Reproductive Toxicology 30 (2010) 147–160

Contents lists available at ScienceDirect

Reproductive Toxicology

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

# Integrating (Q)SAR models, expert systems and read-across approaches for the prediction of developmental toxicity

M. Hewitt, C.M. Ellison, S.J. Enoch, J.C. Madden, M.T.D. Cronin*

*School of Pharmacy and Chemistry, Liverpool John Moores University, Byrom Street, Liverpool, L3 3AF, United Kingdom*

ARTICLE INFO

*Article history:*
Received 8 November 2009
Received in revised form 3 December 2009
Accepted 4 December 2009
Available online 16 December 2009

*Keywords:*
*In silico*
Quantitative structure–activity relationships (QSARs)
Expert system
Toxicity prediction
Weight-of-evidence
Reproductive toxicity
Teratogenicity

ABSTRACT

It has been estimated that reproductive and developmental toxicity tests will account for a significant proportion of the testing costs associated with REACH compliance. Consequently, the use of alternative methods to predict developmental toxicity is an attractive prospect. The present study evaluates a number of computational models and tools which can be used to aid assessment of developmental toxicity potential. The performance and limitations of traditional (quantitative) structure–activity relationship ((Q)SARs) modelling, structural alert-based expert system prediction and chemical profiling approaches are discussed. In addition, the use of category formation and read-across is also addressed. This study demonstrates the limited success of current modelling methods when used in isolation. However, the study also indicates that when used in combination, in a weight-of-evidence approach, better use may be made of the limited toxicity data available and predictivity improved. Recommendations are provided as to how this area could be further developed in the future.



## 1. Introduction

Reproductive and developmental toxicity are relatively poorly understood endpoints within human toxicology. This is because these are general terms that encompass a whole host of individual endpoints. The umbrella term "reproductive toxicity" includes a range of adverse effects induced by substances on sexual function and fertility in adult males and females, developmental toxicity in the offspring and effects on or mediated via lactation [1]. Therefore, a broad spectrum of endpoints is relevant including litter size, neonatal growth, sperm quality, duration of gestation and functional toxicities [2]. Toxicological testing associated with these endpoints is extremely costly both financially and in terms of animal usage [3]. It has been estimated that reproductive and developmental toxicity tests will account for 54% of the testing costs associated with REACH compliance [4]. For example, the two-generation reproductive toxicity test costs in the range of $500,000 to $750,000 per chemical and uses approximately 3200 animals [4,5]. As a result, the use of alternative methods to predict such toxicity is attracting a lot of attention. Given the potential savings which can be made in both money and animal usage, the development of any alterative or non-testing method (e.g. (quantitative) structure–activity relationships ((Q)SARs), *in vitro* assays, etc.) is highly sought after. The recent adoption of the EU REACH legislation is further fuelling this quest, as the implementation of such alternatives is actively encouraged [6].

Despite the urgent requirement for alternatives, the complex and multifaceted nature of reproductive and developmental toxicity makes these endpoints difficult to model. As a result, for computational (*in silico*) approaches such as the use of (Q)SAR, expert systems, category formation and read-across, there is a need to focus on individual endpoints that contribute to the overall toxicity. Unfortunately, given the high costs associated with experimentally obtaining such test results, there are limited data available within the published literature. In addition to the paucity of data complicating the development of *in silico* models, the limited understanding of the mechanisms involved within these processes also hampers progress. A multitude of different mechanisms and modes of action, many of which are yet to be defined [7], may be responsible for eliciting the effects. The diversity and complexity of these endpoints, frequently coupled with limited mechanistic knowledge, poses many problems in the prediction of these effects using (Q)SARs and other modelling approaches. As a result, few computational models are available for reproductive and developmental toxicity in comparison to other toxicological endpoints. One exception is in the area of endocrine disruption, where several models have been developed based upon the binding to the oestrogen and (to a lesser extent) the androgen receptors [7].

* Corresponding author at: School of Pharmacy and Chemistry, Liverpool John Moores University, Byrom Street, Liverpool, L3 3AF, United Kingdom.
Tel.: +44 (0) 151 231 2402; fax: +44 (0) 151 231 2170.
*E-mail address:* m.t.cronin@ljmu.ac.uk (M.T.D. Cronin).

0890-6238/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.reprotox.2009.12.003

Those computational models that are available for reproductive and developmental toxicity have been classified by Cronin and Worth [7] into three categories: (Q)SARs for reproductive/developmental toxicity; structural alert-based expert systems; and (Q)SARs for related absorption, distribution, metabolism and excretion (ADME) properties. These authors discuss the reasons behind the lack of *in silico* model development and general progress in this area, summarising these as follows:

1. A perceived lack of "high quality" toxicity data for modelling.
2. The lack of knowledge of mode and mechanisms of action that is required for modelling.
3. The appreciation that reproductive toxicity is a composite effect comprising a number of endpoints, some with specific (and in certain cases, well defined) mechanisms.
4. A professed difficulty in modelling reproductive toxicology. This is due to a combination of the previous three points.
5. Possibly, that the QSAR community has not viewed reproductive toxicity as an area of concern or interest. This may be because there are not readily available databases for modelling such as there are in some environmental endpoints (e.g. fish acute toxicity) or some other human health effects (e.g. carcinogenicity).

There is now a greater interest in releasing data for these endpoints, allowing improved models to be developed. However, currently available models are clearly not as advanced compared with other areas of toxicology.

The aim of the study described herein was to determine how accurately developmental toxicity could be predicted using available models and software tools individually and in combination. There are a number of available models and software tools to predict endpoints related to developmental toxicity. For the purposes of this study, the following were investigated:

(i) The CAESAR developmental toxicity model—a QSAR based model developed and hosted by the EU CAESAR project [8,9].
(ii) The 'reproductive toxicity super-endpoint' within Derek for Windows—a rule-based expert system (developed by Lhasa Limited, Leeds) which reports structural alerts relating to developmental, teratogenic and testicular toxicity.
(iii) A profiler that determines the potential for oestrogen receptor binding (available within the OECD (Q)SAR Application Toolbox). The rationale for including this endpoint is that oestrogen receptor binding is associated with endocrine disruption and may act as a molecular initiating event, eliciting reproductive or developmental toxicity effects.
(iv) A QSAR model for placental transfer, reported previously in the literature [10]. Ability to cross the placenta is key in determining whether or not an intrinsically toxic compound is likely to elicit a detrimental effect on developing offspring *in vivo*.
(v) A category formation approach, using Toxmatch [11–13] to generate categories of compounds based on structural similarity, from which a read-across estimate of toxicity could be made.

In addition to evaluating the performance of individual models and tools, this study also applied a weight-of-evidence approach in which the information from the individual tools was combined. Using weight-of-evidence, together with expert knowledge an overall prediction of toxicity potential was obtained which made best use of available information.

## 2. Methods

### 2.1. Dataset

Toxicity data for 290 chemicals were obtained from Briggs et al. [14]. This reference source provides the Food and Drug Administration (FDA) teratogenicity classification (A, B, C, D or X) for a database of predominantly pharmaceutical compounds. The five-tiered FDA classification scheme is defined below:

*Risk class A*: Controlled studies in women fail to show any adverse risk to the foetus in the first trimester. Data indicate there is no risk in later trimesters either.
*Risk class B*: Either no evidence from controlled animal studies or controlled studies in women indicate no risk despite animal data indicating foetal risk.
*Risk class C*: Either animal studies indicate foetal risk (with no contrary human data) or there are no controlled studies in either humans or animals. The chemical should only be used if the benefit to the mother justifies the risk to the foetus.
*Risk class D*: Positive evidence of human foetal risk. The chemical should only be used if the benefit to the mother justifies the serious risk to the foetus e.g. if the chemical is being used to treat the mother for a life threatening condition.
*Risk class X*: Positive evidence in humans and animals. The chemical should not be used by pregnant women or those about to become pregnant as the benefit to the mother is unlikely to outweigh the very serious risk to the foetus.

For the purposes of this study the five classifications indicated above were converted to a binary classification scheme, based on the method proposed within the EU CAESAR project [8,9]. Within this scheme any compounds in classes A or B were classified as non-toxicants whereas compounds in classes C, D or X were classified as potential developmental toxicants. The compounds used and their classifications are shown in Table 1. Whilst classifying all compounds in risk class C as toxicants may be debatable, the precautionary principle dictates that erring on the side of caution would be preferable for this endpoint where a false negative prediction could have far more serious consequences than a false positive prediction.

A subset of this database, comprising 57 chemicals, selected previously by Enoch et al. [15], was used as the test chemicals in the current study. These chemicals were randomly selected and comprised 42 developmental toxicants and 15 non-toxicants. (The bias towards toxicants in the test set is reflective of the dataset as a whole for which there are more potential toxicants than non-toxicants.) The remaining 233 compounds were used as the training set of which 156 were classed as developmental toxicants and 77 were classed as non-toxicants. The compound names, FDA classification, binary toxicity classification and assignment to training or test set are given in Table 1.

### 2.2. Models and software tools used in the analysis

As stated previously, relatively few (Q)SAR models have been developed for reproductive and developmental toxicity endpoints, with the majority being in the area of oestrogen receptor binding. However, five models/software tools were selected as being relevant for this study. The approaches were used to determine whether or not the 57 test compounds could be correctly classified as developmental toxicants or non-toxicants, when used alone or in combination.

#### 2.2.1. CAESAR developmental toxicity model

This is a (Q)SAR model developed and hosted by the EU CAESAR project [8,9]. Using a freely available Java-based server application users are able to submit compounds for processing online. The CAESAR model was developed using the same dataset as used in the current study [14]. Although the external predictivity of the CAESAR model could not be investigated here, the current study presented a useful opportunity to evaluate the statistical fit of the model.

A text file containing SMILES notations was used to import the test compounds into the online application. The CAESAR developmental toxicity model generates a text file as an output containing a binary classification of developmental toxicity (i.e. developmental toxicant/non-toxicant). Despite several models being developed for developmental toxicity within the CAESAR project, only one of these models is incorporated within the CAESAR software. The software uses a random forest algorithm developed using the Waikato Environment for Knowledge Analysis (WEKA) open source workbench. In this model 13 chemical descriptors are calculated internally and are used in the final model. See www.casear-project.eu for more details on the methodology.

#### 2.2.2. Derek for Windows (version 11)

In contrast to the CAESAR (Q)SAR model, Derek for Windows [16] is a rule-based expert system developed by Lhasa Limited (Leeds, United Kingdom) which can be used to identify important structural fragments within molecules that are associated with specific toxicological effects. The structural alerts are based either on hypotheses relating to mechanisms of action of a chemical class or on observed empirical relationships.

The likelihood that a chemical will cause toxicity if it contains a structural alert is based on the species in question, as well as some rules associated with bioavailability [17]. Depending on these factors, i.e. the species in question, presence of a structural alert, and bioavailability, Derek for Windows will give one of nine possibilities for the predicted toxicity of a chemical (certain, probable, plausible, equivocal, doubted, improbable, impossible, contradicted, or open). If the compound does not contain any structural alerts and there is no reason based on the physical properties of the compound to predict inactivity, the position is open and the programme will return "nothing to report". For this analysis, predictions of either certain, probable and plausible were taken as positive predictions for toxicity.