P9 Implementation of a Flexible Tool for Automated Literature-Mining and Knowledgebase Development (DevToxMine™)

Singh* AV and Knudsen TB

National Center for Computational Toxicology, Office of Research and Development, United States Environmental Protection Agency and Lockheed Martin*, Research Triangle Park, NC.

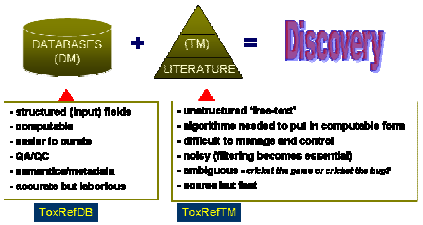
1. ABSTRACT

Deriving novel relationships from the scientific literature is an important adjunct to data-mining activities for complex datasets in genomics and high-throughput screening activities. Automated text-mining algorithms can be used to extract relevant content from the literature and build a thesaurus to convert word relations into concepts. Concept-mining has become an essential knowledge discovery tool to address causal links, associations, relationships, and patterns among vast collections of information. EPA's ToxRefDB database has been built from source data derived from 30-years worth of in vivo animal toxicity studies, mostly rat and rabbit studies. This database includes 751 prenatal developmental toxicity studies on 387 chemicals. For example, large-scale profiling of environmental chemicals for developmental effects with ToxRefDB revealed a species dimorphism of renal-ureteric defects expressed in the rat over rabbit, and a strong correlation between fetal weight reduction and defects of the axial skeleton. In this study, we applied custom text-mining tools to extract the underlying concepts from PubMed. Automated queries were built as <keyword> and <mouse or rat or zebrafish or human> and <embryo or fetal> strings with perl.script to fetch and store facts and information in a MySQL database. Keywords included the ToxCast_320 chemicals and 988 features from an enhanced thesaurus of developmental effects (www.DevTox.org). The raw search returned 186K PubMed abstracts for 82% of the chemicals. Filtering by <keyword> and <species> and <embryo or fetal> narrowed this to 9K abstracts covering 48% of the chemicals. A computational filter applied to find co-occurrences of chemicals, developmental endpoints, and chemical-endpoint linkages returned 4 distinct chemicals and 14 effects at 10 abstracts cutoff value. Although linkages found with the exploratory text-mining tool were conceptualized from relationships mined from ToxRefDB, they included new relationships beyond the ToxRefDB database. This flexible text-mining tool (DevToxMine™), combined with ontology for embryogenesis, is being used to build a knowledgebase (KB) for EPA's Virtual Embryo project.

2. TOOLS AND LIBRARIES

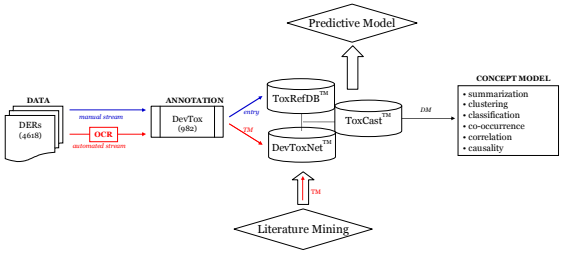
Programming: Perl dynamic programming language was used to develop scripts for different search algorithms. A searchable repository was developed for Data Evaluation Records (DERs), which are agency reviews of registrant submitted toxicity studies (4,618 source DERs of which 1,318 DERs were indexed as 'prenatal developmental toxicity study'), using Optical Character Recognition (OCR). MySQL relational database management system (RDBMS) software was used to develop the database. Gene expressed in the mouse embryo were retrieved from the GXD database (http://genex.hgu.mrc.ac.uk/).

3. APPROACH



DM+TM together can be used to analyze patterns in the information by document summarization and clustering techniques

4. DevToxMine™ Schema



DevToxMine™ is a KB being implemented for EPA's Virtual Embryo to store information such as data, rules and classifications; organize this information into structured concept models (ontologies); and facilitate retrieval of relevant information based on hypothesis-driven queries. The queries may be human readable (manual) or machine-readable (automated).

5. DATA OVERVIEW

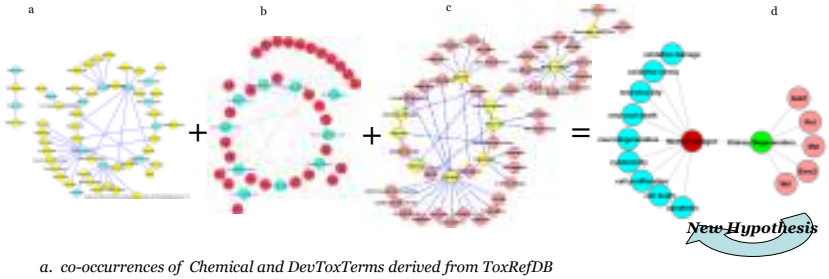
Developmental effects distribution across ToxRefDB dose groups

Results from PubMed automated query

Keyword List	No. of Keywords	No. of Articles	Distinct Articles
Chemicals	48	190	181
Defects	18	1048	979
Genes	120	6705	5303

309 ToxCast™ chemicals x 988 features x Genes in the GXD database based on embryonic expression.

6. EXAMPLE of LITERATURE-BASED DISCOVERY



a. co-occurrences of Chemical and DevToxTerms derived from ToxRefDB
b. co-occurrences of DevToxTerms and Genes (GDX) derived from PubMed
c. co-occurrences of Chemicals and Features derived from Text Mining
d. inference when information from a, b and c are combined

7. SUMMARY

- text- and data mining together form a powerful tool for literature based discovery and knowledge-base development
- limitations: access to full text, need for NLP expertise and experts/collaborators/partners
- DevToxNet™ script built for v-Embryo™ / ToxCast™, but can be generalized to be used for any study area
- literature derived information when combined with structured databases yield novel hypothesis