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.

© 2015 Elsevier Ireland Ltd. All rights reserved.

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/.

software environment for statistical computing and graphics. The number (or percentage) of abstracts assigned to each MOA category is passed to R to generate a MOA profile of the chemical under inspection. The profile is a .zip file including bar plots in pdf format showing the distribution of abstracts, as well as raw data (the number of abstracts under each category) in txt format that could be easily imported into popular software for statistical analysis (e.g., SPSS or excel) if necessary.

The MOA profiles for different (groups of) chemicals are compared, and the statistical significance of differences is tested using chi-squared homogeneity test for each individual MOA category (positive vs, negative), and for each pair of chemicals (using a 2×2 contingency table). This is better preferred than a $2 \times N$ contingency table encapsulating all MOAs, since a single abstract maybe relevant for multiple MOAs. The individual p-values are then adjusted by a Bonferroni correction for the entire profile's p-values.

3. Results

3.1. MOA profiles for DL-PCBs and NDL-PCBs

We compared the MOA profiles for two groups of PCBs: DL-PCBs and NDL-PCBs (Fig. 1). CRAB 2.0 evaluated 1648 PubMed abstracts for a group of ten DL-PCBs (mentioned in Section 2; 788 abstracts) and ten well-studied and abundant NDL-PCBs (mentioned in Section 2; 860 abstracts) (PubMed, March, 2015). Around 40% of these were identified as relevant for MOA classification. The distribution of literature over MOA taxonomy shows statistically significant differences ($p < 0.01$) in the overall profile for the whole group of DL-PCBs as compared to the whole group of NDL-PCBs (Fig. 1). A remarkable difference is seen in the proportions of "*AhR activation*"-related literature for DL- and NDL-PCBs, and as indicated in Section 1, this difference is expected. Other differences are seen in the MOA categories of "*oxidative stress*", "*cell proliferation*", "*cell death*" and "*hormonal receptor*". This suggests that there are systematic differences in literature covering these two groups of congeners, and that the tool has the capacity to detect these differences.

3.2. MOA profiles for TCDD, PCB126 and PCB153

PCB126 is a common and well-studied DL-PCB and PCB153 is equally common and well-studied NDL-PCB. These compounds are often referred to as indicator congeners for respective groups. We used CRAB 2.0 to analyze 538 abstracts for PCB126 and 570 for PCB153 (PubMed, March, 2015) and created a MOA profile for each congener. The profiles were based on 315 abstracts concerning PCB126 and 224 abstracts concerning PCB153 selected by the tool as relevant for MOA classification. When comparing these profiles, we find significant differences ($p < 0.001$) between them (Fig. 2). The differences between PCB126 and PCB153 are mainly reflected in the individual MOAs "*AhR activation*", "*cell death*", and "*strand breaks*", and are thus slightly different from the results shown in Fig. 1. We also analyzed the MOA profile for TCDD, using 8572 abstracts, and compared it to the profiles of the two PCBs. The comparison of PCB126 and TCDD indicates differences regarding "*AhR activation*" and "*oxidative stress*" ($p < 0.001$) while that of PCB153 and TCDD shows differences in "*AhR activation*", "*cell death*", "*oxidative stress*" and "*cytotoxicity*" ($p < 0.001$). In further analysis we compared TCDD with the DL-PCBs and found similar differences as between TCDD and PCB126 (Fig. 3).

3.3. Testing group profiles by comparing them with indicator PCBs

To test whether the two indicator PCBs are representative for the two groups we compared their MOA profiles statistically. The MOA profile of the group DL-PCBs does not exhibit significant differences when compared with that of PCB126 (Fig. 4). Similarly, the profile for the group NDL-PCBs does not significantly differ from that of PCB153 (Fig. 5). However, the reversed comparison (DL-PCBs group MOA profile compared with PCB153, and the NDL-PCBs group MOA profile compared with PCB126) does show significant differences (data not shown).

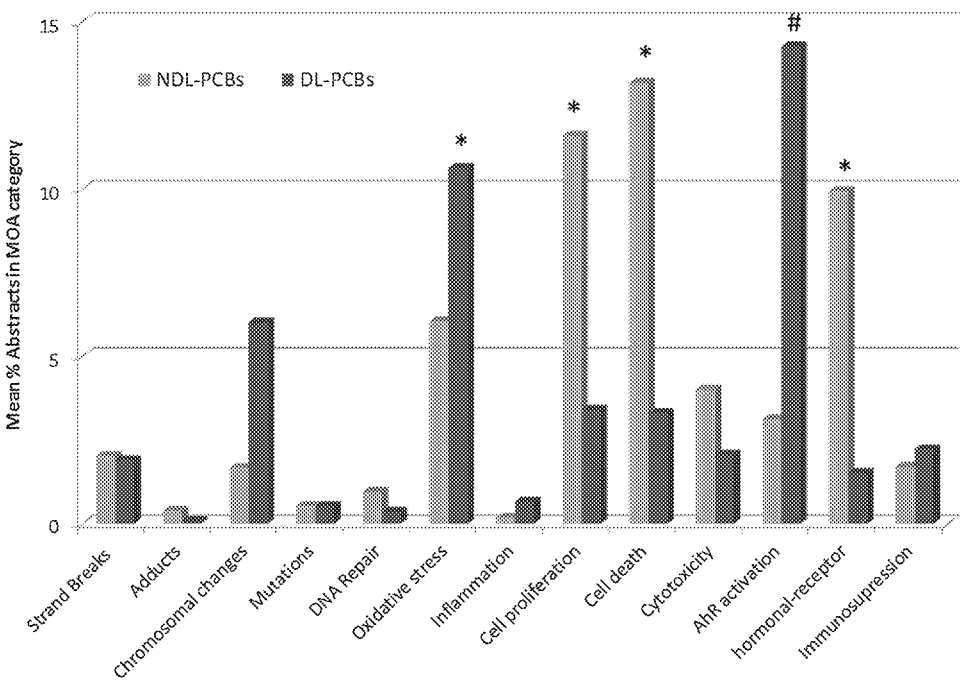

Fig. 1. Literature profiles for DL-PCBs and NDL-PCBs. The group profile for DL-PCBs differs significantly from the group profile for DL-PCBs ($p < 0.01$). (*) denotes ($p < 0.01$), (#) denotes ($p < 0.001$).

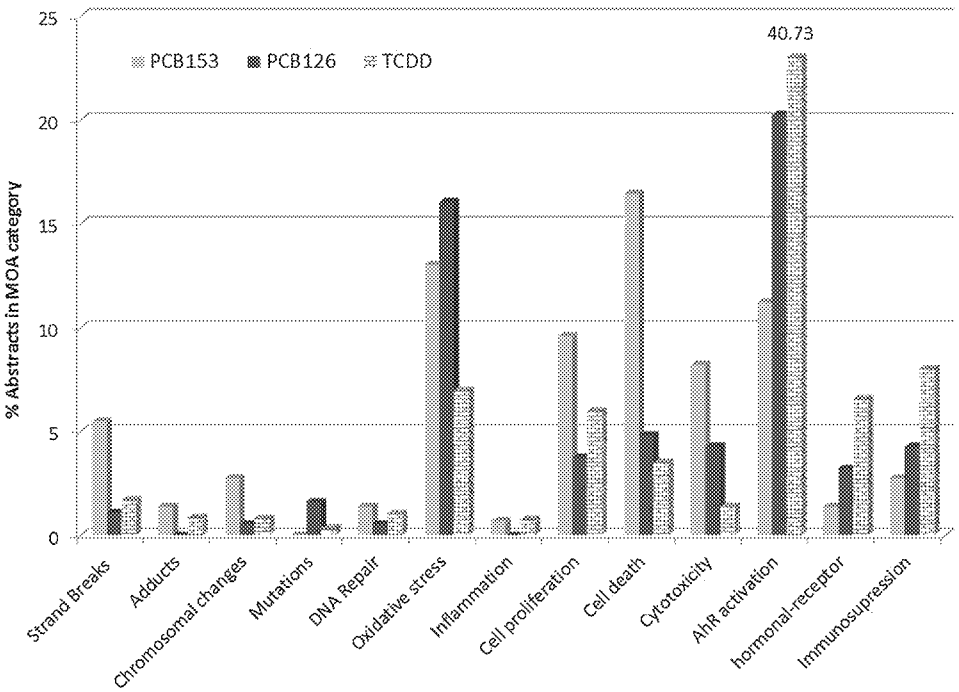

Fig. 2. Literature profiles for PCB126 and PCB153. The profiles are significantly different ($p < 0.01$). The column for TCDD and "*AhR activation*" is truncated, and figure gives% of abstracts.

4. Discussion and Conclusion

Automatic text mining-tool can be used to identify the carcinogenic MOAs for a group of chemicals. By using the recently developed TM tool CRAB 2.0 we have analyzed thousands of PubMed abstracts for comparing the MOA properties of PCBs. We selected the environmental contaminants PCBs for this case study as they constitute 209 similar but chemically distinct congeners, and there is a well-recognized subgroup, DL-PCBs, with a specified MOA. We investigated whether CRAB 2.0 is able to differentiate DL-PCBs from NDL-PCBs in terms of their MOA profiles according to semantically classified scientific literature. We showed that this can be done at a high level of confidence and fine level of detail. This is a highly promising finding, suggesting that our methodology can not only be used to accelerate current risk assessment, as shown by previous studies (Korhonen et al., 2012; Silins et al.,

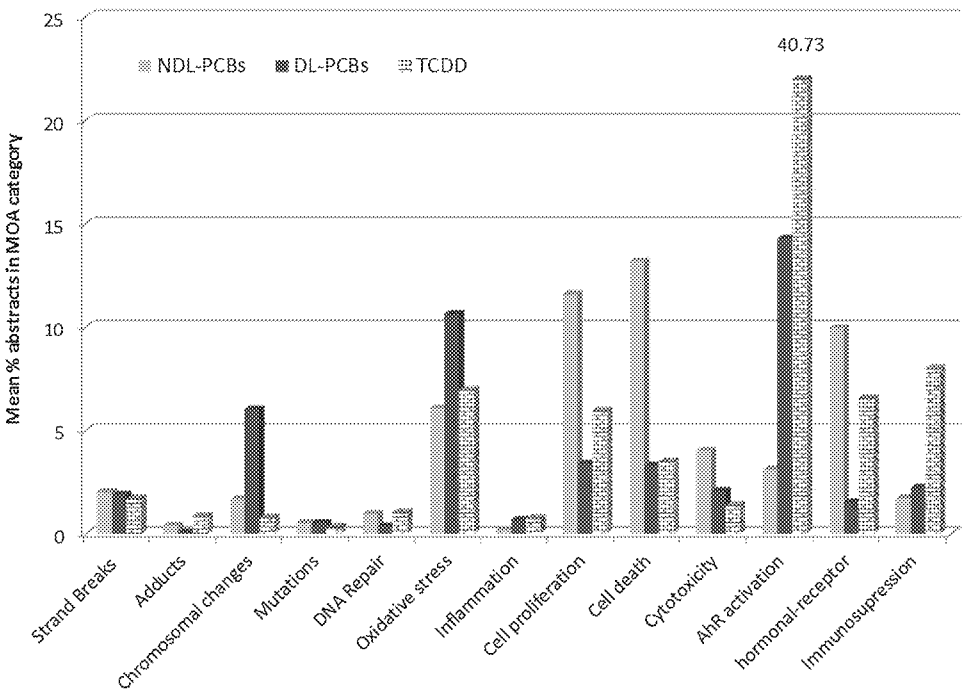

Fig. 3. Literature profiles differ significantly for DL-PCBs, NDL-PCBs and TCDD. The column for TCDD and "*AhR activation*" is truncated, and figure gives% of abstracts.

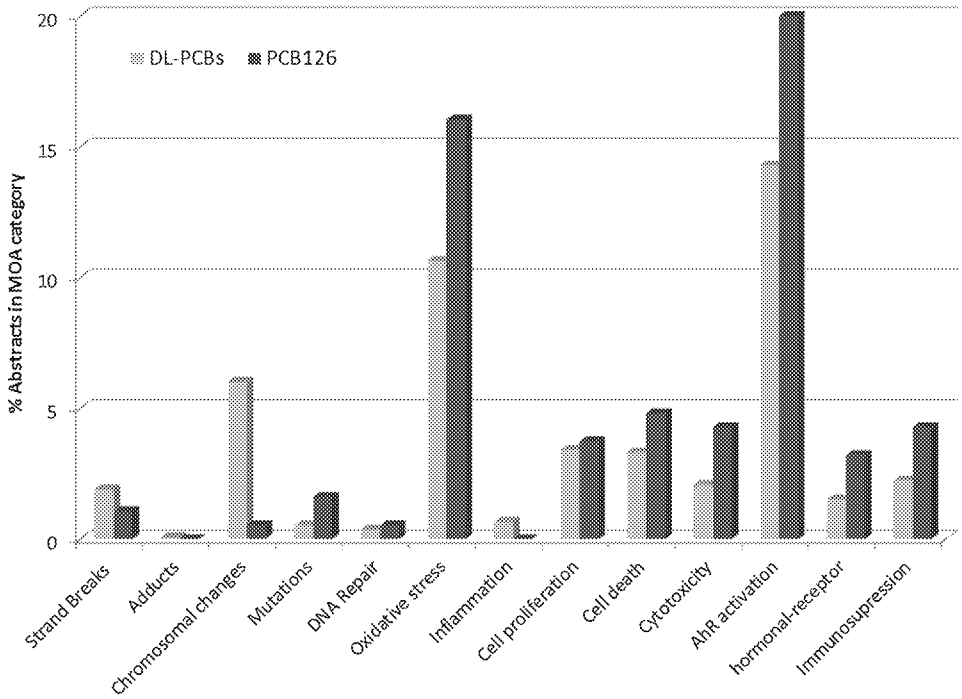

Fig. 4. The indicator compound PCB126 profile is not significantly different from the group profile for DL-PCBs.

2014), but also provide the means for improved, more efficient risk assessment that involves assessing chemicals in semantically meaningful groups.

The time saving aspect is indicated by the large number of abstracts that were analyzed by the tool in few minutes, and the fact that 60% were excluded from analysis as irrelevant. Performed via conventional means, an expert toxicologist may need weeks to complete equivalent analysis. CRAB 2.0 could greatly support expert assessment as it allows assessors to focus on specific properties or data types of chemicals. Although literature profiles do not alone constitute adequate evidence for risk assessment, they can provide a thorough overview of available scientific literature and an excellent starting point for manual checking of details. In order to do complete hazard identification for risk assessment, the classified literature needs to be examined more in details.

Besides allowing for grouping chemicals, the tool might also be used for identifying data gaps (i.e., less well-studied areas) for a group of chemicals. It could also provide material for hypothesis generation. For example, it could be used to analyze why DL-PCBs

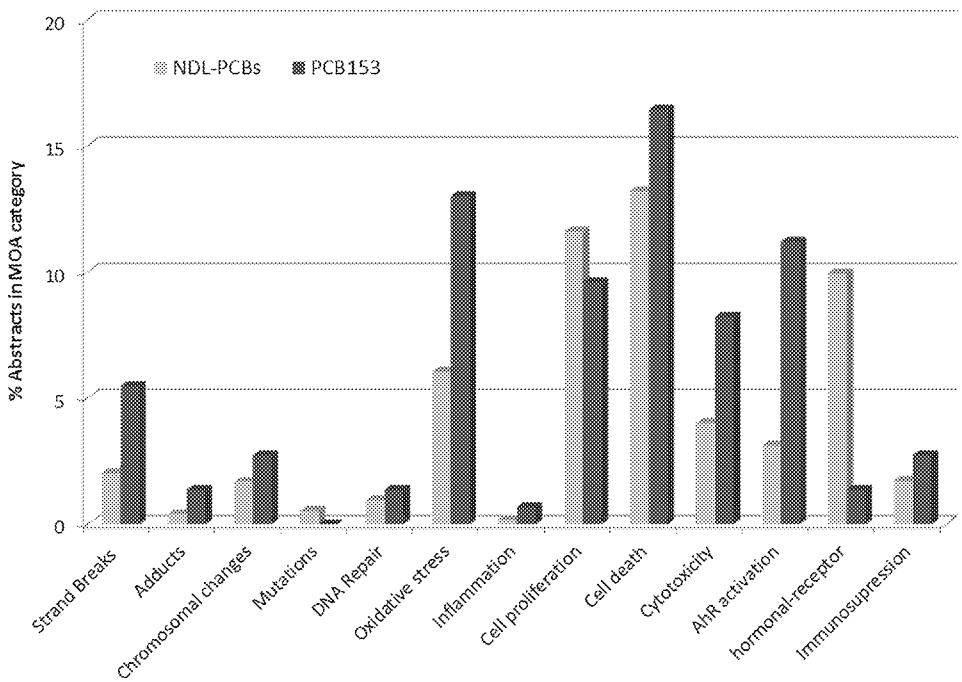

Fig. 5. The indicator compound PCB153 profile is not significantly different from the group profile for NDL-PCBs.

differ from TCDD, and reading the relevant papers may guide risk assessors to important questions about the differences. Or it could be used to shed light on the currently unexplained facts (e.g., why DL-PCBs have a higher percentage of abstracts than NDL-PCBs in the "*chromosomal changes*" and the "*inflammation*" categories). Such facts can be highlighted by the tool and may lead to hypothesis generation. In addition, besides PCBs, there are many other groups of chemicals (e.g., polyaromtic hydrocarbons, endocrine disrupters, flame retardants etc.) that might benefit from risk assessment within chemical groups identified by the tool.

Moreover, in the future the tool could be developed further in various ways, e.g., to take into account journal impact factors, number of citations, and cross references. This may help identifying critical articles and organising the literature for overviewing available information. Although currently focused on cancer, the CRAB tool can be easily adapted to other health risks, provided with a well-defined taxonomy and accordingly, examples of classified literature for machine learning. It might also be of interest to develop TM-based strategies for optimizing subdivisions of chemical groups with many publications.

Acknowledgments

We would like to thank the Royal Society (UK), and Vinnova (Sweden) for financial support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.toxlet.2015.11.003.

References

Al-Anati, L., Hogberg, J., Stenius, U., 2010. Non-dioxin-like PCBs interact with benzo [a]pyrene-induced p53-responses and inhibit apoptosis. Toxicol. Appl. Pharmacol. 249, 166–177.

ATSDR, 2000. Toxicological Profile for Polychlorinated Biphenyls (PCBs). U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES, Agency for Toxic Substances and Disease Registry (ATSDR), Division of Toxicology/Toxicology Information Branch, Atlanta, Georgia-USA.

Borgert, C.J., Quill, T.F., McCarty, L.S., Mason, A.M., 2004. Can mode of action predict mixture toxicity for risk assessment? Toxicol. Appl. Pharmacol. 201, 85–96.

Fernandez-Gonzalez, R., Yebra-Pimentel, I., Martinez-Carballo, E., Simal-Gandara, J., 2015. A critical review about human exposure to polychlorinated dibenzo-p-dioxins (PCDDs), polychlorinated dibenzofurans (PCDFs) and polychlorinated biphenyls (PCBs) through foods. Crit. Rev. Food Sci. 55, 1590–1617.

Guo, Y., Seaghdha, D.O., Silins, I., Sun, L., Hogberg, J., Stenius, U., Korhonen, A., 2014. CRAB 2.0: A text mining tool for supporting literature review in chemical cancer risk assessment. Proc. COLING 2014 25th Int. Conf. Comput. Linguist. Syst. Demonstr., Dublin, Ireland, pp. 76–80.

Hattis, D., Chu, M., Rahmoglu, N., Goble, R., Verma, P., Hartman, K., et al., 2009. A preliminary operational classification system for nonmutagenic modes of action for carcinogenesis. Crit. Rev. Toxicol. 39, 97–138.

Hu, X., Adamcakova-Dodd, A., Lehmler, H.J., Gibson-Corley, K., Thorne, P.S., 2015. Toxicity evaluation of exposure to an atmospheric mixture of polychlorinated biphenyls by nose-only and whole-body inhalation regimens. Environ. Sci. Technol. 49, 11875–11883.

Korhonen, A., Silins, I., Sun, L., Stenius, U., 2009. The first step in the development of text mining technology for cancer risk assessment: identifying and organizing scientific evidence in risk assessment literature. BMC Bioinform. 10, 303.

Korhonen, A., Seaghdha, D.O., Silins, I., Sun, L., Hogberg, J., Stenius, U., 2012. Text mining for literature review and knowledge discovery in cancer risk assessment and research. PLoS One 7, e33427.

Quinete, N., Schettgen, T., Bertram, J., Kraus, T., 2014. Occurrence and distribution of PCB metabolites in blood and their potential health effects in humans: a review. Environ. Sci. Pollut. Res. 21, 11951–11972.

Rappaport, S.M., Smith, M.T., 2010. Environment and disease risks. Science 330, 460–461.

Schwarz, M., Appel, K.E., 2005. Carcinogenic risks of dioxin: mechanistic considerations. Regul. Toxicol. Pharm. 43, 19–34.

Silins, I., Korhonen, A., Stenius, U., 2014. Evaluation of carcinogenic modes of action for pesticides in fruit on the Swedish market using a text-mining tool. Front. Pharmacol. 5, 145.

Stenberg, M., Hamers, T., Machala, M., Fonnum, F., Stenius, U., Lauy, A.A., et al., 2011. Multivariate toxicity profiles and QSAR modeling of non-dioxin-like PCBs—an investigation of in vitro screening data from ultra-pure congeners. Chemosphere 85, 1423–1429.

Van den Berg, M., Birnbaum, L.S., Denison, M., De Vito, M., Farland, W., Feeley, M., et al., 2006. The 2005 world health organization reevaluation of human and mammalian toxic equivalency factors for dioxins and dioxin-like compounds. Toxicol. Sci. 93, 223–241.