European Journal of Pharmaceutical Sciences 52 (2014) 21–25

Contents lists available at ScienceDirect

European Journal of Pharmaceutical Sciences

journal homepage: www.elsevier.com/locate/ejps

CORAL software: Prediction of carcinogenicity of drugs by means of the Monte Carlo method

Alla P. Toropova, Andrey A. Toropov *

IRCCS, Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy

ARTICLE INFO

Article history:
Received 2 August 2013
Received in revised form 1 October 2013
Accepted 12 October 2013
Available online 26 October 2013

Keywords:
QSAR
Monte Carlo method
Carcinogenicity of drug
CORAL software

ABSTRACT

Methodology of building up and validation of models for carcinogenic potentials of drugs by means of the CORAL software is described. The QSAR analysis by the CORAL software includes three phases: (i) definition of preferable parameters for the optimization procedure that gives maximal correlation coefficient between endpoint and an optimal descriptor that is calculated with so-called correlation weights of various molecular features; (ii) detection of molecular features with stable positive correlation weights or vice versa stable negative correlation weights (molecular features which are characterized by solely positive or solely negative correlation weights obtained for several starts of the Monte Carlo optimization are a basis for mechanistic interpretations of the model); and (iii) building up the model that is satisfactory from point of view of reliable probabilistic criteria and OECD principles. The methodology is demonstrated for the case of carcinogenicity of a large set ($n = 1464$) of organic compounds which are potential or actual pharmaceutical agents.

1. Introduction

Quantitative structure–activity relationships (QSAR) are a tool to estimate/predict various endpoints (García et al., 2011; Garro Martinez et al., 2011; Mullen et al., 2011; Toropov et al., 2011; Furtula et al., 2012; Gramatica et al., 2012; Gutman, 2012; Gutman and Furtula, 2012; Ibezim et al., 2012; Vrontaki et al., 2012; Todeschini et al., 2012; Veselinović et al., 2013).

The carcinogenic activity exhibited by chemical substances is a toxicological endpoint of high health interest and worry (Kar and Roy, 2011; Duchowicz et al., 2012). There is a large group of QSAR models for carcinogenicity developed during past years by different researchers. On the other hand, Galvez has gathered the database on the carcinogenic activity in the Discriminant Function (DF) scale (DFcarc) for a wide set of 1815 organic compounds extracted from the Merck index, based on the annual report of carcinogenesis (Galvez, 2000). From this data set, different molecular subsets have been taken to establish QSAR models (Hemmateenejad et al., 2005; Deeb et al., 2007). A recent study (Kar and Roy, 2011) employs for the first time a greater number of carcinogenic compounds, having 1464 molecules from the Galvez data set involving many therapeutic agents. Next QSAR analysis of the same Galvez data has been carried out by other authors group using other approaches (Duchowicz et al., 2012).

The CORAL software is a tool for the QSAR analysis in general (Mullen et al., 2011; Ibezim et al., 2012; Veselinović et al., 2013) and for the QSAR analysis of carcinogenic endpoint in particular (Toropov et al., 2009a,b, 2010, 2011; Toropova et al., 2011a). Consequently, it is interesting task to check up the CORAL software as a tool for the QSAR analysis of the above-mentioned Galves data on the DF_{canc}.

Thus, the aim of the present study is the estimation of QSAR models for carcinogenic potential (DF_{canc}) calculated with the CORAL software.

2. Method

2.1. Data

Numerical data on carcinogenic potentials are available on the Internet (Galvez, 2000). Galves classified the 1815 compounds in 5 classes in the following manner: C = high expectancy of being carcinogenic (>90%); PC = probable carcinogenic activity (between 70% and 90%); I = high expectancy of being non-carcinogenic (>90%); PI = probable non-carcinogenic activity (between 70% and 90%); U = non-classified. The 345 non-classified compounds were removed in order to get robust dataset (Kar and Roy, 2011; Duchowicz et al., 2012). In addition six compounds were excluded owing to their atypical nature (Kar and Roy, 2011).

Numerical data on carcinogenic potentials of the selected 1464 organic compounds (chemical domain which includes hydrocarbons, aliphatic alcohols, phenols, ethers, and esters; anilines,

* Corresponding author. Tel.: +39 0239014805.
E-mail address: andrey.toropov@marionegri.it (A.A. Toropov).

0928-0987/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ejps.2013.10.005

amines, nitriles, nitroaromatics, amides, and carbamates; urea and thiourea derivatives, isothiocyanates, thiols, phosphate esters, and halogenated derivatives) are expressed by DF (Discriminant Function). The range of DF is from −9.91 to 9.86. Positive value of DF is an indicator of carcinogenic compounds, negative value of DF is an indicator of non-carcinogenic compounds. Three splits into the sub-training, calibration, test, and validation sets are examined. These splits are prepared according to the following principles: (i) they are random; (ii) they are different (Table 1); and (iii) each set contains about 25% of the 1464 compounds. Canonic (Weininger 1988, 1990; weininger et al. 1989) for these compounds are prepared with ACD/ChemSketch software (ACD/ChemSketch, 2007).

The roles of these sets are different: sub-training set is the “developer” of the model since correlation weights of compounds from the set are used to build up the model; calibration set is the “critic” of the model since data from this set are used to check whether model is working for compounds which are absent in the sub-training set; the test set is “estimator” of the model in cases of various threshold values; finally, the “invisible” validation set is used for the final estimation of the model with threshold value which gives the best statistical quality for the test set, thus the sub-training, calibration, and test sets are “visible” during building up model, but no information on “invisible” validation set is used in the modeling process (Toropov et al., 2013).

Fig. 1 contains the histogram of distribution of compounds according to DF_{canc} values.

2.2. Optimal descriptor

The model for carcinogenic potential expressed by DF is calculated as the following:

$$DF = C_0 + C_1 \cdot DCW(T, E, SMILES) \quad (1)$$

where *DCW*(*T*, *E*, *SMILES*) is optimal descriptor calculated with formula.

$$\begin{aligned} DCW(T, E, SMILES) = {} & \sum CW(S_k) + \sum CW(SS_k) \\ & + \sum CW(SSS_k) + CW(\mathrm{NOSP}) \\ & + CW(\mathrm{HALO}) + CW(\mathrm{BOND}) \end{aligned} \quad (2)$$

where *CW*(*X*) is correlation weight for a molecular feature extracted from simplified molecular input-line entry system (*SMILES*); S_k, SS_k, and SSS_k are fragments of *SMILES*.

Table 1
Percentage of identity of splits 1–3.

	Set	Split 1	Split 2	Split 3
Split 1	Sub-training	100[a]	25.9	27.2
	Calibration	100	25.0	22.8
	Test	100	26.8	27.6
	Validation	100	25.9	30.3
Split 2	Sub-training		100	26.3
	Calibration		100	24.4
	Test		100	27.4
	Validation		100	29.4
Split 3	Sub-training			100
	Calibration			100
	Test			100
	Validation			100

[a] *Identity* $(\%) = \frac{N_{i,j}}{0.5 * (N_i + N_j)} \times 100$ where $N_{i,j}$ is the number of substances which are distributed into the same set for both *i*-th split and *j*-th split (set = sub-training, calibration, test, validation); N_i is the number of substances which are distributed into the set for *i*-th split; N_j is the number of substances which are distributed into the set for *j*-th split.

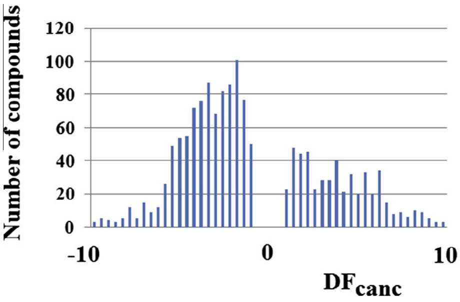


Fig. 1. The histogram of distribution of various DF_{canc} values (the range from −9.91 to 9.86).

For example, in the case of *SMILES* = CCCN

$$S_k = (\text{‘C’}, \text{‘C’}, \text{‘C’}, \text{‘N’})$$

$$SS_k = (\text{‘CC’}, \text{‘CC}, \text{‘CN’})$$

$$SSS_k = (\text{‘CCC’}, \text{‘CCN’})$$

NOSP, HALO, and BOND global molecular descriptors which reflect the presence (absence) of nitrogen, oxygen, sulphur, phosphorus (NOSP), fluorine, chlorine, and bromine (HALO), as well as presence (absence) of double (‘=’), triple (‘#’), and stereochemical (‘@’) covalent bonds (Toropova et al., 2011b). Fig. 2 contains example of calculation of these descriptors.

The Monte Carlo method optimization provides the numerical data on the correlation weights. The “visible” training set contains three subsets with different roles: sub-training set that is “developer” of the model; calibration set that is “critic” of the model; and test set that is “estimator” of the model. The “invisible” validation set contains external compounds which are not involved in the modeling process. *T* and *E* are parameters of the optimization procedure: *T* is threshold for definition of rare (noise) molecular features which should be blocked (i.e., their *CW* = 0) and *E* is the number of epochs of the optimization.

Building up of model by means of the CORAL software for a given split includes three phases. The first phase is selection of preferable T^* and E^* which give best statistic quality of the model *for the test set*. The second phase is calculation of model with *DCW*(T^*, E^*, *SMILES*). Third phase is the checking up of the model with *the validation set*. Fig. 3 gives the graphical representation of this optimization task.

Having carried out several runs of the Monte Carlo optimization, one can get lists of molecular features which are characterized by solely positive correlation weight (these can be interpreted as promoters of endpoint increase) together with features which are characterized by solely negative correlation weight (these can be interpreted as promoters of endpoint decrease). The role of molecular features which have both positive and negative correlation weight is not clear (Toropova et al., 2011b; Toropov et al., 2013).

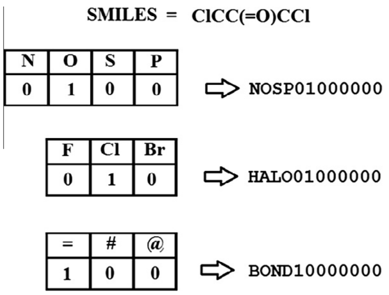


Fig. 2. Example of calculation of global *SMILES*-attributes NOSP, HALO, and BOND.

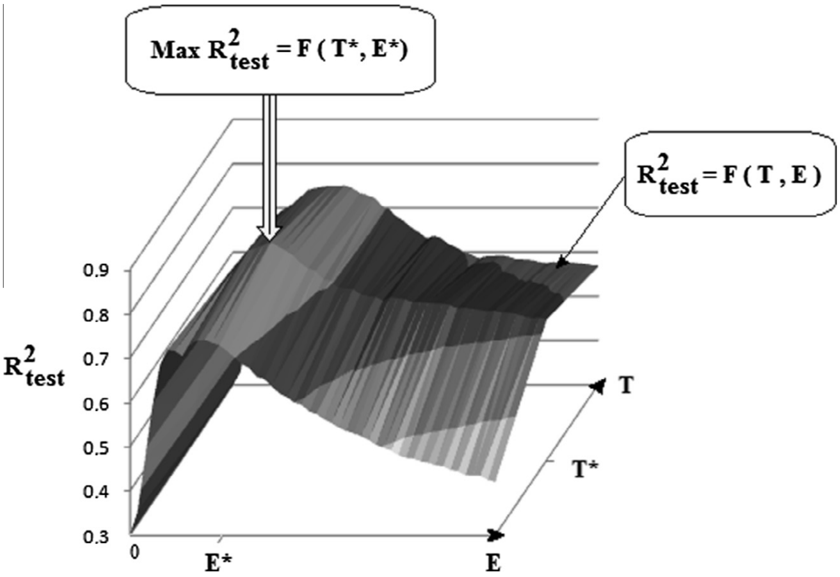


Fig. 3. The graphical representation of the selection (i) the threshold (T^*); and (ii) the number of epochs of the Monte Carlo optimization (E^*).

The percentage of molecular features with defined role for a *k*-th *SMILES* can be calculated as the following:

$$P_k = \frac{N(increase) + N(decrease)}{N(total)} \times 100\% \qquad (3)$$

where $N(increase)$, $N(decrease)$, and $N(total)$ are the number of molecular features extracted from *k*-th *SMILES* which are (i) promoters of increase, (ii) promoters of decrease, and (iii) total number of molecular features extracted from *k*-th *SMILES* (including blocked).

The percentage of molecular features with defined role (promoters of increase or decrease) can be a measure of probability of compounds fall in domain of applicability: if the percentage is large this probability is higher than in the case if this percentage is small. Ideal situation if the percentage is 100. However, in praxis, the percentage is less than 100. We believe that according to the percentage compound can be classified as one falls in the domain of applicability if take place inequality.

$$P_k > |\overline{P}_{TRN} - \Delta P_{TRN}| \qquad (4)$$

where $\overline{P}_{TRN}$ is average percentage of molecular features with defined role for the training set (i.e., for compounds involved in building up model) and ΔP_{TRN} is dispersion of the percentage on the training set.

It is to be noted that average value of the percentage and its dispersion are able to be criteria for the estimation of a split: split is satisfactory if (i) these values are similar for the training and the test set; (ii) the average percentage is as large as possible; and (iii) the dispersion is as small as possible.

3. Results and discussion

Fig. 4 represents data on the percentage of molecular features with defined role. According to above discussed logic all splits are satisfactory, but the split 2 and the split 3 are characterized by larger value of the percentage for both the "visible" set (i.e., the united set of sub-training, calibration, and test) and "invisible" validation set, hence these splits are more successful.

The analysis of data on the threshold (from 1 to 5) and the number of epochs (from 1 to 50) of the optimization gives preferable value $T^* = 2$ and $E^* = 20 \pm 2$ for all three splits. Table 2 contains the statistical quality of models for splits 1–3.

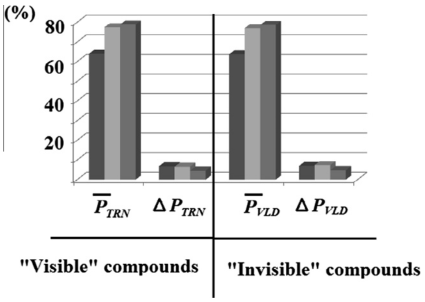


Fig. 4. Percentage of molecular features with defined role for "visible" (sub-training, calibration, and test sets) and "invisible" validation set for three splits.

In the case of the split 1, the model is the following:

$$\mathrm{DF_{canc}} = -0.0108\ (\pm\ 0.0048) + 0.2112\ (\pm\ 0.0003) * DCW(2, 22, SMILES) \qquad (5)$$

Fig. 5 contains the graphical representation for this model.

Table 3 contains comparison of statistical quality of models described in the literature (Kar and Roy, 2011; Duchowicz et al., 2012) and statistical quality suggested in this work for the case of the split 1. The comparison shows that suggested models for $\mathrm{DF_{canc}}$ are comparable with above-mentioned, hence the CORAL software can be estimated as an useful tool for the QSAR analysis of the carcinogenicity of drugs.

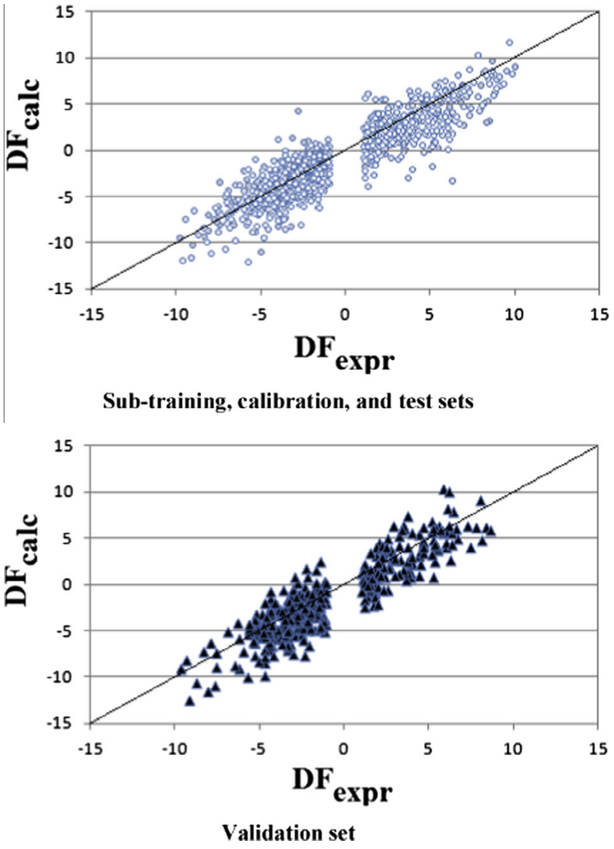
In order to estimate influence of a molecular feature to carcinogenic potential one should take into account two circumstances: (i) whether correlation weight of the feature is solely positive (negative); and (ii) whether the molecular feature has significant prevalence in the training set. According to the logic, one can extract the following promoters of increase (positive correlation weights and considerable prevalence in the training set) for carcinogenic potentials:

1. The presence of aromatic cycles ("cc1", "c2c" fragments in *SMILES*).
2. The presence of nitrogen together with oxygen (NOSP1100).
3. The presence of double bonds ("C=", "O=", BOND100).

Table 2
Statitical quality of models for carcinogenicity: n is the number of compounds in a set; R is correlation coefficient; Q^2 is leave-one-out cross-validated correlation coefficient; $\overline{r_m^2}$ and Δr_m^2 are criteria of predictability described in the literature (Ojha et al., 2011; Roy et al., 2013).

Split	Sub-training set			Calibration set				Test set[a]					Validation set		
	n	R^2	Q^2	s	n	R^2	s	n	R^2	$\overline{r_m^2}$	Δr_m^2	s	n	R^2	s
1	386	0.894	0.893	1.37	336	0.885	1.44	366	0.835	0.81	0.003	1.57	376	0.852	1.53
2	356	0.897	0.896	1.36	376	0.848	1.55	366	0.850	0.82	0.027	1.60	366	0.849	1.57
3	367	0.866	0.857	1.60	312	0.859	1.50	416	0.800	0.77	0.041	1.32	369	0.862	1.45

[a] According to the literature (Ojha et al., 2011; Roy et al., 2013) $\overline{r_m^2}$ should be larger than 0.5 and Δr_m^2 should be less than 0.2.



Sub-training, calibration, and test sets

Validation set

Fig. 5. Graphical representation of the model for split 1 (Eq. (5)).

Table 3
Comparison of different models for carcinogenic potential of the same 1464 compounds.

	Training set			External set			References
	n	R^2	s	n	R^2	s	
1	–	–	–	732	0.713	–	Kar and Roy (2011)
2	732	0.74	2.04	732	0.77	1.91	Duchowicz et al. (2012)
3	1088	0.87	1.46	376	0.852	1.53	This work (split 1)

Similarly, the following molecular features should be interpreted as promoters of decrease for carcinogenic potential (negative correlation weights and considerable prevalence in the training set):

1. The presence of nitrogen (“N”, “NC”, “CN(“, “n”, “nc”, “ncc”, “CN1”).
2. The presence of sulphur (“S”, “SC”).
3. The presence of double and stereochemical bonds (“C=C”, “H@C”, BOND1010).

Table 4
The compliance to the OECD principles.

	QSAR model (for regulatory purposes) should obeys the following five OECD principles:	How a principle is taken into account in this work?
1	A defined endpoint	The carcinogenic activity of organic compounds that is represented by the Discriminant Function (DF) extracted from the Merck index, based on the annual report of carcinogenesis (Duchowicz et al., 2012)
2	An unambiguous algorithm	The Monte Carlo optimization which is represented by the CORAL software available on the Internet
3	A defined domain of applicability	The domain of applicability is defined by means of the percentage of molecular features with defined role (promoters of increase or decrease for the endpoint)
4	Appropriate measures of goodness-of-fit, robustness and predictivity	Large values of correlation coefficient and small root-mean-squared error; $\bar{r}_m^2$ and Δr_m^2 metrics (Roy et al. 2013) (Roy et al. 2013)
5	A mechanistic interpretation, if possible	Lists of stable promoters of endpoint increase and stable promoters of endpoint decrease

Table 4 contains the OECD principles and their realization for the case of the optimal descriptors.

The Supplementary materials section contains (i) the description (representation) of the CORAL method that is examined in this work (Table S1); and (ii) three splits of 1464 compounds into the sub-training, calibration, test, and validation sets (Table S2). These data can be used to reproduce the suggested model with the CORAL software available on the Internet (http://www.insilico.eu/coral).

4. Conclusions

QSAR model for carcinogenic potential of organic compounds which are therapeutic agents is suggested. The domain of applicability for the model are hydrocarbons, aliphatic alcohols, phenols, ethers, and esters; anilines, amines, nitriles, nitroaromatics, amides, and carbamates; urea and thiourea derivatives, isothiocyanates, thiols, phosphate esters, and halogenated derivatives. There is the mechanistic interpretation of the model by lists of statistically significant promoters of increase and statistically significant promoters of decrease for carcinogenic potential. Thus, building up of the model is carried out in accordance with the OECD principles.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ejps.2013.10.005.

References

ACD/ChemSketch Freeware, v. 12.01, 2007. Advanced Chemistry Development, Inc., Toronto, ON, Canada. <http://www.acdlabs.com>, (accessed 25.07.13).

Deeb, O., Hemmateenejad, B., Jaber, A., Garduno-Juarez, R., Miri, R., 2007. Effect of the electronic and physicochemical parameters on the carcinogenesis activity of some sulfa drugs using QSAR analysis based on genetic-MLR and genetic-PLS. Chemosphere 67, 2122–2130.

Duchowicz, P.R., Comelli, N.C., Ortiz, E.V., Castro, E.A., 2012. QSAR study for carcinogenicity in a large set of organic compounds. Curr. Drug Saf. 7, 282–288.

Furtula, B., Gutman, I., Ivanović, M., Vukičević, D., 2012. Computer search for trees with minimal ABC index (2012). Appl. Math. Comput. 219, 767–772.

Galvez, J., 2000. Personal webpage. <http://www.uv.es/~galvez/tablevi.pdf>, (accessed 10.13).

García, J., Duchowicz, P.R., Rozas, M.F., Caram, J.A., Mirífico, M.V., Fernández, F.M., Castro, E.A., 2011. A comparative QSAR on 1,2,5-thiadiazolidin-3-one 1,1-dioxide compounds as selective inhibitors of human serine proteinases. J. Mol. Graph. Model. 31, 10–19.

Garro Martinez, J.C., Duchowicz, P.R., Estrada, M.R., Zamarbide, G.N., Castro, E.A., 2011. QSAR study and molecular design of open-chain enaminones as anticonvulsant agents. Int. J. Mol. Sci. 12, 9354–9368.

Gramatica, P., Cassani, S., Roy, P.P., Kovarich, S., Yap, C.W., Papa, E., 2012. QSAR modeling is not "Push a button and find a correlation": a case study of toxicity of (Benzo-)triazoles on Algae. Mol. Inf. 31, 817–835.

Gutman, I., 2012. Bounds for all graph energies. Chem. Phys. Lett. 528, 72–74.

Gutman, I., Furtula, B., 2012. Vertex-degree-based molecular structure descriptors of benzenoid systems and phenylenes. J. Serb. Chem. Soc. 77, 1031–1036.

Hemmateenejad, B., Safarpour, M.A., Miri, R., Nesari, N., 2005. Toward an optimal procedure for PC-ANN model building: prediction of the carcinogenic activity of a large set of drugs. J. Chem. Inf. Model. 45, 190–199.

Ibezim, E., Duchowicz, P.R., Ortiz, E.V., Castro, E.A., 2012. QSAR on aryl-piperazine derivatives with activity on malaria. Chemometr. Intell. Lab. Syst. 110, 81–88.

Kar, S.K., Roy, K., 2011. Development and validation of a robust QSAR model for prediction of carcinogenicity of drugs. Indian J. Biochem. Biophys. 48, 111–122.

Mullen, L.M.A., Duchowicz, P.R., Castro, E.A., 2011. QSAR treatment on a new class of triphenylmethyl-containing compounds as potent anticancer agents. Chemometr. Intell. Lab. Syst. 107, 269–275.

Ojha, P.K., Mitra, I., Das, R.N., Roy, K., 2011. Further exploring rm 2 metrics for validation of QSPR models. Chemometr. Intell. Lab. Syst. 107, 194–205.

Roy, K., Chakraborty, P., Mitra, I., Ojha, P.K., Kar, S., Das, R.N., 2013. Some case studies on application of "rm 2" metrics for judging quality of quantitative structure-activity relationship predictions: emphasis on scaling of response data. J. Comput. Chem. 34, 1071–1082.

Todeschini, R., Consonni, V., Xiang, H., Holliday, J., Buscema, M., Willett, P., 2012. Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets. J. Chem. Inf. Model. 52, 2884–2901.

Toropov, A.A., Toropova, A.P., Benfenati, E., Manganaro, A., 2009a. QSAR modelling of carcinogenicity by balance of correlations. Mol. Divers. 13, 367–373.

Toropov, A.A., Toropova, A.P., Benfenati, E., 2009b. Additive SMILES-based carcinogenicity models: probabilistic principles in the search for robust predictions. Int. J. Mol. Sci. 10, 3106–3127.

Toropov, A.A., Toropova, A.P., Benfenati, E., 2010. SMILES-based optimal descriptors: QSAR modeling of carcinogenicity by balance of correlations with ideal slopes. Eur. J. Med. Chem. 45, 3581–3587.

Toropov, A.A., Toropova, A.P., Benfenati, E., Gini, G., Leszczynska, D., Leszczynski, J., 2011. SMILES-based QSAR approaches for carcinogenicity and anticancer activity: comparison of correlation weights for identical SMILES attributes. Anti-Cancer Agents Med. Chem. 11, 974–982.

Toropov, A.A., Toropova, A.P., Puzyn, T., Benfenati, E., Gini, G., Leszczynska, D., Leszczynksy, J., 2013. QSAR as a random event: models for nanoparticles uptake in PaCa2 cancer cells. Chemosphere 92, 31–37.

Toropova, A.P., Toropov, A.A., Diaza, R.G., Benfenati, E., Gini, G., 2011a. Analysis of the co-evolutions of correlations as a tool for QSAR-modeling of carcinogenicity: an unexpected good prediction based on a model that seems untrustworthy. Cent. Eur. J. Chem. 9, 165–174.

Toropova, A.P., Toropov, A.A., Benfenati, E., Gini, G., Leszczynska, D., Leszczynski, J., 2011b. CORAL: quantitative structure–activity relationship models for estimating toxicity of organic compounds in rats. J. Comput. Chem. 32, 2727–2733.

Veselinović, A.M., Milosavljević, J.B., Toropov, A.A., Nikolić, G.M., 2013. SMILES-based QSAR model for arylpiperazines as high-affinity 5-HT1A receptor ligands using CORAL. Eur. J. Pharm. Sci. 48, 532–541.

Vrontaki, E., Leonis, G., Papadopoulos, M.G., Simcic, M., Grdadolnik, S.G., Afantitis, A., Melagraki, G., Hadjikakou, S.K., Mavromoustakos, T., 2012. Comparative binding effects of aspirin and anti-inflammatory Cu complex in the active site of LOX-1. J. Chem. Inf. Model. 52, 3293–3301.

Weininger, D., 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. 28, 31–36.

Weininger, D., 1990. Smiles. 3. Depict. Graphical depiction of chemical structures. J. Chem. Inf. Comput. Sci. 30, 237–243.

Weininger, D., Weininger, A., Weininger, J.L., 1989. SMILES. 2. Algorithm for generation of unique SMILES notation. J. Chem. Inf. Comput. Sci. 29, 97–101.