**Summary Of Curation Details For The**

**Comparative Toxicogenomics Database**

Table of Contents

I. INTRODUCTION 1

a. Overview 1

b. Applications 1

c. Curation workflow 2

II. ENCODING METHODS 2

a. Description of data elements curated and controlled vocabularies used 2

Chemicals 2

Genes 2

Diseases 2

Organisms 2

Interactions 2

b. Description of data relationships curated 3

Chemical-Gene Interactions 3

Chemical- and Gene-Disease Relationships 7

Additional Curated Information 7

III. CURATION TOOLS 7

IV. TEXT MINING 8

V. RECENT CTD REFERENCES 8

VI. POSSIBLE DEVELOPMENT PROJECTS 8

# I. INTRODUCTION

## a. Overview

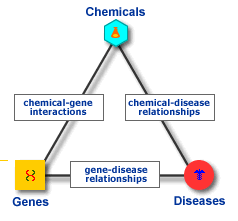
The current goal of CTD is to provide a freely available resource that facilitates understanding of and development of novel hypotheses about the effects of the environment on human health. Data in CTD are manually curated from the literature and comprise:

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

A curator is provided with a selected set of references 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.

# ctd workflow.jpgII. 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.

### Organisms

We use the Eumetazoa portion of the NCBI Taxonomy.

### Interactions

We developed a vocabulary of action terms (Table 2)

Chemical-gene interactions are written by a curator using controlled vocabularies to create a relationship between a chemical and a gene.

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 0 or 1 actor qualifiers (Table 1)
  + Genes can be modified by 0, 1, or 2 actor qualifiers (Table 1)
    - Gene qualifiers are divided into two levels. A 2nd level actor qualifier can only be used if a 1st level actor qualifier is first selected.
  + Every interaction must have at least one C and one G.
* **Action(s)**:
  + Action terms define the nature of an interaction and are represented by a 3-letter mnemonic (Table 2).
  + Action terms can be qualified with an operator (Table 3).
    - The only exceptions are for two action codes: “co-treatment” (w), which cannot be modified, and “binds” (b) which can only be used as either “b” (binds to) or “0b” (does not bind to”); that is, you cannot say “+b” (increased binding) or “-b” (decreased binding).
  + Every action term in an interaction can have only 1 action operator/degree
* **Organism**
* **High-throughput status**
  + 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. These data are not currently displayed, but we will eventually give users the option of specifying or filtering out these data.
* **Interactions**:
  + Interactions can be binary or more complex, represented with nested relationships (examples below).
  + Every interaction must have 2 or more actors; including at least 1 C and 1 G
  + Every interaction must have 1 or more action terms
  + The basic structure of an interaction is: Actor 1/qualifier :: action operatorAction term :: Actor 2/qualifier (see examples below)

Table 1. Actor Qualifier Codes

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Actor** | **Qualifier code** |  | **Qualifier name** | **Translation** | **Notes & Examples** |
| **Chemical** | [blank] |  |  | [nothing] | most commonly used |
| **Chemical** | /n |  | aNalog | analog |  |
| **Chemical** | /y |  | deficiencY | deficiency |  |
| **Chemical** | /b |  | metaBolite | metabolite |  |
|  |  |  |  |  |  |
| **Actor** | **Qualifier code** | **Level** | **Qualifier name** | **Translation** |  |
| **Gene** | [blank] | **1** |  | [nothing] | used rarely (when you simply cannot discern if they assayed mRNA or protein or what) |
| **Gene** | /p | **1** | Protein | protein | commonly used |
| **Gene** | /d | **1** | DNA | gene |  |
| **Gene** | /r | **1** | promoteR | promoter |  |
| **Gene** | /e | **1** | Enhancer | enhancer |  |
| **Gene** | /x | **1** | eXon | exon |  |
| **Gene** | /i | **1** | Intron | intron |  |
| **Gene** | /m | **1** | mRNA | mRNA | commonly used |
| **Gene** | /5 | **1** | 5' UTR | 5' UTR |  |
| **Gene** | /3 | **1** | 3' UTR | 3' UTR |  |
| **Gene** | /a | **1** | polyA | polyA tail |  |
| **Gene** | /f | **2** | modified Form | modified form | e.g., “phosphorylated protein” = G1/p/f |
| **Gene** | /alt | **2** | ALTternative form | alternative form | e.g., “alternative mRNA” = G1/m/alt |
| **Gene** | /mutant | **2** | mutant form | mutant form | if authors call it a mutation |
| **Gene** | /poly | **2** | POLYmorphism | polymorphism | if authors call it a polymorphism |
| **Gene** | /snp | **2** | SNP | SNP | if authors describe a single nucleotide polymorphism |

Table 2. Action Codes

|  |  |  |
| --- | --- | --- |
| **Code** | **Name** | **Definition** |
| w | cotreatment | Involving the use of two or more chemicals and/or genes simultaneously. |
| b | binding | A molecular interaction. |
| rxn | reaction | Any general biochemical or molecular event. |
| act | activity | An elemental function of a molecule. |
| loc | localization | Part of the cell where a molecule resides. |
| exp | expression | The expression of a gene product. |
| abu | abundance | The abundance of a chemical (if chemical synthesis is not known). |
| mut | mutagenesis | The genetic alteration of a gene product. |
| rec | response to chemical | Chemical resistance or chemical sensitivity. (This term can only be used when followed by a chemical – see Table 4) |
| sta | stability | Overall molecular integrity. |
| spl | splicing | The removal of introns to generate mRNA. |
| fol | folding | Bending and positioning of molecule to achieve conformational integrity. |
| trt | transport | Movement of a molecule into or out of a cell. |
| upt | uptake | Movement of a molecule into a cell (by less specific means than import). |
| imt | import | Movement of a molecule into a cell (by more specific means than uptake). |
| sec | secretion | Movement of a molecule out of cell (by less specific means than export) |
| ext | export | Movement of a molecule out of cell (by more specific means than secretion). |
| met | metabolic processing | Biochemical alteration of molecule's structure (does not include changes in expression, stability, folding, localization, splicing, or transport). |
| csy | chemical synthesis | A biochemical event resulting in a new chemical product. |
| deg | degradation | Catabolism or breakdown. |
| ace | acetylation | The addition of an acetyl group. |
| acy | acylation | The addition of an acyl group. |
| alk | alkylation | The addition of an alkyl group. |
| ami | amination | The addition of an amine group. |
| car | carbamoylation | The addition of a carbamoyl group. |
| cox | carboxylation | The addition of a carboxyl group. |
| clv | cleavage | The processing or splitting of a molecule, not necessarily leading to the destruction of the molecule. |
| eth | ethylation | The addition of an ethyl group. |
| gyc | glycation | The non-enzymatic addition of a sugar. |
| gly | glycosylation | The addition of a sugar group. |
| ngl | N-linked glycosylation | The addition of a sugar group to an amide nitrogen. |
| ogl | O-linked glycosylation | The addition of a sugar group to a hydroxyl group. |
| glc | glucuronidation | The addition of a sugar group to form a glucuronide, typically part of an inactivating or detoxifying reaction. |
| hyd | hydrolysis | The splitting of a molecule via the specific use of water. |
| hdx | hydroxylation | The addition of a hydroxy group. |
| lip | lipidation | The addition of a lipid group. |
| ger | geranoylation | The addition of a geranoyl group. |
| far | farnesylation | The addition of a farnesyl group. |
| myr | myristoylation | The addition of a myristoyl group. |
| pal | palmitoylation | The addition of a palmitoyl group. |
| pre | prenylation | The addition of a prenyl group. |
| myl | methylation | The addition of a methyl group. |
| nit | nitrosation | The addition of a nitroso or nitrosyl group. |
| nuc | nucleotidylation | The addition of a nucleotidyl group. |
| oxd | oxidation | The loss of electrons. |
| pho | phosphorylation | The addition of a phosphate group. |
| sul | sulfation | The addition of a sulfate group. |
| sum | sumoylation | The addition of a SUMO group. |
| red | reduction | The gain of electrons. |
| rib | ribosylation | The addition of a ribosyl group. |
| arb | ADP-ribosylation | The addition of a ADP-ribosyl group. |
| ubq | ubiquitination | The addition of an ubiquitin group. |
| glt | glutathionylation | The addition of a glutathione group. |

Table 3. Action Operator Codes

|  |  |  |  |
| --- | --- | --- | --- |
| **Operator Code** | **Operator Name** | **Translation** | **Special uses** |
| + | increase | results in increased…. | never used with "b" (binds) or "w" (co-treatment) |
| - | decrease | results in decreased… | never used with "b" (binds) or "w" (co-treatment) |
| 0 [zero] | not | does not affect the…. |  |
| [blank] | [default] | affects the… |  |

Table 4. “rec” code translations

|  |  |
| --- | --- |
| **Code** | **Translation** |
| -rec | results in chemical resistance to… |
| +rec | results in chemical sensitivity to… |
| rec | affects the chemical susceptibility to… |
| 0rec | does not affect the response to chemical… |

**Example 1: 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.

**Example 2: C1/b b +act G1/p**

C1 = Actor 1

/b = Actor 1 qualifier

b = Action (without an action operator)

+act = Action (with the action operator of “+” = “increased”)

G1 = Actor 2

/p = Actor 2 qualifer

Translation: C1 metabolite binds to and results in increased activity of G1 protein.

**Example 3: [C1 w C2] exp G1/m**

[C1 w C2] = Actor 1

exp = Action (without an action operator)

G1 = Actor 2

/m = Actor 2 qualifier

Translation: [C1 co-treated with C2] affects the expression of G1 mRNA.

**Example 4: C1/n 0rxn [C2 +rxn [[C3 b G1/p] b G2/p]]**

Use of [brackets] to nest reactions creates many different actors within actors.

Working from inside out:

The bracketed reaction [C3 b G1/p]

C3 = Actor 1

b = Action (without an action operator)

G1 = Actor 2

/p = Actor 2 qualifier

This bracketed reaction then becomes its own actor in the next bracketed reaction: [C3 b G1/p] b G2/p

[C3 b G1/p] = Actor 1

b = Action

G2 = Actor 2

/p = Actor 2 qualifier

And likewise, for: C2 +rxn [[C3 b G1/p] b G2/p]

C2 = Actor 1

+rxn = Action (with action operator “+”)

[[C3 b G1/p] b G2/p] = Actor 2

And likewise, for: C1/n 0rxn [C2 +rxn [[C3 b G1/p] b G2/p]]

C1 = Actor 1

/n = Actor 1 qualifier

0rxn = Action (with action operator 0 = “not”)

[C2 +rxn [[C3 b G1/p] b G2/p]] = Actor 2

Translation: C1 does not affect the reaction of [C2 increases the reaction of [[C3 binds to G1 protein] which binds to G2 protein]].

### 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:
  + M: Molecular Mechanism or Marker
  + T: Therapeutic or possible Therapeutic
* An organism

### Additional Curated Information

* **In vitro vs. in vivo status**
  + 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**
  + 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)*.

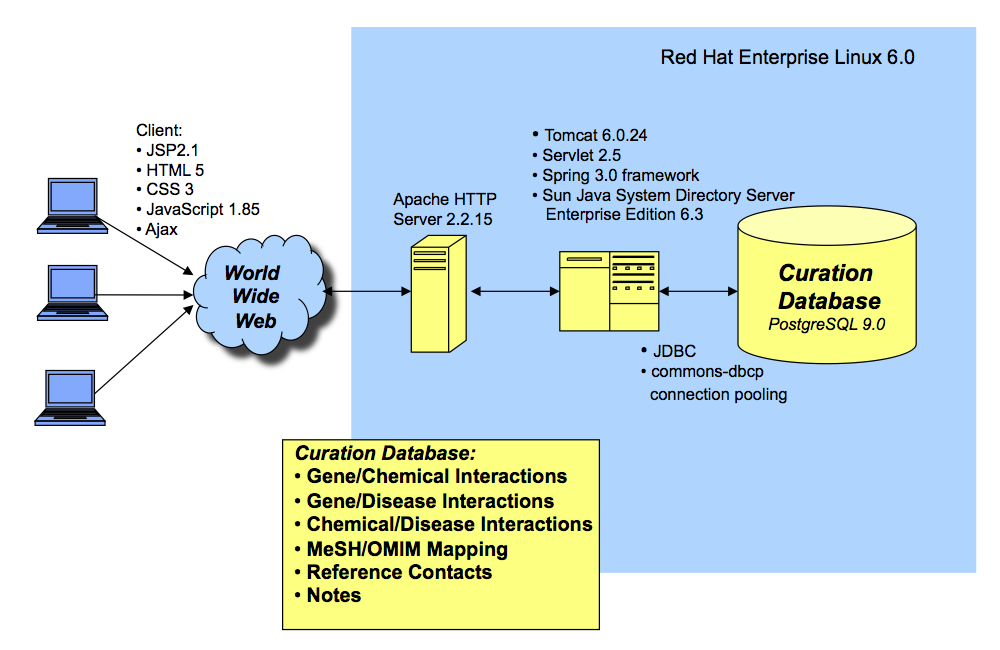


Figure 1. Technical overview of the CTD Curation Tool.

# IV. TEXT MINING

The CTD text mining pipeline is well described in (1) below.

# V. RECENT CTD REFERENCES

1. Davis AP, Wiegers TC, Johnson RJ, Lay JM, Lennon-Hopkins K, Saraceni-Richards C, Sciaky D, Murphy CG, Mattingly CJ. Text mining effectively scores and ranks the literature for improving chemical-gene-disease curation at the Comparative Toxicogenomics Database. PLoS One. 2013 Apr 17;8(4):e58201.
2. Davis AP, Murphy CG, Johnson R, Lay JM, Lennon-Hopkins K, Saraceni-Richards C, Sciaky D, King BL, Rosenstein MC, Wiegers TC, Mattingly CJ. The Comparative Toxicogenomics Database: update 2013. Nucleic Acids Res. 2013 Jan 1;41(D1):D1104-14.
3. Davis AP, Wiegers TC, Rosenstein MC, Mattingly CJ. MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database. Database (Oxford). 2012 Mar 20;2012:bar065.
4. Mattingly CJ. Understanding environment-disease connections: An introduction to the Comparative Toxicogenomics Database (CTD). NCI-Nature Pathway Interaction Database. doi:10.1038/pid.2011.2 (2011).
5. 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.
6. 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.

# VI. POSSIBLE DEVELOPMENT PROJECTS

1. Better identification and ranking of data rich articles for curation of the data described above. Currently we do this based on abstracts, but there would be great value in doing this from the full text of articles.
2. Highlighting of relevant data within abstracts or full text for curation.
3. Extraction of relevant phrases or information from abstracts or full text.