Ecotoxicology and Environmental Safety 124 (2016) 32–36

Contents lists available at ScienceDirect

Ecotoxicology and Environmental Safety

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

Nano-QSAR: Model of mutagenicity of fullerene as a mathematical function of different conditions

Alla P. Toropova [a], Andrey A. Toropov [a,*], Aleksandar M. Veselinović [b], Jovana B. Veselinović [b], Emilio Benfenati [a], Danuta Leszczynska [c], Jerzy Leszczynski [d]

[a] IRCCS-I stituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, Milano, Italy
[b] University of Niš, Faculty of Medicine, Department of Chemistry, Niš, Serbia
[c] Interdisciplinary Nanotoxicity Center, Department of Civil and Environmental Engineering, Jackson State University, 1325 Lynch St, Jackson, MS 39217-0510, USA
[d] Interdisciplinary Nanotoxicity Center, Department of Chemistry and Biochemistry, Jackson State University, 1400 J. R. Lynch Street, P.O. Box 17910, Jackson, MS 39217, USA

ARTICLE INFO

Article history:
Received 6 July 2015
Received in revised form
18 September 2015
Accepted 25 September 2015

Keywords:
Fullerene nanoparticle
Mutagenicity
Quasi-SMILES
Nano-QSAR

ABSTRACT

The experimental data on the bacterial reverse mutation test (under various conditions) on C60 nanoparticles for the cases (i) TA100, and (ii) WP2uvrA/pkM101 are examined as endpoints. By means of the optimal descriptors calculated with the Monte Carlo method a mathematical model of these endpoints has been built up. The models are a mathematical function of eclectic data such as (i) dose (g/plate); (ii) metabolic activation (i.e. with mix S9 or without mix S9); and (iii) illumination (i.e. darkness or irradiation). The eclectic data on different conditions were represented by so-called quasi-SMILES. In contrast to the traditional SMILES which are representation of molecular structure, the quasi-SMILES are representation of conditions by sequence of symbols. The calculations were carried out with the CORAL software, available on the Internet at http://www.insilico.eu/coral. The main idea of the suggested descriptors is the accumulation of all available eclectic information in the role of logical and digital basis for building up a model. The computational experiments have shown that the described approach can be a tool to build up models of mutagenicity of fullerene under different conditions.

1. Introduction

During last decades, a list of traditional substances has expanded and a new group of species–nanomaterials has been introduced both in research laboratories and in the everyday life. Nanomaterials already play significant role in many aspects of modern life, such as chemical and pharmacological industries, medicine, ecology, toxicology. For the case of the majority of organic compounds the estimation of their dangerous and attractive properties (endpoints) can be done by means of well-known quantitative structure–property/activity relationships (QSPRs/QSARs) (Salahinejad and Ghasemi, 2014; Wang et al., 2015; Peric et al., 2015). Thus, endpoints related to nanomaterials are waiting for predictive QSPR/QSAR models (Fourches et al., 2010; Leszczynski, 2010).

However, the simply application of the classic approaches based on the representation of various substances by the structure of their molecules is impossible for majority of nanomaterials. Moreover, the classic QSPR/QSAR models are based on manipulation with considerable large databases, where the number of available substances can be hundreds or even thousands. In the case of nanomaterials as rule their synthesis and identification are very difficult actions and as result, the number of nanomaterials which are examined in some standardized conditions is very small (units or tens, in the best case).

Therefore, in the case of nanomaterials one cannot expect large databases to be available in the near future. Consequently, predictive models for endpoints related to nanomaterials should focus and be developed using small data sets. Moreover, molecular structure of majority of nanomaterials has no convenient representations (for the QSPR/QSAR analyses) such as molecular graph (Toropov and Toropova, 2002, 2003; Toropov and Roy, 2004) or simplified molecular input-line entry system (SMILES) (García et al., 2010; Garro Martinez et al., 2011; Ibezim et al., 2011).

Thus, for the case of nanomaterials, the traditional classic

* Correspondence to: Laboratory of Environmental Chemistry and Toxicology, IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20159 Milano, Italy. Fax: +39 02 3901 4735.
E-mail address: andrey.toropov@marionegri.it (A.A. Toropov).

http://dx.doi.org/10.1016/j.ecoenv.2015.09.038
0147-6513/

QSPR/QSAR theory and praxis should be cardinally modified. International organizations such as, OECD (2006–2010) (Organisation for Economic Co-operation and Development) and REACH (2011) (Registration, Evaluation, Authorisation and Restriction of Chemicals) pay attention on necessity of the further evolution of knowledge related to nanomaterials. Multiple toxicity of nanomaterials have been carried out in the last decade. However, most of these works used non-standardized testing protocols leading to unreproducible and poorly comparable results, which therefore are insufficient for estimation of hazard and risk assessment (Burello and Worth, 2011; Richarz et al., 2015). In addition, the unavailability of consistent physico-chemical characterization data in the same studies makes it difficult to identify which material characteristics determine the documented toxic effects. In other words, the need for standardization and new approaches in the nanosafety area has been recognized (Salahinejad, 2015).

The possible way, to estimate endpoints related to nanomaterials with using so-called optimal descriptors has been suggested (Toropova and Toropov, 2013; Toropova et al., 2013, 2015). The advantage of the approach is the possibility of standardization of eclectic data. In the case of traditional QSPR/QSAR analyses, the paradigm "*Endpoint*=*F*(*structure*)" is the basis to build up a predictive model. In this case, SMILES is the representation of the molecular structure. In the case of the predictive modeling for nanomaterials, an alternative paradigm can be defined as "*Endpoint*=*F*(*eclectic data*)". The eclectic data can be represented by sequence of symbols similar to SMILES, but where symbols are a representation of a list of eclectic conditions. These sequences of symbols should be denoted as "quasi-SMILES" in order to avoid their false interpretation as the "traditional SMILES". Since these models are based on list of eclectic conditions (not on structure), instead of QSPR/QSAR one should use some other terminology for these researches, e.g. "quantitative *conditions*-property/activity relationships" (QCPR/QCARs).

The aim of the present work is to estimate QCAR based on data (conditions) related to (i) dose (g/plate); (ii) metabolic activation (with or without mix S9); and (iii) illumination (dark or irradiation) as a tool to predict two endpoints. The endpoint-1 is the bacterial reverse mutation test that was conducted using *Salmonella typhimurium* strains TA100; and the endpoint-2 is mutagenic effect of fullerene for *Escherichia coli* strain WP2 uvrA/pKM101 (Shinohara et al., 2009).

2. Method

2.1. Data

Data on (i) the bacterial reverse mutation test that was conducted using *S. typhimurium* strains TA100 in the presence and absence of metabolic activation under dark conditions and irradiation are taken in the literature, and (ii) the numerical data on the bacterial reverse mutation test that was conducted using *E. coli* strain WP2 uvrA/pKM101 in the presence and absence of metabolic activation under dark condition and irradiation are taken in the literature (Shinohara et al., 2009). Twenty quasi-SMILES have been defined for the above data. These twenty quasi-SMILES were randomly split into the training, calibration, and validation sets. The training set is basis to build up a model. The calibration set is used to avoid the overtraining (situation where the excellent statistical quality of a model for the training set is accompanied by poor statistical quality of the model for external quasi-SMILES). The validation set is utilized for the final estimation of the predictive potential of the model. For each endpoint three random splits were examined.

Table 1
The list of conditions which have impact upon mutagenicty of fullerene C60 nanoparticles and which were utilized to build up models.

Conditions	Symbols for quasi-SMILES
Dark or Irradiation	"0"=Darkness "1"=Irradiation
Mix S9	"+"=With Mix S9 "−"= Without Mix S9
Dose (g/plate)	"A"=50 "B"=100 "C"=200 "D"=400 "E"=1000

Table 2
The numerical data on endpoints and dispersion of the experimental definition and the scheme of construction of quasi-SMILES.

ID	Darkness or irradiation	Mix S9	Dose	TA100	WP2uvrA/pKM101
1	0	+	A	146 ± 4	113 ± 7
2	0	+	B	141 ± 24	106 ± 19
3	0	+	C	159 ± 8	112 ± 15
4	0	+	D	160 ± 21	115 ± 15
5	0	+	E	177 ± 8	145 ± 13
6	0	−	A	143 ± 14	160 ± 15
7	0	−	B	139 ± 22	162 ± 21
8	0	−	C	169 ± 20	174 ± 26
9	0	−	D	168 ± 7	179 ± 15
10	0	−	E	152 ± 15	220 ± 6
11	1	+	A	129 ± 6	114 ± 10
12	1	+	B	131 ± 4	105 ± 12
13	1	+	C	138 ± 12	113 ± 16
14	1	+	D	137 ± 14	110 ± 18
15	1	+	E	160 ± 6	123 ± 8
16	1	−	A	136 ± 11	127 ± 25
17	1	−	B	136 ± 6	133 ± 12
18	1	−	C	138 ± 7	121 ± 15
19	1	−	D	164 ± 16	117 ± 29
20	1	−	E	172 ± 14	138 ± 19

2.2. Optimal descriptor

The optimal descriptors used in this study are calculated as the following

$$DCW(T, N_{epoch}) = \Sigma CW(S_k) \quad (1)$$

where S_k is a symbol for representation of conditions listed in Table 1.

The $CW(S_k)$ are correlation weights of a symbol from quasi-SMILES. The correlation weights of symbols are calculated with optimization by the Monte Carlo technique. The correlation weights should provide maximal value of the correlation coefficient between the $DCW(T,N_{epoch})$ and experimental data on the TA100 or WP2 uvrA/pKM101. The T and the N_{epoch} are parameters of the optimization: the T (threshold) is coefficient for classification of impacts into two categories: rare and not rare. Correlation weight for rare impact is fixed equal to zero. Therefore rare attributes are not involved in the model. The N_{epoch} is the number of epochs of the Monte Carlo optimization.

Having data on optimal correlation weights, one can:

(1) calculate $DCW(T, N_{epoch})$ for all quasi-SMILES;
(2) calculate (with data on the training set) a model for TA100 or for WP2 uvrA/pKM101:

$$TA100 = C_0 + C_1 \times DCW(T, N_{epoch}) \quad (2)$$