Literature Mining and Knowledge Discovery Tools for Virtual Tissues

Singh AV[1], Knudsen T B[2] and Shah I[2]
[1] Lockheed Martin, Contractor to the USEPA, RTP, NC
[2] National Center for Computational Toxicology, USEPA, RTP, NC

Virtual Tissues (VTs) are *in silico* models that simulate the cellular fabric of tissues to analyze complex relationships and predict multicellular behaviors in specific biological systems such as the mature liver (v-Liver™) or developing embryo (v-Embryo™). VT models require input of biological knowledge about the systems under investigation. We are using VTs to model experimental data, such as ToxCast™ *in vitro* assays, with information about molecular pathways, cellular networks and clinical phenotypes in target organ systems. Knowledgebase development requires a flexible platform to extract and organize relevant facts from the scientific literature and other sources of information. The knowledge discovery workflow starts with information retrieval (IR) by user-defined input on single or multiple keywords to retrieve relevant PubMed abstracts, followed by information extraction (IE) and relationship mapping (RM). Currently, we use the publicly available '@Note'[1] tool for highly customizable named entity recognition (NER). A vocabulary of terms was built to describe pathologically relevant concepts using publicly available ontologies (www.OBOfoundry.org/) including genes, pathways, anatomy, clinical outcomes, and chemicals. The results from @Note are stored in a relational database for statistical analyses to summarize relationships and map them to broader biological concepts. The text-mining (TM) workflow is being implemented as a modular tool that uses open-source libraries. We are using this workflow to extract relevant facts about hepatocarcinogenesis and embryo dysmorphogenesis. This poster will provide specific examples of predicted associations that were data-mined from ToxCast_320 chemicals and 76 ToxRefDB endpoints. Biomedical literature mining aims to identify and extract plausible patterns for explicit (IE) and implicit (TM) concepts that are previously known and unknown, respectively and that can be used to better understand the inferred associations in predictive modeling. [*This work has been reviewed by EPA and cleared for presentation, but does not reflect official Agency policy*].

1 http://sysbio.di.uminho.pt/anote/wiki/index.php/Main_Page.