Chemosphere 124 (2015) 40–46

Contents lists available at ScienceDirect

Chemosphere

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

Quasi-QSAR for mutagenic potential of multi-walled carbon-nanotubes

Andrey A. Toropov *, Alla P. Toropova

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

HIGHLIGHTS

- Quasi-QSAR for MWCNTs is suggested.
- The model of mutagenicity is a mathematical function of conditions.
- The statistical quality of the quasi-QSAR is quite good.

ARTICLE INFO

Article history:
Received 22 July 2014
Received in revised form 15 October 2014
Accepted 18 October 2014
Available online 20 November 2014

Handling Editor: Tamara S. Galloway

Keywords:
MWCNT
Mutagenicity TA100
Monte Carlo method
Quasi-QSAR

ABSTRACT

Available on the Internet, the CORAL software (http://www.insilico.eu/coral) has been used to build up quasi-quantitative structure–activity relationships (quasi-QSAR) for prediction of mutagenic potential of multi-walled carbon-nanotubes (MWCNTs). In contrast with the previous models built up by CORAL which were based on representation of the molecular structure by simplified molecular input-line entry system (SMILES) the quasi-QSARs based on the representation of conditions (not on the molecular structure) such as concentration, presence (absence) S9 mix, the using (or without the using) of preincubation were encoded by so-called quasi-SMILES. The statistical characteristics of these models (quasi-QSARs) for three random splits into the visible training set and test set and invisible validation set are the following: (i) split 1: $n = 13$, $r^2 = 0.8037$, $q^2 = 0.7260$, $s = 0.033$, $F = 45$ (training set); $n = 5$, $r^2 = 0.9102$, $s = 0.071$ (test set); $n = 6$, $r^2 = 0.7627$, $s = 0.044$ (validation set); (ii) split 2: $n = 13$, $r^2 = 0.6446$, $q^2 = 0.4733$, $s = 0.045$, $F = 20$ (training set); $n = 5$, $r^2 = 0.6785$, $s = 0.054$ (test set); $n = 6$, $r^2 = 0.9593$, $s = 0.032$ (validation set); and (iii) $n = 14$, $r^2 = 0.8087$, $q^2 = 0.6975$, $s = 0.026$, $F = 51$ (training set); $n = 5$, $r^2 = 0.9453$, $s = 0.074$ (test set); $n = 5$, $r^2 = 0.8951$, $s = 0.052$ (validation set).

1. Introduction

Quantitative structure–property/activity relationships (QSPRs/QSARs) based on various descriptors (Gutman et al., 2005; Furtula and Gutman, 2011) are a tool of investigation of physicochemical (Torrens and Castellano, 2012; Nesmerak et al., 2013, 2014; Achary, 2014a), biological (Rallo et al., 2005; Torrens and Castellano, 2014; Achary, 2014b), and therapeutical (Afantitis et al., 2011; Veselinović et al., 2013a,b; Comelli et al., 2014; Deng et al., 2014a,b) behavior of various substances.

There are the considerable number of attempts to carry out the QSPR/QSAR analyses of various nanomaterials (Kar et al., 2014; Toropov et al., 2013; Pathakoti et al., 2014; Singh and Gupta, 2014). However, the main problem of this fresh field of natural sciences is deficit of available experimental data on physicochemical parameters and biological activity of these substances (González-Díaz et al., 2013; Winkler et al., 2013).

Traditional QSAR approaches were used to build up nano-QSAR in recent work (Ibrahim et al., 2010; Fourches et al., 2010; Shahlaei et al., 2014). It is paradox, but quantum mechanics descriptors often are involved as a tool to build up a model for nanomaterials (Ibrahim et al., 2010; Shahlaei et al., 2014), in spite of large size of their molecules. Finally, the prediction of endpoints related to nanomaterials as a mathematical function of their physicochemical properties can be prepared if these data are available (Sayes and Ivanov, 2010).

So-called optimal descriptors calculated with the Monte Carlo technique are an attractive alternative for the above-mentioned approaches, because (i) these descriptors can be calculated from arbitrary eclectic information (Toropova and Toropov, 2013; Toropov and Toropova, 2014); and (ii) these descriptors can be easily modified for fresh experimental data if these will become available (Toropov et al., 2010; Toropova et al., 2010).

Analysis of the state-of-art situation with QSAR theory for nanomaterials has shown: large standardized databases on the nanomaterials and their endpoints is absent, but there are small data related sometimes to the same endpoint, such as cytotoxicity

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

http://dx.doi.org/10.1016/j.chemosphere.2014.10.067
0045-6535/

of nano metal oxides, membrane damage implicated by various nanoparticles.

Under such circumstances, one can define the two principles for building up and validation of QSAR related to various nanomaterials: (i) the optimal descriptor is able be a translator of eclectic information into prediction for an endpoint; (ii) each QSPR/QSAR model is a random event. In other words, in advance (Toropova and Toropov, 2013; Toropov and Toropova, 2014), (i) one can establish local quasi-QSAR which are based on available data on phenomena related to nanomaterials, which often are not molecular architecture in classic meaning; and (ii) the local quasi-QSAR should be checked up for different distributions into visible training and invisible validation sets. Finally, it is to be noted that building up of quasi-QSPR/QSAR models for substances which are not nanomaterials also can be useful from practical point of view.

The basic principles of the design of the optimal descriptors can be formulated as the following:

(i) One should collect information on all circumstances which are able to influence upon biochemical behavior of the complex system. The complex system can be e.g. interaction of nanomaterial with membrane (Toropova and Toropov, 2013); interaction of peptides (Vishnepolsky and Pirtskhalava, 2014); skin whitening cosmetics (Alqadami et al., 2013) and any other situation where one can control by behavior of system which united a large group of elements.
(ii) One should define preliminary hierarchy of the above elements.
(iii) The traditional paradigm of the QSPR/QSAR analysis "Endpoint = Mathematical function of the molecular structure" must be replaced by paradigm "Endpoint = Mathematical function of all available eclectic data". It is to be noted, that the "molecular structure" can be a "specific eclectic data". Quasi-SMILES (Toropova and Toropov, 2013; Toropov and Toropova, 2014) which are representation of behavior of the complex system in contrast to traditional SMILES (Weininger, 1988; Weininger et al., 1989; Weininger, 1990) which are representation of the molecular structure should be prepared and distributed into the visible training and invisible validation sets.
(iv) Rational discrimination of elements (attributes) of the complex system in accordance with their the frequencies (in the visible training set and invisible validation set) and their the information contributions (correlation weights) should be carried out by means of comparison of series of Monte Carlo optimizations.
(v) The validation of the predictive potential of the preferable rational model should be carried out with quasi-SMILES of the validation set which are invisible during building up the model.

The aim of the present work is the estimation of ability of the optimal descriptors to be a tool to build up quasi-QSAR for the mutagenic potential of multi-walled carbon-nanotubes (MWCNTs).

2. Method

2.1. Data

The numerical data on mutagenic potential of MWCNTs taken from the literature (Wirnitzer et al., 2009). Mean mutant counts after incubation of Salmonella strains TA100 without and with metabolic activation (S9 mix) in the plate incorporation and in the preincubation (tube) part of the *Salmonella* microsome test. These eclectic data (conditions) related to biologic behavior of MWCNTs can be represented by sequences of symbols similar to well-known SMILES (Weininger, 1988; Weininger et al. 1989; Weininger, 1990). However, since the traditional SMILES are representation of the molecular structure, whereas the SMILES used in this work are representation of conditions, these last ones should be named "quasi-SMILES".

The negative value of decimal logarithm of the TA100 (pTA100) has been examined as the endpoint for the quasi-QSARs. Table 1 contains the available eclectic data used in this work for the MWCNTs. Table 2 contains quasi-SMILES used to represent the data. These twenty four quasi-SMILES are distributed into the training, test, and validation sets three times. These distributions have been done according two principles (i) these are random; and (ii) the ranges in sets (training, test and validation) are comparable. Table 3 represents these distributions. Data collected in the training set are involved to build up model. Data collected in test set are used in order to check up and avoid the overtraining. Data collected in validation set are used to estimate predictive potential of these models.

2.2. Optimal descriptors

The optimal descriptors have been calculated with attributes (A_k) of quasi-SMILES which are representing conditions of the biological acting of MWCNTs. The calculation of descriptors is carried out by means of formula:

$$\mathrm{DCW}(T,N) = \Sigma \mathrm{CW}(A_k) \quad (1)$$

Table 1 contains A_k used in this work. The $\mathrm{CW}(A_k)$ is correlation weight for an attribute A_k, that is extracted from quasi-SMILES (Table 2); the T is the threshold to divide attributes into two categories rare (noise) or not rare: Correlation weights are calculated for not rare attributes by the Monte Carlo optimization that gives maximum of correlation coefficient between $\mathrm{DCW}(T,N)$ and pTA100. The N is the number of epochs of the Monte Carlo optimization. These parameters (T and N) have considerable influence for the statistical quality of a model (Toropov et al., 2013). Consequently, one should define the preferable T^* and N^* which give maximal correlation coefficient between the pTA100 and $\mathrm{DCW}(T,N)$ for the test set (Toropov et al., 2013). Having $\mathrm{CW}(A_k)$ which give maximum of the correlation coefficient one can define using data from the training set the following model:

$$\mathrm{pTA100} = C_0 + C_1 * \mathrm{DCW}(T^*, N^*) \quad (2)$$

The statistical quality of the model calculated with Eq. (2) for the test set (not for the training set) is the real criterion for the preliminary estimation of the predictive potential of the model: the data on the test set give possibility to avoid the overtraining of the model (situation where excellent quality for the training set is accompanied by the poor statistics for test set). The data

Table 1
The available eclectic data used in this work for the MWCNTs together with codes used to represent the data.

Attribute	Codes of attributes (A_k) and their meaning
Preincubation	'*y*' = with preincubation '*n*' = without preincubation
Mix S9	'+' = with Mix S9 '–' = without Mix S9
Dose (μg/plate)	'*A*' = 0 '*B*' = 50 '*C*' = 158 '*D*' = 500 '*E*' = 1581 '*F*' = 5000

Table 2
The available eclectic data used in this work for MWCNTs together with quasi-SMILES used to represent conditions of acting of the MWCNTs.

No.	Quasi-SMILES	TA100	pTA100
1	*A* – *n*	121	−2.083
2	*B* – *n*	130	−2.114
3	*C* – *n*	121	−2.083
4	*D* – *n*	124	−2.093
5	*E* – *n*	111	−2.045
6	*F* – *n*	94	−1.973
7	*A* – *y*	132	−2.121
8	*B* – *y*	133	−2.124
9	*C* – *y*	126	−2.100
10	*D* – *y*	130	−2.114
11	*E* – *y*	128	−2.107
12	*F* – *y*	123	−2.090
13	*A* + *n*	136	−2.134
14	*B* + *n*	143	−2.155
15	*C* + *n*	136	−2.134
16	*D* + *n*	127	−2.104
17	*E* + *n*	117	−2.068
18	*F* + *n*	114	−2.057
19	*A* + *y*	188	−2.274
20	*B* + *y*	188	−2.274
21	*C* + *y*	190	−2.279
22	*D* + *y*	182	−2.260
23	*E* + *y*	175	−2.243
24	*F* + *y*	173	−2.238

Table 3
Distribution of available data into the visible training and test sets, and invisible validation sets.

	Training set	Test set	Validation set
Split 1	1 2 3 4 8 10 11 12 16 17 19 21 23	6 9 14 18 24	5 7 13 15 20 22
Split 2	1 2 3 4 6 8 10 11 12 16 17 19 23	5 13 14 18 21	7 9 15 20 22 24
Split 3	1 2 6 7 8 9 10 11 12 14 15 16 17 22	3 5 20 21 24	4 13 18 19 23

on the both training and test sets are visible during building up the model. Therefore, the predictability of the model should be additionally checked up with the external "invisible" validation set (Toropov et al., 2013).

The measure of the statistical (probabilistic) quality of a features which are extracted from quasi-SMILES can be calculated as the following:

$$\mathrm{Defect}(A_k) = \begin{cases} \frac{|P_{\mathrm{TRN}}(A_k) - P_{\mathrm{TST}}(A_k)|}{N_{\mathrm{TRN}}(A_k) + N_{\mathrm{TST}}(A_k)}, & \text{if } N_{\mathrm{TST}}(A_k) > 0 \\ 1, \text{otherwise} \end{cases} \quad (3)$$

where the $P_{\mathrm{TRN}}(A_k)$ is the probability of presence of the A_k in SMILES of the training set, i.e.

$P_{\mathrm{TRN}}(A_k) = N_{\mathrm{TRN}}(A_k)/N_{\mathrm{TRN}}$.

The $P_{\mathrm{TRN}}(A_k)$ is the probability of presence of the A_k in SMILES of the test set, i.e.

$P_{\mathrm{TST}}(A_k) = N_{\mathrm{TST}}(Ak)/N_{\mathrm{TST}}$.

The $N_{\mathrm{TRN}}(A_k)$ is the number (frequency) of SMILES which contain A_k in the training set.

The N_{TRN} is the total number of SMILES in the training set.

The $N_{\mathrm{TST}}(A_k)$ is the number (frequency) of SMILES which contain Ak in the test set.

The N_{TST} is the total number of SMILES in the test set.

The logic: if the probability of A_k in the training set is equal to the probability of A_k in the test set it is the ideal situation and the defect is zero. However, this situation is not typical, i.e. the difference between the probability of A_k in the training set and the probability of A_k in the test set is not zero. Under such circumstances, the frequency of A_k in the sub-training set and in the test set also should be taken into account: if these are small then the defect of A_k must be larger. Finally, if A_k is absent in the test set, the Defect(A_k) is maximal. Thus, the measure calculated with Eq. (3) can be used for the classification of the active (not blocked) attributes.

2.3. The selection of quasi-SMILES into the domain of applicability

Having the numerical data on the Defect(A_k) one can compare reliability of the prediction for an quasi SMILES, using the following criterion Defect(quasiSMILES):

$$\mathrm{Defect(quasiSMILES)} = \Sigma \mathrm{Defect}(A_k) \quad (4)$$

The domain of applicability can be defined as the following: quasi-SMILES falls into the domain of applicability if its Defect(quasiSMILES) obeys the condition:

$$\mathrm{Defect(quasiSMILES)} < 2 * \overline{\mathrm{Defect(quasiSMILES)}} \quad (5)$$

where $\overline{\mathrm{Defect(quasiSMILES)}}$ is average for visible set (training and test sets).

Thus the DefectSMILES gives possibility to define the domain of applicability for quasi-SMILES.

3. Results and discussion

The quasi-QSAR for mutagenic potentials of MWCNTs for three random splits into the visible training and test sets and invisible external validation set (Table 3) are characterized by the following statistical parameters:

Split 1

$$\mathrm{pTA100} = -2.2606(\pm 0.0060) + 0.0390(\pm 0.0019) * \mathrm{DCW}(2,9) \quad (6)$$

$n = 13$, $r^2 = 0.8037$, $^{c}R_p^2 = 0.7493$, $q^2 = 0.7260$, $s = 0.033$, $F = 45$ (training set).

$n = 5$, $r^2 = 0.9102$, $^{c}R_p^2 = 0.7268$, $s = 0.071$ (test set).

Validation set:

$n = 6$, $r^2 = 0.7627$, $s = 0.044$.

$$\frac{r^2 - r_0^2}{r^2} = 0.0875 < 0.1$$

$$\frac{r^2 - r_0'^2}{r^2} = 0.0001 < 0.1$$

$$k = 1.0034 (0.85 < k < 1.15)$$

$$k' = 0.9963 (0.85 < k' < 1.15)$$

Split 2

$$\mathrm{pTA100} = -2.3041(\pm 0.0165) + 0.0672(\pm 0.0057) * \mathrm{DCW}(3,7) \quad (7)$$

$n = 13$, $r^2 = 0.6446$, $^{c}R_p^2 = 0.6341$, $q^2 = 0.4733$, $s = 0.045$, $F = 20$ (training set).

$n = 5$, $r^2 = 0.6785$, $^{c}R_p^2 = 0.5159$, $s = 0.054$ (test set).

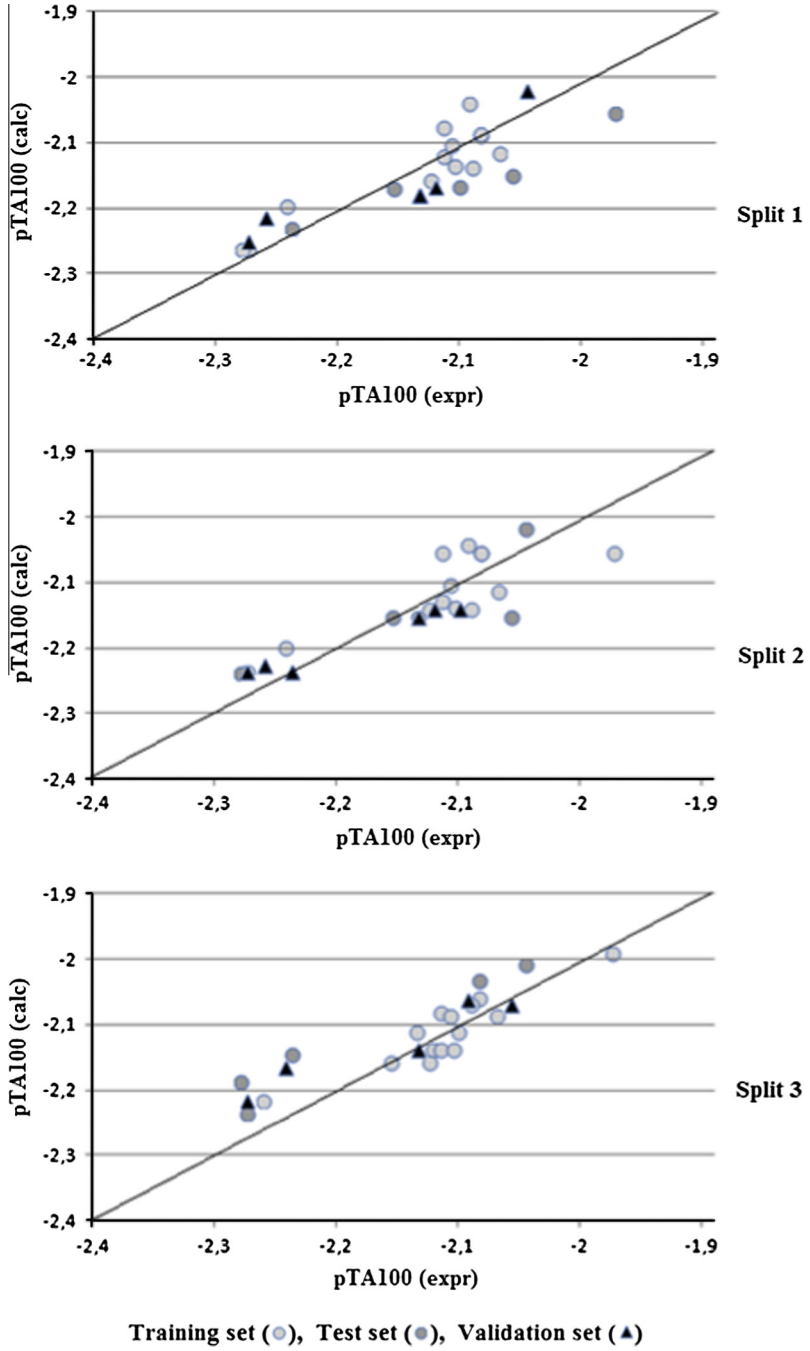

Fig. 1. Graphical representation of quasi-QSARs for mutagenicity of MWCNTs (pTA100) calculated with Eqs. (6)–(8).

Validation set:

$n = 6$, $r^2 = 0.9593$, $s = 0.032$.

$$\frac{r^2 - r_0^2}{r^2} = 0.3943 < 0.1$$

$$\frac{r^2 - r_0'^2}{r^2} = 0.1294 < 0.1$$

$$k = 0.9999(0.85 < k < 1.15)$$

$$k' = 0.9999(0.85 < k' < 1.15)$$

Split 3

$$\mathrm{pTA100} = -6.6503(\pm 0.2944) + 1.4829(\pm 0.0959) * \mathrm{DCW}(1,3) \quad (8)$$

$n = 14$, $r^2 = 0.8087$, ${}^{c}R_p{}^2 = 0.7927$, $q^2 = 0.6975$, $s = 0.026$, $F = 51$ (training set).
$n = 5$, $r^2 = 0.9453$, ${}^{c}R_p{}^2 = 0.7245$, $s = 0.074$ (test set).

Validation set:

$n = 5$, $r^2 = 0.8951$, $s = 0.052$.

$$\frac{r^2 - r_0^2}{r^2} = 0.0698 < 0.1$$

Table 4
Domain of applicability (DA) for three random splits.

No.	Quasi-SMILES	Split 1		Split 2		Split 3	
		Defect(quasi-SMILES)	DA	Defect(quasi-SMILES)	DA	Defect(quasi-SMILES)	DA
1	*A* – *n*	1.0369	No	0.0730	YES	1.0332	No
2	*B* – *n*	0.0523	YES	0.0730	YES	0.0368	YES
3	*C* – *n*	0.0523	YES	0.0730	YES	0.0975	YES
4	*D* – *n*	1.0369	No	1.0730	No	1.0332	No
5	*E* – *n*	1.0369	No	0.0807	YES	0.0522	YES
6	*F* – *n*	0.0369	YES	0.0730	YES	0.0522	YES
7	*A* – *y*	1.0369	No	0.0866	YES	1.0321	No
8	*B* – *y*	0.0523	YES	0.0866	YES	0.0356	YES
9	*C* – *y*	0.0523	YES	0.0866	YES	0.0964	YES
10	*D* – *y*	1.0369	No	1.0866	No	1.0321	No
11	*E* – *y*	1.0369	No	0.0943	YES	0.0511	YES
12	*F* – *y*	0.0369	YES	0.0866	YES	0.0511	YES
13	*A* + *n*	1.0423	No	0.0853	YES	1.0415	No
14	*B* + *n*	0.0577	YES	0.0853	YES	0.0450	YES
15	*C* + *n*	0.0577	YES	0.0853	YES	0.1058	YES
16	*D* + *n*	1.0423	No	1.0853	No	1.0415	No
17	*E* + *n*	1.0423	No	0.0930	YES	0.0605	YES
18	*F* + *n*	0.0423	YES	0.0853	YES	0.0605	YES
19	*A* + *y*	1.0423	No	0.0989	YES	1.0404	No
20	*B* + *y*	0.0577	YES	0.0989	YES	0.0439	YES
21	*C* + *y*	0.0577	YES	0.0989	YES	0.1046	YES
22	*D* + *y*	1.0423	No	1.0989	No	1.0404	No
23	*E* + *y*	1.0423	No	0.1066	YES	0.0594	YES
24	*F* + *y*	0.0423	YES	0.0989	YES	0.0594	YES
		$\overline{Defect(quasiSMILES)} = 0.4889$	$N_{DA} = 12$[a] $r^2_{ALL} = 0.75$ $r^2_{DA} = 0.80$	$\overline{Defect(quasiSMILES)} = 0.2521$	$N_{DA} = 20$ $r^2_{ALL} = 0.74$ $r^2_{DA} = 0.75$	$\overline{Defect(quasiSMILES)} = 0.3195$	$N_{DA} = 16$ $r^2_{ALL} = 0.80$ $r^2_{DA} = 0.82$

r^2_{ALL} = correlation coefficient between pTA100(expr) and pTA100(calc) for all quasi-SMILES; r^2_{DA} = correlation coefficient between pTA100(expr) and pTA100(calc) for the domain of applicability.
[a] N_{DA} = the number of quasi-SMILES which fall into domain of applicability.

Table 5
Correlation weights, frequencies, and defects of attributes (A_k) for three random splits.

Correlation weights CW(A_k)				Frequencies of A_k in training (TRN) and in test (TST) sets						Defect(A_k)		
A_k	1	2	3	1		2		3		1	2	3
				TRN	TST	TRN	TST	TRN	TST			
+	0.24913	0.37795	1.02140	5	3	4	4	5	3	0.0269	0.0615	0.0304
–	2.64903	1.80147	1.07314	8	2	9	1	9	2	0.0215	0.0492	0.0221
A	−0.80185	0.0	0.99914	2	0	2	1	2	0	1.0000	0.0000	1.0000
B	−0.52772	0.0	0.98511	2	1	2	1	3	1	0.0154	0.0000	0.0036
C	−0.79547	0.0	1.01724	2	1	1	1	2	2	0.0154	0.0000	0.0643
D	0.39704	0.17037	0.99845	3	0	3	0	3	0	1.0000	1.0000	1.0000
E	0.87459	0.55239	1.03364	3	0	3	1	2	1	1.0000	0.0077	0.0190
F	0.0	0.0	1.04531	1	3	2	1	2	1	0.0000	0.0000	0.0190
n	2.60048	1.90277	1.02293	6	3	7	4	7	2	0.0154	0.0238	0.0111
y	0.52998	0.62728	0.97069	7	2	6	1	7	3	0.0154	0.0374	0.0100

$$\frac{r^2 - r_0^2}{r^2} = 0.2686 < 0.1$$

$$k = 0.9852 (0.85 < k < 1.15)$$

$$k' = 1.0148 (0.85 < k' < 1.15)$$

in Eqs. (6)–(8): the n is the number of quasi-SMILES in set (i.e. in training, test, validation); the r^2 is determination coefficient; the ${}^cR_p{}^2$ is parameter of Y-randomization (Ojha and Roy, 2011): a model has predictive potential if ${}^cR_p{}^2$ is larger than 0.5; the q^2 is cross-validated r^2; the s is root-mean-square error; the F is Fischer F-ratio. The estimation of statistical quality for the validation sets has been done with special criteria (Golbraikh and Tropsha, 2002; Melagraki and Afantitis, 2013). In the case of split 2 (Eq. (7)), the above-mentioned criteria partially are not satisfied, but in all other cases the statistical quality of models for the validation sets is good.

Fig. 1 contains the graphical representation of these models.

Unfortunately, the distribution of available data into the above-mentioned visible and invisible sets have considerable influence on the predictability of these models as well as on the domain of applicability, defined according to distribution of the attribute over the training and test sets (Table 4). However, the criterion to select the domain applicability is confirmed by calculations for the three random splits into the training set, test set, and validation set (Table 4).

Having data on several attempts to build up model for pTA100, one can extract A_k which have stable positive correlation weight value or vice versa stable negative correlation weight value

(Toropova and Toropov, 2013). The stable positive value is an indicator of the promoter of endpoint increase vice versa stable negative value is an indicator of the promoter of endpoint decrease. Thus, the models based on the optimal descriptors can has a mechanistic interpretation. However, this interpretation becomes statistically significant if the number of analyzed substances is large enough. Unfortunately, large standardized databases on nanomaterials still are not available. In the case of models calculated with Eqs. (6)–(8), only stable promoters of pTA100 increase can be detected. These are "–" (absence of S9 mix) and "*n*" (process without preincubation). The same approach for the larger number of the quasi-SMILES, can be considerably more informative and can provide not trivial mechanistic interpretation for the future models (quasi-QSARs). Table 5 shows that frequencies of attributes in the training and test sets are dramatically changing owing to the small number of available quasi-SMILES (apparently, this influences also upon external validation set). However, one can expect that situation will be more stable if the same approach will be used for larger dataset (decrease of changes of the frequencies of attributes in the training set and in the test set).

Thus, the approach gives models in accordance with the OECD principles (OECD, 2007). In the future, having large databases on various endpoints related to nanomaterials the optimal descriptors based on the eclectic data can be an attractive alternative of classic QSAR, but at present reliable statistical checking up of these descriptors is limited by availability of only small experimental datasets. It is to be noted, the CORAL software has been recognized as a possible tool for the mathematical modeling of endpoints related to nanomaterials (Panneerselvam and Choi, 2014).

4. Conclusions

Optimal descriptors calculated with eclectic data gives statistically significant model for mutagenic potentials of MWCNTs under various conditions. These models (quasi-QSARs) are built up in accordance with the OECD principles for validation of a QSAR. However, it is to be noted that the distribution into visible training and test set and invisible external validation set has influence upon the predictive potential of these models.

Acknowledgments

We thank the EU project PROSIL funded under the LIFE program (Project LIFE12 ENV/IT/000154), the EC Project NANOPUZZLES (Project Reference: 309837) and EC project PreNanoTox (Contract 309666). We also express our gratitude to Dr. L. Cappellini, Dr. G. Bianchi and Dr. R. Bagnati for valuable consultations on the computer science.

References

Achary, P.G.R., 2014a. QSPR modelling of dielectric constants of π-conjugated organic compounds by means of the CORAL software. SAR QSAR Environ. Res. 25, 507–526.

Achary, P.G.R., 2014b. Simplified molecular input line entry system-based optimal descriptors: QSAR modelling for voltage-gated potassium channel subunit Kv7.2. SAR QSAR Environ. Res. 25, 73–90.

Afantitis, A., Melagraki, G., Koutentis, P.A., Sarimveis, H., Kollias, G., 2011. Ligand – based virtual screening procedure for the prediction and the identification of novel b-amyloid aggregation inhibitors using Kohonen maps and Counterpropagation Artificial Neural Networks. Eur. J. Med. Chem. 46, 497–508.

Alqadami, A.A., Abdalla, M.A., Alothman, Z.A., Omer, K., 2013. Application of solid phase extraction on multiwalled carbon nanotubes of some heavy metal ions to analysis of skin whitening cosmetics using ICP-AES. Int. J. Environ. Res. Public Health 10 (1), 361–374.

Comelli, N.C., Ortiz, E.V., Kolacz, M., Toropova, A.P., Toropov, A.A., Duchowicz, P.R., Castro, E.A., 2014. Conformation-independent QSAR on c-Src tyrosine kinase inhibitors. Chemometr. Intell. Lab. Syst. 134, 47–52.

Deng, F., Xie, M., Zhang, X., Li, P., Tian, Y., Zhai, H., Li, Y., 2014a. Combined molecular docking, molecular dynamics simulation and quantitative structure–activity relationship study of pyrimido[1,2-c][1,3]benzothiazin-6-imine derivatives as potent anti-HIV drugs. J. Mol. Struct. 1067, 1–13.

Deng, F.-F., Xie, M.-H., Li, P.-Z., Tian, Y.-L., Zhang, X.-Y., Zhai, H.-L., 2014b. Study on the antagonists for the orphan G protein-coupled receptor GPR55 by quantitative structure–activity relationship. Chemometr. Intell. Lab. Syst. 131, 51–60.

Fourches, D., Pu, D., Tassa, C., Weissleder, R., Shaw, S.Y., Mumper, R.J., Tropsha, 2010. A quantitative nanostructure–activity relationship modelling. ACS Nano 4, 5703–5712.

Furtula, B., Gutman, I., 2011. Relation between second and third geometric – arithmetic indices of trees. J. Chem. 25, 87–91.

Golbraikh, A., Tropsha, A., 2002. Beware of q2! J. Mol. Graph. Model 20, 269–276.

González-Díaz, H., Arrasate, S., Gómez-San, A.J., Sotomayor, N., Lete, E., Besada-Porto, L., Ruso, J.M., 2013. General theory for multiple input–output perturbations in complex molecular systems. 1. Linear QSPR electronegativity models in physical, organic, and medicinal chemistry. Curr. Top. Med. Chem. 13, 1713–1741.

Gutman, I., Toropov, A.A., Toropova, A.P., 2005. The graph of atomic orbitals and its basic properties. 1. Wiener index. MATCH Commun. Math. Comput. Chem. 53, 215–224.

Ibrahim, M., Saleh, N.A., Hameed, A.J., Elshemey, W.M., Elsayed, A.A., 2010. Structural and electronic properties of new fullerene derivatives and their possible application as HIV-1 protease inhibitors. Spectrochim. Acta – Part A: Mol. Biomol. Spectrosc. 75, 702–709.

Kar, S., Gajewicz, A., Puzyn, T., Roy, K., 2014. Nano-quantitative structure–activity relationship modeling using easily computable and interpretable descriptors for uptake of magnetofluorescent engineered nanoparticles in pancreatic cancer cells. Toxicol. Vitro 28, 600–606.

Melagraki, G., Afantitis, A., 2013. Enalos KNIME nodes: exploring corrosion inhibition of steel in acidic medium. Chemometr. Intell. Lab. Syst. 123, 9–14.

Nesmerak, K., Toropov, A.A., Toropova, A.P., Kohoutova, P., Waisser, K., 2013. SMILES-based quantitative structure–property relationships for half-wave potential of N-benzylsalicylthioamides. Eur. J. Med. Chem. 67, 111–114.

Nesmerak, K., Toropov, A.A., Toropova, A.P., 2014. SMILES-based quantitative structure–retention relationships for RP HPLC of 1-phenyl-5-benzylsulfanyltetrazoles. Struct. Chem. 25, 311–317.

OECD environment health and assessment No. 69, 2007 OECD Environment Health and Safety Publications Series on Testing and Assessment No. 69, 2007. Guidance Document on the Validation of (Quantitative) Structure–Activity Relationship [(Q)SAR] Models. <http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono(2007)2&doclanguage=en> (accessed 22.07.14).

Ojha, P.K., Roy, K., 2011. Comparative QSARs for antimalarial endochins: importance of descriptor-thinning and noise reduction prior to feature selection. Chemometr. Intell. Lab. 109, 146–161.

Panneerselvam, S., Choi, S., 2014. Nanoinformatics: emerging databases and available tools. Int. J. Mol. Sci. 15, 7158–7182.

Pathakoti, K., Huang, M.-J., Watts, J.D., He, X., Hwang, H.-M., 2014. Using experimental data of *Escherichia coli* to develop a QSAR model for predicting the photo-induced cytotoxicity of metal oxide nanoparticles. J. Photochem. Photobiol. B: Biol. 130, 234–240.

Rallo, R., Espinosa, G., Giralt, F., 2005. Using an ensemble of neural based QSARs for the prediction of toxicological properties of chemical contaminants. Process Safe. Environ. Protect. 83, 387–392.

Sayes, C., Ivanov, I., 2010. Comparative study of predictive computational models for nanoparticle-induced cytotoxicity. Risk Anal. 30, 1723–1734.

Shahlaei, M., Nowroozi, A., Khodarahmi, R., 2014. A combined DFT and QSAR calculations to study substituted biphenyl imidazoles as bombesin receptor subtype-3 agonists. Lett. Drug Des. Discovery 11, 665–676.

Singh, K.P., Gupta, S., 2014. Nano-QSAR modeling for predicting biological activity of diverse nanomaterials. RSC Adv. 4, 13215–13230.

Toropov, A.A., Toropova, A.P., 2014. Optimal descriptor as a translator of eclectic data into pTA100 prediction: mutagenicity of fullerene as a mathematical function of conditions. Chemosphere 104, 262–264.

Toropov, A.A., Toropova, A.P., Benfenati, E., Leszczynska, D., Leszczynski, J., 2010. SMILES-based optimal descriptors: QSAR analysis of fullerene-based HIV-1 PR inhibitors by means of balance of correlations. J. Comput. Chem. 31, 381–392.

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., 2013. Optimal descriptor as a translator of eclectic information into the prediction of membrane damage by means of various TiO_2 nanoparticles. Chemosphere 93, 2650–2655.

Toropova, A.P., Toropov, A.A., Benfenati, E., Leszczynska, D., Leszczynski, J., 2010. QSAR modeling of measured binding affinity for fullerene-based HIV-1 PR inhibitors by CORAL. J. Math. Chem. 48, 959–987.

Torrens, F., Castellano, G., 2012. QSPR prediction of retention times of phenylurea herbicides by biological plastic evolution. Curr. Drug Safe. 7, 262–268.

Torrens, F., Castellano, G., 2014. QSPR prediction of chromatographic retention times of pesticides: partition and fractal indices. J. Environ. Sci. Health – Part B 49, 400–407.

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

Veselinović, A.M., Milosavljević, J.B., Toropov, A.A., Nikolić, G.M., 2013b. SMILES-based QSAR models for the calcium channel-antagonistic effect of 1,4-dihydropyridines. Arch. Pharm. 346, 134–139.
Vishnepolsky, B., Pirtskhalava, M., 2014. Prediction of linear cationic antimicrobial peptides based on characteristics responsible for their interaction with the membranes. J. Chem. Inf. Model. 54 (5), 1512–1523.
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.
Winkler, D.A., Mombelli, E., Pietroiusti, A., Tran, L., Worth, A., Fadeel, B., McCall, M.J., 2013. Applying quantitative structure–activity relationship approaches to nanotoxicology: current status and future potential. Toxicology 313, 15–23.
Wirnitzer, U., Herbold, B., Voetz, M., Ragot, J., 2009. Studies on the in vitro genotoxicity of baytubes®, agglomerates of engineered multi-walled carbon-nanotubes (MWCNT). Toxicol. Lett. 186, 160–165.