Toxicology in Vitro 40 (2017) 102–114

Contents lists available at ScienceDirect

Toxicology in Vitro

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

In silico prediction of the mutagenicity of nitroaromatic compounds using a novel two-QSAR approach

Yi-Lung Ding [a], You-Chen Lyu [a], Max K. Leong [a,b,*]

[a] Department of Chemistry, National Dong Hwa University, Shoufeng, Hualien 97401, Taiwan
[b] Department of Life Science and Institute of Biotechnology, National Dong Hwa University, Shoufeng, Hualien 97401, Taiwan

ARTICLE INFO

Article history:
Received 31 July 2016
Received in revised form 13 November 2016
Accepted 21 December 2016
Available online 24 December 2016

Keywords:
Nitroaromatic compounds
Mutagenicity
2-QSAR
Partial least square
Hierarchical support vector regression

ABSTRACT

Certain drugs are nitroaromatic compounds, which are potentially toxic. As such, it is of practical importance to assess and predict their mutagenic potency in the process of drug discovery. A classical quantitative structure-activity relationship (QSAR) model was developed using the linear partial least square (PLS) scheme to understand the underline mutagenic mechanism and a non-classical QSAR model was derived using the machine learning-based hierarchical support vector regression (HSVR) to predict the mutagenicity of nitroaromatic compounds based on a series of mutagenicity data (TA98 − S9). It was observed that HSVR performed better than PLS as manifested by the predictions of the samples in the training set, test set, and outlier set as well as various statistical validations. A mock test designated to mimic real challenges also confirmed the better performance of HSVR. Furthermore, HSVR exhibited superiority in predictivity, generalization capabilities, consistent performance, and robustness when compared with various published predictive models. PLS, conversely, revealed some mechanistically interpretable relationships between descriptors and mutagenicity. Thus, this two-QSAR approach using the predictive HSVR and interpretable PLS models in a synergistic fashion can be adopted to facilitate drug discovery and development by designing safer drug candidates with nitroaromatic moiety.

1. Introduction

Nitroaromatic compounds (NACs), which are also termed aromatic nitros or nitroarenes, are a class of compounds with one or more nitro substituents attached to mono- or polycyclic aromatic hydrocarbons, viz. arenes, and can be derivatized from polycyclic aromatic hydrocarbons (PAHs) (Pederson and Siak, 1981). Many NACs show mutagenic and carcinogenic properties since they can be converted to highly reactive electrophilic nitroanion radical, nitroso intermediates, and *N*-hydroxy derivative, which, in turn, can form adducts with DNA, tissue proteins, and serum albumins and hemoglobins (Sabbioni, 1994), leading to various forms of toxicity such as mutagenicity (Debnath et al., 1991), immunotoxicity (Li et al., 1999), hepatotoxicity (Aßmann et al., 1997), skin sensitization (Cronin et al., 1998), nephrotoxicity (Travlos et al., 1996), neurotoxicity (Philbert et al., 2000), teratogenicity (Kovacic and Somanathan, 2006), or carcinogenicity (Debnath et al., 1991).

Generally, NACs are considered structural alerts in the process of drug discovery and development due to their potential toxic risks (Benigni and Bossa, 2006; Stepan et al., 2011) despite the fact that an appreciable number of drugs contain the nitroaromatic structure (Boelsterli et al., 2006). For example, nitroaromatic flutamide, which is a therapeutic agent for treating metastatic prostate cancer by blocking both endogenous and exogenous testosterone, is considered potentially hepatotoxic (Martelli et al., 2000). It has been attempted to replace the nitro group by cyano to reduce its toxicity while maintaining its efficacy (Coe et al., 2007).

Furthermore, it is of great importance to accurately and reliably assess toxicity of drug candidates during drug discovery and development (Dobo et al., 2012; Escobar et al., 2013; McCarren et al., 2011; Valerio, 2009; Valerio et al., 2007; Venkatapathy et al., 2009). Of various assay systems (Rao et al., 2004), the in vitro Ames test, which uses a histidine-free medium with an engineered *Salmonella typhimurium* bacteria to detect the sensitivity of mutagenics, is the most predictive and prevalent one (Ames et al., 1975; Benigni et al., 2010; Hornberg et al., 2014). Additionally, it has been shown that mutagenicity in *S. typhimurium* is closely related to carcinogenicity in rodent and human bioassays (Mortelmans and Zeiger, 2000). In fact, Ames tests have been extensively adopted by the European Union's Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH) legislation for predicting toxicity of novel chemicals (Claxton et al., 2010). In addition, different *S. typhimurium* strains have been engineered to test for different types of mutations and metabolic pathways. The TA100 and TA98 strains, for instance, are developed to detect base-pair substitution

* Corresponding author at: Department of Chemistry, National Dong Hwa University, Shoufeng, Hualien 97401, Taiwan.
E-mail address: leong@mail.ndhu.edu.tw (M.K. Leong).

http://dx.doi.org/10.1016/j.tiv.2016.12.013
0887-2333/

mutations and frame-shift mutations, respectively. Of various bacterial strains, TA100 and TA98 are the most effective for mutagenicity testing because they are more sensitive to mutagenic activity (Hornberg et al., 2014).

The mechanisms of nitroaromatic mutagenesis are rather complex, involving in penetration into the cellular systems, diffusion and binding to the active site of the specific enzyme, reduction reaction in the presence of certain enzymes to form an aromatic hydroxylamine intermediate, which, in turn, can further produce a nitrenium ion, viz. an electrophilic intermediate per se, to further react with nucleophiles such as proteins or DNA to form an adduct (Hakimelahi and Khodarahmi, 2005). Aromatic amines also go through the same mechanism of adduct reaction except that the aromatic hydroxylamine intermediates arise from oxidation. Thus, the formation of nitrenium intermediate can take place at nitro or amine moiety in those amine-containing NACs. This can yield a complicated situation since there is no compelling evidence to support the predominance of both functional groups (Fu, 1990). Nevertheless, it has been indicated that the TA98 strain consists of a full complement of nitroreductases required to activate the reduction reaction of NACs, whereas aromatic amines demand the presence of an exogenous metabolic activation system, viz. S9 mix, to initiate the oxidation reaction, suggesting that mutagenicity can only take place at the nitro moiety in nitroaromatic amines in *S. typhimurium* strain TA98 without S9 mix (Fu, 1990). Accordingly, such ambiguity can be eliminated once the mutagenicity study of NACs is carried out using *S. typhimurium* strain TA98 without S9 microsomial activation.

In addition to in vitro and in vivo assay systems, *in silico* approaches play an increasing role in regulatory toxicology and drug safety assessment (Benigni et al., 2013; Cherkasov et al., 2014; Fioravanzo et al., 2012; Greene and Pennie, 2015; Simon-Hettich et al., 2006; Valerio, 2013). In fact, various quantitative structure-activity relationship (QSAR) techniques, which are a mathematic means to establish the relationship between biological activity and chemical characteristics, have been adopted to develop predictive models, and more importantly, they have been extensively employed by REACH, the Seventh Amendment (of the Cosmetics Directive), and the Screening Information Data Set (SIDS) to predict toxicity of untested compounds as a supplement to experimental assessments (Bakhtyari et al., 2013). It has been suggested that *in silico* approach can be a good alternative to Ames tests (Naven et al., 2012). In fact, various packages and theoretical models have been proposed to quantitatively predict mutagenicity of Ames test (Hillebrecht et al., 2011; Xu et al., 2012). Also, a number of comparative molecular field analysis (CoMFA), molecular similarity indices analysis (CoMSIA), genetic function approximation (GFA), hologram QSAR (HQSAR), and classical QSAR models have been developed to predict mutagenicity of NACs (Caliendo et al., 1995; Debnath et al., 1991; Fan et al., 1998; King et al., 1996; Klein et al., 2000c; Lopez de Compadre et al., 1998; Lopez de Compadre et al., 1990; Lopez De Compadre et al., 1988; Maynard et al., 1986; Nair and Sobhia, 2008; Wang et al., 2005).

Linear predictive models, such as linear partial least square (PLS), are useful to interpret the relationship between descriptor and biological activity (Hasegawa and Funatsu, 2010). Such advantage, however, can become an insurmountable difficulty for developing a sound predictive model in case that structure-activity data are not linear per se as illustrated by the study conducted by Debnath et al. (1991), in which the nonlinear relationship between log *P* and log TA98 was observed. Machine learning (ML) methods, conversely, can extract the nonlinear relationship between input and output that are otherwise often mishandled by linear approaches since it has been demonstrated that ML-based models generally perform better than their conventional linear counterparts (Hou et al., 2009). Nevertheless, ML-based models are normally considered as "black box" approaches since they are hard to interpret the relationship between input and output. In another word, the advantages of PLS can be the limitations of ML-based schemes and vice versa. Nevertheless, both seemingly contradictory approaches can be carried out in a synergistic fashion to produce high predictivity while the interpretability can still be maintained at the same time as recently proposed by Fujita and Winkler (Fujita and Winkler, 2016). It has been suggested that linear SVM can be an alternative to resolve the conflict between predictivity and interpretability (Cherkasov et al., 2014). In reality, linear SVM is not the ultimate solution to this issue since it has been shown that linear SVM models cannot properly address the nonlinear relationship between input and output as compared with SVM (Ren et al., 2007).

Herein, the objective of this investigation was to use the novel two-QSAR approach by combing the predictive ML-based model and the interpretable linear model to accurately predict the mutagenicity of NACs and to understand the underline complex mechanism based on the most comprehensive data collection (TA98 − S9) from the literature, respectively. The former was carried out by the hierarchical support vector regression (HSVR) scheme (Leong et al., 2009) since it can simultaneously possess the advantageous characteristics of local model and global model, viz. broader coverage of applicability domain (AD) and higher level of predictivity, respectively, and the latter was executed by the PLS scheme to offer straightforward interpretation.

2. Materials and methods

2.1. Data compilation

The sample data play a crucial role in determining the quality of a predictive model (Cherkasov et al., 2014). As such, it is of necessity to collect a great number of samples with broad ranges of chemical structures and biological activities to develop a predictive model. A comprehensive literature search was carried out to compile data from a variety of sources (André et al., 1997; André et al., 1995; Debnath et al., 1991; El-Bayoumy et al., 1981; Hooberman et al., 1994; Juneja et al., 1991; Juneja et al., 2001; Jung et al., 1991; Klein et al., 2000b; Klein et al., 2000c; LaVoie et al., 1981; Ludolph et al., 2001; Rosenkranz et al., 1985; Rosenkranz and Mermelstein, 1985; Takamura-Enya et al., 2006; Tokiwa et al., 1981; Tokiwa et al., 2003; Vance et al., 1987; von Tungeln et al., 1994; Watanabe et al., 1994; You et al., 1993; Yu et al., 1992), whose mutagenicity values were expressed by the logarithm of the number of revertants per nanomole and assayed by *S. typhimurium* TA98 − S9 microsomial preparation. All chemical structures were carefully scrutinized to ensure their integrity and certainty since compounds with different chirality can exert different activities. The SMILES strings, CAS registry numbers, TA98 values, and literature references of all molecules collected in this study are listed in Table S1 (Supporting information).

2.2. Molecular descriptors

All selected molecules were first subjected to full geometry optimization using the density functional theory (DFT) B3LYP method with the basis set 6-31G(d,p) by the Gaussian 09 package (Gaussian, Wallingford, CT) in the dimethyl sulfoxide (DMSO) solvent using the polarizable continuum model (PCM) (Cammi and Tomasi, 1995; Miertuš et al., 1981) to mimic the real assay conditions. The atomic charges of optimized structures were calculated using the molecular electrostatic potential (ESP) model of Merz and Kollman (Besler et al., 1990). Furthermore, the energies of frontier orbitals, viz. HOMO energy (E_{HOMO}) and LUMO energy (E_{LUMO}), and dipole moment, were retrieved from optimization calculation. The energy-related descriptors, namely hardness (η), softness (σ), chemical potential (μ), and electrophilic index (ω), were also determined because of their implication in the formation of electrophile-nucleophile adduct (LoPachin et al., 2012).

$$\eta = (E_{LUMO} - E_{HOMO})/2 \quad (1)$$

$$\sigma = 1/\eta \quad (2)$$