Toxicology 370 (2016) 127–137

Contents lists available at ScienceDirect

Toxicology

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

Integrating *in silico* models to enhance predictivity for developmental toxicity

Marco Marzo[a,*], Sunil Kulkarni[b], Alberto Manganaro[a], Alessandra Roncaglioni[a], Shengde Wu[c], Tara S. Barton-Maclaren[b], Cathy Lester[c], Emilio Benfenati[a]

[a] *Laboratory of Environmental Chemistry and Toxicology, Department of Environmental Health Sciences, IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy*
[b] *Existing Substances Risk Assessment Bureau, Health Canada, Ottawa, Ontario, Canada*
[c] *The Procter & Gamble Company, 8700 Mason-Montgomery Rd, Mason, OH 45040, USA*

ARTICLE INFO

Article history:
Received 13 July 2016
Received in revised form 8 September 2016
Accepted 27 September 2016
Available online 28 September 2016

Keywords:
Developmental toxicity
QSAR
In vitro and alternatives
VEGA
In silico

ABSTRACT

Application of *in silico* models to predict developmental toxicity has demonstrated limited success particularly when employed as a single source of information. It is acknowledged that modelling the complex outcomes related to this endpoint is a challenge; however, such models have been developed and reported in the literature. The current study explored the possibility of integrating the selected public domain models (CAESAR, SARpy and P&G model) with the selected commercial modelling suites (Multicase, Leadscope and Derek Nexus) to assess if there is an increase in overall predictive performance. The results varied according to the data sets used to assess performance which improved upon model integration relative to individual models. Moreover, because different models are based on different specific developmental toxicity effects, integration of these models increased the applicable chemical and biological spaces. It is suggested that this approach reduces uncertainty associated with *in silico* predictions by achieving a consensus among a battery of models. The use of tools to assess the applicability domain also improves the interpretation of the predictions. This has been verified in the case of the software VEGA, which makes freely available QSAR models with a measurement of the applicability domain.

1. Introduction

The assessment of the potential for developmental and reproductive toxicity (DART) of a chemical is one of the important safety considerations incorporated by international regulatory agencies globally (ECHA, 2006; EPA, 2014; CEPA, 1999). The European REACH legislation requires specific assessment of DART (ECHA, 2006) for substances present in the European market. Similarly, in the US, assessment of reproductive or developmental effects is part of a Work Plan for chemical review by the EPA (2014). Canada's Chemicals Management Plan, under the Canadian Environmental Protection Act, 1999 (CEPA, 1999), takes a risk-based approach to chemical risk assessment which involves the review and evaluation of available scientific information from a variety of sources and considers multiple lines of evidence including the reproductive and developmental endpoints. Since the majority of these substances requiring assessment has no empirical DART data, *in silico* predictive models based on the concept of (Quantitative) Structure Activity Relationships ((Q)SAR) may be used as a supporting approach to address the data gaps. (Q)SAR models are built upon a training set of chemicals that have empirical data on DART effects. However, modelling DART is a challenge because of the complexity of the systems and processes that contribute to normal development and reproductive function. Furthermore, DART is not a singular endpoint; in fact it is composed of several different effects or sub-endpoints and encompasses several adversities including those of fetal growth (fetal growth retardation, decrease in fetal weight), fetal survival (fetal death, pre- and post-implantation loss), structural dysmorphogenesis and visceral organ toxicity. Maternal toxicity considerations add another level of complexity. Another key challenge to building such models is the limited availability of empirical DART data. In this study, the data on developmental toxicity on different species was combined together. Using a conservative approach we have designated a compound as positive for one or more of these endpoints if it was flagged as positive for developmental toxicity

* Corresponding author.
E-mail address: marco.marzo@marionegri.it (M. Marzo).

http://dx.doi.org/10.1016/j.tox.2016.09.015
0300-483X/

for any one of the species irrespective of dose and route. This is mainly due to the complexity, high operational costs, long durations and laboratory animal testing requirements associated with developmental and reproductive toxicity guideline experiments (OECD, 2015). Some programs offer specific models for each type of developmental or reproductive adverse effect resulting in a more local rather than global assessment of this endpoint (Lo Piparo and Worth, 2010).

The current study compared and integrated publicly available predictive models for developmental toxicity only, such as CAESAR (Cassano et al., 2010), present in VEGA platform, (Benfenati et al., 2013) and the decision tree by Wu et al. (2013) using all categories present in the original model but testing only training set compounds with data for developmental toxicity. These models are built upon different data sets, use different ways to process the chemical information, and are based on different algorithms. Furthermore, we used a third modeling approach, SARpy (Ferrari et al., 2013), which can automatically extract rules for toxicity (expressed as fragments) from data sets of chemicals to build categorical models based on the presence in the molecule of fragments associated with the effect.

The purpose of the current analysis was: (i) to compare different predictive models for developmental toxicity; (ii) to make accessible new predictive models (P&G and SARpy) that will help the assessment of chemicals present in the market; and (iii) to evaluate if the integration of independent *in silico* models improves the predictive performance.

2. Materials and methods

2.1. Data sets and related toxicity properties

Three independent data sets were used which included: 1) the data set used by Wu et al. (2013) referred to here as the P&G data set; 2) a data set obtained from the Leadscope database (Leadscope, 2014); and 3) the data set used to build the CAESAR developmental toxicity model (Arena et al., 2004; Cassano et al., 2010). The data sources were different, and we compared them (as under) by evaluating the possible overlap and conflict. Because experimental data are few, all available data were used irrespective of species, route and dose. Also data sets were composed of heterogenic compounds rather than compounds belonging to similar series or structural classes.

The P&G data set is composed of 716 compounds and for each compound there are reproductive toxicity and/or developmental toxicity effects. We took into consideration only developmental toxicity effects for model development. The assessment for developmental toxicity in this data set is organized as shown in Table 1.

Compounds that did not present any kind of toxicity (No Evidence) were treated as negative (or inactive). On the other hand, compounds that were flagged as “D”, “D(MT)”, “DTer” and “DTer (MT)” were considered positive (or active) for developmental toxicity. Chemicals without any experimental value (“no data” in the table) were excluded. Before using the predictive models on this data set, all SMILES representations for individual structures were examined to identify and exclude multi-constituent compounds using the publically available internet databases ChemIDplus (MLN, 2016) or ChemSpider (RSC, 2014); compounds that did not have the SMILES as well as those for which it was impossible to find the correct structure by using the CAS number were also excluded. At the end of this process of curation of the original data set, there were 639 chemicals in total: 585 positive and 54 negative chemicals, respectively.

The second data set was obtained from the Leadscope database. It comprised more than 2000 compounds with experimental data for different endpoints of developmental toxicity on rats, mice and rabbits. Some endpoints were measured as specific toxicity effects, as reported in Table 2.

We took into consideration only compounds that have experimental results for all endpoints listed in Table 2, both for positive and negative compounds. If a compound was positive for one or more of these endpoints it was flagged as positive for developmental toxicity. The final Leadscope data set was composed of 1320 compounds; 786 positive and 534 negative.

The last data set considered in this study was the training set used within the CAESAR predictive model. This data set comprises data taken from Arena et al. (2004). These experimental data have been classified in the original paper according to FDA categories for pregnancy, and their quality has been additionally checked by toxicologists within the CAESAR consortium (IRFMN, 2009), as described in the paper by Cassano et al. (2010). This data set of 292 chemicals predominantly consisted of potential teratogens with 201 active and 91 inactive compounds.

A key consideration regarding the outcome of the analysis is that the specific endpoints as defined within all three data sets are not fully overlapping, and thus there may be chemicals defined as toxic according to different criteria. This can affect the results of the predictions, as discussed below.

2.2. In silico models

In total, this study considered six different developmental toxicity model algorithms as described below:

2.2.1. P&G model

The original P&G model is a decisional tree built upon a data set of compounds with data displaying a known precedent for developmental and reproductive toxicity (Wu et al., 2013). The P&G model decision tree has six nodes representing different chemical features. For example, a compound that has a cyclic structure would form a node. In this study there are 25 categories as further splitting of these six nodes. Each category represents groups of compounds with a determined biological activity or common chemical feature. The 25 categories are further divided into 129 subcategories or “rules” defined by a structural backbone or scaffold with substituents. Considering all possible substituent

Table 1
Definition of toxicology categories in the P&G data set.

Code	Definition
D	For adverse effects on the fetus during pregnancy or those occurring during the perinatal period
D(MT)	Developmental effects only occur in the presence of maternal toxicity
DTer	The spectrum of developmental toxicity endpoints includes structural malformations
DTer (MT)	Structural malformation only occurred in the presence of maternal toxicity
No Evidence	No adverse effects on the fetus during pregnancy or during the perinatal period reported in studies evaluating those endpoints
No Data	No relevant studies identified