Ecotoxicology and Environmental Safety 110 (2014) 280–287

Contents lists available at ScienceDirect

Ecotoxicology and Environmental Safety

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

Computational models to predict endocrine-disrupting chemical binding with androgen or oestrogen receptors

Yingjie Chen, Feixiong Cheng, Lu Sun, Weihua Li, Guixia Liu, Yun Tang *

Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China

ARTICLE INFO

Article history:
Received 19 February 2014
Received in revised form
2 August 2014
Accepted 5 August 2014
Available online 3 October 2014

Keywords:
Endocrine-disrupting chemicals
Androgen receptor
Oestrogen receptors
Machine learning
Substructure alert

ABSTRACT

Rapidly and correctly identifying endocrine-disrupting chemicals (EDCs) is an important issue in environmental risk assessment. Major EDCs are associated with the androgen receptor (AR) and oestrogen receptors (ERs). Because of the high cost and time-consuming nature of experimental tests, *in silico* methods are valuable alternative tools for the identification of EDCs. In this study, a large dataset related to EDCs was constructed. Each molecule was represented with seven fingerprints, and computational models were subsequently developed to predict AR and ER binders *via* machine learning methods including *k*-nearest neighbour (*k*NN), C4.5 decision tree (C4.5 DT), naïve Bayes (NB), and support vector machine (SVM) algorithms. The best model for predicting AR binders was PubChem Fingerprint-SVM, which exhibited an accuracy of 0.84. For ER binders, the best method was Extended Fingerprint-SVM with an accuracy of 0.79. Moreover, several representative substructure alerts for characterizing EDCs, such as phenol, trifluoromethyl, and annelated rings, were identified using the combination of information gain and substructure frequency analysis. Our study involved a systematic computational assessment of EDCs related to AR and ERs, and provides significant information on the structural characteristics of these chemicals, which are a great help in identifying EDCs.

1. Introduction

Many environmental chemicals can interfere with human endocrine system, resulting in adverse effects on the developmental, reproductive, neurological and immune systems (Colborn, 1995). These chemicals are referred to as endocrine-disrupting chemicals (EDCs). Bisphenol A is one such compounds, and it has multiple effects on many endocrine-related signalling pathways (Rubin, 2011). In fact, EDCs seriously threaten human health and have been demonstrated to be related to such phenomena as the global increase of testicular cancer, the regional decline of sperm counts, the decline and altered sex ratios in some regions, the increase in the incidence of breast cancer and endometriosis (Liu et al., 2007). Thus, EDCs have attracted both scientific and public attention, and EDCs are recognized as substances of very high concern (SVHC) (Li and Gramatica, 2010). Recently, the problem was addressed by the new European regulation REACH (Registration, Evaluation, Authorization & Restriction of Chemicals, http://ec.europa.eu/enterprise/sectors/chemicals/reach/index_en.htm), which set out the most demanding steps for regulating the use of such substances and requires a plan for safer alternatives. EDCs impact the endocrine system through a variety of complex mechanisms, primarily by binding to the receptors that are closely related to the balance of endocrine hormones, such as androgen receptor (AR) and oestrogen receptors (ERs).

Experimental assays for screening of the biological activity of large libraries of EDCs are time-consuming and expensive. Accordingly, the benefits of QSAR (Quantitative Structure–Activity Relationship) techniques to identify possible EDCs become obvious. Using QSAR methods, biological activity or classification can be predicted based on chemical structures and properties, which can decrease the number of animal tests. Such behaviour is in line with EU recommendations in the new REACH system for chemical regulation (Liu et al., 2007). In recent decades, some QSAR models were developed to identify potential EDCs related to AR or ERs. Li used *k*-nearest neighbours (*k*NN), local lazy IB1, ADTree methods and the consensus approach with DRAGON descriptors to build models for predicting the AR binders out of 625 chemicals (Li and Gramatica, 2010). The consensus model improved the external sensitivity from 57.1% to 76.4% compared with the results of Vinggaard (Vinggaard et al., 2008). In another paper (Li and Gramatica, 2010), the authors used similar methods to build models for predicting ER binders with 838 compounds. The prediction accuracy of the best model was 0.86. Panaye (Panaye et al., 2008) attempted to classify 202 chemicals as potential AR

* Corresponding author. Tel.: +86 21 6425 1052; fax: +86 21 6425 1033.
E-mail address: ytang234@ecust.edu.cn (Y. Tang).

http://dx.doi.org/10.1016/j.ecoenv.2014.08.026
0147-6513/

binders using recursive partitioning trees with several important descriptors. They proposed a multi-step classification procedure to detect the androgenic activity of chemicals. Stojić (Stojić et al., 2010) developed a model for predicting oestrogen-active endocrine disruptors based on 188 chemicals using counter-propagation artificial neural networks (CPANN) with DRAGON descriptors. The R^2 of the training set was equal to 0.85 and the R^2 of the test set was equal to 0.74. The authors analysed the mechanistic interpretation of the model thoroughly as well.

Nevertheless, most of these models were built by statistic methods with limited compounds and molecular descriptors. On one hand, the endocrine system is unusually complex and interferes with a large number of potential targets in human body directly or indirectly (Li and Gramatica, 2010). However, the aforementioned models only focused on either AR or ER binders, not both together. Thus, it is necessary to systematically study EDCs with two or more related receptors. On the other hand, molecular descriptor selection is an unavoidable process before building QSAR models. Descriptor selection is intricate, and the selected descriptors significantly impact the prediction accuracy of the QSAR models. It is also difficult to explain the models and the underlying mechanism using the selected individual or several simple chemical descriptors. Therefore, new molecular features or mixing multiple features to build models is more frequently used in the recent literature.

In this study, high-quality diverse data were collected from the literature and databases. Next, seven fingerprints were used to represent the chemicals, and four machine learning methods were applied to build binary classification models for the prediction of EDCs that bind to AR or ERs. Five-fold cross validation and external set validation were used to determine the predictive ability of the models. The chemical diversity of the datasets was also investigated. Substructure alerts (Kruhlak et al., 2007) of EDCs were analysed by the information gain and substructure frequency analysis methods, and several important patterns were obtained.

2. Materials and methods

2.1. Dataset construction and analysis

2.1.1. The dataset of AR binders and non-binders

A total of 1157 chemicals in the training set were extracted from three publications (Gunde Egeskov, 2012; Jensen et al., 2011; Li and Gramatica, 2010) and the external validation set containing 121 chemicals was based on data collected from another literature (Vinggaard et al., 2008). The AR binding affinity was expressed as IC_{25}, which means the concentration of a test compound displaying 25% inhibition of the activity induced by 0.1 nM R1881. The compounds were classified as AR binders if the IC_{25} is lower than 10 μM and non-binders if the IC_{25} is greater than 10 μM or there is no active value (Gunde Egeskov, 2012).

2.1.2. The dataset of ER binders and non-binders

A total of 333 chemicals serving as the training set were collected from the literature (Liu et al., 2007;Stojić et al., 2010). The external validation set was composed of 4648 chemicals from the Estrogenic Activity Database (EADB) (Shen et al., 2013), which is publicly available from http://www.fda.gov/ScienceResearch/BioinformaticsTools/EstrogenicActivityDatabaseEADB/default.htm. The database incorporates an extensive collection of chemicals obtained from *in vitro* and *in vivo* assays. We used the chemicals in this database after removing the chemicals with discordant ER binding activity data and the duplicated substances within the training set. The compounds of the training set were classified into ER binders and non-binders according to the original literature (Liu et al., 2007; Stojić et al., 2010). For the external validation set, chemicals with concordant positive results in all the tested assays are labelled as ER binders, and chemicals with concordant negative results in all the tests are labelled as ER non-binders. The detailed data concordance analysis method can be found in the original literature (Shen et al., 2013).

To obtain global models with chemical diversity, the data was merged from different sources with duplicates removed. Inorganic and metal ion-contained compounds were omitted. Salt chemicals were transformed to the corresponding acid or base. The detailed statistical descriptions of the entire AR and ERs datasets are listed in Table 1. EDCs (receptor binders) were represented as +1 and non-EDCs (receptor non-binders) as −1 when building binary classification models. The SMILES strings and classification of all chemicals can be found in the Supporting Information Table SI1.

Table 1
Statistic data of chemicals used in the training sets and the external validation sets of AR and ERs.

	Total number	Training set		External validation set	
		Binder	Non-binder	Binder	Non-binder
AR	1278	350	807	43	78
ERs	4981	188	145	4070	578

2.1.3. Chemical space and similarity analysis of the datasets

The chemical space distribution of the dataset was defined by MW (Molecular Weight) and Ghose–Crippen LogKow (ALogP). The structural diversity of the dataset was assessed by the average Tanimoto similarity indexes, based on MDL Public Keys. The "Calculate Diversity Metrics" protocol in Discovery Studio (version 3.5, Accelrys Software Inc., San Diego, 2010) was used to calculate the average molecular similarity of the datasets.

2.2. Calculation of molecular fingerprints

PaDEL-Descriptor (Yap, 2011) was used to calculated seven molecular fingerprints for each molecule, including a CDK Fingerprint (FP, 1024 bits), CDK Extended Fingerprint (Ext, 1024 bits), Estate Fingerprint (Est, 79 bits), MACCS keys (Mac, 166 bits), PubChem Fingerprint (Pub, 881 bits), Substructure Fingerprint (FP4, 307 bits), and Klekota–Roth Fingerprint (KR, 4860 bits). Detailed descriptions of these fingerprints can be found in the original literature (Klekota and Roth, 2008; Yap, 2011).

2.3. Model building methods

Four machine learning methods were used to build the models, including *k*NN, C4.5 decision tree (C4.5 DT), naïve Bayes (NB), and support vector machine (SVM). The first three methods were performed in Orange Canvas 2.0 (freely available at the following website: http://www.ailab.si/orange/). The SVM algorithm was performed in the LIBSVM 3.16 package (Chang and Lin, 2011) (freely available at the following website: http://www.csie.ntu.edu.tw/~cjlin/libsvm/).

2.3.1. k-nearest neighbours (kNN)

*k*NN predicts a classification for test cases on the basis of the majority voting of its *k* nearest neighbours in the feature space (Kauffman and Jurs, 2001). The nearness is measured by the Euclidian distance metrics, and the parameter of *k* was set to five in the present work.