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.

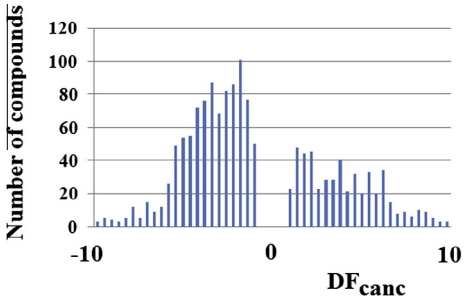

Fig. 1. The histogram of distribution of various DF_{canc} values (the range from −9.91 to 9.86).

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.

$$DCW(T,E,SMILES) = \sum CW(S_k) + \sum CW(SS_k) + \sum CW(SSS_k) + CW(\text{NOSP}) + CW(\text{HALO}) + CW(\text{BOND}) \quad (2)$$

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*.

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).

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.

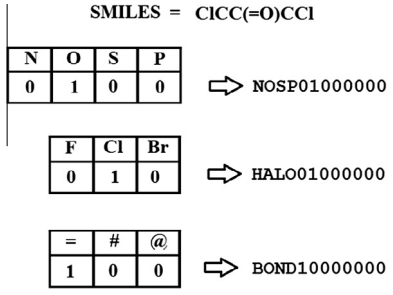

Fig. 2. Example of calculation of global *SMILES*-attributes NOSP, HALO, and BOND.