Mol Divers (2009) 13:367–373
DOI 10.1007/s11030-009-9113-4

SHORT COMMUNICATION

QSAR modelling of carcinogenicity by balance of correlations

A. A. Toropov · A. P. Toropova · E. Benfenati · A. Manganaro

Received: 19 August 2008 / Accepted: 12 January 2009 / Published online: 4 February 2009

Abstract Optimal descriptors based on the simplified molecular input line entry system (SMILES) have been utilized in modeling of carcinogenicity. Carcinogenicity of 401 compounds has been modeled by means of balance of correlations for the training (n = 170) and calibration (n = 170) sets. The obtained models were evaluated with an external test set (n = 61). Comparison of models based on the balance of correlations and models which were obtained on the basis of the total training set (i.e., both training and calibration sets as the united training set) has shown that the balance of correlations improves the statistical quality for the external test set.

Keywords QSAR · SMILES · Optimal descriptor · Carcinogenicity · Balance of correlations · Applicability domain

Introduction

Quantitative structure–property/activity relationships (QSPR/QSAR) are designed to predict the physicochemical and/or biochemical behaviour of substances which have not been examined experimentally. This can be done by analyzing parameters related to the molecular structures.

Carcinogenicity is an extremely complex biochemical phenomenon involving processes at the cellular level. The carcinogenicity of a substance depends on its molecular structure and a certain number of phenomena which are only partially known. Typically, the carcinogenic process involves one or more processes, showing a relationship with the mutagenic potential of a substance, but other processes are possible for carcinogens which are non mutagenic. Constructing a robust quantitative model based on information about the molecular structure is possible [1–5], and quantitative models for carcinogenicity have been reported [2–4]. However, more typically carcinogenicity evaluations aim to classify substances as active or inactive [5–7].

Thus, for carcinogenicity there have been both QSAR and SAR (structure–activity relationship) studies. In some cases the input of the models have been classical descriptors, but most typically in SAR studies chemical fragments have been used. Thus, in one case general features are responsible for the activity, while in the second case the toxic effect is due by the presence of a certain fragment responsible for the genotoxic effect. Historically the first approach is more related to the chemical QSAR models, which were addressing ecotoxicological endpoints such as aquatic toxicity. In these cases general features such as the partition coefficient between octanol and water have been more commonly used. The second approach, based on fragments, originates from the finding of toxicologists that some chemical resides, such as those present in nitrosoamines and epoxides, are at the basis of genotoxicity. Ashby et al. [8] listed a series of these fragments. Some models are formally a codification of fragments, such as by the programs HazardExpert [9] and Oncologic [10]. However, other approaches, still based on fragments, relied on software to identify fragments associated

Electronic supplementary material The online version of this article (doi:10.1007/s11030-009-9113-4) contains supplementary material, which is available to authorized users.

A. A. Toropov (✉) · A. P. Toropova
Institute of Geology and Geophysics, Khodzhibaev St. 49, 100041, Tashkent, Uzbekistan
e-mail: aatoropov@yahoo.com

A. A. Toropov · A. P. Toropova · E. Benfenati · A. Manganaro
Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156, Milano, Italy

to genotoxicity. This is the case of MULTICASE and of the model developed by Bursi et al. [11] for mutagenicity. Indeed, in case of data sets of thousands of compounds, such as in the case mutagenicity, computers may provide useful help to manage the huge amount of possible fragments and toxic effects. Results of models based on human expertise codified in rules and those from automatic learning gave similar results [12].

Some models, actually, involve a combination of tools, based on hazard on fragments or general features, such as with HazardExpert [9].

Our approach is somehow between the approach fragments and that on general parameters. The present study assessed optimal descriptors calculated with the simplified molecular input line entry system (SMILES) for modeling carcinogenic potentials by correlation balance between training and calibration sets.

Carcinogenicity data

Experimental values for carcinogenicity were taken from Ref. [7]. Carcinogenicity is expressed as the potency dose that induces cancer in rats (TD50, in mg/kg body weight). These values have been converted into mmol/kg body weight. The log(TD50) was examined as endpoint for modeling. We used all substances for which TD50 numerical values in mg/kg body weight for the rat have been reported and CAS numbers are given. Substances classified as not positive (NP) were not examined. In total 401 compounds were used.

We randomly split them into training ($n = 170$), calibration ($n = 170$) and test ($n = 61$) sets. The range of log(TD_{50}) values for these sets was very similar (about from -2 to 5).

Method

SMILES is a representation of the molecular structure by a sequence of symbols [13–15]. These are images of chemical elements (e.g., 'C', 'Br', 'Cl') and indicators of molecular attributes such as double ('=') or triple ('#') bonds, chiral center ('@'), *cis*- and *trans*-isomerism (by '/' and '\') and others [16,17].

Optimal descriptors calculated with molecular graphs are used for models of an assigned endpoint by one-variable correlation [18–20]. Databases available from internet contain SMILES [21,22] stimulated constructing SMILES based descriptors in general [23], and SMILES-based optimal descriptors [24,25], in particular. SMILES notations used in this study were obtained by the ChemSketch software [17], that generates canonical SMILES.

A SMILES-based optimal descriptors of correlation weights (DCW) used in this study were calculated as

$$\mathrm{DCW(LimS)} = \mathrm{CW(dC)}\,\Pi\,\mathrm{CW}(^{1}s_{k})\,\Pi\,\mathrm{CW}(^{2}s_{k})\,\Pi\,\mathrm{CW}(^{3}s_{k}) \tag{1}$$

where $^{1}s_{k}$, $^{2}s_{k}$ and $^{3}s_{k}$ are SMILES attributes (SA_{k}) of one or two or three elements. The element SMILES can be a symbol of the SMILES notation or two symbols: list of the SMILES elements is the following: "@@", "Br", "Cl", "#", "(", "+", "−", "/", "1", "2", "3", "4", "5", "=", "@", "C" (capital letter), "F", "H", "N", "O", "P", "S", "[", "\", "c" (lowercase letter), "n", "o", "s". It should be noted that the ")" is replaced by "(" since these symbols are indicators of the same phenomenon (branching). The SMILES attribute of the $^{1}s_{k}$ type contains only one SMILES element; attribute of the $^{2}s_{k}$ type contains two SMILES elements (e.g., "C___C___", "C___N___", etc.); the attribute of the $^{3}s_{k}$ type contains three SMILES elements (e.g., "C___N___1___", "C___O___2___", "Br__c___2___", etc).

For instance SMILES of "O = CC" is represented by the following SMILES attributes:

$^{1}s_{k}$ type (C____________, C____________, = ____________, O____________);

$^{2}s_{k}$ type (C___C________, C___ = ________, O___ = ________);

$^{3}s_{k}$ type (C___C___ = ___, O___ = ___C___);

dC (!-02_________).

CW(x) is the correlation weight for the SMILES attribute x. CWs are calculated by the Monte Carlo method optimization procedure [21,22] that provides CWs values which, used in Eq. 1, give a maximum correlation coefficient between the descriptor and carcinogenicity. There is an analogy between the three level separation of the SMILES notation in a $^{1}s_{k}$, $^{2}s_{k}$ and $^{3}s_{k}$ attributes and the extended connectivity of zero-(vertex), first- (edge), and second order (path of length 2) defined in a molecular graph [26–28]. We used the range of the SMILES elements according to ASCII codes of the symbols. In other words, each 'AB' composition can only have one representation (not both 'AB' and 'BA', and only 'ABC' not 'CBA').

Finally, the dC is the difference of the number of 'C' (capital letter) in the given SMILES notation minus the number of 'c' (lowercase letter) in the given SMILES notation. For example, this global SMILES attribute is denoted as '!001', if dC = N('c') – N('C') = 1, and as '!-02' if the dC = −2. The CW(dC) is the correlation weight of the dC. The symbol "C" (capital letter) is the representation of a carbon atom in the sp^{3} configuration. The symbol "c" (lowercase letter) is the representation of a carbon atom in sp^{2} configuration. Thus the dC is a some measure of presence of rigid and flexible fragments in molecular architecture. The examined substances contain chlorine that gives an additional 'C'. The chlorine is not rigid fragment in molecular system and we have calculated the dC taking into account the 'C' from chlorine atoms.

There are 747 SMILES attributes for the 401 substances under consideration. Some of the attributes are rare and some are absent in the training set. Rare attributes being used in this approach can lead to overtraining (i.e., the good statistical characteristics for the training and the calibration sets, but poor statistical characteristics for the test set). Hence detecting of rare SMILES attributes and blocking their influence are necessary.

In Ref. [23], to define the rare attributes a special scheme was used. The total number of SMILES attributes in the training set is the sum of all attributes from each SMILES notation of the training set. According to Ref. [23], one can use threshold (lim) to define rare attributes. If the lim = 3, then each SMILES attribute that takes place in the training set less than 3 times should be defined as rare. However, in this case two attributes which take place 3 times should be classified as not rare, even if the first takes place in the only one SMILES notation whereas the second in three SMILES notations. Taking into account possibility of described situation, a more adequate criterion to define rare attributes should be defined as the number of SMILES notations (in the training set) which contain the given attribute (SAk). In this study this criterion was used. This number is denoted as limS. If to use this criterion, then for the first above mentioned attribute the limS = 1, and for the second abovementioned attribute the limS = 3.

The optimization of the correlation weights used in Refs. [22–26] is based on the maximization of the correlation coefficient between the DCW-like descriptor and property/activity of interest. In the present study a novel target function has been studied. The target function is defined as

$$\text{Target Function} = R_T + R_C - \text{ABS}(R_T - R_C) * 0.1 \quad (2)$$

where R_T and R_C are correlation coefficients between the DCW and carcinogenicity for the training and calibration sets, respectively. In other words an attempt to obtain correlation weights which satisfy two conditions: first, R_T and R_C are as large as possible, and second, $R_T - R_C \to 0$ has been accomplished. Thus, the composition of the training set determinates the list of SA_k and the calibration set controls the balance of the correlation coefficient for the training and calibration set.

Thus, the role of the training set is a generator of the model (list of SMILES attributes and their optimal correlation weights), whereas the role of the calibration set is a provider of the robustness of the model (i.e., a tool to avoid the overtraining).

An informative characteristic of the distribution of the SA_k is

$$dP12(SA_k) = P_{TRN}(SA_k) - P_{CLB}(SA_k) \quad (3)$$

$$dP13(SA_k) = P_{TRN}(SA_k) - P_{TST}(SA_k) \quad (4)$$

$$dP23(SA_k) = P_{CLB}(SA_k) - P_{TST}(SA_k) \quad (5)$$

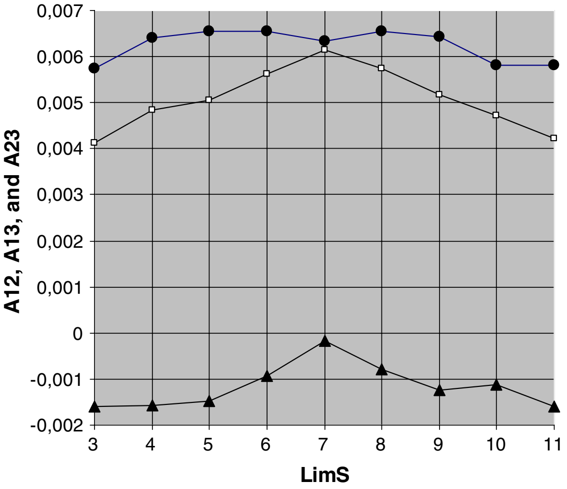

Fig. 1 Plot of LimS versus the A12 (circles), A13 (white squares), and A23 (triangles) indexes

where $P_{TRN}(SA_k)$, $P_{CLB}(SA_k)$, and $P_{TST}(SA_k)$ are the probability that the SMILES of the training, calibration, and test sets contain the given SA_k. The ideal situation is

$$dP12(SA_k) = dP13(SA_k) = dP23(SA_k) = 0 \quad (6)$$

The SMILES-based model can be characterized by the average values dP12, dP13, and dP23:

$$A12 = (1/N_{act}) \cdot \sum_{\text{all active SA}} dP12(SA_k) \quad (7)$$

$$A13 = (1/N_{act}) \cdot \sum_{\text{all active SA}} dP13(SA_k) \quad (8)$$

$$A23 = (1/N_{act}) \cdot \sum_{\text{all active SA}} dP23(SA_k) \quad (9)$$

These values give the possibility to compare different splits into training, calibration, and test sets: the preferable situation is when the values are close to zero. It is to be noted that the calculation with Eqs. 2–4 convey information about the relative prevalence of the SMILES attributes (SA_k) in the training, calibration, and test sets. For instance, if the probability of the presence of SA_k in the calibration set is more than the probability of the presence of SA_k in training set, then dP12 is less than zero. Calculations with Eqs. 7–9 are performed over all active SMILES attributes (not blocked for the given LimS). The blocked SA_k have not been taken into account calculating the A12, A13, and A23.

Results

The plot of LimS versus the A12, A13, and A23 (Fig. 1) indicates that LimS = 7 gave a minimum value for the average A12 and maximums for the A13 and A23. It will be seen below, the most robust prediction takes place near LimS = 7

Table 1 Statistical characteristics of the carcinogenicity models with LimS from 3 to 11

		Training set (n = 170)			Calibration set (n = 170)			Test set (n = 61)					
LimS	N_{act}	r^2	s	F	r^2	s	F	r^2	s	F	A12	A13	A23
3	341	0.8206	0.604	769	0.8206	0.616	769	0.5012	1.102	59	0.00573	0.00413	−0.00159
4	293	0.7861	0.660	618	0.7861	0.658	618	0.6072	0.859	93	0.00640	0.00484	−0.00157
5	261	0.7696	0.685	561	0.7689	0.680	559	0.6422	0.834	108	0.00654	0.00506	−0.00148
6	239	0.7498	0.714	504	0.7486	0.712	500	**0.7348**	0.676	164	0.00655	0.00562	−0.00093
7	212	0.7140	0.763	420	0.7138	0.758	419	0.7036	0.679	140	0.00633	0.00615	−0.00018
8	195	0.6922	0.792	378	0.6920	0.788	378	**0.7305**	0.643	161	0.00655	0.00575	−0.00080
9	176	0.6674	0.823	337	0.6673	0.821	337	0.7147	0.660	148	0.00642	0.00516	−0.00125
10	166	0.6597	0.833	326	0.6597	0.836	326	0.7068	0.679	142	0.00581	0.00471	−0.00111
11	153	0.6400	0.856	299	0.6401	0.859	299	0.6902	0.696	132	0.00581	0.00421	−0.00159

N_{act}, the number of active (i.e., not blocked) SMILES attributes
Bold indicates best limS values

(i.e., LimN = 6 and 8), but not in the same place. Our hypothesis is the following: minimums and maximums of the A12, A13, and A23 curves are indicators for searching robust versions of the DCW(LimS) models: in other words, the robust model might be near to these maximum and minimum values. However, this hypothesis must be checked with other substances and endpoints for a wider application to other models.

These indexes A12, A13, and A23 can be used for the definition of the applicability domain. In particular, one can hope to obtain reasonable prediction for some additional compounds in the test set, if their insertion does not considerably decrease the A13, because a considerable decrease of the A13 indicates that these additional compounds have SMILES attributes which are rare or absent in the training (see Eqs. 4, 8).

Table 1 shows the statistical quality of the models obtained with LimS from 3 to 11. LimN = 6 gave the best prediction for the test set. Figure 2 shows graphically the correlation coefficients for the training, calibration, and test sets in relation to LimS from 3 to 11.

One can see from Table 2 that the statistical characteristics of the DCW(6)-model are almost identical for the three probes of the Monte Carlo optimization.

The one-variable model for carcinogenicity, obtained in the first probe of the Monte Carlo optimization for DCW(6), is the following:

$$\begin{aligned}\log(TD_{50}) &= -45.2316(\pm 0.1667)\\ &\quad +44.3883(\pm 0.1588) * DCW(6)\\ n &= 170, r^2 = 0.752,\\ s &= 0.711 (\text{training set})\\ n &= 170, r^2 = 0.751,\\ R_{pred}{}^2 &= 0.745, SDEP = 0.648 (\text{calibration set})\end{aligned}$$

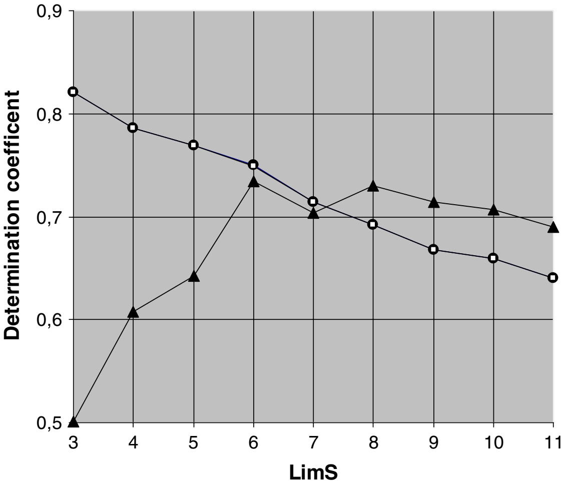

Fig. 2 Statistical quality of the models for the training (circles), calibration (white squares) and test (triangles) sets on different LimS Averages of three probes of the Monte Carlo optimization values of the determination (i.e., R^2) coefficients for the training, calibration, and test sets were used

$$\begin{aligned}n &= 61, r^2 = 0.723,\\ R_{pred}{}^2 &= 0.706, SDEP = 0.707 (\text{test set})\\ SDEP &= L\{\Sigma_{k=1}^{n}(y_{obs}[k] - y_{pred}[k])^2/n\}^{0.5} \qquad (10)\end{aligned}$$

where, the $y_{obs}[k]$, $y_{pred}[k]$ are values of the $\log(TD_{50})$ observed and predicted, respectively; n is the number of calibration or test set compounds.

The experimental values and those calculated with Eq. 10 for carcinogenicity are presented in the *electronic supplementary material*. Figure 3 shows graphically the regression curves of the predicted versus experimental values for the training, calibration, and test sets.

The best DCW(6) model obtained with the general training set (i.e., the set combining both the training and calibration, n = 340) has the following statistical characteristics

Table 2 Statistical characteristics of the carcinogenicity models with LimS = 6 for three probes of the Monte Carlo optimization

	Training set (n = 170)			Calibration set (n = 170)			Test set (n = 61)		
Probe	r^2	s	F	r^2	s	F	r^2	s	F
1	0.7518	0.711	509	0.7505	0.709	505	0.7234	0.696	154
2	0.7457	0.720	493	0.7454	0.714	492	0.7275	0.665	157
3	0.7519	0.711	509	0.7501	0.712	504	0.7535	0.666	180
Average	0.7498	0.714	504	0.7486	0.712	500	0.7348	0.676	164

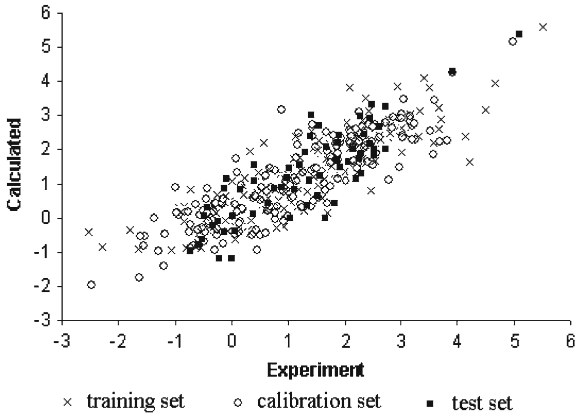

Fig. 3 Graphical representation of the model calculated with Eq. 10

n = 340, $r^2 = 0.772$, s = 0.680, F = 1143 (training set) and n = 61, $r^2 = 0.704$, s = 0.720, F = 140 (test set). One can see that the statistical characteristics for the united training set are better, but the statistical characteristics for the external test set are not better. Thus, the correlation balance gives an improvement for the carcinogenicity prediction (i.e., improvement for the external test set).

Discussion

The classification approach is more typically used in research dedicated to carcinogenicity [1]. However, quantitative models have more heuristic significance. In other words, there is a motivation to construct QSAR model carcinogenicity [29–31] because it can be useful for a general assessment of the chemical compounds. Indeed, the risk relative to a certain compound has to be established comparing the exposure level and the effect. Thus, information on the carcinogenicity expressed as a dose can be useful as it gives an indication of the carcinogenic potential which can be compared with the level of exposure for the human population.

The correlation coefficient for the carcinogenicity model described in Ref. [2] is about $r^2 = 0.85$, but the value is a measure of 5–7 variable models for 35 compounds belonging to the same chemical class of nitroso compounds. Vice versa our model refers to a much more demanding tasks, which is a prediction of a ten-times larger data set containing chemicals belonging to a wide variety of chemical classes, and acting through many different biochemical mechanisms. Artificial neural networks (ANN) approach has also been used in QSAR analysis of the carcinogenicity [3,4]. ANN models gave statistical quality similar to quality of the model calculated with Eq. 10. According to Ref. [4] the correlation coefficient of the QSAR for carcinogenicity of 45 benzene derivatives is about $r^2 = 0.7$. Recently suggested QSAR models for rodent carcinogenicity nitroso compounds obtained by multiple linear regression analysis (MLR) [31] reported the following statistical characteristics: n = 48, $r^2 = 0.859$, s = 0.361, F = 42 (training set) and n = 6, $Q^2 = 0.71$, s = 0.488 (test set). We have carried out QSAR analysis of this data: the correlation balance for these nitroso compounds (represented with SMILES generated by ChemSketch [17]) gave a model, characterized by the following statistical parameters: n = 23, $r^2 = 0.692$, s = 0.547, SDEP = 0.587, (training set); n = 23, $r^2 = 0.876$, $R_{pred}{}^2 = 0.831$, s = 0.892, (calibration set); and n = 8, $r^2 = 0.791$, $R_{pred}{}^2 = 0.689$,s = 0.389 (test set). Details of the QSAR model is presented in the *electronic supplementary material.*

QSAR models of toxicity towards rats for benzene derivatives built by MLR approach [32] have shown that the increase of the number of variable increases of the statistical quality of the toxicity model for the *training* set, but even three variable model has statistical characteristics for the *external test set* lower that two variable model. Cross validation criterions, without model estimation with an external test set, are not enough and can lead to wrong conclusions [33,34].

There are other studies in the literature reported very good performance for the QSAR models of carcinogenicity, but these addressed much more homogeneous data sets [35,36].

An important characteristic of our model is the attention to the validation. We used an external test set only for the validation purpose. The statistical quality of the model calculated with Eq. 10 is satisfactory, because this model gives results which are reasonable good for the *external test set* .

Thus, our model gives reasonably good results considering that they are checked with an external test set, and the model is based on a wide heterogeneous data set. Results with r^2 close to 0.7 are not suitable to be used as substitute of experimental models. However, these models can still be used for at least two purposes: (1) as additional information, because they can highlight the presence of possible carcinogenic risk, and (2) as screening tools. In both cases the use can be for priority setting, in order to identify chemicals to be evaluated first; as a method to identify more promising compounds, because with a lower potential toxicity. A chemical industry which wants to develop some chemicals, out of a wider series, may find this tool useful.

The probabilistic selection of the SMILES attributes (i.e., blocking of rare SA_k) gives two important features: (1) a robust model, calculated as an mathematical function of the SMILES characters; (2) a list of statistically significant SA_k, that can be used as a support for mechanistic interpretation of carcinogenicity as a chemical phenomenon. This is only part of the overall phenomenon, since carcinogenicity is a complex biological phenomenon.

In the *electronic supplementary material* we include represent a logical scheme for the estimation of different molecular fragment and different dC as potential promoters of increasing or decreasing of carcinogenicity. Results of these probabilistic estimations depend on the splits into the training, calibration, and test sets. Total list of the SA_k together with correlation weights for the three probes of the Monte Carlo optimization are given in the *electronic supplementary material*.

Generally the analysis for all compounds indicates that nitrogen (i.e., 'N' and 'n' symbols in the SMILES), together with the double bonds (i.e., '='), has a maximum prevalence in the group of attributes related to an increase of the carcinogenicity. This can be understood considering the fact that many aromatic amines and related compounds transformed into amines shows carcinogenic activity. Nitrosoamines are also often carcinogenic [37,38]. The presence of cycles (i.e., symbols of '1', '2', '3'), oxygen atoms, or halogens also promotes higher carcinogenicity values for the substances under consideration. There are well-known cases of carcinogens such as polycyclic aromatic compounds (presence of the 'c' in SMILES), epoxides, and dioxins.

While there are computer programs that encode human knowledge on these carcinogens, such as Oncologic [10,29] and DEREK [30], our approach automatically extracts the molecular features responsible of the activity. Our approach is a contribution to the QSPR/QSAR studies using an automatic process which starts from a simple chemical representation [39–42].

Conclusions

We developed a predictive model for carcinogenicity of chemical compounds based on SMILES format. The correlation balance (using preliminary checking of QSAR model with the calibration set) gave more robust prediction for carcinogenicity values on the external test set than a QSAR approach based on training and test sets (without calibration set). An important component in our SMILES-based optimal descriptors approach is removing (blocking) rare SMILES attributes, which, being used, can lead to the overtraining. The list of statistically significant SMILES attributes (promoters of both an increase and decrease carcinogenicity) is an heuristically useful component of this model.

This approach gives possibility to define the applicability domain. The applicability domain may be defined as chemicals with a SMILES containing the SMILES attributes which are not rare. To define the rare SMILES attributes, one can use the limS criterion.

Acknowledgements Authors are grateful to the EC project CAESAR (contract SSPI-022674-CAESAR) and the Marie Curie Fellowship for financial support (contract 39036, CHEMPREDICT).

References

1. Richardson SD, Plewa MJ, Wagner ED et al (2007) Occurrence, genotoxicity, and carcinogenicity of regulated and emerging disinfection by-products in drinking water: a review and roadmap for research. Mutat Res-Rew Mutat 636:178–242
2. Helguera AM, González MP, Cordeiro MNDS et al (2007) Quantitative structure carcinogenicity relationship for detecting structural alerts in nitroso-compounds. Toxicol Appl Pharm 221:189–202
3. Gini G, Lorenzini M, Benfenati E et al (1999) Predictive carcinogenicity: a model for aromatic compounds, with nitrogen-containing substituents, based on molecular descriptors using an artificial neural network. J Chem Inf Comput Sci 39:1076–1080
4. Vracko M (1997) A study of structure-carcinogenic potency relationship with artificial neural networks. The using of descriptors related to geometrical and electronic structures. J Chem Inf Comput Sci 37:1037–1043
5. Contrera JF, Kruhlak NL, Matthews EJ et al (2007) Comparison of MC4PC and MDL-QSAR rodent carcinogenicity predictions and the enhancement of predictive performance by combining QSAR models. Regul Toxicol Pharm 49:172–182
6. Richard AM (1994) Application of SAR methods to non-congeneric data bases associated with carcinogenicity and mutagenicity: issues and approaches. Mutat Res 305:73–97
7. <http://www.epa.gov/ncct/dsstox/sdf_cpdbas.html>
8. Ashby J (1994) Two million rodent carcinogens? The role of SAR and QSAR in their detection. Mutat Fund-Mol M 305:3–12
9. http://www.compudrug.com/
10. http://www.epa.gov/oppt/newchems/tools/oncologic.htm
11. Kazius J, McGuire R, Bursi RJ (2005) Derivation and validation of toxicophores for mutagenicity prediction. J Med Chem 48: 312–320
12. Benigni R, Netzeva T, Benfenati E et al (2007) The expanding role of predictive toxicology: an update on the (Q)SAR models for mutagens and carcinogens. J Environ Sci Heal Part C 25:53–97
13. 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
14. Weininger D, Weininger A, Weininger JL (1989) SMILES. 2. Algorithm for generation of unique SMILES notation. J Chem Inf Comput Sci 29:97–101
15. Weininger D (1990) SMILES. 3. DEPICT. Graphical depiction of chemical structures. J Chem Inf Comput Sci 30:237–243

16. <http://www.daylight.com>
17. ACD/ChemSketch Freeware (2007) version 11.00, Advanced Chemistry Development, Inc., Toronto, ON, Canada, http://www.acdlabs.com
18. Randic M, Dobrowolski JC (1998) Optimal molecular connectivity descriptors for nitrogen-containing molecules. Int J Quant Chem 70:1209–1215
19. Randic M, Basak SC (2000) Construction of high-quality structure–property–activity regressions: the boiling points of sulfides. J Chem Inf Comput Sci 40:899–905
20. Toropov AA, Toropova AP (1998) Optimization of correlation weights of the local graph invariants: use of the enthalpies of formation of complex compounds for the QSPR modeling. Russ J Coord Chem 24:81–85
21. <http://chem.sis.nlm.nih.gov/chemidplus/>
22. <http://webbook.nist.gov/chemistry/>
23. Vidal D, Thormann M, Pons M. (2005) LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. J Chem Inf Model 45:386–393
24. Toropov AA, Benfenati E. (2007) Optimisation of correlation weights of SMILES invariants for modelling oral quail toxicity. Eur J Med Chem 42:606–613
25. Toropov AA, Toropova AP, Mukhamedzhanova D et al (2005) Simplified molecular input line entry system (SMILES) as an alternative for constructing quantitative structure–property relationships (QSPR). Indian J Chem Sec A 44:1545–1552
26. Toropov AA, Toropova AP (2003) QSPR modeling of alkanes properties based on graph of atomic orbitals. J Mol Struct (Theochem) 637:1–10
27. Toropov AA, Benfenati E (2006) Correlation weighting of valence shells in QSAR analysis of toxicity. Bioorg Med Chem 14:3923–3928
28. Raska IJr, Toropov A (2006) Comparison of QSPR models of octanol/water partition coefficient for vitamins and non vitamins. Eur J Med Chem 41:1271–1278
29. <http://www.epa.gov/oppt/cahp/pubs/can.htm>
30. <http://www.drtak.org/develops/RTK/>
31. Helguera AM, Cordeiro MNDS, Perez MAC et al (2008) QSAR modeling of the rodent carcinogenicity of nitrocompounds. Bioorg Med Chem 16:3395–3407
32. Toropov AA, Rasulev BF, Leszczynski J (2007) QSAR modeling of acute toxicity for nitrobenzene derivatives towards rats: comparative analysis by MLRA and optimal descriptors . QSAR Comb Sci 26:686–693
33. Golbraikh A, Tropsha A (2002) Beware of q2!. J Mol Graph Model 20:269–276
34. Johnson SR (2008) The trouble with QSAR (or how i learned to stop worrying and embrace fallacy). J Chem Inf Model 48:25–26
35. Mayer J, Cheeseman MA, Twaroski ML (2008) Structure–activity relationship analysis tools: validation and applicability in predicting carcinogens. Regul Toxicol Pharm 50:50–58
36. Morales AH, Duchowicz PR, Pérez MÁC et al (2006) Application of the replacement method as a novel variable selection strategy in QSAR. 1. Carcinogenic potential. Chemom Intell Lab 81:180–187
37. Andrzejewski P, Kasprzyk-Hordern B, Nawrocki J (2005) The hazard of N-nitrosodimethylamine (NDMA) formation during water disinfection with strong oxidants. Desalination 176:37–45
38. <http://ecb.jrc.it>
39. Duchowicz PR, Mercader AG, Fernández FM et al (2008) Prediction of aqueous toxicity for heterogeneous phenol derivatives by QSAR. Chemom Intell Lab 90:97–107
40. Casanola-Martın GM, Marrero-Ponce Y, Hassan Khan MT et al (2007) TOMOCOMD-CARDD descriptors-based virtual screening of tyrosinase inhibitors: evaluation of different classification model combinations using bond-based linear indices. Bioorg Med Chem 15:1483–1503
41. Toropov AA, Rasulev BF, Leszczynski J (2008) QSAR modeling of acute toxicity by balance of correlations. Bioorg Med Chem 16:5999–6008
42. Roy PP, Leonard JT, Roy K (2008) Exploring the impact of size of training sets for the development of predictive QSAR models. Chemom Intell Lab 90:31–42