Toxicology and Applied Pharmacology 272 (2013) 465–475

Contents lists available at ScienceDirect

Toxicology and Applied Pharmacology

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

Predicting carcinogenicity of diverse chemicals using probabilistic neural network modeling approaches

Kunwar P. Singh [a,b,*], Shikha Gupta [a,b], Premanjali Rai [a,b]

[a] Academy of Scientific and Innovative Research, Council of Scientific & Industrial Research, New Delhi, India
[b] Environmental Chemistry Division, CSIR-Indian Institute of Toxicology Research, Post Box 80, Mahatma Gandhi Marg, Lucknow 226 001, India

ARTICLE INFO

Article history:
Received 11 June 2013
Accepted 22 June 2013
Available online 13 July 2013

Keywords:
Carcinogenicity
Diversity
Probabilistic neural network
Generalized regression neural network
Interspecies model
Molecular descriptors

ABSTRACT

Robust global models capable of discriminating positive and non-positive carcinogens; and predicting carcinogenic potency of chemicals in rodents were developed. The dataset of 834 structurally diverse chemicals extracted from Carcinogenic Potency Database (CPDB) was used which contained 466 positive and 368 non-positive carcinogens. Twelve non-quantum mechanical molecular descriptors were derived. Structural diversity of the chemicals and nonlinearity in the data were evaluated using Tanimoto similarity index and Brock–Dechert–Scheinkman statistics. Probabilistic neural network (PNN) and generalized regression neural network (GRNN) models were constructed for classification and function optimization problems using the carcinogenicity end point in rat. Validation of the models was performed using the internal and external procedures employing a wide series of statistical checks. PNN constructed using five descriptors rendered classification accuracy of 92.09% in complete rat data. The PNN model rendered classification accuracies of 91.77%, 80.70% and 92.08% in mouse, hamster and pesticide data, respectively. The GRNN constructed with nine descriptors yielded correlation coefficient of 0.896 between the measured and predicted carcinogenic potency with mean squared error (MSE) of 0.44 in complete rat data. The rat carcinogenicity model (GRNN) applied to the mouse and hamster data yielded correlation coefficient and MSE of 0.758, 0.71 and 0.760, 0.46, respectively. The results suggest for wide applicability of the inter-species models in predicting carcinogenic potency of chemicals. Both the PNN and GRNN (inter-species) models constructed here can be useful tools in predicting the carcinogenicity of new chemicals for regulatory purposes.

© 2013 Elsevier Inc. All rights reserved.

Introduction

A large number of need-based synthetic chemicals are added everyday to the existing list. Several of these chemicals have been identified as potentially toxic to the humans. The regulatory agencies emphasize for the safety assessment of the existing as well as the new chemicals prior to their manufacture and use. Carcinogenicity and mutagenecity are among the major issues to be addressed in safety assessment of the chemicals. Chemicals that induce tumors, increase tumor incidence, or shorten the time to tumor occurrence are termed as carcinogens (Fjodorova et al., 2010a). Depending on the mechanism of carcinogenesis, the chemicals may be categorized as genotoxic or non-genotoxic. In general, the human epidemiological studies, long-term bioassays in experimental animals, transgenic assays, toxicokinetics and cancer mechanism studies, including the in vitro methods are considered the data sources for identification of the carcinogens (Bernauer et al., 2005). In the absence of human carcinogenicity data, long-term animal bioassays for carcinogenicity are regularly used to determine whether chemical agents are capable of inducing cancer in humans (NRC, 1983). A carcinogenic dose–response assessment of a chemical yields two widely used measures of carcinogenic potency, which is the dose at which chemicals cause carcinogenicity in a test animal; (a) tumor dose (TD_{50}) and (b) oral slope factor (OSF) (Venkatapathy et al., 2009). TD_{50} is defined as that chronic dose-rate in mg of chemical per kg body weight per day (mg/kg-bw/d), which would induce tumors in half the test animals at the end of its standard life span with respect to the control animals (Peto et al., 1984). Moreover, for regulatory purposes, it has been accepted that without human data, the animal bioassays are acceptable as definite evidence of carcinogenicity and substances that induce tumors in animals are considered as suspected human carcinogens until convincing evidence to the contrary is presented (IARC, 2006). At present, our knowledge on carcinogenicity relies on the data generated from rodent's carcinogenicity assays. Accordingly, several on-line database are available on rodent carcinogenicity which prominently include, the US National Toxicology Program (NTP) database (http://ntp-apps.niehs.nih.gov/ntp_tox/index.cfm), the Carcinogenic Potency Database (CPDB) (http://potency.berkeley.edu/cpdb.html), Istituto

* Corresponding author. Fax: +91 522 2628227.
E-mail addresses: kpsingh_52@yahoo.com, kunwarpsingh@gmail.com (K.P. Singh).

0041-008X/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.taap.2013.06.029

Superiore di Sanita, Chemical Carcinogens: "Structures and Experimental Data" (ISSCAN) (http://www.epa.gov/ncct/dsstox/sdf_isscan_external.html), and Pesticides Action Network (PAN) database (http://www.pesticideinfo.org).

However, the experimental approach for the carcinogenicity testing of the chemicals is very expensive; time consuming and unethical, emphasizing for the need of the computational modeling methods capable of predicting the carcinogenicity of the chemicals using their structural properties. Several quantitative structure–activity relationship (QSAR) models, describing mathematical relationship between the structural features and carcinogenicity of various categories of chemicals have been proposed. The chemical class-based QSAR models (Benigni et al., 2000; Franke et al., 2001; Gini et al., 1999a; Helguera Morales et al., 2006; Richard and Woo, 1990; Villemin et al., 1994; Zhang et al., 1992) although have strong mechanistic ground leading to better interpretation of predictions, they suffer with a limited applicability domain. Models for non-congeneric chemicals based on heterogeneous databases are reported (Contrera et al., 2003; Loew et al., 1985; Vracko, 1997) along with several expert systems (Klopman et al., 2004; Lagunin et al., 2005; Matthews and Contrera, 1998; Woo and Lai, 2005). However, none of these approaches resulted in highly satisfactory accuracies with structurally diverse compounds. All of these approaches consider different types of molecular descriptors as estimators. Selection and computation of relevant descriptors to extract information from compound structures are the major limitations of this research field. Moreover, modeling approaches based on linear relationship have limitations as the structural properties used as estimators may have complex non-linear dependence. The artificial intelligence (AI) approach based models, in general are capable to capture the complex nonlinear relationships between the relevant descriptors (or properties) and observed responses (Singh and Gupta, 2012). Among these, ANNs have emerged as unbiased tools of prediction of response using a set of independent estimators, and subsequently these methods have been applied in toxicity prediction of the organic chemicals (Panaye et al., 2006; Wang et al., 2010). Probability density function (PDF) based neural networks such as probabilistic neural networks (PNNs), and generalized regression neural networks (GRNNs) have successfully been used in various classification and regression modeling (Adeli and Panakkat, 2009; Panaye et al., 2006; Singh et al., 2012a, 2013). These methods provide high throughput and rapid adaptation and do not require any iterative training and thus can learn quite quickly. Moreover, they produce reproducible outputs and without any risk for local minimum of error surface (Walzack and Massart, 2000). Moreover, selection of the numerical descriptors of the compounds is one of the most critical steps in AI modeling approaches. Since, no single descriptor can capture all properties of a compound and it is not known in advance as which of the descriptors are relevant to a particular problem, it is highly desirable to select the ones which catch maximum possible chemical and structural properties (Consonni et al., 2002).

In this study, we constructed the probability based neural network models (PNN, GRNN) for classification and regression to predict the carcinogenic potency of diverse non-congeneric chemicals (www.epa.gov/ncct/dsstox/sdf_epafhm.html; http://www.pesticideinfo.org) with special attention being paid to the calculation and selection of molecular descriptors. Classification of the chemicals was performed to identify the positive and non-positive carcinogens; and regression was performed to predict the carcinogenicity of the chemicals ($-\log TD_{50}$) using a set of selected descriptors. The predictive and generalization abilities of the models constructed here (PNN, GRNN) were evaluated using several statistical criteria. The predictive models constructed for rat's carcinogenicity were also applied to the mouse and hamster carcinogenicity data (CPDB) and to the pesticide data (PAN).

The proposed probability function based neural network models can be used for identifying the carcinogen/non-carcinogen compounds and their prioritization for carcinogenic potency.

Materials and methods

Data set

For developing predictive models for carcinogenicity potency of non-congeneric chemicals, we considered the Lois Gold CPDB (Carcinogenic potency database) (http://potency.berkeley.edu/) reporting the rodent's carcinogenicity studies (http://www.epa.gov/ncct/dsstox/sdf_cpdbas.html). The CPDB is a single standardized resource of information on many chronic long term bioassays. It contains a large diversity of chemical structures (1547 substances), and reports tumor data in rodents. Here we have considered the data reporting carcinogenic potency of chemicals in rodents (rats, mouse, and hamster). For these chemicals, the carcinogenic potency is expressed as tomourgenic dose (TD_{50}). The TD_{50} value for a given target site(s) in the absence of tumors in control animals was taken to be the chronic dose (in mg/kg-bw/day) which induced tumors in half of the test animals at the end of a standard life span for the species (Gold et al., 1999). Rat driven data have been considered more suitable for human carcinogenicity prediction. In this study, a total of 834 chemicals (out of 1241 chemicals) for rat carcinogenicity, 292 chemicals for mouse carcinogenicity (positive carcinogens), and 57 chemicals for hamster carcinogenicity were selected, excluding the remaining ones due to non-availability of their complete set of properties. The list of selected chemicals used for carcinogenic potency modeling along with original rat, mouse and hamster-derived rodent carcinogenic potency expressed as discrete end point is provided in Supporting Information (Table 1SI, 2SI). Among the selected 834 chemicals, 466 were positive carcinogens and the remaining 368 were non-positive in rat, whereas, in mouse all 292 positive carcinogens were considered. In the case of hamster, among the total 57 chemicals, 38 were positive and 19 were non-positive carcinogens. For classification (positive, non-positive), all the 834 chemicals for rat, 632 chemicals for mouse, and 57 chemicals for hamster were taken, whereas, for regression modeling set of 457 chemicals for rat, 292 for mouse, and 38 for hamster were selected.

Molecular descriptors and feature selection

Molecular descriptors map the structure of the compound into a set of numerical or binary values representing various molecular properties that are deemed to be important for explaining activity. A set of 50 different molecular descriptors (physico-chemical, constitutional, geometrical, and topological) of each of the 834 chemicals considered here was selected initially. These molecular descriptors were calculated using chemspider (www.chemspider). The physico-chemical properties were computed by molecular structures, whereas, the constitutional, geometrical and topological descriptors were calculated by 2D structures of the molecules, which were taken in the form of SMILES (simplified molecular input line entry system). Since, all the descriptors may not be relevant to the classification and regression modeling, elimination of less significant descriptors can improve the accuracy of prediction, and facilitate the interpretation of the model through focusing on the most relevant variables. Here, the initial feature selection for classification and regression modeling was performed using the PNN and GRNN approaches trained by a Gaussian kernel function. For optimal values of the kernel function parameter σ, the PNN and GRNN models were trained by using the complete set of features computing the respective scoring functions to rank the contribution of features in the current set. The lowest ranked features were then removed (Xue et al., 2006). The PNN and GRNN systems were retained by using the remaining set of features, and the corresponding misclassification rate (MR) in classification and mean squared error (MSE) of prediction were computed by means of 10-fold cross validation. Finally selected descriptors could be gathered in four categories: physico-chemical (octanol–water partition coefficient, Log P; density, melting point, half

Download English Version:

https://daneshyari.com/en/article/5846669

Download Persian Version:

https://daneshyari.com/article/5846669

Daneshyari.com