Improved Cancer Risk Assessment Using Text Mining

2448

Ilona Silins[1], Anna Korhonen[2], Johan Högberg[1], Lin Sun[2] and Ulla Stenius[1]

[1]Institute of Environmental Medicine, Karolinska Institutet, Sweden [2]Computer Laboratory, University of Cambridge, UK

Cancer Risk Assessment examines existing scientific evidence to determine the relationship between exposure to a chemical and the likelihood of developing cancer from that exposure.

Text mining (TM) is a growing field of computer science. It involves the discovery (by computer) of new, previously ”unknown” information, by automatically extracting information from different written resources. It has become increasingly popular due to the need to provide access to the tremendous body of texts available in biomedical sciences.

Aim:

To develop a text mining tool that can aid risk assessors manage the existing literature to improve quality, consistency and effectiveness of the entire cancer risk assessment process.

Why?

Extensive data gathering and evaluation for cancer risk assessment is challenging and time-consuming!

- Data are scattered across thousands of journal articles from different areas of biomedicine.
- Rapid development of molecular biology techniques.
- Increased knowledge of mechanisms involved in cancer development.
- Exponentially growing volume of literature.

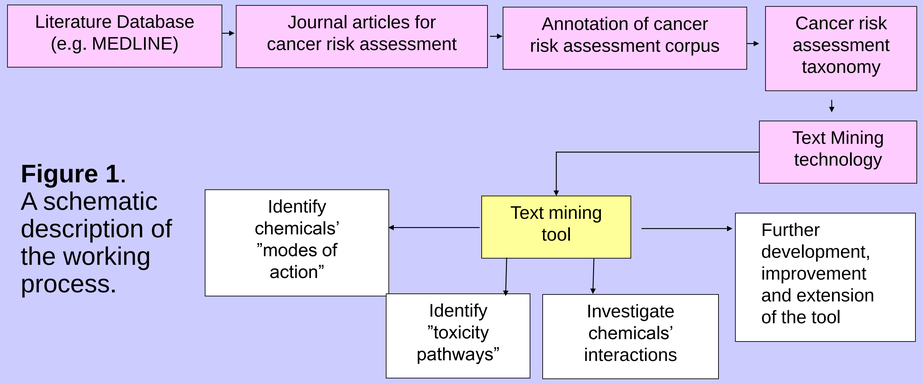


Figure 1. A schematic description of the working process.

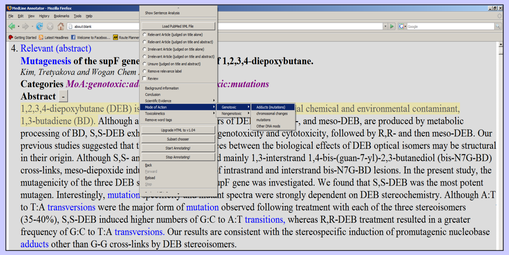


Figure 2. A screenshot of the annotation tool (used for the development of taxonomies) and an annotated abstract.

Test-chemicals used for annotation: Diethylnitrosamine, 1,3-Butadiene, Benzo(a)pyrene, Styrene, Diethylstilbestrol, Chloroform, Phenobarbital, Fumonisin B1. In total1297 Medline abstracts.

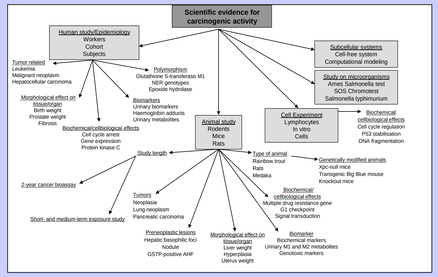
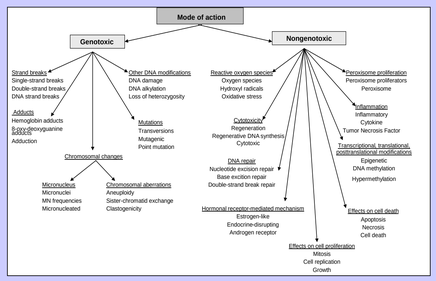


Figure 3. A preliminary taxonomy ”Scientific evidence for carcinogenic activity”, with example keywords.

Figure 4. A preliminary ”Mode of Action” taxonomy (with example keywords) which specifies some of the scientific evidence needed for cancer risk assessment.



USER TEST

Table 1. Results from user test. A user test examine the practical usefulness of the classification in a near real-world scenario. Experts were asked to judge if the system classified the relevant abstracts correctly in the taxonomies (as shown in figure 3 and 4).

Name	Mode of Action	Σ	Precision
Aflatoxin B1	Genotoxic	189	0.95
Benzene	Genotoxic	461	0.99
PCBs	Non-genotoxic	761	0.89
Tamoxifen	Non-genotoxic	382	0.96
TCDD	Non-genotoxic	641	0.96

Future goals and applications

- Further improvements, extentions and development of our tool to aid risk assessors managing the existing literature
- Identify chemicals' ”Modes of Action”
- Identify cellular toxicity pathways that can be used to develop new *in vitro* tests for carcinogenic substances
- Identify chemicals' interactions for risk assessment purposes

Contact:
Ilona Silins, Institute of Environmental Medicine, Karolinska Institutet
Nobels väg 13, 171 77 Stockholm, Sweden
Email: Ilona.Silins@ki.se
Project's homepage: CRAB
http://www.cl.cam.ac.uk/~alk23/crab/crab.html