Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

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

Automated and reproducible read-across like models for predicting carcinogenic potency

Elena Lo Piparo [a,*,1], Andreas Maunz [b,1], Christoph Helma [b], David Vorgrimmler [b], Benoît Schilter [a]

[a] *Chemical Food Safety Group, Nestlé Research Center, Lausanne, Switzerland*
[b] *In Silico Toxicology GmbH, Basel, Switzerland*

ARTICLE INFO

Article history:
Received 8 April 2014
Available online 15 July 2014

Keywords:
Alternative method
Risk assessment
Quantitative structure activity relationship (QSAR)
Toxicity
Cancer potency (TD_{50})
Genotoxicity
Read-across

ABSTRACT

Several qualitative (hazard-based) models for chronic toxicity prediction are available through commercial and freely available software, but in the context of risk assessment a quantitative value is mandatory in order to be able to apply a Margin of Exposure (predicted toxicity/exposure estimate) approach to interpret the data. Recently quantitative models for the prediction of the carcinogenic potency have been developed, opening some hopes in this area, but this promising approach is currently limited by the fact that the proposed programs are neither publically nor commercially available. In this article we describe how two models (one for mouse and one for rat) for the carcinogenic potency (TD_{50}) prediction have been developed, using lazar (Lazy Structure Activity Relationships), a procedure similar to read-across, but automated and reproducible. The models obtained have been compared with the recently published ones, resulting in a similar performance. Our aim is also to make the models freely available in the near future thought a user friendly internet web site.

1. Introduction

Given increasing pressure to reduce animal testing, alternative methods relating chemical structure to toxicity have been increasingly valued in many regulatory organisations (ECHA, 2011; ICCR, 2012; U.S. EPA, 2008; EFSA, 2010; Arvidson et al., 2010). The contribution of computational toxicology to the future of regulatory decisions in public health has been addressed (NAS, 2007; Rusyn and Daston, 2010; U.S. EPA, 2012) and nowadays computational tools are widely promoted to support regulatory assessments and decision making in the field of food safety (Lo Piparo et al., 2011). In this context read-across has been mentioned as the most actionable short term strategy for reducing animal use.

From a food sector perspective, the application of such approaches may bring significant benefits not only in terms of saving time, cost, and with respect to reduction of use of laboratory animals, but also will open new horizons of risk assessment, giving the possibility of establishing levels of safety concern associated with human exposure to toxicologically uncharacterized chemicals. This is very relevant for both fast decision making (management of emergency safety issues) and priority setting (safety by design in research and development, R&D). Indeed new molecules are continuously identified and quantified in products as a consequence of the impressive improvement of analytical methods, and therefore companies need often to face and manage cases of emerging issues associated with chemicals for which no or little toxicological data are available. Moreover fast preliminary safety evaluations are increasingly required at the beginning of R&D projects for priority setting of potential new ingredients and to design intrinsically safe chemicals (safety by design).

In silico strategies are already integrated in the preclinical screening scheme of pharmaceutical discovery pipelines where an early identification of unacceptable toxicological hazard is a clear competitive advantage (Benfenati et al., 2009). Unfortunately it is difficult to directly transfer and use this expertise to food safety. Indeed the need of the food sector is different, where the most likely application of computational toxicology models would be in the establishment of the level of safety concern associated with the inadvertent/accidental presence of chemicals in finished products. This requires not only qualitative information on the potential hazardous properties of the chemical (e.g. probability that a compound is carcinogenic) but also quantitative information (e.g. carcinogenic potency) allowing a comparison with estimated exposure to establish the level of concern (Schilter et al., 2014).

Several qualitative (hazard-based) models for carcinogenicity prediction are available through commercial and free software, but only few tools are currently available for quantitative prediction. Carcinogenicity has often been considered as a too complex

[*] Corresponding author.
E-mail address: elena.lopiparo@rdls.nestle.com (E. Lo Piparo).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.yrtph.2014.07.010
0273-2300/© 2014 Elsevier Inc. All rights reserved.

end point (many mechanisms of action involved and little structural commonality) to be adequately modelled and quantitatively predicted.

In this contest, guidance for genotoxic impurities (GTI) was developed by US-FDA. The guidance suggests calculating cancer risk based on carcinogenic potency from a structural similar known carcinogen (USFDA, 2008). In addition recent efforts in the (Q)SAR field have resulted in the development of local quantitative models for the prediction of carcinogenic potency, opening some hopes in this area. Indeed these models provide reasonable predictions with errors within the same order of magnitude than the estimated variability of experimental data. This promising approach is currently limited by the fact that the proposed models are neither publically (Bercu et al., 2010 and Toropov et al., 2009) nor freely available (Contrera, 2011).

In contrast with the use of QSAR tools, generally the application of read-across is a more *ad hoc* approach involving a range of subjective choices in terms of similarity metrics and criteria for analogues selection. In this paper we describe two quantitative models (one for rat and one for mouse) to predict carcinogenic potency of genotoxic compounds by an alternative, automated and reproducible read-across like procedure. The models have been developed using *Lazar* (shortcut for *la*zy structure–*a*ctivity *r*elationships), a modular framework for predictive toxicology (Maunz and Helma, 2008; Maunz et al., 2013). The lazar models have been compared with the recently published ones by Bercu et al. (2010) and Contrera (2011), resulting in a similar performance.

Furthermore to provide transparency and meet regulatory demands the models have been submitted to QMRF (QSAR Model Reporting Format) Database (http://ihcp.jrc.ec.europa.eu/our_labs/predictive_toxicology/qsar_tools/QRF) and will be made freely available online through a user friendly platform that will provide detailed supporting information to the predicted toxicity values, such as the identification of the similar compounds used to build the model and the prediction confidence.

2. Materials and methods

2.1. Lazar similarity search

Lazar searches a database with chemical structures and experimental data (training set) for compounds similar to the query structure (*neighbours*) and calculates a prediction from the experimental measurements of the neighbours. Therefore it provides predictions for a given query compound in a three-step process (Maunz et al., 2013):

- Identification of similar compounds in the training dataset (neighbours).
- Creation of a local or read-across model for predictions based on structures and experimental activities of these neighbours.
- Application of the local or read-across model to predict the activity of the query compound.

For the determination of toxicity-related chemical similarities it is important to consider only descriptors, or features, that are relevant for the toxic endpoint under investigation. The crucial task is therefore to identify these features. *Lazar* relies on data mining algorithms to identify relevant features automatically from the training data. This procedure is reproducible and saves expensive expert work.

2.2. Statistical learning

In statistical learning theory, overfitting occurs when a statistical model describes noise instead of the underlying relationship. Machine Learning (ML) algorithms, for example Support Vector Machines (SVM) and Random Forests (RF) support strategies to limit the fit to the training data.

SVMs are a class of algorithms where data points are treated as vectors. For classification and regression, the data points are usually mapped to a high-dimensional feature space through kernel functions. SVMs support regularization via an internal cost function (Vapnik and Cortes, 1995).

The RF algorithm incorporates a general strategy for regularization known as bagging (short for bootstrap aggregation) (Breiman, 2001). In bagging, the training data is not processed as a whole by the learning algorithm, but n so-called bootstrap samples are drawn with replacement and trained upon individually. For increasing n, the instances that where not selected for each sample, termed OOB (out-of-bag), will cover around 36% of the data, on average. RF builds a decision tree model for each bootstrap sample to predict the dependent variable, and predicts the OOB data with it to estimate the error rate of the model (Liaw and Wiener, 2002). A RF model consists of a set of n such trees. A prediction for an unknown data point (query compound) is derived by averaging over the individual tree predictions of the forest. RF can fit arbitrarily shaped dependent variables, especially non-linear and non-continuous ones, and is able to handle large amounts of features.

Lazar was designed to handle high-dimensional, numerically unconstrained feature spaces, while maintaining its instance-based approach, i.e. a separate model is trained for each query structure in a time efficient manner. Technically, this work presents:

- Instance-based SVM learners with regularization.
- Feature selection services, controlled by bootstrapping.

These are employed for:

- Feature selection from more than 300 freely available, non-proprietary, physico-chemical descriptors (Steinbeck et al., 2006; O'Boyle et al., 2011; Wegner, 2004) using a Random Forest approach.
- Several regression and derived classification models for predicting numeric TD_{50} values and categories for potency.

2.3. Data set

A measure of carcinogenic potency is given by TD_{50}, defined by the daily dose in mg/kg/day that causes a tumor type in 50% of the exposed animals that otherwise would not develop the tumor in a standard lifetime (Gold et al., 2001). The datasets were composed from CPDB entries by Bercu et al., available in supplementary material for download. They consist of two datasets, one for rat and one for mouse, each being split into 90% training and 10% test data. The split was done by selecting every tenth compound from the full data, sorted on TD_{50} values, which allowed full coverage of training TD_{50} values in the test set. Moreover, Bercu et al. converted TD_{50} values to pTD_{50} for data normalization by the following equation:

$$pTD_{50} = -\log\left(\frac{TD_{50}}{1000 * \text{molecular weight}}\right)$$

Dividing by molecular weight transforms the cancer potency value on a molar basis. This study made no changes to the data whatsoever, neither to compounds nor to activity values. Therefore the dataset employed by this article, such as the one from Bercu et al., contains a total of 460 training set plus 51 test set compounds for rat, and 362 training set plus 40 test set compound for mouse.