Toxicology Letters 241 (2016) 32–37

Contents lists available at ScienceDirect

Toxicology Letters

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

Grouping chemicals for health risk assessment: A text mining-based case study of polychlorinated biphenyls (PCBs)

Imran Ali[a,*], Yufan Guo[b], Ilona Silins[a], Johan Högberg[a], Ulla Stenius[a], Anna Korhonen[b]

[a] *Institute of Environmental Medicine, Karolinska Institutet, Stockholm SE-171 77, Sweden*
[b] *Department of Theoretical and Applied Linguistics, University of Cambridge, Cambridge CB3 9DA, UK*

HIGHLIGHTS

- Literature based MOA profiles of PCBs confirms the existing knowledge.
- Modes of action profile for DL-PCBs differs significantly from that of NDL-PCBs.
- Text mining-based CRAB tool could significantly improve the risk assessment process.

ARTICLE INFO

Article history:
Received 4 September 2015
Received in revised form 4 November 2015
Accepted 4 November 2015
Available online 10 November 2015

Keywords:
Polychlorinated biphenyls
Chemical risk assessment
Mode of action
Text-mining
Classification of literature
CRAB 2.0

ABSTRACT

As many chemicals act as carcinogens, chemical health risk assessment is critically important. A notoriously time consuming process, risk assessment could be greatly supported by classifying chemicals with similar toxicological profiles so that they can be assessed in groups rather than individually. We have previously developed a text mining (TM)-based tool that can automatically identify the mode of action (MOA) of a carcinogen based on the scientific evidence in literature, and it can measure the MOA similarity between chemicals on the basis of their literature profiles (Korhonen et al., 2009, 2012). A new version of the tool (2.0) was recently released and here we apply this tool for the first time to investigate and identify meaningful groups of chemicals for risk assessment.

We used published literature on polychlorinated biphenyls (PCBs)—persistent, widely spread toxic organic compounds comprising of 209 different congeners. Although chemically similar, these compounds are heterogeneous in terms of MOA. We show that our TM tool, when applied to 1648 PubMed abstracts, produces a MOA profile for a subgroup of dioxin-like PCBs (DL-PCBs) which differs clearly from that for the rest of PCBs. This suggests that the tool could be used to effectively identify homogenous groups of chemicals and, when integrated in real-life risk assessment, could help and significantly improve the efficiency of the process.

1. Introduction

The need for assessment of human health risks posed by environmental chemicals is growing. Huge efforts are being invested in identification of suspected carcinogens in particular. To establish carcinogenic effects of a chemical (or a mixture of chemicals) in humans, multiple epidemiological studies showing correlations between exposure and health outcomes are needed. These have to be supported by a plausible "mode of action" (MOA) based on experimental studies in various model systems (IARC: http://monographs.iarc.fr/) (Rappaport and Smith, 2010; Borgert et al., 2004).

A MOA refers to a sequence of key events that result in cancer development, capturing the current understanding of different processes leading to carcinogenesis. Identification of a chemical's MOA is a heavily literature-dependent task which could greatly benefit from text mining (TM) support. MOA analysis requires a thorough review of literature available for each chemical under inspection. Since the scientific data used for MOA assessment is highly varied and well-studied chemicals may have tens of thousands of publications, literature review can be extremely time consuming when conducted via conventional means, i.e.,

[*] Corresponding author at: Institute of Environmental Medicine, Karolinska Institutet, Box 210, 171 77 Stockholm, Sweden. Fax: +46 8 34 38 49.
E-mail addresses: imran.ali@ki.se, epa_ali@yahoo.com (I. Ali).

http://dx.doi.org/10.1016/j.toxlet.2015.11.003
0378-4274/© 2015 Elsevier Ireland Ltd. All rights reserved.

typically a keyword-based search via PubMed search interface followed by manual expert judgment (Korhonen et al., 2009).

We have recently introduced and released CRAB 2.0—a powerful, fully-integrated TM-based tool designed to assist the entire process of literature review in real-life cancer risk assessment (Korhonen et al., 2012; Guo et al., 2014). The CRAB tool classifies PubMed literature on a given chemical according to the taxonomy based on currently established carcinogenic MOAs (Korhonen et al., 2009). The distribution of classified literature for individual MOAs referred to as "MOA profile" below have proved highly accurate in intrinsic evaluations and have also been used to confirm known properties of chemicals without human input (Korhonen et al., 2012). However, no study aimed at improving real-life chemical risk assessment has been reported using this new version of the tool yet.

Here we focus on this, and in particular the potential of the tool in enabling simultaneous study of the carcinogenic effects of several cancer causing agents through an extensive analysis of existing PubMed literature. We investigate whether the tool could be used to identify groups of chemicals similar in their MOA. If yes, it could enable more efficient risk assessment in the future.

Polychlorinated biphenyls (PCBs) are man-made products that have been used in technical applications since 1929. Although their production was terminated in many countries during the 1970s, due to the persistent nature and high lipid solubility, the general population is exposed to PCBs mainly via food and to some extent from indoor air (ATSDR, 2000). The toxicity of PCBs is still studied in many laboratories (Fernandez-Gonzalez et al., 2015; Hu et al., 2015; Quinete et al., 2014), including our own (Al-Anati et al., 2010). The literature on PCBs is huge, and the risk assessment is complicated by the fact that they comprise of 209 different congeners with variable toxicity. Some are established or suspected human carcinogens (IARC), while others may have other conspicuous effects and some might be of negligible concern.

PCBs are often divided into two subgroups: dioxin-like (DL-PCBs) and non-dioxin-like (NDL-PCBs). This division is based on the positions of the chlorine atoms, which determine the affinity for and activation of the aryl hydrocarbon receptor (AhR). Activation of AhR is considered the MOA of DL-PCBs and AhR activation is also the MOA of the known human carcinogen dioxin 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD)—hence the term "dioxin-like". In health risk assessment of DL-PCBs (or mixtures of them) a relative toxicity factor (toxic equivalency factor, TEF) is used to compare a DL-PCB with TCDD. The use of TEF values is based on the assumption that DL-PCBs and TCDD act via the same MOA. In the current WHO-TEF concept, TCDD has a value of 1 and most of the DL-PCBs have TEF values varying from 1×10^{-1} to 10^{-5}. NDL-PCBs do not bind AhR, and therefore other MOAs are assumed (Schwarz and Appel 2005; Van den Berg et al., 2006).

In this study we investigated and analyzed the TM-generated MOA profiles of DL-PCBs and NDL-PCBs. Each profile revealed a distinct distribution of the literature over different MOA categories, indicating that CRAB 2.0 can detect the MOA differences at a fine level of detail and thus identify homogenous groups of chemicals. This suggests that the tool has the potential to assist the development of protocols for assessing groups of chemicals, which might lead to improved efficiency of risk assessment.

2. Methods

We used the newly developed CRAB 2.0 tool[1]—to classify PubMed literature of different chemicals according to their carcinogenic MOAs. The tool supports gathering of literature via PubMed query interface, semantic classification according to MOA, and automated statistical analysis of the classified literature.

2.1. Gathering literature

For comparative analysis we collected PubMed literature on a group of DL-PCBs (PCB 126, 77, 81, 169, 105, 114, 118, 123, 156, and 157) with focus on PCB126, a reference chemical TCDD to which the toxicity of DL-PCBs are compared and a group of NDL-PCBs (PCB 52, 74, 101, 118, 122, 128, 138, 153, 170, and 180) with focus on PCB153 (Stenberg et al., 2011). CRAB 2.0 interacts with E-utilities[2]—the PubMed query interface. As shown in CRAB tool interface (Supplementary Fig. 1), a query for a particular chemical (e.g., PCB153) is forwarded to PubMed, and the relevant abstracts resulting from the query are downloaded on the CRAB 2.0 server in XML format.

2.2. Text mining-based MOA analysis of literature

The collected abstracts are automatically classified according to a taxonomy which covers different types of scientific data used for cancer risk assessment (Korhonen et al., 2012). The taxonomy is based on current understanding of the processes leading to cancer and includes two main categories: genotoxic and non-genotoxic MOA, and is further organized into more specific sub-categories according to the classification by Hattis et al. (2009) (Korhonen et al., 2009, 2012). The CRAB tool downloads all PubMed abstracts for a given chemical for automatic analysis of the abstracts according to the evidence mentioned for different carcinogenic MOA sub-categories. Thus based on the literature data and classification pattern, a publication profile is generated (displayed as percent of the total number of MOA abstracts). The tool does not exclude abstracts with no-effect results; however such results are rarely published. A possible exception is data on mutagenicity, an endpoint that might require manual inspection.

In semantic classification of literature, each abstract downloaded from PubMed is turned into a vector of "bags of words" features, whose value equals 1 if the corresponding word/MeSh term is observed in the abstract, and 0 otherwise. Abstracts represented by feature vectors are then assigned to relevant taxonomy class(es) using supervised machine learning: by support vector machines (SVM) with the Jensen–Shannon divergence (JSD) kernel trained in advance on a set of manually classified MOA abstracts (not necessarily focused on any specific chemical). The output of semantic classification is a taxonomy structure, where the number of abstracts assigned to each category is shown alongside the link to the relevant abstracts (Supplementary Fig. 2). Evaluation of the classifier reported in (Korhonen et al., 2012) shows that it is highly accurate at an *F*-score of 0.78. The processing time depends on data size, ranging from a few minutes to a few hours (memory: 5,859,372 kB, CPU: Quad-Core AMD Opteron(tm) Processor 2347 HE).

2.3. Statistical analysis of classified literature

In evaluation of the first version of CRAB (Korhonen et al., 2012), post-hoc statistical analysis of the classifier output (e.g., calculating and visualizing the distribution of abstracts over taxonomy classes) proved highly useful for obtaining a broad overview of the data in literature and identifying the data gaps. CRAB 2.0 allows viewing statistics of classified literature with a single click (Supplementary Fig. 3). The system interacts with R[3]—a free

[1] http://omotesando-e.cl.cam.ac.uk/CRAB/request.html.

[2] http://www.ncbi.nlm.nih.gov/books/NBK25501/.

[3] http://www.r-project.org/.