The Comparative Toxicogenomics Database (CTD): A public resource for building chemical-gene networks

Carolyn J. Mattingly[1], Michael C. Rosenstein[1], Allan P. Davis[1],
John N. Forrest, Jr.[2], James L. Boyer[2]
[1]Mount Desert Island Biological Laboratory, Salisbury Cove, ME 04672;
[2]Yale University School of Medicine, New Haven, CT 06520

The etiology of most chronic diseases involves interactions between environmental factors and genes that modulate important physiological processes[3]. Despite the prevalence of chemicals in the environment and their importance in the etiology of most human diseases, their mechanisms of action and effects on human health are poorly understood. To promote understanding of these mechanisms and provide insights into the molecular basis of differential susceptibility to chemical exposures, we are developing an authoritative public database of scientifically reviewed information on environmental chemicals, significant genes, and their interactions. In the Comparative Toxicogenomics Database[2] (CTD; http://ctd.mdibl.org/), we present chemical-gene associations from the published literature, a set of curated toxicologically important genes and their proteins, and visualization capabilities to facilitate cross-species sequence comparisons of these genes and proteins. Manual curation of interactions between chemicals and genes and proteins in diverse species is underway and these data will be integrated with CTD in the coming year. These interactions will be key to predicting complex chemical-gene interaction networks.

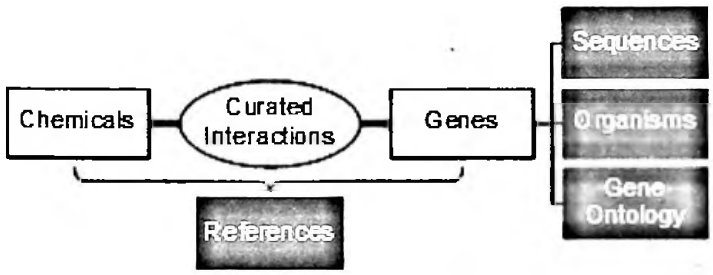

Figure 1. High Level View of the Primary Data Types in CTD.

The major types of data integrated in CTD are: 1) nucleotide and protein sequences; 2) published references; 3) curated genes and Gene Sets (sets of curated genes); 4) a hierarchical vocabulary of chemicals; 5) the Gene Ontology (hierarchical vocabulary of biological processes, cellular components, and molecular functions); and 6) a hierarchical vocabulary of organisms (Figure 1).

CTD currently contains 1.2 million nucleotide and protein sequences for vertebrates and invertebrates, which enable broad cross-species sequence comparisons. Nucleotide and protein sequences are included for all vertebrates and invertebrates to enable broad visitor-driven cross-species sequence comparisons. Nucleotide sequences and annotations are acquired from the National Center for Biotechnology Information (NCBI). In order to minimize nucleotide sequence redundancy, we include only Reference Sequences for *Homo sapiens* (human), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Drosophila melanogaster* (fruit fly), and *Caenorhabditis elegans* (nematode). Amino acid sequences and annotations are acquired from the European Bioinformatics Institute's Swiss-Prot and TrEMBL databases.

CTD provides a filtered set of over 70,000 references that are specifically relevant to chemical-gene interactions. References are acquired from PubMed and are first identified as relevant using a text-mining strategy that iteratively searches abstracts and titles for co-occurrences of chemical and gene terms from CTD vocabularies. Articles are subsequently manually curated for chemical-gene interactions (see below).

Several controlled vocabularies were integrated to ensure consistency in data integration, annotation, access, and interpretation including portions of hierarchical vocabularies for organism taxonomy, chemicals, and Gene Ontology (GO). Our organism vocabulary consists of the Eumetazoa portion (vertebrates and invertebrates) of the NCBI Taxonomy vocabulary. Our chemical vocabulary was adapted from the National Library of Medicine's Medical Subject Headings and Supplementary Concepts. We integrated the GO vocabulary, allowing visitors to search for genes and proteins by biological processes, cellular components, and molecular functions. This capability greatly increases the complexity of questions that can be asked (e.g., "What kinases [GO molecular function term] are affected by arsenic [chemical term]?"). The organism and chemical vocabularies include synonyms, which allow visitors to retrieve the same data with different but related terms (e.g., "zebrafish" and "*Danio rerio*" or "TCDD" and "2,3,7,8-tetrachlorodibenzo-*p*-dioxin"). The hierarchical structure of the vocabularies also gives visitors the flexibility to search by specific or general terms (e.g., "zebrafish" vs. "teleostei" or "mercury" vs. "metals," respectively).

Cross-species genes are defined in CTD by their constituent nucleotide and protein sequences from vertebrates and invertebrates. Cross-species genes are constructed using sequence analysis methods in combination with literature review. Curated genes will provide users with direct access to toxicologically relevant gene, protein, and cross-species sequence data that will facilitate user-driven comparative sequence analyses. Gene Sets group closely related curated genes, such as those that have undergone duplication events in specific species (e.g., CYP1A4, CYP1A5) or are members of large families (e.g., ABC) and provide visitors with a broader perspective about their gene of interest. For example, the CYP1A Gene Set, which includes the curated genes CYP1A1, CYP1A2, CYP1A3, CYP1A4, and CYP1A5, combines information about mammalian-, teleostei-, and avian-specific genes.

The user interface in CTD provides access to gene, sequence, reference, and chemical data from a range of perspectives. Where possible, data is presented in a cross-species context. From the CTD home page visitors can find information using gene or reference query forms or by browsing the chemical vocabulary.

Visitors may retrieve information about genes using the gene query form by gene, Gene Set, chemical, GO term, organism, or sequence ID. Vocabulary browsers are provided for chemical, organism, and GO terms to allow visitors to browse for specific or general query terms. Results may also be restricted to those with associated microarray data. Currently, CTD data is linked to microarray results in the Environment, Drugs and Gene Expression database[1], which is the only robustly populated, publicly available microarray database devoted to toxicology-related gene expression information. Complex queries using combinations of these fields on the gene query form are possible. For example, a query for genes with receptor activity (GO term) that are affected by 2,3,7,8-tetrachlorodibenzo-p-dioxin (chemical term) retrieves a list of 18 unique genes. Each gene may be selected for supplementary information on a gene detail page (Figure 2). Detail pages for individual genes allow visitors to navigate between categories of data including basic information, references and cited chemicals, GO annotations, vertebrate and invertebrate sequences, and associated microarray data. As curated genes and Gene Sets are integrated with data in CTD links will be provided for Gene Set detail pages that will place genes in a broader cross-species context (see above).

The CTD Reference Query form allows visitors to search for articles by chemical, gene, organism, author, or citation information. In addition to citation information, reference results display genes and chemicals cited in the reference. These associations are compiled using a combination of NLM MeSH annotation of PubMed articles and a reference retrieval text-mining strategy developed for CTD.

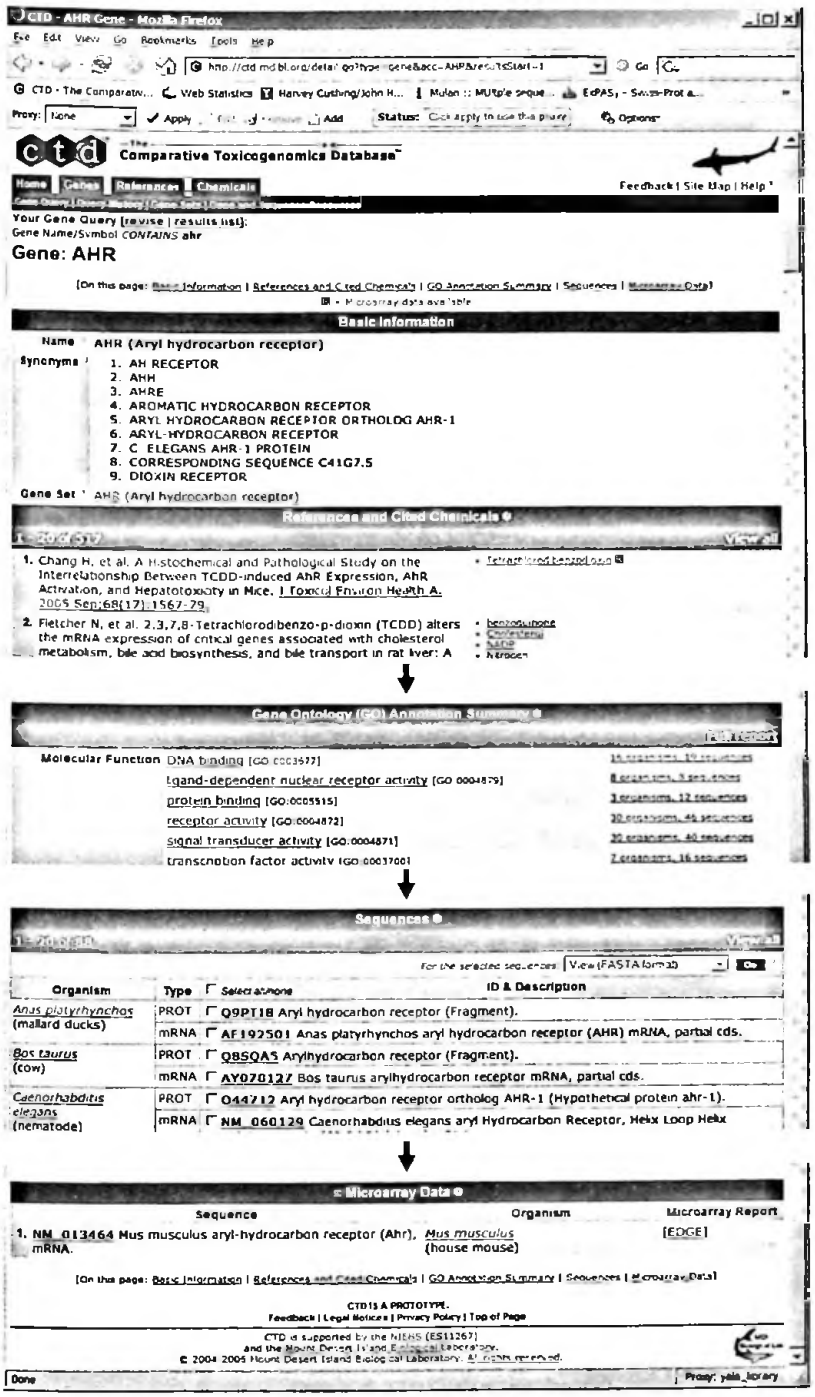

Figure 2. CTD gene detail pages centralize cross-species and Toxicology data related to genes and proteins.

Visitors may query CTD from a chemical perspective using the chemical browser. A chemical query provides users with access to chemical detail pages that provide chemical structures, links to associated sequences and references in CTD, and links to relevant chemical and microarray databases. These chemical detail pages provide an important synthesis of molecular and traditional toxicology information from otherwise disconnected sources.

Genes and proteins function together in complex networks rather than in isolation. Understanding mechanisms of chemical actions requires knowledge of these networks and constituent chemical-gene interactions, which may be direct (e.g., "chemical binds to protein") or indirect (e.g., "chemical results in activated transcription of a gene" via intermediate events). In order to help elucidate the mechanisms of chemical action, we have begun manually curating specific chemical-gene interactions in diverse species from our text-mined set of references. To date, CTD has manually curated over 15,000 interactions involving more than 1,500 chemicals and 2,300 genes in 75 different species. In 2006, we will begin integrating these data with information in the CTD web interface. These data will allow visitors to ask sophisticated questions by querying with numerous parameters (e.g., "LPS [chemical term] affects transport [CTD interaction] of which proteins?"; "which protein kinases [GO molecular function term] play a role in chemical resistance [CTD interaction term] to tamoxifen [chemical term]?"; "in which organisms does tetrachlorodibenzodioxin [chemical term] bind the aryl hydrocarbon receptor [gene term]?"). Collectively, these interactions will be critical for building chemical-gene interaction networks and

supporting hypothesis-driven research that aims to elucidate the mechanisms by which chemicals modulate toxicity and disease (Figure 3).

Development of CTD is supported by NIEHS (R33 ES011267), NCRR (P20 RR-016463), and the Mount Desert Island Biological Laboratory.

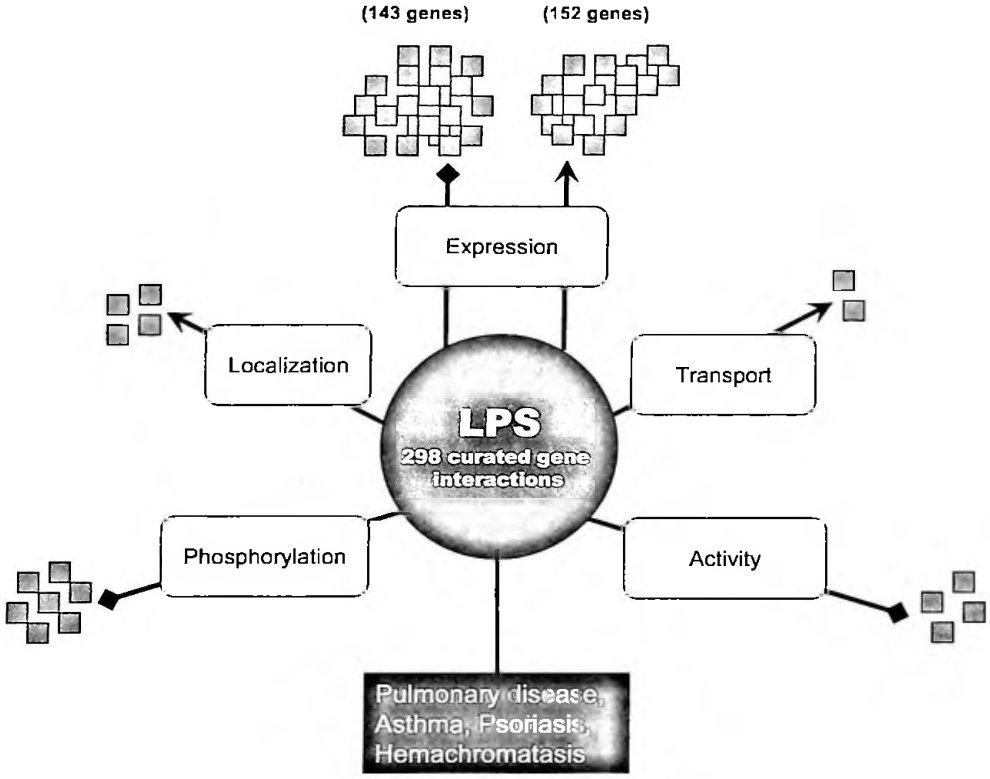

Figure 3. Manually curated chemical-gene interactions will provide information that is key to building chemical-gene interaction networks. For example, interactions have been curated between lipopolysaccharide (LPS) and 298 genes or proteins. Interactions include affects on expression, localization, phosphorylation, activity and transport. When integrated in CTD, these interactions will be presented with other annotated data such as diseases associated with chemicals and genes or proteins.

1. **Hayes, K.R., A.L. Vollrath, G.M. Zastrow, B.J. McMillan, M. Craven, S. Jovanovich, D.R. Rank, S. Penn, J.A. Walisser, J.K. Reddy, R.S. Thomas, and C.A. Bradfield** EDGE: a centralized resource for the comparison, analysis, and distribution of toxicogenomic information. *Mol Pharmacol.* 67: 1360-1368, 2005.
2. **Mattingly C.J., G.T. Colby, M.C. Rosenstein, J.N. Forrest, Jr., and J.L. Boyer.** Promoting comparative molecular studies in environmental health research: an overview of the comparative toxicogenomics database (CTD). *Pharmacogenomics J.* 4(1):5-8, 2004.
3. **Schwartz D.A., J.H. Freedman, and E.A. Linney.** Environmental genomics: a key to understanding biology, pathophysiology and disease. *Hum Mol Genet.* 13: Spec No 2:R217-224, 2004.