Toxicology in Vitro 36 (2016) 197–209

Contents lists available at ScienceDirect

Toxicology in Vitro

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

Consensus of classification trees for skin sensitisation hazard prediction

D. Asturiol *, S. Casati, A. Worth

Joint Research Centre, Via Enrico Fermi 2749, Ispra 21027, VA, Italy

ARTICLE INFO

Article history:
Received 18 December 2015
Received in revised form 8 July 2016
Accepted 21 July 2016
Available online 22 July 2016

Keywords:
QSAR
Skin sensitisation
In vitro
In silico
Prediction
Decision tree

ABSTRACT

Since March 2013, it is no longer possible to market in the European Union (EU) cosmetics containing new ingredients tested on animals. Although several *in silico* alternatives are available and achievements have been made in the development and regulatory adoption of skin sensitisation non-animal tests, there is not yet a generally accepted approach for skin sensitisation assessment that would fully substitute the need for animal testing.

The aim of this work was to build a *defined approach* (*i.e.* a predictive model based on readouts from various information sources that uses a fixed procedure for generating a prediction) for skin sensitisation hazard prediction (sensitiser/non-sensitiser) using Local Lymph Node Assay (LLNA) results as reference classifications. To derive the model, we built a dataset with high quality data from *in chemico* (DPRA) and *in vitro* (KeratinoSens™ and h-CLAT) methods, and it was complemented with predictions from several software packages.

The modelling exercise showed that skin sensitisation hazard was better predicted by classification trees based on *in silico* predictions.

The defined approach consists of a consensus of two classification trees that are based on descriptors that account for protein reactivity and structural features. The model showed an accuracy of 0.93, sensitivity of 0.98, and specificity of 0.85 for 269 chemicals. In addition, the defined approach provides a measure of confidence associated to the prediction.

1. Introduction

The assessment of skin sensitisation potential represents a key requirement within several pieces of chemicals' regulations in the EU. For example, the REACH regulation (EC, 2006) foresees that chemicals produced or marketed in quantities of one tonne or more *per annum* be assessed for their potential to cause allergic contact dermatitis in humans, and within the Cosmetics Regulation (EC, 2009) skin sensitisation is one of the toxicological endpoints that require particular focus. The REACH regulation demands that testing on vertebrate animals should be considered only as last resort. The Cosmetics Regulation banned the animal testing of cosmetic ingredients in 2009 and the marketing of cosmetics containing new ingredients tested on animals in 2013 (EC, 2009).

The main chemical and biological mechanisms underpinning skin sensitisation are established (Karlberg et al., 2008; Martin, 2015; Martin et al., 2011) and have been described in the form of an adverse outcome pathway (AOP) (OECD, 2012a,b). Within this AOP, four key events (KE) are considered necessary for the acquisition of skin sensitisation: the covalent binding to skin proteins (KE-1) – also considered to be the molecular initiating event (MIE) –, the activation of keratinocytes (KE-2), the maturation of dendritic cells (KE-3), and the activation and proliferation of memory T-cells.

Progress has been made over the past ten years in the development of non-testing and testing methods addressing the key events of the skin sensitisation AOP. Three animal-free methods that account for KEs 1, 2, and 3 have been formally assessed by the European Union Reference Laboratory for Alternatives to Animal Testing (EURL ECVAM). These methods are: the direct peptide reactivity assay (DPRA) (EURL ECVAM, 2013; Gerberick et al., 2004), KeratinoSens™ (Emter et al., 2010; EURL ECVAM, 2014; Natsch and Emter, 2008), and the human cell-line activation test (h-CLAT) (Ashikaga et al., 2006; EURL ECVAM, 2015; Sakaguchi et al., 2006).

The DPRA, KeratinoSens™, and h-CLAT have been adopted by the Organisation for Economic Cooperation and Development (OECD) as Test Guidelines 442C (OECD, 2015a) and 442D (OECD, 2015b) and 442E (not yet published), respectively. Despite these methods predict LLNA responses with an accuracy of about 80% they are not proposed to be used as standalone alternatives. One of the reasons put forward for this is that they model specific KEs of the AOP and not the final adverse effect.

Progress has been made in the integration of results from *in silico*, *in chemico* and *in vitro* methods in defined approaches (OECD, 2016a, 2016b) to improve skin sensitisation hazard/potency prediction with respect to the individual methods. The first approach of this kind was developed by Natsch et al. (Natsch et al., 2009). The authors made a

* Corresponding author at: European Commission; Directorate General Joint Research Centre; Directorate F – Health, Consumers and Reference Materials; Chemicals Safety and Alternative Methods, VA, Italy.
E-mail address: david.asturiol-bofill@ec.europa.eu (D. Asturiol).

http://dx.doi.org/10.1016/j.tiv.2016.07.014

proof of concept of the prediction model based on scores proposed by Jowsey et al. (Jowsey et al., 2006), which was intended for predicting skin sensitisation potency. The model did not predict LLNA potency successfully, but a good performance was achieved in predicting skin sensitisation hazard for 116 chemicals (sensitivity = 0.86, specificity = 0.94, and accuracy = 0.88). Since then, a number of other approaches integrating non-animal data which use the AOP as a framework and are proposed for skin sensitisation hazard and/or potency assessments have been published. These range from simple weight-of-evidence (WoE) approaches (*e.g.* Bauch et al., 2012; Guyard-Nicodème et al., 2015; Macmillan et al., 2016; Urbisch et al., 2015), tiered approaches involving interim decision steps at the end of each tier (*e.g.* Takenouchi et al., 2015; van der Veen et al., 2014), and multiple regression models (Natsch et al., 2015) to more complex mathematical models (MacKay et al., 2013), artificial neural networks (*e.g.* Hirota et al., 2013, 2015; Tsujita-Inoue et al., 2014), and support vector machine-based approaches (Strickland et al., 2016). Another model integrating data from various sources and developed for LLNA potency prediction is the one based on a Bayesian Network (Jaworska, 2011; Jaworska et al., 2013, 2015). Bayesian networks are probabilistic models that can work with data gaps and can guide additional testing by quantifying the additional test information value before performing the testing.

Some authors have analysed in detail the performance of various *in silico* methods and expert systems when predicting skin sensitisation potential (Teubner et al., 2013; van der Veen et al., 2014). They showed that in general this kind of skin sensitisation methods had sensitivities above 0.70 and specificities below 0.65, even when some of them were combined. They concluded that the methods evaluated were not sufficiently accurate to be broadly used for skin sensitisation prediction. Alves et al. (Alves et al., 2015) recently showed that random forest models built from *in silico* descriptors obtained from the 2D structure of chemicals can have higher accuracy and larger applicability domains than the *in silico* methods reviewed by Teubner et al. and van der Veen et al. Alves et al. developed a series of consensus random forest models that predict skin sensitisation hazard (sensitiser *vs.* non sensitiser) using LLNA results as reference data. Their models used descriptors calculated with Dragon (Talete Srl, 2010) and SiRMS (Muratov et al., 2010) and were applied to a total of 406 chemicals, the largest skin sensitisation dataset published to date. The authors finally used a model based on a consensus of random forests that showed an accuracy of 0.82, sensitivity of 0.79, and specificity of 0.85 for the training set. These predictive performance values were obtained for 82% of the chemicals of the training set (chemical space coverage = 82%) as the predictions of the remaining 18% of chemicals were discarded because the two forests had contradictory outputs and the overall prediction was considered equivocal. The corresponding statistics for the validation set are not reported here as they are not representative because the validation set was highly unbalanced, *i.e.* contained 152 sensitisers and only 5 non-sensitisers. It is worth mentioning that the coverages of the validation sets of the different models developed by Alves et al. were significantly lower than those of the training sets, being of 50% the highest amount of chemicals of the validation tests that could be predicted.

The aim of our work was to build a model for predicting skin sensitisation hazard (sensitiser/non-sensitiser) that was simple, accurate, highly sensitive, and if possible integrating data from different sources, *i.e.* a defined approach. In order to develop the best model possible we have built a high quality database of 269 chemicals with LLNA data and skin sensitisation results obtained from DPRA, KeratinoSens™, and h-CLAT. The dataset has been quality checked by EURL ECVAM in collaboration with the test developers, and has been completed with a number of descriptors predicted with several free and licensed software packages yielding about 4500 descriptors for each of the 269 chemicals. This database has been used to build different classification trees to predict skin sensitisation hazard using LLNA results as reference. The two trees with the highest specificities and accuracies against LLNA classifications were used in a conservative consensus approach as final prediction model. In addition, a qualitative confidence measure on the prediction was added to the model by taking into account the leaves that were used in each individual tree to obtain the final consensus prediction.

2. Materials and methods

2.1. Dataset compilation

A dataset of 269 organic chemicals (170 sensitisers and 99 non sensitisers) identified by their chemical name and SMILES codes with *in chemico, in vitro*, and *in vivo* skin sensitisation data (LLNA and human) was built to develop a model to predict skin sensitisation hazard.

The initial collected dataset contained a total of 315 substances with human and/or LLNA data. Of these, 22 substances with only human data available were not considered. 16 inorganic chemicals and two mixtures (Pepperwood and Kathon CG) were discarded since they could not be calculated with most *in silico* software packages. Ammonium peroxodisulphate was also discarded because it was considered an inorganic chemical by TIMES (Dimitrov et al., 2005b), and 1,6-diisocyanatohexane, methylisoeugenol, 4-methylcatechol, diphenylmethane-4,4′-diisocyanate, and 4-nitrobenzyl chloride were discarded because they were considered as sensitisers or respiratory sensitisers in the sources but had no associated LLNA EC3 values, which was interpreted as an indication of lower quality data. The remaining 269 chemicals were used for modelling.

The dataset can be found in the Supporting Information (SI_Dataset.xls) and contains: name, SMILES, human skin sensitisation classification (1 to 6 categories), NOEL values (μg/cm^2) (Basketter et al., 2014), human GHS derived classifications (1A, 1B, NS), the LLNA EC3 values obtained from the different sources with a corresponding final call made by the authors for those cases in which multiple LLNA studies were available for the same chemical, and the *in chemico* and *in vitro* readouts that are explained in the next section. Binary descriptors indicating positive or negative predictions for each of the methods and the LLNA skin sensitisation hazard are also included in the dataset. In addition, the values of DRAGON and TIMES-SS descriptors used in the consensus model, a column indicating the use given to each chemical for each tree (*i.e.* training set, test set, or external test set), and the final consensus model predictions with the corresponding qualitative confidence measures are reported.

2.2. In chemico and in vitro data

The non-animal data included in the dataset were those generated with the three validated and OECD adopted methods, *i.e.* DPRA, KeratinoSens™, and h-CLAT, and were obtained from the validation study reports (EURL-ECVAM, 2012, 2015; EURL-ECVAM, 2014) and the scientific literature (Bauch et al., 2012; Emter et al., 2010; Gerberick et al., 2004, 2007; Natsch and Emter, 2008; Natsch et al., 2013; Nukada et al., 2013; Takenouchi et al., 2013).

The DPRA (OECD, 2015a) is an *in chemico* method which addresses peptide reactivity, considered to be the Molecular Initiating Event (MIE) or Key Event (KE)-1 in the skin sensitisation AOP (OECD, 2012a), by measuring the depletion of synthetic heptapeptides containing either cysteine or lysine following 24 hour incubation with a single concentration of the test substance. Depletion of the peptide in the reaction mixture is measured by HPLC using UV detection. Average peptide depletion data for cysteine and lysine are interpreted using a classification model in which chemicals classified as having minimal reactivity are considered to lack skin sensitisation potential whereas chemicals classified as having low, moderate, or high reactivity are considered to be skin sensitisers. DPRA data included in the dataset were: a) the % cysteine and b) the % lysine depletion values, c) average of cysteine and lysine depletion values, d) the DPRA positive or negative prediction, and