Computers in Biology and Medicine 99 (2018) 161–172

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Multivariate statistical analysis for selecting optimal descriptors in the toxicity modeling of nanomaterials

Sunil Kr. Jha[a,*], T.H. Yoon[b], Zhaoqing Pan[a]

[a] School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China
[b] Nanoscale Characterization & Environmental Chemistry Lab, Department of Chemistry, College of Natural Sciences, Hanyang University, Seoul, 133-791, South Korea

ARTICLE INFO

Keywords:
Nanomaterials
Descriptors selection
Principal component analysis
Toxicity prediction

ABSTRACT

The present study is based on the application of a multivariate statistical analysis approach for the selection of optimal descriptors of nanomaterials with the objective of robust qualitative modeling of their toxicity. A novel data mining protocol has been developed for the selection of an optimal subset of descriptors of nanomaterials by using the well-known multivariate method principal component analysis (PCA). The selected subsets of descriptors were validated for qualitative modeling of the toxicity of nanomaterials in the PC space. The analysis and validation of the proposed schemes were based on five decisive nanomaterial toxicity data sets available in the published literature. Optimal descriptors were selected on the basis of the maximum loading criteria and using a threshold value of cumulative variance ≤ 90% on PC directions. A maximum inter-class separation(B) and the minimum intra-classes separation(A) were obtained for toxic vs. nontoxic nanomaterials in the PC space with the selected subsets of optimal descriptors compared to their other combinations for each of the datasets.

1. Introduction

Nanomaterials are the nanometer size substances [1] typically synthesized from metal oxides: TiO_2. ZnO, Fe_2O_3, CuO, SiO_2, WO_3, etc. [2–5]; metals: Ag, Au, Pt, Co, Ni, Cu, etc. [6–9]; carbon [10]; and other chemical compounds such as cadmium chloride, silver carbonate, sodium molybdate, and aluminum chloride [11]. The application of numerous commercial products based on nanomaterials is expanding in our daily practice, including in food, health, cosmetics, and households [12]. Superfluous applications of nanomaterials result in not only the constructive imprints for the advancement of human lives but also some harmful impacts on health and the environment [12,13]. Consequently, it is essential to assess the injurious impacts of nanomaterials on living beings as well the environment to acquire their maximum benefits with the minimum hazard. In this regard, a computational model to measure the effect of nanoparticle deposition in the nasal cavity, nasopharynx, oropharynx, larynx, and trachea was developed by Ghalati et al. [39]. Toropov et al. [40] have presented a brief review of databases and software that can be used in toxicity prediction of drugs and nanomaterials.

In most of the studies, the impacts of nanomaterials were measured by assessing their biological reactions, such as cell viability, cellular uptake, mitochondrial activity, cellular ROS, oxidative stress, DNA damage, and LC_{50}, in two ways: directly through *in vivo* (on human, rat, fish, insects, pig, and other animals) experiments or indirectly via *in vitro* (cultured cells of human and animals) experiments [2,6,8,11,14,15]. Novel QSTR-Perturbation models [33–36], structure–activity relationships [37], theoretical descriptors [41], and the Quasi-SMILES-based nano-quantitative structure–activity model [42] have been developed and discussed in the prediction of ecotoxicity and cytotoxicity of nanoparticles. Due to the availability of multiple cellular responses (endpoints) of nanomaterials, it is a challenging task to identify the significant and collective toxicity measures that necessitate further research in this direction. Moreover, thorough information on mutual relationships among the multiple endpoints needs to be addressed. At the same time, both the *in vivo* and *in vitro* experiments are slow and expensive and contain ethical concerns that note the requirement of robust toxicity prediction models by using the available databases of nanomaterials in published literature and using advanced analysis procedures.

1.1. Literature survey and motivation of the present study

In the latter trend, some reports can be found in the literature relating to the physicochemical descriptors of nanomaterials with their cellular toxicity measures using the multivariate statistical analysis methods (summarized in Ref. [16]). The related studies can be broadly characterized into three main categories: (i) qualitative classification of

* Corresponding author.
E-mail address: 002891@nuist.edu.cn (S.K. Jha).

https://doi.org/10.1016/j.compbiomed.2018.06.012
Received 25 April 2018; Received in revised form 13 June 2018; Accepted 13 June 2018

nanomaterials (toxic vs. nontoxic) using novel features, such as principal components (PCs), e.g., nanomaterials grouping using their descriptors by Sayes and Ivanov [17] and Lynch et al. [18], and clustering methods, e.g., nanomaterial biological activity interpretation using hierarchical clustering by Shaw et al. [19]; (ii) quantitative classification of toxic vs. nontoxic nanomaterials using supervised pattern recognition methods, such as linear discriminant analysis (LDA) [17], artificial neural network (ANN) modeling of reorganizational energy of nanotubes by Taherpour et al. [20], support vector machine (SVM) [21], naïve Bayes, k nearest neighbor [22] in nanostructure activity relationship modeling; and (iii) quantitative prediction of toxicity measures using regression analysis methods (discussed as quantitative structure–activity relationships (QSARs) models [21–23]. In some recent studies, data imputation and PChem score-based screening approaches [43] and generalized linear model, SVM, random forest, and ANN methods [44] have been implemented in toxicity prediction and the classification of nanomaterials.

The recognition of an optimal set of significant physicochemical descriptors of nanomaterials is essential for the improved efficiency of each of the earlier mentioned models of three categories in toxicity prediction. The physicochemical descriptors of nanomaterials can be arranged into three groups as intrinsic (fundamental) descriptors, such as the number of atoms, position in the periodic table (both period and group), mass of atoms, mass of the compound, band gap, density, and electronegativity, between intrinsic and extrinsic (measurable) descriptors, such as surface charge, size, hydrodynamic size, and metal dissolution, and extrinsic descriptors, such as concentration. The success of toxicity prediction modeling of nanomaterials also requires an understanding of (i) chemical associations among the measurable physicochemical descriptors, and (ii) the biological mechanisms of cellular responses and their relationship with the intrinsic as well as extrinsic physicochemical descriptors [24,25]. Moreover, similar weights of intrinsic and extrinsic descriptors were considered in the majority of studies for the toxicity modeling, which causes loss of toxicity information of nanomaterials. Though the weight of each of the physicochemical descriptors must be identified, thereafter, significant descriptors should be selected in the toxicity modeling of nanomaterials. For that reason, there is an urgent need for a collective procedure for selecting the significant descriptors in nanomaterial toxicity modeling in diverse experimental conditions. The selection of an optimal subset of nanomaterial descriptors by using the multivariate analysis methods can be a feasible solution. There are few published studies in this direction based on different experiments. Sayes and Ivanov [17] have proposed a procedure to determine the independent physicochemical descriptors of nanomaterials by grouping them using PCA and correlation analysis; Lynch [18] used a similar approach for the identification of interrelated physicochemical descriptors in toxicity modeling. However, we hardly found any reports that present a common physicochemical descriptor selection and validation protocol for dissimilar nanomaterials toxicity datasets measured in diverse conditions. In the present study, a simple and common procedure for the selection and validation of significant physicochemical descriptors of nanomaterial in their qualitative toxicity modeling is proposed. Moreover, the contribution of each of the descriptors (whether intrinsic or extrinsic) can be explained by the proposed qualitative model. Five different nanomaterial toxicity datasets from the published literature have been included in the analysis and validation of the proposed approach. In addition to the qualitative discrimination of toxic vs. nontoxic nanomaterials in the PC space with the selected optimal set of physicochemical descriptors, a quantitative class separability measure was used in the validation.

2. Experimental nanomaterials datasets

Five benchmark experimental nanomaterials datasets available in the recently published literature were collected and used in the present analysis. A short description of each of the datasets is as follows.

2.1. Nanomaterial dataset-I

The dataset was collected from published research of Liu et al. [4] (Table 1 and Table 2), in which nanostructure activity-based toxicity classification model is described. It contains fifteen physicochemical descriptors (G_{Me}: group of metal nanoparticles, Z_w: zeta potential in mV, isoelectric point, C_s: surface area concentration in m^2L^{-1}, C_n: number concentration in L^{-1}, d: size in nm, C_m: mass concentration in mg/L, ρ: density in g/cm^3, θ_v: ratio of C_m and ρ, E_{MeO}: atomization energy in $kcal/eqv^{-1}$, N_{me}: number of metal atoms, N_o: number of oxygen atoms, m_{Me}: mass of metal, m_{Meo}: molecular weight of the metal oxide and P_{me}: period of metal) of nine metal oxide nanoparticles (Al_2O_3, CeO_2, Co_3O_4, TiO_2, ZnO, CuO, SiO_2, Fe_3O_4, and WO_3). A total of eighty-three exposure concentrations of

Table 1
Loading value of nanomaterial descriptors on principal components (PCs) in the analysis of dataset-I.

Principal components (PCs)	Physicochemical descriptors														
	N_{Me}	N_o	m_{Me}	m_{MeO}	G_{Me}	Z_W	IEP	C_s	C_n	d	C_m	ρ	Θ_v	P_{Me}	E_{MeO}
PC1	−0.07	−0.16	−0.36	−0.36	0.29	0.07	0.04	0.31	0.07	0.32	0.18	−0.39	0.28	−0.38	0.14
PC2	−0.32	−0.24	0.02	−0.25	−0.04	−0.17	−0.14	−0.23	−0.48	0.28	−0.43	−0.12	−0.35	0.01	0.20
PC3	−0.37	−0.14	0.32	−0.03	−0.25	−0.36	−0.34	0.28	0.14	0.19	0.29	0.09	0.30	0.30	0.16
PC4	0.35	0.51	−0.05	0.27	0.05	−0.41	−0.45	0.08	−0.08	0.13	−0.13	−0.29	−0.02	−0.16	0.10
PC5	−0.15	−0.20	0.05	−0.06	0.51	−0.22	−0.35	0.02	0.04	−0.21	−0.06	0.04	−0.03	−0.05	−0.67
PC6	0.01	0.12	0.20	0.16	0.09	0.33	0.06	0.52	−0.63	0.11	−0.11	0.08	0.23	0.11	−0.17
PC7	−0.15	0.22	0.41	0.12	0.29	0.44	−0.10	−0.27	0.26	0.50	0.09	−0.18	−0.13	0.07	−0.05
PC8	0.14	−0.17	−0.36	0.19	−0.44	−0.02	−0.01	−0.06	0.00	0.55	0.08	0.10	−0.04	0.00	−0.52
PC9	−0.03	0.13	−0.03	0.01	0.17	−0.16	0.06	−0.43	−0.49	0.01	0.68	0.16	0.08	−0.11	0.03
PC10	0.15	−0.27	0.23	0.01	−0.32	0.14	−0.09	−0.11	−0.16	−0.29	0.19	−0.73	0.03	0.07	−0.17
PC11	0.16	−0.07	−0.38	−0.14	0.04	0.43	−0.62	0.01	−0.05	−0.11	0.15	0.16	−0.13	0.34	0.19
PC12	−0.02	0.02	0.06	0.01	0.00	−0.08	0.11	0.47	0.01	−0.01	0.37	−0.03	−0.78	−0.06	0.00
PC13	0.31	−0.16	0.42	−0.26	−0.21	0.14	−0.28	0.00	0.00	0.03	0.00	0.33	0.00	−0.62	0.02
PC14	−0.62	0.44	−0.18	−0.01	−0.35	0.22	−0.17	0.00	0.00	−0.23	0.00	−0.04	0.00	−0.33	−0.16
PC15	−0.19	−0.44	−0.08	0.75	0.12	0.09	−0.11	0.00	0.00	−0.07	0.00	0.02	0.00	−0.30	0.26

Note → N_{me} (number of metal atoms), N_o (number of oxygen atoms), m_{Me} (atomic mass of metal), m_{Meo} (molecular weight of metal oxide), G_{Me} (group of metal in periodic table), Z_w (zeta potential in mV), IPE (isoelectric point), C_s (concentration in $m^2\ L^{-1}$), C_n (concentration in L^{-1}), d (size in nm), C_m (concentration in mg/L), ρ (density in g/cm^3), Θ_v (volume fraction), P_{me} (period of metal in periodic table), E_{MeO} (atomization energy in $kcal/eqv^{-1}$).