Ecotoxicology and Environmental Safety 112 (2015) 39–45

Contents lists available at ScienceDirect

Ecotoxicology and Environmental Safety

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

Optimal descriptor as a translator of eclectic data into prediction of cytotoxicity for metal oxide nanoparticles under different conditions

Alla P. Toropova [a], Andrey A. Toropov [a,*], Robert Rallo [b], Danuta Leszczynska [c], Jerzy Leszczynski [d]

[a] IRCCS, Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy
[b] Departament d'Enginyeria Informatica i Matematiques, Universitat Rovira i Virgili, Av. Països Catalans, 26, 43007 Tarragona, Catalunya, Spain
[c] Interdisciplinary Nanotoxicity Center, Department of Civil and Environmental Engineering, Jackson State University, 1325 Lynch Street, Jackson, MS 39217-0510, USA
[d] Interdisciplinary Nanotoxicity Center, Department of Chemistry and Biochemistry, Jackson State University, 1400 JR Lynch Street, P.O. Box 17910, Jackson, MS 39217, USA

ARTICLE INFO

Article history:
Received 14 July 2014
Received in revised form
1 October 2014
Accepted 3 October 2014
Available online 5 November 2014

Keywords:
QSAR
Quasi-SMILES
Quasi-QSAR, Nano-QSAR
Monte Carlo method
Cytotoxicity
Metal oxide nanoparticle

ABSTRACT

The Monte Carlo technique has been used to build up quantitative structure–activity relationships (QSARs) for prediction of dark cytotoxicity and photo-induced cytotoxicity of metal oxide nanoparticles to bacteria *Escherichia coli* (minus logarithm of lethal concentration for 50% bacteria pLC50, LC50 in mol/L). The representation of nanoparticles include (i) in the case of the dark cytotoxicity a simplified molecular input-line entry system (SMILES), and (ii) in the case of photo-induced cytotoxicity a SMILES plus symbol '^'. The predictability of the approach is checked up with six random distributions of available data into the visible training and calibration sets, and invisible validation set. The statistical characteristics of these models are correlation coefficient 0.90–0.94 (training set) and 0.73–0.98 (validation set).

1. Introduction

Nanomaterials become important components of modern everyday life. This requires studies that would reveal their characteristics and provide guidelines to facilitate their safe applications. Predictive models for nanomaterials can be useful for theoretical and practical reasons (Randic, 1991; Cosentino et al., 2000; Balaban et al., 2005; Ivanciuc et al., 2006; Tetko et al., 2008; Bhhatarai et al., 2010; Das and Trinajstic, 2010; Mitra et al., 2010; Duchowicz et al., 2011; Furtula and Gutman, 2011; Afantitis et al., 2011; Toropov et al. 2012b,c; Liu et al., 2013; Cohen et al., 2013; Toropova and Toropov, 2014) to the same extend as models for "classic" substances (organic, inorganic, organometallic) have been used.

Many of suggested approaches which are aimed to build up quantitative structure–property/activity relationships (QSPRs/QSARs) for nanomaterials were obtained with "classic" descriptors (Fourches et al., 2010; Petrova et al., 2011), tested for "classic" substances. However, (owing to the uncertainty of molecular architecture that is related to nanomaterials), the development of fresh "nanodescriptors" (Leszczynski 2010; Toropova and Toropov, 2013) becomes a necessary task of modern computational approaches focusing on the problem. An attractive and innovative alternative to "classic" descriptors are optimal descriptors calculated using available eclectic data (Toropova et al., 2013; Toropova and Toropov, 2013).

Optimal descriptors (Toropova et al., 2010, 2011, 2012a, Toropov et al., 2010a,b, 2013a,b), could be considered as a transitional step between "classic" and "nanodescriptors". On the one hand, these descriptors can be calculated with data on the molecular structure (i.e. just as "classic" descriptors); but on the other hand, these descriptors can be computed using eclectic information about a substance, even without detailed data on its molecular structure (Toropov et al., 2007; Toropova and Toropov, 2013).

However, data on various nanoparticles can be represented by special strings which are encoded data on physicochemical and biochemical conditions of impact of the nanoparticles. These SMILES-like strings can be named "quasi SMILES", since they represent conditions in contrast of traditional SMILES which represent solely the molecular structures.

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

http://dx.doi.org/10.1016/j.ecoenv.2014.10.003
0147-6513/

The paradigm for traditional QSPR/QSAR analyses could be expressed as:

Endpoint = F(Molecular Structure)

In the case of the nanomaterials the paradigm can be modified as follows:

Endpoint = F(Available Eclectic Data)

The available eclectic data can be (i) the molecular structure of substances which are involved in phenomenon under consideration; (ii) presence/absence of photo-inducing; and (iii) any other circumstances which are able to have influence on the phenomenon under consideration (Toropova and Toropov, 2013; Toropov and Toropova, 2014).

Consequently, one can define the following hybrid paradigm:

Endpoint = F(Molecular Structure and ***Available Eclectic Data)***

Since the above mentioned quasi SMILES are basis for establishing of correlation between impacts (these are not only data on the molecular structure, but any available eclectic data with influence upon nanoparticles) which are defining the behavior of metal oxide nanoparticles, these correlations can be named as "quasi-QSARs" or "nano-QSARs". In the present work, the only eclectic factor is the presence or vice versa absence of photo-inducing, however the number of eclectic components for the quasi-QSAR or nano-QSAR can be larger (Toropova and Toropov, 2013; Toropov and Toropova, 2014).

The aim of the present study is an attempt to build up united QSAR model for dark cytotoxicity and photo-induced cytotoxicity of metal oxide nanoparticles to bacteria *Escherichia coli*, using optimal descriptors which are a mathematical function of atomic composition and the conditions (i.e. the dark or the photo-inducing).

2. Method

2.1. Data

The numerical data on cytotoxicity of metal oxide nanoparticles to bacteria *E. coli* (the concentration of the nanoparticles that proved to be fatal to 50% of the bacteria *E. coli* LC50, in mol/L) have been taken from the literature (Pathakoti et al., 2014). The negative decimal logarithm of the LC50 (pLC50) has been examined as the endpoint. Six random distributions of the available data into training and calibration sets (these metal oxide nanoparticles are used to build up the model) and validation set (these metal oxide nanoparticles are not involved to build up the model, they are used to check up predictability of the model) are examined. All these splits are prepared according to the following principles: (i) they are random; (ii) the range of endpoints in each sub-set is similar to ranges for other sub-sets; and (iii) these splits are not identical (Table 1). The dark cytotoxicity and photo-induced cytotoxicity are examined as an united endpoint, owing to application of the model which is a mathematical function of atomic composition and conditions (presence or absence of photo-inducing).

2.2. Optimal descriptors

In order to take into account the photo-induction, the symbol '^' is used. Thus, SMILES used in this work are not equivalent traditionally used ones (Weininger, 1988, 1990; Weininger et al., 1989). Under such circumstances, the term 'quasi-SMILES' is used to define the name for the used representation of metal oxide nanoparticles, because the quasi-SMILES is the representation of data on molecular structure together with condition: presence or absence of photo-inducing. The presence of photo-inducing indicated by symbol '^' that is added at the end of traditional SMILES (Table 2).

Table 1
Upper triangle of percentages of identity for random splits.

	Set	Split 1	Split 2	Split 3	Split 4	Split 5	Split 6
Split 1	Training	100.0[a]	72.3	72.7	65.1	57.8	69.8
	Calibration	100.0	16.7	0.0	33.3	0.0	16.7
	Validation	100.0	16.7	33.3	15.4	16.7	15.4
Split 2	Training		100.0	76.2	58.5	69.8	58.5
	Calibration		100.0	42.9	0.0	30.8	28.6
	Validation		100.0	16.7	30.8	33.3	30.8
Split 3	Training			100.0	53.7	65.1	68.3
	Calibration			100.0	0.0	30.8	28.6
	Validation			100.0	30.8	0.0	15.4
Split 4	Training				100.0	52.4	70.0
	Calibration				100.0	15.4	0.0
	Validation				100.0	30.8	42.9
Split 5	Training					100.0	61.9
	Calibration					100.0	15.4
	Validation					100.0	30.8
Split 6	Ttraining						100.0
	Calibration						100.0
	Validation						100.0

Where $N_{i,j}$ is the number of substances distributed into the same set for both the i-th split and the j-th splits (set = training, calibration, and validation); N_i is the number of substances distributed into the set for the i-th split; N_j is the number of substances distributed into the set for the j-th split.

[a] $Identity\,(\%) = N_{i,j}/0.5*(N_i + N_j) \times 100$

Thus the optimal descriptors have been calculated as follows:

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

where A_k is an attribute of the quasi-SMILES that comprises one symbol (e.g. 'O', 'V', etc.) or two symbols which should be examined as one (e.g. 'Cu', 'Al', etc.). In the case of dark cytotoxicity, nanoparticles are represented by SMILES of ACD/ChemSketch software (ACD/I-LAB, 2014), in the case of photo-induced cytotoxicity, nanoparticles are represented by the SMILES of ACD/ChemSketch software (ACD/I-LAB, 2014) plus symbol '^' (Table 2).

The $CW(x)$ is correlation weight for an attribute x, that is extracted from a quasi-SMILES; the T is the threshold to divide attributes into two categories rare (noise) or not rare; the N is the number of epochs of the Monte Carlo optimization. Correlation weights are calculated for not rare attributes by the Monte Carlo optimization that gives maximum of determination coefficient between $DCW(T,N)$ and pLC50 for the calibration set. The preferable values for the T^* and N^* which provides best statistics for the calibration set should be defined at the preliminary phase of the QSAR analysis (Toropova et al., 2011). Having T^*, N^*, and $CW(x)$ which give maximum of the determination coefficient for the calibration set, one can define (using data from the training set) the following model:

$$\mathrm{pLC50} = C_0 + C_1{*}DCW(T^*, N^*) \quad (2)$$

The predictability of the model should be checked up with external validation set.

Table 3 contains the numerical data on the correlation weights of different attributes involved in the modeling process. These are (i) various chemical elements represented traditionally by one (e.g. 'O', 'V') or by two symbols (e.g. 'La', 'Ni'). The symbol '=' represents double bonds. The symbol '^' represents the photo-inducing. The symbols '[' and ']' are used in the classic SMILES for encoding special group or metal (Weininger, 1988, 1990; Weininger et al. 1989). Thus, all attributes have transparent interpretation. The correlation