Regulatory Toxicology and Pharmacology 67 (2013) 468–485

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

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

Computer models versus reality: How well do *in silico* models currently predict the sensitization potential of a substance

Wera Teubner [a,*], Anette Mehling [b], Paul Xaver Schuster [c], Katharina Guth [c], Andrew Worth [d], Julien Burton [d], Bennard van Ravenzwaay [c], Robert Landsiedel [c]

[a] BASF Schweiz AG, Basel, Switzerland
[b] BASF Personal Care and Nutrition GmbH, Düsseldorf, Germany
[c] BASF SE, Ludwigshafen, Germany
[d] European Commission, Joint Research Centre, Institute for Health and Consumer Protection, Ispra, Italy

ARTICLE INFO

Article history:
Received 18 April 2013
Available online 30 September 2013

Keywords:
Case Ultra
GHS
OECD toolbox
Adverse outcome pathway
(Q)SAR
Skin sensitization
TIMES-SS
TOPKAT
Toxtree
Vega

ABSTRACT

National legislations for the assessment of the skin sensitization potential of chemicals are increasingly based on the globally harmonized system (GHS). In this study, experimental data on 55 non-sensitizing and 45 sensitizing chemicals were evaluated according to GHS criteria and used to test the performance of computer (*in silico*) models for the prediction of skin sensitization. Statistic models (Vega, Case Ultra, TOPKAT), mechanistic models (Toxtree, OECD (Q)SAR toolbox, DEREK) or a hybrid model (TIMES-SS) were evaluated. Between three and nine of the substances evaluated were found in the individual training sets of various models. Mechanism based models performed better than statistical models and gave better predictivities depending on the stringency of the domain definition. Best performance was achieved by TIMES-SS, with a perfect prediction, whereby only 16% of the substances were within its reliability domain. Some models offer modules for potency; however predictions did not correlate well with the GHS sensitization subcategory derived from the experimental data. In conclusion, although mechanistic models can be used to a certain degree under well-defined conditions, at the present, the *in silico* models are not sufficiently accurate for broad application to predict skin sensitization potentials.

1. Introduction

Allergic contact sensitization (ACD), often referred to as delayed-type hypersensitivity, is the clinical manifestation of an allergy to a topically applied substance and a significant contributor to both occupational and consumer dermatitis. Currently, it is estimated that 15–20% of the Western world population will suffer from allergic contact dermatitis at some point during the course of his or her lives. The clear social and economic impact of ACD is reflected by the requirement for the evaluation of the sensitization potential of a substance placed on the market found in much current legislation world-wide, amongst others, the European Chemicals Regulation concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH; Regulation (EC) No 1907/2006). REACH requires an assessment of the intrinsic potential of a substance to cause sensitization (hazard) as part of the base set of toxicological endpoints to be evaluated for the vast majority of substances to be registered. Unless arguments for waiving or adaptation of the standard testing regime can be applied, REACH requires testing in mice (LLNA; OECD testing guidelines 429 and 442) or in exceptional cases in guinea pigs (GPMT, Buehler; OECD 406). At least in the EU, registrants must share animal data to avoid unnecessary animal suffering and so classification and labeling of industrial chemicals is frequently derived from a single study. According to the globally harmonized system for classification and labeling of chemicals (GHS, ST/SG/AC.10/30/Rev.4), a skin sensitizing substance can be assigned the label of a strong sensitizer (Cat 1A) or "other" (low to moderate) sensitizer (Cat 1B) if sufficient data is available to make such a distinction based on frequency of occurrence in humans and/or potency in animals. For the hazard class assessment of finished products such as a cosmetic formulation, GHS allows two options: Either the finished product can be tested or the experimental data of the ingredients or a comparable mixture can be used. Reliable human data such as the repeated insult patch test is certainly the most

Abbreviations: ACD, allergic contact dermatitis; AOP, adverse outcome pathway; ECHA, European Chemicals Agency; GHS, globally harmonized system of classification and labeling of chemicals; LLNA, local lymph node assay; LMW, low molecular weight; OECD, Organisation for Economic Cooperation and Development; (Q)SAR, quantitative structure–activity relationship; REACH, Registration Evaluation Authorisation and Restriction of Chemicals.

* Corresponding author. Address: BASF Schweiz AG, Product Safety Switzerland, Klybeckstr. 141, 4057 Basel, Switzerland. Fax: +41 61 63 61304.
E-mail address: wera.teubner@basf.com (W. Teubner).

0273-2300/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.yrtph.2013.09.007

relevant information for a finished product. Repeated insult patch testing has been applied by cosmetic and consumer product firms to ensure the safety of consumer products, but discussions are ongoing on the ethics and human testing should only be conducted to confirm the absence of a sensitization potential. For the purposes of classification and labeling, new tests on humans shall not be performed as stated in article 7 of EC directive 1272/2008. For the overall hazard labeling of the finished product based on the ingredients, there are two concentration limits: A formulation will need to be labeled as a skin sensitizer if it contains more than 0.1% of strongly sensitizing substances (CAT 1A). In case of moderately sensitizing substances (Cat 1B), the threshold is 1%.

An allergic reaction consists of the sensitization phase and the elicitation phase. Sensitization represents the initial priming phase of the immune system without which an allergen would not be recognized. While sensitization itself does not yet lead to an allergic response repeated exposure to sufficient amounts of the allergen is the needed to elicit the allergic reaction found in ACD. The concept of adverse outcome pathways (AOP) has recently been described by the OECD in which mechanistic modes of toxicity, i.e. cause and steps leading to effects, are identified and used to develop new toxicological tests and test strategies (ENV/JM/MONO(2011)8). There are a number of key steps to the sensitization process which outlined in the OECD AOP for skin sensitization (OECD ENV/JM/MONO(2012)10). Among these, haptenization, i.e. binding of low molecular weight substances to protein entities within the skin either directly, or following metabolic (prohapten) or non-metabolic (prehapten) conversion is in most cases one of the primary and essential steps (Jaeckh et al., 2012; Oesch et al., 2007). The correlation between skin protein reactivity and skin sensitization potential has been known for many years (Dupuis and Benezra, 1982; Landsteiner and Jacobs, 1936; Lepoittevin, 2006) and the majority of haptens identified of causing allergic responses are electrophilic low molecular weight chemicals or their reactive metabolites that then form covalent bonds with nucleophilic centers on proteins. Although most allergens exhibit electrophilic reactivity and/or are of low molecular weight (LMW), the converse conclusion cannot be made that all electrophilic or LMW substances are automatically sensitizers – further steps along the AOP are needed. Non-electrophilic protein interactions have also been reported and can involve mechanisms such as haptenization via disulfide linkages or formation (chemical thiol haptenization), coordinate covalent binding of metals to protein structures and direct interactions of molecules with the T-cell receptor complex and/or major histocompatibility complex (reviewed in Chipinda et al., 2011). Several reaction mechanistic domains can currently be defined for skin sensitizers: (pro) Michael-type receptors, S_NAr electrophiles, SN2 electrophiles, Schiff-base formers and acylating agents (Aptula and Roberts, 2006; Aptula et al., 2007). Mechanisms such as these can be incorporated into *in silico* models as possible determinants to define a mechanistic basis for a mode of action for sensitization.

In the last decade, much effort has been placed in developing alternatives to animal testing (Mehling et al., 2012). Among the approaches used, is the development of *in silico* prediction models, in particular those based on (quantitative) structure activity relationships [(Q)SARs]. These models combine the use of physical chemical or structural properties (descriptors) and computational methods to assign a molecule to a certain category or biological activity, e.g. based on chemical class or mechanistic reactivity domain, relevant for the prediction of a certain toxicological endpoint. In the case of sensitization, the molecular weight and reactivity of electrophiles are often used to classify skin penetration properties and mechanistic domains. Computational searches and algorithms compare the molecular properties and functional similarities with those found in the database and used as the training set. (Q)SAR models can either include current mechanistic knowledge such as electrophilicity (e.g., DEREK, Toxtree, TIMES-SS) or be fully based on statistical evaluation of structure fragments determined from the model training set (e.g., Case Ultra, Vega, TOPKAT) (Patlewicz and Worth, 2008). The OECD (Q)SAR toolbox provides mechanistic information along the adverse outcome pathway from which substance-specific (Q)SAR models can be created. In general, a (Q)SAR model should have clearly defined endpoints, an unambiguous algorithm, sufficient data on robustness, a high degree of predictivity and the applicability domains should also be clearly defined. A number of models and expert systems contain modules for the prediction of skin sensitization potentials are available and include DEREK Nexus (LHASA, Leeds, UK), TIMES-SS (University of Bourgas, Bulgaria), Case Ultra (MULTICASE, Cleveland, USA), TOPKAT (Accelrys, San Diego, USA), VEGA (Vega), Toxtree and the OECD (Q)SAR toolbox. Comprehensive reviews on the use of (Q)SARs and their use in integrated testing strategies have recently been published by the European Center for the Validation of Alternative Methods (Kinsner-Ovaskainen et al., 2012) and by the UK Government Departments and Agencies (IGHRC, 2013).

When evaluating *in silico* systems, a number of points should be considered. Among the first yet often not considered point is for what purpose the *in silico* models are used, e.g. the models can be used to obtain scientific and/or mechanistic information or they can be used for regulatory purposes. They can also be used to group chemicals into categories based on mechanistic principles or structural properties. When used for regulatory purposes, it should also be kept in mind that legislation in different countries varies as do the criteria for classification. Other aspects to consider include the availability, ease of use, accuracy of prediction, applicability domain, the quality of the data set used to develop the models, incorporation of metabolic activity, transparency, the quality of the reports generated and how much the model depends on expert knowledge.

In this study, the sensitization potential of 100 substances was studied using seven (Q)SAR models, namely: Case Ultra, TOPKAT, DEREK, VEGA v2.1.3, TIMES-SS v2.27, Toxtree and the OECD (Q)SAR toolbox version 3.1. The (Q)SAR predictions were then compared to the experimental data primarily obtained from animal studies and the classification according to GHS. In a smaller-scale study, ToxWiz was investigated with 34 sensitizing and 19 non-sensitizing substances for which animal, human and in vitro data had recently been published and which was used to establish an in vitro skin sensitization battery (Bauch et al., 2012). The (Q)SAR tools were assessed with respect to predictive capacity, availability and ease of use both for the single models and their combination.

2. Methods

The test set consisted of 45 and 55 substances which were sensitizing and non-sensitizing in animal studies, respectively. An overview of the substances including the study type and the GHS classification is provided in Table 1. For the choice of substances, the following criteria were applied: (i) The substances had to have reliable and adequate experimental data for the purpose of classification and labeling according to UN GHS. (ii) It should not be expected to be present in the training set of the (Q)SAR models. (iii) The molecular weight should be in a range for which skin permeability cannot be excluded, i.e. 97 of the 100 substances had molecular weights ranging from 53 to 500 Da. Three substances with the CAS numbers 693-36-7, 41672-81-5 and 16470-24-9 had higher molecular weights of 683, 608 and 1073 g/mol, respectively. (iv) Substances had to be mono-constituent and to cover a range of industrial chemicals and functional groups. With respect