Summary of Curation Details for the
Comparative Toxicogenomics Database

Table of Contents

I. INTRODUCTION
a. Overview

The current goal of CTD is to provide a freely available resource that facilitates understanding of and development of novel hypotheses about the effects of the environment on human health (http://ctd.mdibl.org). Data in CTD are manually curated from the literature and comprise:

1. chemical-gene interactions
2. chemical-disease relationships
3. gene-disease relationships

These interactions/relationships are then integrated to form the chemical-gene-disease triad:
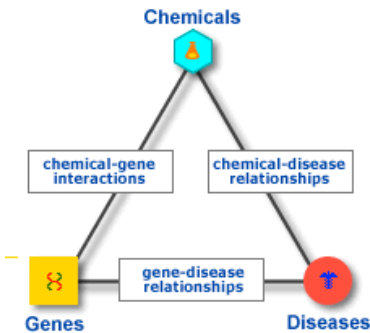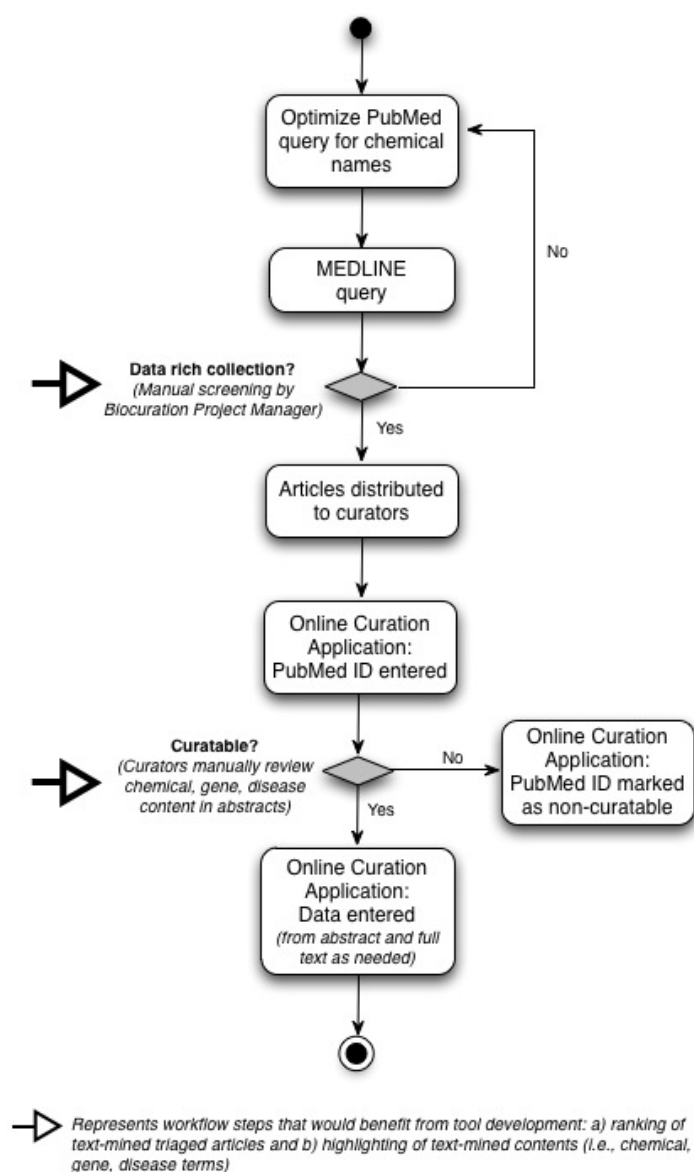

b. Applications

CTD is intended for use by biomedical researchers at academic, research, and government institutions who are interested in understanding how factors in the environment influence human health. Unique integration of chemical, gene and protein, and disease data in combination with novel analysis tools support development of testable hypotheses that may advance identification of exposure and disease biomarkers, mechanisms of chemical actions, and the complex etiologies of chronic diseases.

c. Curation workflow
The biomedical science literature is preflitered by performing a PubMed search on priority chemicals and their synonyms. Depending on the size of resulting data sets, the lead Biocurator may limit the number of articles to be

curated. Articles are then distributed to curators for examination. The abstracts are read, and if necessary, a curator may access the full-text for additional information. Relevant data from the paper is coded using controlled vocabularies using a web-based curation application. The data are loaded and available via the public web application on a monthly basis.

Represents workflow steps that would benefit from tool development: a) ranking of text-mined triaged articles and b) highlighting of text-mined contents (i.e., chemical, gene, disease terms)

II. ENCODING METHODS

a. Description of data elements curated and controlled vocabularies used

Chemicals
We use the MeSH "Chemical and Drugs" [D] hierarchy, with some modifications; we've trimmed this extensive tree a bit to remove terms that we do not consider to be chemicals of interest to CTD (e.g., the "Amino Acids, Peptides, and Proteins" branch or the "Nucleic Acids, Nucleotides, and Nucleosides" branch, etc.).

Genes
We use CTD gene pages, which are based upon imported gene pages from Entrez-Gene; however, unlike Entrez-Gene, a gene page in CTD represents the gene for all species.

Diseases
We use a mix of OMIM terms and the MeSH "Disease" [C] and "Mental Disorders" [F03] hierarchies. For future curation purposes, most disease terms will be from MeSH.

Interactions
We developed a vocabulary of chemical-gene interaction types (http://ctd.mdibl.org/downloads/). Chemical-gene interactions are written by a curator using controlled vocabularies to create a relationship between a chemical and a gene.

Organisms
We use the Eumetazoa portion of the NCBI Taxonomy.

Chemical-disease and gene-disease relationships are captured using the appropriate disease term conjoined to a qualifier code of either M (for a marker/molecular mechanism relationship) or T (for a therapeutic relationship) to the disease.

b. Description of data relationships curated

Chemical-Gene Interactions
Chemical-gene interactions must include:
- **Actors**:
  - Actors comprise Chemicals (C) and Genes (G).
  - Chemicals can be modified by actor qualifiers.
  - Genes can be modified by actor qualifiers.

- o Every interaction must have at least one C and one G.
- **Action(s)**:
  - o Action terms define the nature of an interaction and are represented by a 3-letter mnemonic
  - o Action terms can be qualified with an operator
  - o Every action term in an interaction can have only 1 action operator/degree
- **Organism**
- **High-throughput status**
  - o When the chemical-gene interaction(s) derive(s) from a high-throughput experiment (e.g., microarray), this is noted by selecting a check box in the curation application.
- **Interactions**:
  - o Interactions can be binary or more complex, represented with nested relationships (examples below).
  - o Every interaction must have 2 or more actors; including at least 1 C and 1 G
  - o Every interaction must have 1 or more action terms
  - o The basic structure of an interaction is: Actor 1/qualifier :: action operator→Action term :: Actor 2/qualifier.

    Example: C1 +exp G1/m
    C1= Actor 1
    +exp= Action (with an action operator of + to indicate an "increase")
    G1= Actor 2
    /m= Actor 2 qualifier
    Translation: C1 results in increased expression of G1 mRNA.

Chemical- and Gene-Disease Relationships
Disease relationships include:
- A chemical or a gene
- A disease (D)
- One of two possible qualifiers that describe the nature of the relationship:
  - o M: Molecular Mechanism or Marker
  - o T: Therapeutic or possible Therapeutic
- An organism

Additional Curated Information
- **In vitro vs. in vivo status**
  - o Using check boxes in the curation application, curators specify whether the data they curated were generated in an in vitro or in vivo system.
- **Abstract vs full text**
  - o Using check boxes in the curation application, curators specify whether the data they curated were identified in the abstract or full text of an article. This information may assist future text mining projects.

III. CURATION TOOLS
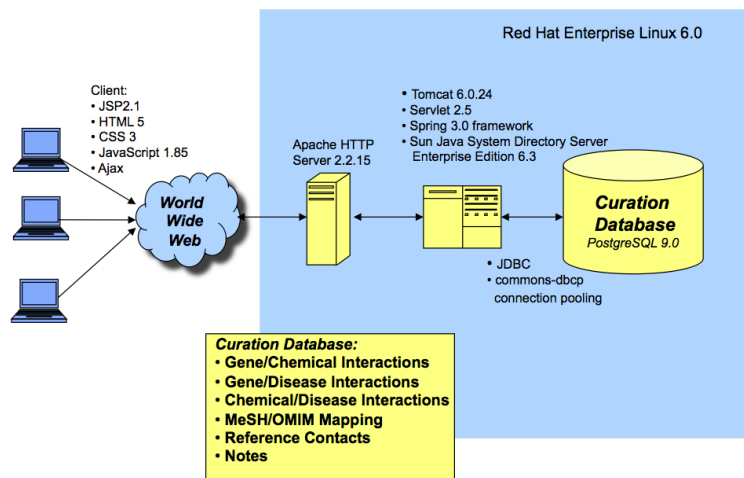
The CTD Curation Tool is internet-based and integrates *JSP2.1 /Servlet 2.5, HTML5, CSS3, JavaScript 1.85*, and *AJAX*, in the context of an MVC architecture, and in conjunction with an *Apache HTTP Server 2.2.15* and *Tomcat 6.0.24*. Data is stored in a *PostgreSQL 9.0* database management system and is accessed using *commons-dbcp* connection pooling in conjunction with *JDBC*. The operating environment is *Red Hat Enterprise Linux 6.0*. Security is managed using the *Spring 3.0 Framework* in conjunction *LDAP* via *Sun Java System Directory Server Enterprise Edition 6.3 (Figure 1)*.

Figure 1. Technical overview of the CTD Curation Tool.

IV. TEXT MINING

a. Summary

In 2009, a prototype a Java-based text-mining application was developed and evaluated using a CTD data set consisting of manually curated molecular interactions and relationships from 1,600 documents. *OSCAR 3* and *MetaMap* were used for chemical term recognition, *ABNER* and *MetaMap* were used for gene/protein term recognition, and *MetaMap* was used for disease term recognition; these recognition tools were integrated into the Java-based prototype. *Lucene* was used for document ranking, as was a custom rule-based formula developed in-house.

A variety of criteria and accompanying weights were used to rank documents with the rule-based and Lucene-based ranking algorithms described in this report. Those that yielded the best results when compared against our manually curated data were used for the study. These criteria were based on the curators' experience and intuitions, but there was no formal method used for recalibrating or deriving these weights (see PMID: 19814812 for details).

V. RECENT CTD REFERENCES

1. Davis AP, Wiegers TC, Murphy CG, and Mattingly CJ. 2011. The curation paradigm and application tool used for manual curation of the scientific literature at the Comparative Toxicogenomics Database. Database (2011) doi: 10.1093/database/bar034.
2. Mattingly CJ. 2011. The Comparative Toxicogenomics Database (CTD): A resource for building hypotheses about environment-disease connections. NCI-Nature Pathway Interaction Database. In press.
3. Davis AP, King BL, Mockus S, Murphy CG, Saraceni-Richards C, Rosenstein M, Wiegers T, and Mattingly CJ. 2011. The Comparative Toxicogenomics Database: update 2011. Nucleic Acids Res 39(Database issue): D1067-1072.
4. Wiegers TC, Davis AP, Cohen KB, Hirschman L, and Mattingly CJ. (2009) Using text mining to enhance manual curation of chemical-gene-disease networks for the Comparative Toxicogenomics Database (CTD) BMC Bioinformatics. Oct 8;10:326. PMC2768719.