NanoImpact 1 (2016) 60–64

Contents lists available at ScienceDirect

NanoImpact

journal homepage: www.journals.elsevier.com/nanoimpact

Research paper

Quasi-SMILES as a tool to utilize eclectic data for predicting the behavior of nanomaterials

Alla P. Toropova *, Andrey A. Toropov, Serena Manganelli, Caterina Leone, Diego Baderna, Emilio Benfenati, Roberto Fanelli

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

ARTICLE INFO

Article history:
Received 22 February 2016
Received in revised form 29 March 2016
Accepted 18 April 2016
Available online 21 April 2016

Keywords:
Nano-QSAR
Nanoparticles
Cytotoxicity
HaCaT
Escherichia coli
Quasi-SMILES

ABSTRACT

Nowadays, nanomaterials are often considered a scientific hit. However, despite the immense advantages of nanomaterials, there are studies, which have shown that these materials can also harmfully impact both human health and the environment. A preliminary evaluation of the hazards related to nanomaterials can be performed using predictive models. The aim of the present study is building up a single QSAR model for predicting cytotoxicity of metal oxide nanoparticles on (i) *Escherichia coli* (*E. coli*) and (ii) human keratinocyte cell line (HaCaT) based on the representation of the available eclectic data, encoded into quasi-SMILES. Quasi-SMILES are an analog and an attractive alternative of traditional simplified molecular input-line entry systems (SMILES). In contrast to traditional SMILES quasi-SMILES are a tool to represent not only molecular structures, but also different conditions, such as physicochemical properties and experimental conditions. The statistical quality of the models is average correlation coefficient (r^2) and root mean squared error (RMSE) for the training set 0.79 and 0.216; the average r^2 and RMSE for validation set are 0.90 and 0.247, respectively.

1. Introduction

Human exposure to NPs has been in existence for many years. It involves public and occupational health exposure to ultrafine particulate air pollution. A broader source of exposure is related to nanoparticles which are abundant in nature, as they are produced in many natural processes, including photochemical reactions, volcanic eruptions, forest fires, and simple erosion, and by plants and animals (Buzea et al., 2007).

In more recent years, due to the rapid expansion of nanotechnology, environmental and human exposure to engineered nanoparticles has also become unavoidable (Ray et al., 2009).

For this reason, the need to gain knowledge about safety and potential hazards of nanoparticles is dramatically increasing. Within this context, nanotoxicology has become an emerging discipline. However, while the number of nanoparticle types and their applications continues to increase, studies to characterize their effects after exposure and to address their potential toxicity are few in comparison. In the medical field in particular, nanoparticles are being utilized in diagnostic and therapeutic tools to better understand, detect, and treat human diseases. Exposure to nanoparticles for medical purposes involves intentional contact and control; therefore, understanding the properties of nanoparticles and their effect on the body is crucial before clinical use can occur. The first step towards understanding how an agent will react in the body often involves cell-culture studies. Compared to animal studies, cellular testing is less ethically ambiguous, is easier to control and reproduce, and is less expensive (Lewinski et al., 2008).

Building up predictive models for endpoints related to nanomaterials is an important task of modern natural sciences (Singh and Gupta, 2014). Likely, the traditional quantitative structure — property/activity relationships (QSPRs/QSARs) (Melagraki and Afantitis, 2013; Scotti et al., 2014; Toropov et al., 2014; Kleandrova et al., 2015; Speck-Planche and Cordeiro, 2015; Duchowicz et al., 2015; Ibezim et al., 2012; Veselinović et al., 2015a; Veselinović et al., 2015b) based on the molecular structure is not able to solve this task.

The problem with nanomaterials is that a chemical structure is not sufficient to describe them so that a range of other unique properties needs to be considered, including particle size, shape and surface (Toropov et al., 2012).

A model for endpoints related to nanomaterials can be organized in the following form: the measured calculated endpoint is a mathematical function of all available eclectic information, which may be (i) chemical structure, (ii) atom compositions, (iii) conditions of synthesis/preparation of the nanomaterial, (iv) the features of nanomaterials related to their manufacture. This list can be easily extended (size, porosity, symmetry, electromechanical properties, etc.). To define a predictive model for an endpoint related to nanomaterials the traditional paradigm for QSAR modeling, 'Endpoint = F (molecular structure)', can be replaced

* Corresponding author at: Laboratory of Environmental Chemistry and Toxicology, IRCCS - Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy.
E-mail address: alla.toropova@marionegri.it (A.P. Toropova).

http://dx.doi.org/10.1016/j.impact.2016.04.003
2452-0748/

by 'Endpoint = F (eclectic information)' (Toropov and Toropova, 2015a; Manganelli et al., 2016; Toropov and Toropova, 2015b; Toropov et al., 2015; Toropova and Toropov, 2015a).

The aim of the present work is an attempt to build up united predictive model for two endpoints: (i) cytotoxicity to *Escherichia coli* and (ii) human keratinocyte cell line (HaCaT) for metal nanoparticles using optimal descriptors based on quasi-SMILES. Quasi-SMILES are a modification of the traditional simplified molecular input-line entry systems (SMILES) (Weininger, 1988; Weininger et al., 1989; Weininger, 1990) representing eclectic data using a string of characters, encoding particular conditions, not of the molecular structure. In fact, the aim of the present work can be also defined as an attempt to answer question: "How one should organize databases related to nanomaterials in order to extract from these databases satisfactory prediction of the behavior for nanomaterials, which were not examined in experiment?"

2. Method

2.1. Data

The endpoint considered for the QSAR analysis was cytotoxicity of metal oxide nanoparticle on *E. coli* (*E. coli*) (Puzyn et al., 2011) and human keratinocyte cell line (HaCaT) (Gajewicz et al., 2015), expressed as the negative logarithm of half maximal effective concentration (pEC_{50}). pEC_{50} data (mol/L) were taken from the literature (see Table 1). Fig. 1 shows the toxicity data for nano-sized metal oxides against *E. coli* and HaCaT cells: pEC_{50} values on HaCaT are higher in comparison to those obtained from *E. coli*. This trend of toxicity is reversed only for In_2O_3, SnO_2, and TiO_2, which are more toxic to HaCaT than to *E. coli* (Kar et al., 2016).

The total set of available data has been split (three times) into the training (n = 22), calibration (n = 5), and validation (n = 5) sets. These splits are built up according to principles: (i) these splits are random; (ii) the ranges of endpoints are similar for each sub-set (i.e. for the training, calibration, and validation set); and (iii) these splits are different. It is possible to notice that there is a good balance of cytotoxicity data between the two sets of values. Furthermore, the cytotoxicity ranges are also quite similar going from 1.76 to 3.32 in the case of line cell line and in the case of *E. coli* from 1.74 to 3.45. These values are given as pEC_{50} where EC50 is the cytotoxicity effect observed the dose which produces effect on 50% of the cells.

In fact these endpoints are a mathematical function of the same conditions (same structures of nano-oxides) and two additional codes (%11 and %12) give possibility to attempt to build up united model for these endpoints. The similar approach was used in work (Toropova and Toropov, 2015b) for united model of mutagenicity for fullerene and multi-walled carbon nanotubes (MWCNTs) under different conditions.

Table 1
Numerical data on the toxicity to *Escherichia coli* and human keratinocyte cell line (HaCaT).

No.	Nano-oxide	Traditional SMILES	Additional codes: HaCaT = %11 *E. coli* = %12	pEC_{50} in molar scale
1.	Al_2O_3	O[Al]O[Al]O	%11	1.85
2.	Bi_2O_3	O[Bi]O[Bi]O	%11	2.5
3.	CoO	[Co]O	%11	2.83
4.	Cr_2O_3	O[Cr]O[Cr]O	%11	2.3
5.	Fe_2O_3	O[Fe]O[Fe]O	%11	2.05
6.	In_2O_3	O[In]O[In]O	%11	2.92
7.	La_2O_3	O[La]O[La]O	%11	2.87
8.	NiO	[Ni]O	%11	2.49
9.	Sb_2O_3	O[Sb]O[Sb]O	%11	2.31
10.	SiO_2	O[Si]O	%11	2.12
11.	SnO_2	O[Sn]O	%11	2.67
12.	TiO_2	O[Ti]O	%11	1.76
13.	V_2O_3	O[V]O[V]O	%11	2.24
14.	Y_2O_3	O[Y]O[Y]O	%11	2.21
15.	ZnO	O[Zn]	%11	3.32
16.	ZrO_2	O[Zr]O	%11	2.02
17.	Al_2O_3	O[Al]O[Al]O	%12	2.49
18.	Bi_2O_3	O[Bi]O[Bi]O	%12	2.82
19.	CoO	[Co]O	%12	3.51
20.	Cr_2O_3	O[Cr]O[Cr]O	%12	2.51
21.	Fe_2O_3	O[Fe]O[Fe]O	%12	2.29
22.	In_2O_3	O[In]O[In]O	%12	2.81
23.	La_2O_3	O[La]O[La]O	%12	2.87
24.	NiO	[Ni]O	%12	3.45
25.	Sb_2O_3	O[Sb]O[Sb]O	%12	2.64
26.	SiO_2	O[Si]O	%12	2.2
27.	SnO_2	O[Sn]O	%12	2.01
28.	TiO_2	O[Ti]O	%12	1.74
29.	V_2O_3	O[V]O[V]O	%12	3.14
30.	Y_2O_3	O[Y]O[Y]O	%12	2.87
31.	ZnO	O[Zn]	%12	3.45
32.	ZrO_2	O[Zr]O	%12	2.15

2.2. Optimal descriptor

Optimal descriptors also called 'quasi-SMILES', of nanoQSAR analysis were calculated with CORAL software (http://www.insilico.eu/coral). These were built and optimized starting from the coding of an experimental condition (in vitro test): HaCaT and *E. coli* were encoded as "%11" and "%12" respectively. These codes were combined with the traditional SMILES of nano-oxides (see Table 1). The 32 resulting combined systems (traditional SMILES-in vitro test) were randomly split into training, calibration and validation sets, with similar distribution of endpoint values.

Optimal descriptors were calculated as follows:

$$DCW(T, N) = \sum CW(S_k) \quad (1)$$

where $CW(S_k)$ are the correlation weights for each fragment S_k contained in the quasi-SMILES (Table 2).

The correlation weights are calculated using the Monte Carlo optimization method (Veselinović et al., 2015a; Veselinović et al., 2015b; Toropov et al., 2012; Toropov and Toropova, 2015a; Manganelli et al., 2016; Toropov and Toropova, 2015b; Toropov et al., 2015; Toropova and Toropov, 2015a). The optimization process makes use of two parameters: (i) the threshold (T), which is a tool for classifying codes as either rare (and thus likely less reliable features, probably introducing noise into the model) or not rare features, which are used by the model and labeled as active; and (ii) the number of epochs (N), which is the number of cycles (sequence of modifications of correlation weights for all codes involved in model development) for the optimization (Toropov and Toropova, 2015a; Manganelli et al., 2016; Toropov and Toropova, 2015b; Toropov et al., 2015). The target function of the optimization procedure is the correlation coefficient between cytotoxicity and descriptors calculated with Eq. 1 for the training set. However, the process should be stopped when the correlation coefficient for the calibration set reach maximum. If the process will be continued after this maximum, the model most probably will give the overtraining (i.e. excellent statistical quality for the training set, but poor quality for the calibration and for the validation set).

Thus, the model should be optimized using condition the $T = T^*$ and $N = N^*$ which give the maximum of the correlation coefficient for the calibration set. These T^* and N^* should be defined from computational calculations with T from range $\{T_1, T_2, \ldots, T_n\}$ and N from range {1, 2, …, N}. Having the correlation weights obtained by described manner, one can calculate by using Eq. 1 the optimal descriptor for any system of eclectic conditions and by utilizing the systems of the training set build up a model:

$$pEC50 = C_0 + C_1 * DCW(T*, N*) \quad (2)$$

The model should be checked up with the calibration set and if the statistical quality is satisfactory, then the obtained model should have