteroid derivatives was downloaded from PubChem database [13] and were used as input data in this analysis. The name, PubChem identification numbers along with the value of logP are given in Table 2.

The molecules geometry was optimized in HyperChem program at semi-empirical PM3 level of theory. The resulted log-files with the data collection were extracted using the utility program JSChem [14].

Table 2. The forty derivatives of 7β-Hydroxysteroids

No.	ID	logP	No.	ID	logP	No.	ID	logP	No.	ID	logP
1	12760132	10.2	11	76310266	8.2	21	57390981	5.2	31	76322257	10.7
2	70682679	7.1	12	56663807	6.4	22	16758147	8.5	32	76325907	3.8
3	70682680	6.5	13	56847117	6.2	23	22213946	6.2	33	76327928	3.8
4	70688976	6.2	14	70686910	6	24	16759984	5.9	34	76333144	4.2
5	70693211	6.2	15	70691082	7.1	25	16758161	4.2	35	371617	6.1
6	70697302	6.5	16	11647965	8.4	26	76336739	11.2	36	313039	8
7	12836861	4.7	17	52947587	4.9	27	57396177	3	37	9922115	4.2
8	24867469	4.2	18	24982302	3.8	28	57399636	3	38	9924252	5.4
9	16082386	8.1	19	49823443	5.9	29	57401396	3	39	11551321	3
10	12358742	5.2	20	22216291	3	30	76322252	9.8	40	11957457	4.2

A series of topological descriptors [15] were calculated by TopoCluj program [16] on the matrices: adjacency (AD), connectivity (CON), distance (DI), D3D - 3D (three-dimensional) distance, detour (DE), Cluj distance (CjDi), Cluj detour (CjDe), Cluj indices (on distance CFDi and on detour CFDe), the results being given in Table 3.

All the molecular structures were superposed to draw a hypermolecule by using Nano Studio program [17]. The resulted hypermolecule mimics the configuration or shape of the biological receptor to which the ligands have to bind [18].

Table 3. Calculated topological indices for the 7 beta-Hydroxysteroids in Table 2

Mol.	AD	CON	DI	D3D	DE	CjDi	CjDe	CFDi	CFDe
1	35	35	2522	4138.48	7313	4573.5	1750.5	5062	1809
2	34	34	2670	3980.77	6965	4467	1923.5	4918	1966.5
3	33	33	2369	3510.96	6451	4047	1665	4487	1706.5
4	33	33	2342	3522.35	6424	4019.5	1638	4460	1679.5
5	33	33	2342	3355.66	6424	4019.5	1638	4460	1679.5
6	33	33	2369	3520.97	6451	4047	1665	4487	1706.5
7	25	25	926	1354.38	3297	1729.5	546.5	1983	572.5
8	24	24	802	1163.2	2969	1504.5	463	1739	486.5
9	33	33	2335	3350.73	6511	4029	1648.5	4550.5	1693
10	26	26	1052	1868.14	3627	1956.5	632	2229	660.5
11	35	35	2237	3659.41	8399	4455.5	1302	5281	1361
12	33	33	2345	3436.73	6348	3996	1647.5	4362	1687.5
13	33	33	2342	3309.13	6424	4019.5	1638	4460	1679.5
14	32	32	2098	3117.36	5967	3657.5	1436.5	4086	1476.5
15	34	34	2560	3759.51	6855	4349.5	1813.5	4808	1856.5
16	36	36	2435	3417.31	8962	4849.5	1424.5	5718	1485
17	24	24	796	1153.47	2949	1490.5	461	1732	488
18	27	27	1149	1681.56	3971	2130	715.5	2485	747.5
19	23	23	699	988.57	2662	1310	402	1516	421.5
20	25	25	887	1279.08	3279	1655	535.5	1966	562.5
21	26	26	1052	1530.66	3627	1956.5	632	2229	660.5
22	36	36	2436	3497	9036	4811.5	1459.5	5703.5	1521.5
23	33	33	2342	3355.66	6424	4019.5	1638	4460	1679.5
24	23	23	699	1023	2662	1310	402	1516	421.5
25	24	24	802	1158.41	2969	1504.5	463	1739	486.5

Mol.	AD	CON	DI	D3D	DE	CjDi	CjDe	CFDi	CFDe
26	35	35	2944	4219.49	7659	5068.5	2080.5	5527	2130
27	25	25	887	1279.08	3279	1655	535.5	1966	562.5
28	25	25	887	1279.08	3279	1655	535.5	1966	562.5
29	25	25	887	1279.08	3279	1655	535.5	1966	562.5
30	34	36	2668	3785.72	7103	4609	1864.5	5048.5	1911.5
31	34	34	2668	3828.26	7103	4609	1864.5	5048.5	1911.5
32	27	27	1172	1646	3967	2178	728	2483	763
33	27	27	1172	1654.23	3967	2177.5	728	2482.5	762.5
34	24	24	802	1169.48	2969	1504.5	463	1739	486.5
35	24	24	732	1086.05	2660	1437	483.5	1707.5	501.5
36	32	32	2276	3204.88	6075	3832	1619	4173	1656.5
37	24	24	802	1169.48	2969	1504.5	463	1739	486.5
38	28	30	1412	2008.12	4395	2535	914	2829	944.5
39	25	25	887	1279.08	3279	1655	535.5	1966	562.5
40	24	24	802	1169.48	2969	1504.5	463	1739	486.5

The hypermolecule description was made by means of mass fragments with respect to the logP as the modelled property in order to identify a model able to both estimate and predict the logP on a series of 7β-hydroxysteroid derivatives. This has been demonstrated as an efficient and helpful method in prediction of molecular property and/or bioactivities [11, 18, 19].

Superposition of the forty ligands over the hypermolecule (Figure 2) resulted in a binary vector, of value 1 for those positions with existing atoms and value 0 otherwise.

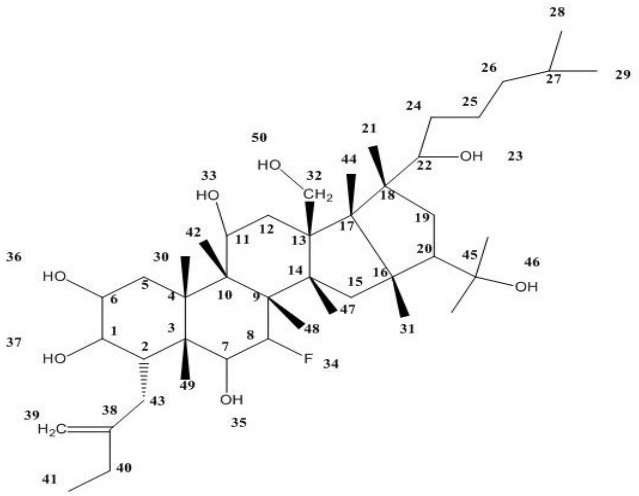

Figure 2. Representation in ChemBioDraw of the hypermolecule comprising all common and different structural features of each molecule and its atoms positions

Next, the value 1 was replaces by the corresponding mass fragment weight. After a primary correlation, only statistically significant positions were retained (Table 4).

Table 4. Statistically significant positions, calculated with the mass fragment weight

Mol.	p17	p18	p26	p28	p33	p34	p35	p36	p37	p40	p43	p50
1	12.011	12.011	12.011	0	0	0	0	0	17.007	0	12.011	0
2	12.011	12.011	12.011	12.011	0	0	17.007	0	17.007	0	0	0
3	12.011	12.011	12.011	12.011	0	0	17.007	0	17.007	0	0	0
4	12.011	12.011	12.011	0	0	0	17.007	0	17.007	0	0	0
5	12.011	12.011	12.011	0	0	0	17.007	0	17.007	0	0	0
6	12.011	12.011	12.011	12.011	0	0	17.007	0	17.007	0	0	0
7	12.011	17.007	0	0	0	0	0	0	17.007	0	0	0
8	12.011	17.007	0	0	0	0	0	0	17.007	0	0	0
9	12.011	12.011	12.011	0	0	19	17.007	0	17.007	0	0	0
10	12.011	12.011	0	0	0	0	0	0	17.007	0	0	0
11	12.011	12.011	0	0	0	0	0	0	17.007	0	12.011	0
12	12.011	12.011	12.011	0	0	0	0	0	17.007	0	0	0
13	12.011	12.011	12.011	0	0	0	17.007	0	17.007	0	0	0
14	12.011	12.011	12.011	0	0	0	17.007	0	17.007	0	0	0
15	12.011	12.011	12.011	12.011	0	0	17.007	0	17.007	0	0	0
16	12.011	12.011	0	0	0	0	0	0	17.007	0	12.011	0
17	12.011	0	0	0	0	0	0	17.007	17.007	0	0	0
18	12.011	12.011	0	0	17.007	0	0	0	17.007	0	0	0
19	12.011	0	0	0	0	0	0	0	17.007	0	0	0
20	12.011	17.007	0	0	17.007	0	0	0	17.007	0	0	0
21	12.011	12.011	0	0	0	0	0	0	17.007	0	0	0
22	12.011	12.011	0	0	0	0	17.007	0	0	0	0	0
23	12.011	12.011	12.011	0	0	0	17.007	0	17.007	0	0	0
24	12.011	0	0	0	0	0	0	0	17.007	0	0	0
25	12.011	17.007	0	0	0	0	0	0	17.007	0	0	0
26	12.011	12.011	12.011	0	0	0	0	0	17.007	12.011	12.011	0
27	12.011	17.007	0	0	17.007	0	0	0	17.007	0	0	0
28	12.011	17.007	12.011	0	17.007	0	0	0	17.007	0	0	0
29	12.011	17.007	0	0	17.007	0	0	0	17.007	0	0	0
30	12.011	12.011	12.011	0	0	0	0	0	17.007	0	12.011	0
31	12.011	12.011	12.011	0	0	0	0	0	17.007	12.011	12.011	0
32	12.011	12.011	0	0	0	0	0	0	17.007	0	0	17.007
33	12.011	12.011	0	0	0	0	0	0	17.007	0	0	17.007
34	12.011	17.007	0	0	0	0	0	0	17.007	0	0	0
35	0	0	0	0	0	0	17.007	17.007	17.007	0	0	0
36	12.011	12.011	12.011	17.007	0	0	0	0	17.007	0	0	0
37	12.011	17.007	0	0	0	0	0	0	17.007	0	0	0
38	12.011	12.011	0	0	0	0	0	0	17.007	0	0	0
39	12.011	17.007	0	0	17.007	0	0	0	17.007	0	0	0
40	12.011	17.007	0	0	0	0	0	0	17.007	0	0	0

LogP property was modeled using mass fragments for weighting the superposition vectors. The model was validated by the leave-one-out and training vs. test procedures [20,21].

The whole sample was randomly split [22] in training and test sets, with ~2/3 of compounds in training set. The compounds in the training set was used to identify the model while the compounds in test set was used to validate this model.

ACKNOWLEDGMENTS

This paper is a result of a doctoral research being financially supported by the Sectoral Operational Programme for Human Resources Development 2007-2013. co-financed by the European Social Fund under the project POSDRU/187/1.5/S/155383 - “Doctoral and postdoctoral programs for scientific research support”.

R E F E R E N C E S

1. D.H. Peterson, H.C. Murray, *Journal of the American Chemical Society*, **1952**, *74*, 5933.
2. S.B. Mahato, S. Garai, *Steroids*, **1997**, *62(4)*, 332.
3. A. Malaviya, J. Gomes, *Bioresource Technology*, **2008**, *99*, 6725.
4. H. Pellissier, M. Santelli, *Organic Preparations and Procedure International*, **2001**, *33(1)*, 1.
5. H. Hong, H. Fang, Q. Xie, R. Perkins, D.M. Sheehan, W. Tong, *SAR QSAR Environmental Research*, **2003**, *14(5-6)*, 373.
6. T.G. Gantchev, H. Ali, J.E. Van Lier, *Journal of Medicinal Chemistry*, **1994**, *37*, 4164.
7. W. Tong, D.R. Lowis, R. Perkins, Y. Chen, W.J. Welsh, D.W. Goddette, T.W. Heritage, D.M. Sheehan, *Journal of Chemical Information and Computer Sciences,* **1998**, *38(4)*, 669.
8. N. Frimayanti, M.L. Yam, H.B. Lee, R. Othman, S.M. Zain, N.A. Rahman, *International Journal of Molecular Sciences*, **2011**, *12*, 8626.
9. A. Golbraikh, A. Tropsha, *Journal of Computer-Aided Molecular Design,* **2002**, *5*, 231.
10. S.D. Bolboacă, L. Jäntschi, *BIOMATH. International Conference on Mathemathical Methods and Models in Biosciences,* **2013**, *2(1)*, 1309089
11. M. Goodarzi, B. Dejaegher, Y.V. Heyden, *Journal of AOAC International*, **2012**, *95(3)*, 636.
12. A.M. Harsa, T.E. Harsa, S.D. Bolboacă, M.V. Diudea, *Current Computer-Aided Drug Design*, **2014**, *10(2)*, 115.
13. PubChem [online] Available from: https://pubchem.ncbi.nlm.nih.gov/compound/439865.
14. C.L. Nagy, M.V. Diudea, JSCHEM 1.0. Babeş-Bolyai University, Cluj, **2004**.
15. M.V. Diudea, I. Gutman, L. Jäntschi, *Molecular Topology*, NOVA: New York, **2002**.
16. O. Ursu, M.V. Diudea, TOPOCLUJ software program, Babeş-Bolyai University, Cluj, **2005**.
17. C.L. Nagy, M.V. Diudea, Nano Studio. 1.0., Babeş-Bolyai University, Cluj, **2009**.
18. A.M. Harsa, T.E. Harsa, M.V. Diudea, *Journal of Enzyme Inhibition and Medicinal Chemistry*, **2011**, *26(4)*, 1.
19. T.E. Harşa, A.M. Harşa, M.V. Diudea, B. Szefler, *Revue Roumaine de Chimie*, **2015**, *60(7-8)*, 727.
20. S.D. Bolboacă, L. Jäntschi, M.V. Diudea, *Current Computer-Aided Drug Design*, **2013**, *9(2)*, 195.
21. S.D. Bolboacă, L. Jäntschi, *Environmental Chemistry Letters*, **2008**, *6*, 175.
22. S.D. Bolboacă, *Applied Medical Informatics*, **2010**, *28(2)*, 9.