ANTICANCER RESEARCH *38*: 6189-6194 (2018)
doi:10.21873/anticanres.12972

Predicting Cytotoxicity of 2-Phenylindole Derivatives Against Breast Cancer Cells Using Index of Ideality of Correlation

ANDREY A. TOROPOV and ALLA P. TOROPOVA

Laboratory of Environmental Chemistry and Toxicology,
Mario Negri Institute of Pharmacological Research IRCCS, Milan, Italy

Abstract. *Background: Breast cancer is one of the leading types of cancer in women worldwide. Quantitative structure–activity relationship (QSAR) methods play an important role in the search for new anticancer agents. A QSAR model for cytotoxicity against the breast cancer cell line MCF7, based on hybrid optimal descriptors, has been suggested. A modified version of the hybrid descriptor is suggested. Materials and Methods: A QSAR model for the anticancer activity of 2-phenylindole derivatives was built using the Index of Ideality of Correlation (IIC), which is a new criterion for predictive potential. The calculation can be carried out with a modified version of the CORAL software. Results: The model for the anticancer activity suggested here is better than the one described in the literature. Conclusion: Taking into account the data on molecular rings together with the use of new criterion of predictive potential (IIC), the QSAR improves the prediction for anticancer activity.*

Anticancer drug discovery is a complex and important field of natural sciences. Quantitative structure–activity relationships (QSARs) are not able to provide complete data on molecular architecture required for new anticancer agents, but QSARs can help reduce the time needed and cost of the search for such agents (1).

The design of new chemical compounds that are active against the breast cancer cell line MCF7 has several conceptually different approaches. These are the well-known ADMET approach (absorption, distribution, metabolism, excretion, and toxicity) (2); general virtual screening based on comparison of different molecular features (3), molecular docking (4); various chemosensitization effects (5); 2D-QSAR (6, 7); 3D-QSAR (8) with analysis of stereoselectivity (9); study of the molecular C-skeleton architecture (10); and finally, comparative analysis of different classes of molecules with anticancer potential (11-16), for example as in the above-mentioned work (1). It is to be noted, that QSAR analysis should obey principles suggested by the Organisation for Economic Co-operation and Development (OECD) (17) and recommendations of the EU chemical control regulation in the European Union (Registration, Evaluation, Authorisation and Restriction of Chemicals, REACH) (18).

The aim of the present study was to improve the CORAL model described in (1) by means of using two approaches, namely (i) using of the Index of Ideality of Correlation (IIC), which is a new criterion for the predictive potential of QSAR models (19-22); and (ii) using correlation weights, which are related to the presence of different rings in the molecular structure (23-25).

Materials and Methods

Dataset. The dataset of 102 2-phenylindole derivatives having cytotoxicity against the MCF7 breast cancer cell line was taken from the literature (1). The molecular structure of these 2-phenylindole derivatives are represented by simplified molecular input-line entry system (SMILES) (26) and the concentration of these compounds producing 50% *in vitro* MCF7 cellular toxicity (IC_{50}, in nM) was transformed into the corresponding negative logarithm (pIC_{50}). These compounds were randomly split into training, invisible training, calibration, and validation sets and were studied here. Each of the sets has a special role. The training set is the builder of the model. The invisible training set is the inspector of the model (checking whether model is satisfactory for molecules absent from the training set). The calibration set must detect the start of overtraining. The validation set is the estimator of the predictive potential of the model.

Optimal descriptor. The optimal descriptor used here was calculated as the following:

$$DCW(T^*;N^*) = \sum_{k=1}^{NA} CW(S_k) + \sum_{k=1}^{NA-1} CW(SS_k) + \sum_{k=1}^{NA-2} CW(SSS_k) + CW(HARD) + \alpha\,[\,CW(C5) + CW(C6)] \quad \text{(Eq. 1)}$$

This article is freely accessible online.

Correspondence to: Alla P. Toropova, Laboratory of Environmental Chemistry and Toxicology, Mario Negri Institute of Pharmacological Research IRCCS, Via La Masa 19, 20156 Milan, Italy. Tel: +39 0239014595, Fax: +39 0239014735, e-mail: alla.toropova@marionegri.it

Key Words: QSAR, anticancer activity, Monte-Carlo method, CORAL software, Index of Ideality of Correlation.

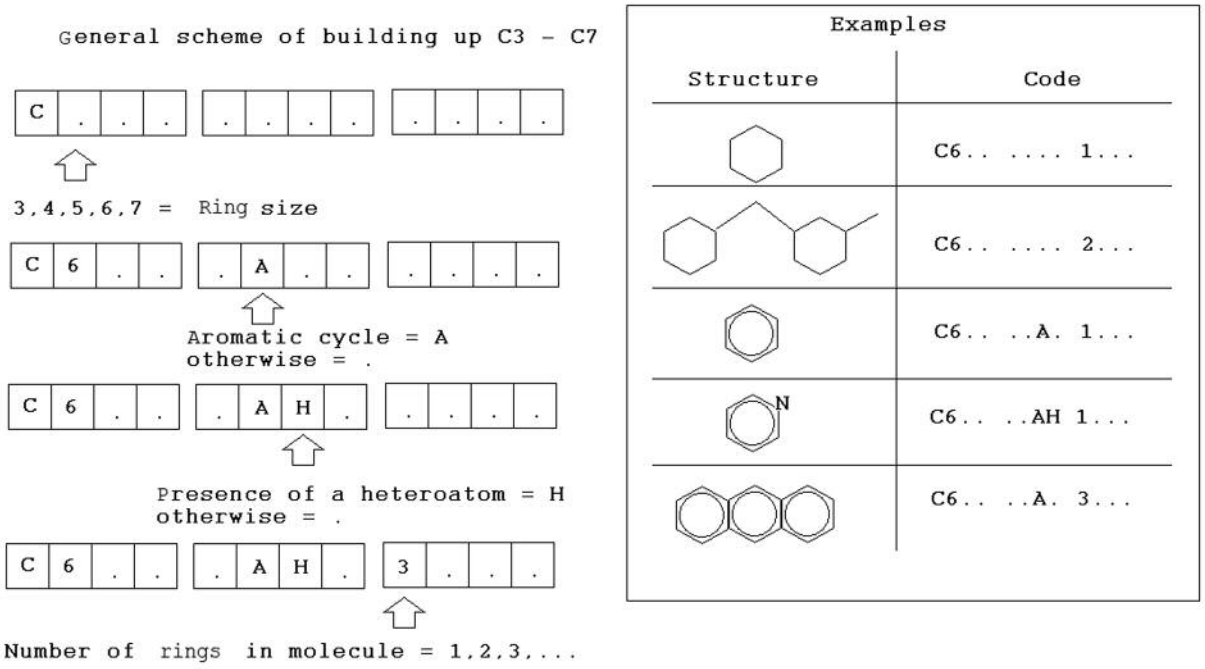


Figure 1. *The general scheme of definition of codes reflecting the presence of different rings in a molecular structure.*

whereby Sk is the SMILES atom *i.e.* one elemental symbol (*e.g.* C, N, and O) or two symbols which cannot be examined separately (*e.g.* Cl, and Si); SS_k is a combination of two SMILES atoms; similarly SSS_k is a combination of three SMILES atoms; $CW(S_k)$, $CW(SS_k)$, and $CW(SSS_k)$ are the correlation weights of the above-mentioned attributes of SMILES; NA is the number of attributes in SMILES; α is 1, *i.e.* the presence of rings is involved in building the model, or 0 *i.e.* presence of rings is not involved in building model). HARD is a descriptor characterizing SMILES as a whole (24). The C5 and C6 are descriptors characterizing rings in the molecular structure. These descriptors are calculated with the molecular graph (23-25). C5 and C6 are codes sensitive to the number of corresponding rings in the molecular structure, the presence (absence) heteroatoms, and the presence (absence) of aromaticity. Figure 1 shows the general scheme for the definition of these codes.

The numerical data on correlation weights of these features of the molecular structure extracted from SMILES and graph were calculated with the Monte-Carlo method, *i.e.* the optimization procedure that gives maximal value of a target function (TF). QSAR models were calculated with the Monte-Carlo optimization based on two kinds of target functions TF_1 and TF_2:

$$TF_1=R_{TRN} + R_{iTRN} - |R_{TRN} - R_{iTRN}|\times 0.1 \quad \text{(Eq. 2)}$$

$$TF_2=TF_1+IIC_{CLB}\times 0.1 \quad \text{(Eq. 3)}$$

where by R_{TRN} and R_{iTRN} are correlation coefficients between the observed and predicted endpoints for the training and invisible training sets respectively. IIC_{CLB} is calculated with data on the calibration (CLB) set as the following:

$$\mathrm{IIC}_{CLB} = r_{CLB}\frac{\min({}^{-}MAE_{CLB};{}^{+}MAE_{CLB})}{max({}^{-}MAE_{CLB};{}^{+}MAE_{CLB})} \quad \text{(Eq. 4)}$$

$${}^{-}MAE_{CLB} = \frac{1}{{}^{-}N}\sum_{k=1}^{{}^{-}N}|\Delta_k|;$$
$$\Delta_k < 0; {}^{-}N \text{ is the number of } \Delta_k < 0 \quad \text{(Eq. 5)}$$

$${}^{+}MAE_{CLB} = \frac{1}{{}^{+}N}\sum_{k=1}^{{}^{+}N}|\Delta_k|;$$
$$\Delta_k \geqslant 0; {}^{+}N \text{ is the number of } \Delta_k \geqslant 0 \quad \text{(Eq. 6)}$$

$$\Delta_k = observed_k - calculated_k \quad \text{(Eq. 7)}$$

The observed and calculated values are corresponding values of the endpoint.

Having the numerical data on the correlation weights the predictive model is calculated by the least squares method with compounds from the training set:

$$pIC_{50}=C_0+C_1\times DCW(T^{*};N^{*}) \quad \text{(Eq.8)}$$

The predictive potential of this model should be checked with an external validation set.

Table I. *The statistical characteristics of the CORAL models for three random splits.*

Split	TF	α	Set	n	R^2	CCC	Q^2	IIC	MAE
1	TF_1	1	TRN	26	0.8435	0.9151	0.8206	0.6735	0.376
			iTRN	25	0.8658	0.8108	0.8430	0.3340	0.463
			CLB	26	0.7777	0.8752	0.7243	0.8316	0.222
			VLD	25	0.9017				0.211
	TF_2	**1**	**TRN**	**26**	**0.8037**	**0.8912**	**0.7755**	**0.7684**	**0.452**
			iTRN	**25**	**0.8047**	**0.8091**	**0.7658**	**0.3587**	**0.461**
			CLB	**26**	**0.9100**	**0.9530**	**0.8816**	**0.9538**	**0.148**
			VLD	**25**	**0.9685**				**0.128**
	TF_2	0	TRN	26	0.8151	0.8981	0.7852	0.7738	0.461
			iTRN	25	0.8167	0.8111	0.7577	0.2986	0.480
			CLB	26	0.8552	0.9239	0.8149	0.9247	0.202
			VLD	25	0.9227				0.224
2	TF_1	1	TRN	25	0.8360	0.9107	0.8111	0.5143	0.343
			iTRN	25	0.7728	0.8705	0.7329	0.6604	0.418
			CLB	26	0.8104	0.8918	0.7722	0.6560	0.291
			VLD	26	0.8231				0.305
	TF_2	**1**	**TRN**	**25**	**0.7919**	**0.8839**	**0.7599**	**0.6992**	**0.378**
			iTRN	**25**	**0.7671**	**0.8698**	**0.7264**	**0.8681**	**0.480**
			CLB	**26**	**0.8850**	**0.9322**	**0.8609**	**0.9407**	**0.234**
			VLD	**26**	**0.9179**				**0.218**
	TF_2	0	TRN	25	0.7928	0.8844	0.7592	0.6996	0.385
			iTRN	25	0.7383	0.8501	0.6940	0.6572	0.482
			CLB	26	0.8495	0.9138	0.8289	0.9216	0.266
			VLD	26	0.8908				0.234
3	TF_1	1	TRN	26	0.8723	0.9318	0.8510	0.6849	0.273
			iTRN	25	0.8315	0.8558	0.8028	0.6668	0.439
			CLB	25	0.8777	0.9307	0.8604	0.7621	0.315
			VLD	26	0.6087				0.326
	TF_2	**1**	**TRN**	**26**	**0.8236**	**0.9033**	**0.7912**	**0.7779**	**0.378**
			iTRN	**25**	**0.7707**	**0.8321**	**0.7334**	**0.7456**	**0.549**
			CLB	**25**	**0.9107**	**0.9497**	**0.8978**	**0.9543**	**0.268**
			VLD	**26**	**0.8871**				**0.147**
	TF_2	0	TRN	26	0.8428	0.9147	0.8132	0.7869	0.342
			iTRN	25	0.8056	0.8660	0.7741	0.7400	0.468
			CLB	25	0.8996	0.9411	0.8860	0.9485	0.297
			VLD	26	0.8436				0.186

TRN, iTRN, CLB, and VLD are training, invisible training, calibration, and validation sets, respectively; n: number of compounds in a set; R^2: correlation coefficient; CCC: concordance correlation coefficient; Q^2: cross-validated correlation coefficient; IIC: index of ideality of correlation; MAE is mean absolute error. The best models are indicated in bold.

Results

The CORAL models for the pIC_{50} in the case of using target function TF1 for three random splits were the following:

pIC_{50}=4.392 (±0.028)+0.1393 (±0.0021) × DCW(1,2) (Eq. 9)

pIC_{50}=3.279 (±0.042)+0.1050 (±0.0017) × DCW(1,2) (Eq. 10)

pIC_{50}=3.735 (±0.049)+0.1009 (±0.0016) × DCW(1,2) (Eq. 11)

The CORAL models for the pIC_{50} in the case of using target function TF_2 for three random splits were the following:

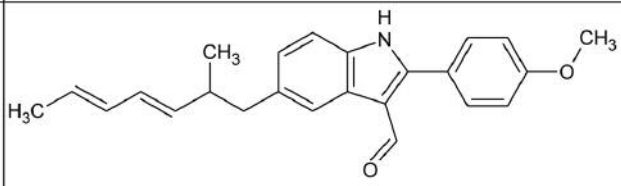
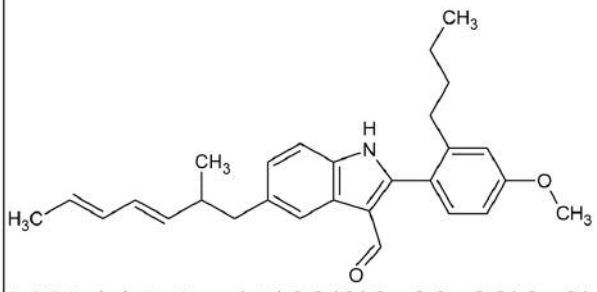
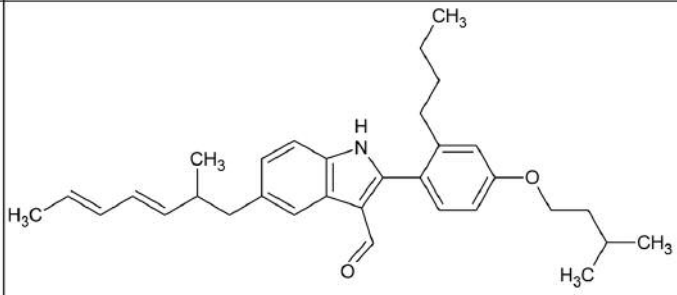
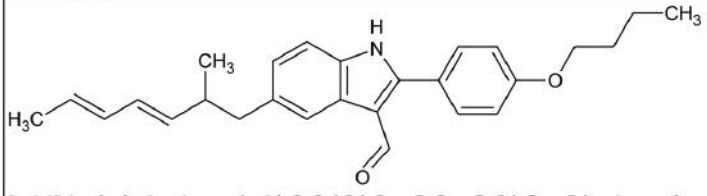
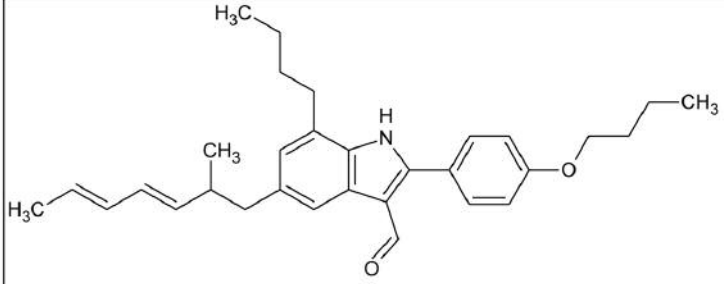
pIC_{50}=4.446 (±0.036)+0.08419 (±0.0014) × DCW(1,15) (Eq. 11)

pIC_{50}=4.029 (±0.034)+0.09477 (±0.0018) × DCW(1,15) (Eq. 12)

pIC_{50}=3.978 (±0.062)+0.09811 (±0.0021) × DCW(1,15) (Eq. 13)

Table I presents the statistical characteristics of these models. Target function TF_2 gave better models for all three random splits in comparison with optimization with TF_1. The Monte-Carlo optimization without correlation weights for C5 and C6 gave models characterized by reduced predictive potential in comparison with models where correlation weights for C5 and C6 were taken into account (Table I).

Table II. *List of possible anticancer agents according to described models.*

	Structure and SMILES	pIC_{50}
1	 [nH]1c(c(c2c1ccc(c2)CC(C)C=CC=CC)C=O)c1ccc(cc1)OC	7.9157 (Eq. 11) 7.9837 (Eq. 12) 8.1042 (Eq. 13)
2	 [nH]1c(c(c2c1ccc(c2)CC(C)C=CC=CC)C=O)c1c(CCCC)cc(cc1)OC	8.1816 (Eq. 11) 8.7779 (Eq. 12) 8.5370 (Eq. 13)
3	 [nH]1c(c(c2c1ccc(c2)CC(C)C=CC=CC)C=O)c1c(CCCC)cc(cc1)OCCC(C)C	8.3780 (Eq. 11) 9.2520 (Eq. 12) 9.2166 (Eq. 13)
4	 [nH]1c(c(c2c1ccc(c2)CC(C)C=CC=CC)C=O)c1ccc(cc1)OCCCC	8.0863 (Eq. 11) 8.3703 (Eq. 12) 8.5750 (Eq. 13)
5	 [nH]1c(c(c2c1c(CCCC)cc(c2)CC(C)C=CC=CC)C=O)c1ccc(cc1)OCCCC	8.3521 (Eq. 11) 9.1645 (Eq. 12) 9.0079 (Eq. 13)

pIC_{50}: Negative logarithm of the concentration of compound producing 50% *in vitro* MCF7 cellular toxicity.

Discussion

Having data on several runs of the Monte-Carlo optimization allows the possibility to detect SMILES attributes, which have solely positive correlation weights. These attributes can be qualified as promoters of increase for pIC_{50}. Corresponding computational experiments have confirmed that there are molecular features, which are promoters of pIC_{50} increase. These are: (i) features of five-member and six-member rings; (ii) branching of the molecular skeleton; and (iii) the presence of double bonds.

Table II presents the molecular structures of potential effective anticancer agents against the MCF7 breast cancer cell line defined according to the above-mentioned conditions, *i.e.* presence of one five-member ring, two six-member aromatic rings, presence of double bonds, and the bifurcations of molecular skeleton.

The statistical characteristics of models calculated with Eq. 11-13 are better than the statistical characteristics of the CORAL models suggested in the original work (1), where the best model was characterized by r^2=0.8603, and mean absolute error=0.225 (validation set). Thus, using the correlation weights for C5 and C6 together with modified target function TF_2 improves the model for cytotoxicity of 2-phenylindole derivatives against the MCF7 breast cancer cell line.

Conclusion

The IIC is a new criterion for the predictive potential of a QSAR model. The use of the index as a component of the target function for the Monte-Carlo optimization improves the predictive potential of models for the cytotoxicity of 2-phenylindole derivatives against the MCF7 breast cancer cell line. The use of global SMILES codes C5 and C6, which are sensitive to the presence and quality of rings, provides the possibility of improving QSAR models for this endpoint.

Acknowledgements

The Authors are grateful for the contribution of the project LIFE-CONCERT contract (LIFE17 GIE/IT/000461) for financial support. The Authors would also like to thank the Editorial Board for the help in preparation of the article.

References

1 Gaikwad R, Ghorai S, Amin SA, Adhikari N, Patel T, Das K, Jha T and Gayen S: Monte-Carlo based modelling approach for designing and predicting cytotoxicity of 2-phenylindole derivatives against breast cancer cell line MCF7. Toxicol In Vitro *52*: 23-32, 2018.

2 Yadav DK, Kumar S, Saloni, Singh H, Kim M-H, Sharma P, Misra S and Khan F: Molecular docking, QSAR and ADMET studies of withanolide analogs against breast cancer. Drug Des Devel Ther *11*: 1859-1870, 2017.

3 Konova V, Lagunin A, Pogodin P, Kolotova E, Shtil A and Poroikov V: Virtual screening of chemical compounds active against breast cancer cell lines based on cell cycle modelling, prediction of cytotoxicity and interaction with targets. SAR QSAR Environ Res *26(7-9)*: 595-604, 2015.

4 Murahari M, Kharkar PS, Lonikar N and Mayur YC: Design, synthesis, biological evaluation, molecular docking and QSAR studies of 2,4-dimethylacridones as anticancer agents. Eur J Med Chem *130*: 154-170, 2017.

5 Gade DR, Makkapati A, Yarlagadda RB, Peters GJ, Sastry BS and Rajendra Prasad VVS: Elucidation of chemosensitization effect of acridones in cancer cell lines: Combined pharmacophore modeling, 3D QSAR, and molecular dynamics studies. Comput Biol Chem *74*: 63-75, 2018.

6 Georgey H: Synthesis, cytotoxic activity and 2D-QSAR study of some imidazoquinazoline derivatives. Molecules *19(3)*: 3777-3792, 2014.

7 Ismail NSM, George RF, Serya RAT, Baselious FN, El-Manawaty M, Shalaby EM and Girgis AS: Rational design, synthesis and 2D-QSAR studies of antiproliferative tropane-based compounds. RSC Adv *6(104)*: 101911-101923, 2016.

8 Alam S and Khan F: 3D-QSAR studies on maslinic acid analogs for anticancer activity against breast cancer cell line MCF-7. Sci Rep *7(1)*: art no 6019, 2017.

9 George RF: Stereoselective synthesis and QSAR study of cytotoxic 2-(4-oxo-thiazolidin-2-ylidene)-2-cyano-*N*-arylacetamides. Eur J Med Chem *47(1)*: 377-386, 2012.

10 Prakash O, Khan F, Sangwan RS and Misra L: ANN-QSAR model for virtual screening of androstenedione c-skeleton containing phytomolecules and analogues for cytotoxic activity against human breast cancer cell line MCF-7. Comb Chem High Throughput Screen *16(1)*: 57-72, 2013.

11 Girgis AS, Stawinski J, Ismail NSM and Farag H: Synthesis and QSAR study of novel cytotoxic spiro[3*H*-indole-3, 2'(1'*H*)-pyrrolo[3,4-c]pyrrole]-2,3',5'(1*H*,2'a*H*, 4'*H*)-triones. Eur J Med Chem *47(1)*: 312-322, 2012.

12 Fortuna CG, Barresi V, Berellini G and Musumarra G: Design and synthesis of trans 2-(furan-2-yl)vinyl heteroaromatic iodides with antitumour activity. Bioorg Med Chem *16(7)*: 4150-4159, 2008.

13 Ballistreri FP, Barresi V, Benedetti P, Caltabiano G, Fortuna CG, Longo ML and Musumarra G: Design, synthesis and *in vitro* antitumor activity of new trans 2-[2-(heteroaryl)vinyl]-1,3-dimethylimidazolium iodides. Bioorg Med Chem *12(7)*: 1689-1695, 2004.

14 Girgis AS, Panda SS, Farag ISA, El-Shabiny AM, Moustafa AM, Ismail NSM, Pillai GG, Panda CS, Hall CD and Katritzky AR: Synthesis, and QSAR analysis of anti-oncological active spiro-alkaloids. Org Biomol Chem *13(6)*: 1741-1753, 2015.

15 Jung H, Shin SY, Jung Y, Tran TA, Lee HO, Jung K-Y, Koh D, Cho SK and Lim Y: Quantitative relationships between the cytotoxicity of flavonoids on the human breast cancer stem-like cells MCF7-SC and their structural properties. Chem Biol Drug Des *86(4)*: 496-508, 2015.

16 AlQudah DA, Zihlif MA and Taha MO: Ligand-based modeling of diverse aryalkylamines yields new potent P-glycoprotein inhibitors. Eur J Med Chem *110*: 204-223, 2016.

17 Kar S and Roy K: First report on development of quantitative interspecies structure-carcinogenicity relationship models and exploring discriminatory features for rodent carcinogenicity of diverse organic chemicals using OECD guidelines. Chemosphere *87(4)*: 339-355, 2012.
18 E Wedebye EB, Dybdahl M, Nikolov NG, Jónsdóttir ST and Niemelä JR: QSAR screening of 70,983 REACH substances for genotoxic carcinogenicity, mutagenicity and developmental toxicity in the ChemScreen project. Reprod Toxicol *55*: 64-72, 2015.
19 Toropov AA and Toropova AP: The index of ideality of correlation: A criterion of predictive potential of QSPR/QSAR models? Mutat Res Genet Toxicol Environ Mutagen *819*: 31-37, 2017.
20 Toropova AP and Toropov AA: The index of ideality of correlation: A criterion of predictability of QSAR models for skin permeability? Sci Total Environ *586*: 466-472, 2017.
21 Toropov AA, Carbó-Dorca R and Toropova AP: Index of Ideality of Correlation: new possibilities to validate QSAR: A case study. Struct Chem *29(1)*: 33-38, 2018.
22 Stoičkov V, Stojanović D, Tasić I, Šarić S, Radenković D, Babović P, Sokolović D and Veselinović AM: QSAR study of 2;4-dihydro-3H-1;2;4-triazol-3-ones derivatives as angiotensin II AT1 receptor antagonists based on the Monte-Carlo method. Struct Chem *29(2)*: 441-449, 2018.
23 Toropova MA, Toropov AA, Raška I and Rašková M: Searching therapeutic agents for treatment of Alzheimer disease using the Monte-Carlo method. Comput Biol Med *64*: 148-154, 2015.
24 Toropova AP and Toropov AA: CORAL: Binary classifications (active/inactive) for drug-induced liver injury. Toxicol Lett *268*: 51–57, 2017.
25 Islam MA and Pillay TS: Simplified molecular input line entry system-based descriptors in QSAR modeling for HIV-protease inhibitors. Chemometr Intell Lab Syst *153*: 67-74, 2016.
26 Weininger D: SMILES; a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci *28*: 31-36, 1988.

Received October 8, 2018
Revised October 19, 2018
Accepted October 23, 2018