OXFORD

Data and text mining

AOP-helpFinder webserver: a tool for comprehensive analysis of the literature to support adverse outcome pathways development

Florence Jornod[1], Thomas Jaylet[1], Ludek Blaha[2], Denis Sarigiannis[3], Luc Tamisier[4] and Karine Audouze (ID) **[1,*]**

[1]Université de Paris, T3S, Inserm UMR-S1124, Paris F-75006, France, [2]RECETOX, Faculty of Science, Masaryk University, Brno CZ62500, Czech Republic, [3]HERACLES Research Center on the Exposome and Health, Aristotle University of Thessaloniki, Center for Interdiciplinary Research and Innovation, Thessaloniki 57001, Greece and [4]Université de Paris, SPPIN CNRS UMR 8003,Paris F-75006, France

*To whom correspondence should be addressed.
Associate Editor: Jonathan Wren

Abstract

Motivation: Adverse outcome pathways (AOPs) are a conceptual framework developed to support the use of alternative toxicology approaches in the risk assessment. AOPs are structured linear organizations of existing knowledge illustrating causal pathways from the initial molecular perturbation triggered by various stressors, through key events (KEs) at different levels of biology, to the ultimate health or ecotoxicological adverse outcome.

Results: Artificial intelligence can be used to systematically explore available toxicological data that can be parsed in the scientific literature. Recently, a tool called AOP-helpFinder was developed to identify associations between stressors and KEs supporting thus documentation of AOPs. To facilitate the utilization of this advanced bioinformatics tool by the scientific and the regulatory community, a webserver was created. The proposed AOP-helpFinder webserver uses better performing version of the tool which reduces the need for manual curation of the obtained results. As an example, the server was successfully applied to explore relationships of a set of endocrine disruptors with metabolic-related events. The AOP-helpFinder webserver assists in a rapid evaluation of existing knowledge stored in the PubMed database, a global resource of scientific information, to build AOPs and Adverse Outcome Networks supporting the chemical risk assessment.

Availability and implementation: AOP-helpFinder is available at http://aop-helpfinder.u-paris-sciences.fr/index.php

Contact: karine.audouze@u-paris.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Structured organization of toxicological and ecotoxicological data is now feasible using the adverse outcome pathways (AOP) framework (Ankley *et al.*, 2010). An AOP is defined by a linear combination of biological events, started from a molecular initiating event (MIE) triggered by stressors (pollutants, ionizing radiations, nanomaterials or climate stressors) connected through a series of key events (KEs) occurring at various levels of the biological organization, to an adverse outcome (AO). Biological events (MIE, KE and AO) are not linked to a unique AOP, but can be shared, allowing the establishment of Adverse Outcome Network (AON) that reflect better the true complexity of the biology. Combined with new approach methodologies (Parish *et al.*, 2020), AOPs and AONs are extremely

useful in establishing integrated approaches to testing and assessment (IATA) for environmental and risk assessment, and they aid to the development of novel nonanimal toxicity testing strategies (Delrue *et al.*, 2016).

With advances in technologies, huge amounts of data have become available, compiled in well-structured toxicological databases (e.g. CTD, CompTox), in AOP-oriented webservers (AOP-wiki, sAOP, AOP4EUpest) and scientific publications (Williams *et al.*, 2017). Innovative data mining tools are needed to identify sparse but complementary data such as Abstract Sifter allowing to have a view of the toxicological information landscape for a set of entities as chemicals (Baker *et al.*, 2017) or ComptoxAI (https://comptox.ai/index.html). Artificial intelligence (AI) technology, that uses natural language processing (NLP), is an interesting way to facilitate the

identification of links between relevant information that can be used to build novel AOPs (Song *et al.*, 2020), and identify knowledge gaps and research needs (Zgheib *et al.*, 2021). Several tools use text mining (TM), an AI method to transform unstructured into structured text. For example, Limtox provides a biomedical search for adverse hepatobiliary reactions (Cañada *et al.*, 2017). Recently, the AOP-helpFinder tool, based on TM and graph theory was proposed to identify stressor-KE relationships by examining large collections of scientific abstracts, and was applied to bisphenol A substituents and pesticides (Carvaillo *et al.*, 2019; Jornod *et al.*, 2020; Rugard *et al.*, 2020).

Here, we present the AOP-helpFinder webserver, which uses an updated version of the tool, to provide an easy but effective resource for identifying and compiling existing knowledge from the scientific literature. The main optimized features are (i) the capability to choose to search in full abstracts or without considering the introductory parts, (ii) the possibility to perform a refined search using machine learning and (iii) an automatic update of the PubMed database before each search. A case study with endocrine disruptors (ED) and metabolism-related events is provided, illustrating the capacity of the tool to collate quickly an overview of the existing information.

2 Materials and methods

2.1 The AOP-helpFinder webserver
The proposed webserver is easy to use, and requires only the user email to access the upload page and to receive information when the results are available for download. This simple procedure is in line with digital sobriety that aims to reduce the environmental impact by limiting computing use. Two input files are needed: one with the stressors of interest and the second with biological events (i.e. MIE, KE and/or AO). Before running the tool to identify if knowledge connecting stressors and biological events exists, the user can choose between two options: reduce search and refinement filter (see the following section and Supplementary Material), as well as the output format (date, title, PMID, etc.).

2.2 The AOP-helpFinder tool
To increase the performance of the previously developed version, several methods were tested using a set on ED and biological events related to metabolism, and two were kept through the process (see Supplementary Material, https://github.com/jornod/aop-helpFinder):(i) 'reduced search': searches are performed in the full abstracts or without considering the introductory part, which appears to be covered usually by the first 20% of the abstracts. This option allows avoiding too many false positives, as the introduction often reflects a working hypothesis instead of the conclusions of the publication and (ii) 'refinement filter': after the preprocessing that uses a stemming process (Carvaillo *et al.*, 2019), the tool can refine the searches by combining a deletion of sentences containing context words with a lemmatization process. Lemmatization is a machine learning method for text normalization used in NLP that considers the context and converts the word to its meaningful base form. This option is very useful when terms have common stems (e.g. tests, testis □ test) leading to incorrect meanings and spelling errors. Further, an automatic daily update of the PubMed database was newly implemented using the NCBI API to screen the full existing knowledge.

The current version of the AI tool mined the PubMed database, that is a global source for scientific literature. Nevertheless, the developed method screens text-based knowledge, and therefore the AOP-helpFinder server could be improved for mining multiple sources (databases, literature), including studies reporting negative findings, to accelerate information gathering when data are limited and present in diverse sources (Carvaillo *et al.*, 2019).

The advantage of the proposed method, is its capacity to be adapted for literature searches in general, independent of AOP development, in order to identify interconnections between the query
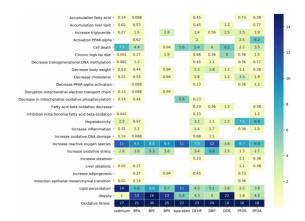

Fig. 1. Example of ED comentioned in PubMed scientific abstracts with biological events related to metabolism, identified by the AOP-helpFinder webserver. The numbers correspond to the % of retrieved abstracts mentioning both the stressor (column) and the event (line) among all identified abstracts (the colors are according to the percentage for better visualization). For example, among all identified abstracts that comentioned bisphenol S (BPS) and at least one event from the list, 13% of the abstracts were comentioning BPS (fourth column) and obesity (the second line from the bottom)

Event	cadmium	BPA	BPF	BPS	bparaben	DEHP	DBP	DDE	PFOS	PFOA
Accumulation fatty acid	0.14	0.088				0.45			0.73	0.38
Accumulation liver lipid	0.02	0.53				0.45				0.77
Increase triglyceride	0.27	1.6	2.8			1.4	0.56	2.5	2.5	1.9
Activation PPAR-alpha		0.62				2			2.5	6.2
Cell death	7.5	4.4		0.94	5.6	5.4	4	6.2	2.2	3.5
Chronic high fat diet	0.061	0.27	1.9			0.68	0.56	5	0.36	1.5
Decrease transgenerational DNA methylation	0.082	1.2				0.45	1.1		0.36	0.77
Decrease body weight	0.53	0.44		0.94		2.3	2.8	1.2	1.1	0.38
Decrease cholesterol	0.25	0.53		0.94		1.8		1.2	3.3	1.9
Decrease PPAR-alpha activation		0.088				0.23			0.36	1.2
Disruption mitochondrial electron transport chain	0.14	0.088		0.94						
Decrease in mitochondrial oxidative phosphorylation	0.14	0.44			5.6	0.23				
Fatty acid beta oxidation decrease						0.23	0.56	1.2		0.38
Inhibition mitochondrial fatty acid beta oxidation	0.041					0.23				1.2
Hepatotoxicity	2.5	0.97				3.2	1.1	2.5	7.6	6.9
Increase inflammation	0.31	1.2				1.4	1.7		0.36	1.5
Increase oxidative DNA damage	0.14	0.088				0.68	1.1			
Increase reactive oxygen species	11	9.5	8.8	9.4	11	7.9	12	3.8	8.7	6.9
Increase oxidative stress	2.8	3.8	5.3	3.8		3.4	6.8	2.5	1.5	2.7
Increase steatosis						0.23			1.1	0.38
Liver steatosis	0.02	0.27		0.94		0.45			1.1	0.38
Increase adipogenesis		0.27				0.45			0.73	
Induction epithelial mesenchymal transition	0.02	0.18							0.36	
Lipid peroxidation	14	6.8	8.8	5.7	11	4.5	5.1	3.8	2.5	3.8
Obesity	1	13	16	13	5.6	4.3	4	22	1.8	4.2
Oxidative Stress	27	21	30	25	22	23	24	16	16	18

keywords, as it was successfully done to decipher nonvalidated test methods for ED (Zgheib *et al.*, 2021).

2.3 Case study on endocrine disruptors and metabolism
The AOP-helpFinder webserver was used for a case study aiming at automatically identifying existing relationships and knowledge gaps between 10 ED (Supplementary Table S1) and 294 biological events related to metabolism (Supplementary Table S2). The webserver was launched using 'reduced search' (omitting searches in the first 20% of the abstracts) and 'refinement filter'. Among the 83 970 abstracts retrieved in the PubMed database as of May 10, 2021 related to at least one ED (Supplementary Table S1), a total of 4622 were retained (comentioning ED and event). Among the 294 events, 108 were identified as comentioned with at least one ED (Supplementary Table S2). Figure 1 illustrates the large disparity of knowledge for the 10 selected ED in the area of metabolism (see Supplementary Fig. S1 for all results). For example, cadmium, bisphenol A and di(2-ethylhexyl) phthalate (DEHP) are well studied chemicals as the webserver retrieved scientific articles for almost all biological events of interest. Other chemicals (bisphenol F, bisphenol S, butyl-paraben) appear to be less studied (Supplementary Table S1), and the information were essentially identified for extensively studied biological events such as oxidative stress or obesity.

3 Conclusion
The AOP-helpFinder webserver uses an automatic AI screening to rapidly retrieve existing knowledge on links between stressors and biological events to build AOPs and AONs. This webserver allows highly effective searches in PubMed as it considerably reduces the time of finding relevant scientific articles. The comprehensive AI-based analyses of existing literature support various needs of the risk assessment such as establishment of causality between chemicals and AOs through AOPs and AONs, identification of gaps or prioritization and design of future experimental and epidemiological studies.

Funding

References

Ankley,G.T. *et al.* (2010) Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. *Environ. Toxicol. Chem.*, **29**, 730–741.

Baker,N. *et al.* (2017) Abstract Sifter: a comprehensive front-end system to PubMed. *F1000Research*, **6**, 2164.

Cañada,A. *et al.* (2017) LimTox: a web tool for applied text mining of adverse event and toxicity associations of compounds, drugs and genes. *Nucleic Acids Res.*, **45**, W484–W489.

Carvaillo,J.-C. *et al.* (2019) Linking bisphenol S to adverse outcome pathways using a combined text mining and systems biology approach. *Environ. Health Perspect.*, **127**, 47005.

Delrue,N. *et al.* (2016) The adverse outcome pathway concept: a basis for developing regulatory decision-making tools. *Altern. Lab. Anim.*, **44**, 417–429.

Jornod,F. *et al.* (2020) AOP4EUpest: mapping of pesticides in Adverse Outcome Pathways using a text mining tool. *Bioinformatics*, **36**, 4379–4381.

Parish,S.T. *et al.* (2020) An evaluation framework for new approach methodologies (NAMs) for human health safety assessment. *Regul. Toxicol. Pharmacol.*, **112**, 104592.

Rugard,M. *et al.* (2020) Deciphering adverse outcome pathway network linked to bisphenol F using text mining and systems toxicology approaches. *Toxicol. Sci.*, **173**, 32–40.

Song,J. *et al.* (2020) Upregulation of angiotensin converting enzyme 2 by shear stress reduced inflammation and proliferation in vascular endothelial cells. *Biochem. Biophys. Res. Commun.*, **525**, 812–818.

Williams,A.J. *et al.* (2017) The CompTox chemistry dashboard: a community data resource for environmental chemistry. *J. Cheminform.*, **9**, 61.

Zgheib,E. *et al.* (2021) Identification of non-validated endocrine disrupting chemical characterization methods by screening of the literature using artificial intelligence and by database exploration. *Environ. Int.*, **154**, 106574.