LITERATURE MINING AND KNOWLEDGE DISCOVERY TOOLS FOR VIRTUAL TISSUES

Amar V Singh*, Thomas B Knudsen and Imran Shah

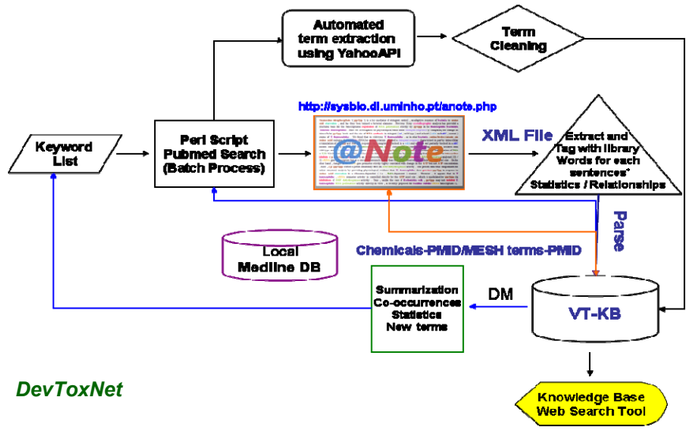
U.S EPA, ORD, Computational Toxicology Research Program and *Lockheed Martin, Research Triangle Park, NC.

research&development

Science Question

Deriving novel relationships from information in the scientific literature is an important adjunct to data-mining of complex datasets such as ToxCast™ and ToxRefDB. Concept-mining is a knowledge discovery tool to help address causal links, associations, relationships, and patterns among collections of relevant terms and keywords.

The question addressed here pertains to the applications of text-mining algorithms to extract relevant information from the literature and find key relationships.

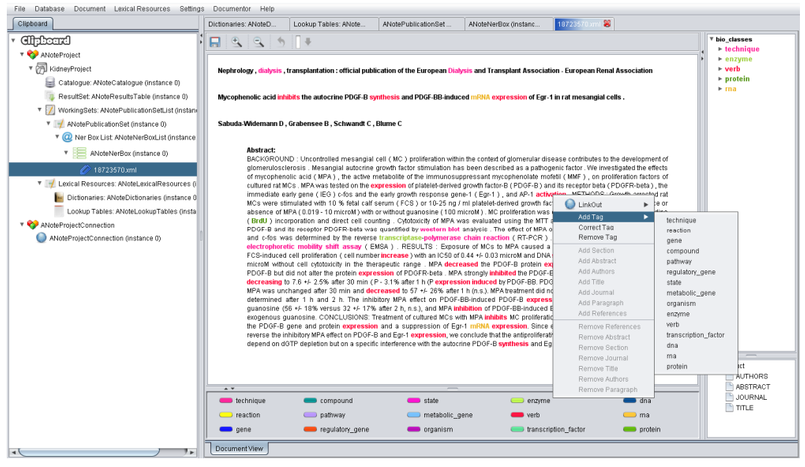


VT-KB workflow: Implementation to store information such as data, rules and classifications; organize this information into structured concept models (ontologies); and facilitate retrieval of relevant information based on hypothesis-driven queries. The queries may be human readable (manual) or machine-readable (automated).

Research Goals

1. Build text- and data-mining tools to extract and organize relevant facts from the scientific literature and other sources of information.
2. Build knowledgebase for Virtual Tissues (VT), *in silico* models to simulate the cellular fabric of tissues to predict multicellular behaviors in specific biological systems such as the adult liver (v-Liver™) or developing embryo (v-Embryo™).

Named Entity Recognition (NER) Based Search



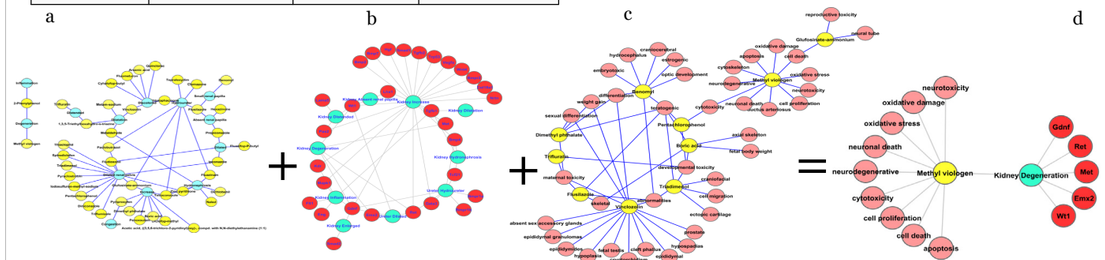
Snapshot of @Note Application

Methods/Approach

Hypothesis Generation

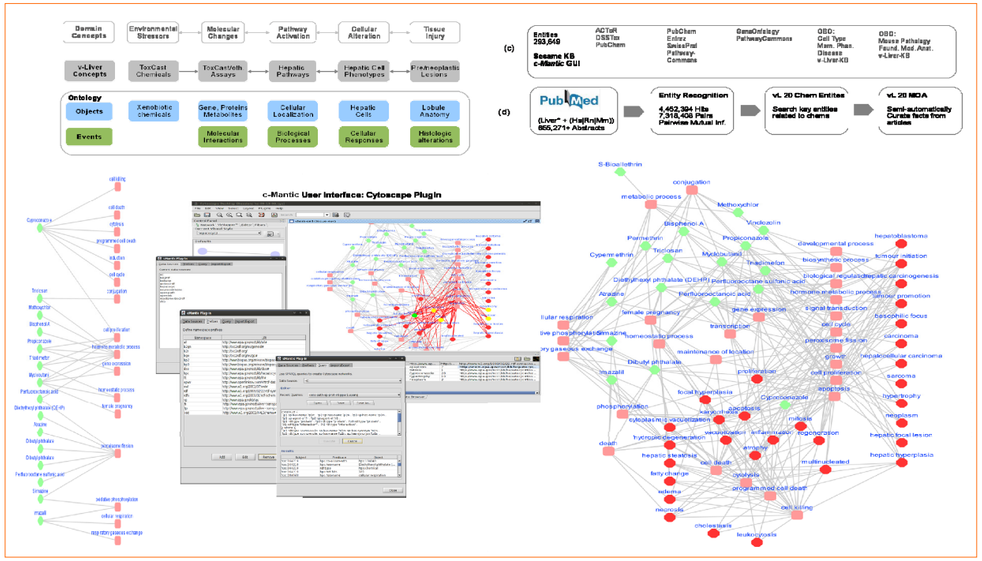
Keyword List	No. of Keywords	No. of Articles	Distinct Articles
Chemicals	48	190	181
Defects	18	1048	979
Genes	120	6705	5303

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



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 for chemical-specific hypothesis when information from a, b and c are combined

Knowledge Acquisition



Results/Conclusions

- Text- and data-mining together form a powerful tool for knowledgebase development
- ToxCast_320 chemicals- raw search returned 186K PubMed abstracts covering 82% of chemicals; filtering by <embryo or fetal> - 9K abstracts covering 48% of the chemicals
- Literature derived information when combined with structured data yield novel hypotheses

Impact and Outcomes

- Semantic integration of unstructured, distributed, heterogeneous data to structured data
- Knowledgebase to support generalized, dynamic model development
- Flexible-extensible strategy to support decision-making in risk assessment

Future Directions

- Expand knowledgebase to other domains (eg, vasculogenesis, stem cell biology, alternative models)
- Web-interface to query the database for literature based discoveries
- Connect to ACToR database for open access

References

1. Singh AV, Yang C, Kavlock RJ and Richard AM (2010) Development Toxicology Research Strategies: Computational Toxicology. Comprehensive Toxicology, 2nd edition (Editors Daston GP and Knudsen TB) Elsevier:NY
2. Lourenco A et al.@Note: A workbench for Biomedical Text Mining (2009) Journal of Biomedical Informatics Volume 42, Issue 4, Pages 710-720
3. Singh Av, Shah I and knudsen TB. (2009) Literature mining and knowledgebase development for v-Embryo (in Preparation)