The Comparative Toxicogenomics Database (CTD)

Carolyn J. Mattingly,[1,2,3] Glenn T. Colby,[1,2,3] John N. Forrest,[2,3,4] and James L. Boyer[2,3,4]

[1]Department of Bioinformatics, [2]Center for Membrane Toxicity Studies, and [3]Center for Marine Functional Genomic Studies, Mount Desert Island Biological Laboratory, Salsbury Cove, Maine, USA; [4]Department of Medicine, Yale University School of Medicine, New Haven, Connecticut, USA

The Mount Desert Island Biological Laboratory in Salsbury Cove, Maine, USA, is developing the Comparative Toxicogenomics Database (CTD), a community-supported genomic resource devoted to genes and proteins of human toxicologic significance. CTD will be the first publicly available database to *a*) provide annotated associations among genes, proteins, references, and toxic agents, with a focus on annotating data from aquatic and mammalian organisms; *b*) include nucleotide and protein sequences from diverse species; *c*) offer a range of analysis tools for customized comparative studies; and *d*) provide information to investigators on available molecular reagents. This combination of features will facilitate cross-species comparisons of toxicologically significant genes and proteins. These comparisons will promote understanding of molecular evolution, the significance of conserved sequences, the genetic basis of variable sensitivity to environmental agents, and the complex interactions between the environment and human health. CTD is currently under development, and the planned scope and functions of the database are described herein. The intent of this report is to invite community participation in the development of CTD to ensure that it will be a valuable resource for environmental health, molecular biology, and toxicology research. ***Key words:*** **aquatic, comparative, database, environmental health, fishes, genomic, health, toxicogenomics, toxicology.** ***Environ Health Perspect*** **111:793–795 (2003). doi:10.1289/txg.6028 available via *http://dx.doi.org/* [Online 13 February 2003]**

Approximately 75,000 chemicals are currently listed in the U.S. Environmental Protection Agency (U.S. EPA) Toxic Substances Control Act Chemical Substances Inventory (U.S. EPA 2003); however, the toxic potential and the molecular mechanisms underlying the action of many of these chemicals are not well understood. Scientists have long exploited diverse experimental models to understand the complexity of gene–environment interactions. With the rising number of publicly available sequences and completely sequenced genomes, comparative studies are proving to be essential for elucidating biological systems (Koonin et al. 2000) and annotating accumulating genomic and proteomic data (Whelan et al. 2001). Comparisons of more distantly related vertebrate and invertebrate species may be of particular value for identifying conserved genetic and molecular mechanisms (Wittbrodt et al. 2002). It is on this premise that the Comparative Toxicogenomics Database (CTD) is being developed.

CTD will facilitate comparisons of sequences and functions of toxicologically significant genes and proteins from diverse organisms, with an emphasis on aquatic and mammalian species. The goal is to provide unique insights into the significance of conserved sequences and polymorphisms, the genetic basis of variable sensitivity, molecular evolution, and adaptation. The potential value of such comparisons is demonstrated by studies of the aryl hydrocarbon receptor (AhR) (Hahn 2002; Thomas et al. 2002), which modulates the toxic action of the environmental contaminant 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD) (Poland and Knutson 1982; Schmidt and Bradfield 1996). Mammals, fishes, and aquatic invertebrates exhibit different toxicity profiles (Hahn 2002). Studies of AhR in these organisms identified duplication events in fishes and differences in sequence identity, TCDD-binding capacity, and activation of downstream targets (Hahn 2002; Thomas et al. 2002). Although the physiologic roles of AhR are still not well understood, correlations between AhR sequences and functions in distantly related organisms may provide valuable information about the evolutionary impact on this gene, possible insights into the genetic basis of toxicity, and directions for future research.

There is a strong precedent for comparative studies with aquatic organisms. The recent sequencing of the pufferfish (*Fugu rubripes*) genome has resulted in the discovery of nearly 1,000 human genes not described previously in the public domain (Aparicio et al. 2002). The anticipated sequences for zebrafish (*Danio rerio*) and spotted green pufferfish (*Tetraodon nigroviridis*) genomes will likely make additional contributions to the annotation of the human genome. Evolutionarily diverse aquatic organisms have become important models for studying human disease. For example, membrane transporters that are the sites of action of diuretic drugs, including the bumetanide-sensitive Na-K-Cl cotransporter and the thiazide-sensitive NaCl cotransporter, were first cloned from specialized organs in marine species (Gamba et al. 1993; Xu et al. 1994). Mutagenesis studies in teleosts have generated a spectrum of biologically relevant and nonoverlapping phenotypes (Wittbrodt et al. 2002). Large-scale genetic screens have produced more than 500 zebrafish mutants, many with phenotypes similar to human disorders (Dooley and Zon 2000). Medaka (*Oryzias latipes*) are routinely used for studies in carcinogenesis and environmental health (Wittbrodt et al. 2002). The more distantly related elasmobranchs have provided unique insight into conserved functional domains of genes associated with human liver function (Ballatori and Villalobos 2002; Cai et al. 2001, 2002) and cystic fibrosis (Aller et al. 1999).

The growing body of genomic information available to the scientific community has led to an increase in the number and scope of biological databases. A recent review (Baxevanis 2002) estimated a total of 335 existing databases in 2002, an increase from 281 in 2001. These databases address a range of complex challenges for biologists, such as managing comprehensive repositories of genomic and proteomic data (Benson et al. 2002; O'Donovan et al. 2002), annotating species-specific genomes (Blake et al. 2002; Sprague et al. 2001), and identifying protein families and conserved domains (Baxevanis 2002). Existing toxicology databases have cataloged chemical and physical properties of toxic agents, mutagenicity data, environmental health and regulatory information, ecologic data, and scientific references (Russom 2002; Wexler 2001; Young 2002). It is impotant to note that there is no existing publicly available resource that provides toxicologic

Address correspondence to C.J. Mattingly, Dept. of Bioinformatics, Mount Desert Island Biological Laboratory, Salsbury Cove, ME 04672 USA. Telephone: (207) 288-3605. Fax: (207) 288-2130. E-mail: cmattin@mdibl.org

We thank N. Ballatori, J. Blake, J. Eppig, B. Forbush, and D. Towle for insightful feedback and support. This project is funded by National Institute of Environmental Health Sciences grants ES11267-02 and ES03828-17.

Inquiries about CTD may be sent to ctd@mdibl.org

Received 30 September 2002; accepted 12 February 2003.

annotation of genomic and proteomic data from diverse species. In addition to CTD, another public toxicogenomic database is being developed by the National Center for Toxicogenomics at the National Institute of Environmental Health Sciences (NIEHS). The Chemical Effects in Biological Systems (CEBS) Knowledge Base will capture and integrate global molecular expression data with pathway and regulatory network information related to toxicology and human disease (Waters et al. 2003). It is the goal of both development groups that CTD and CEBS be complementary in focus and functionally compatible.

Scope

Biologic features and strategic plan. CTD is being developed at the Mount Desert Island Biological Laboratory (MDIBL) in Salsbury Cove, Maine, USA, in collaboration with investigators at NIEHS Marine and Freshwater Biomedical Sciences (MFBS) centers and other scientists with expertise in molecular biology, toxicology, and bioinformatics. CTD will include curated information about nucleotide and protein sequences, associated references, toxic agents, reagents, and taxonomy. Tools for data analysis, manipulation, and visualization for comparative studies will also be provided. This scope of features dictates a phased implementation approach that will combine automated and manual curation strategies. The first year (September 2002–August 2003) will include three implementation phases.

Phase I will focus on the acquisition and integration of sequences, references to the scientific literature, and toxic agents. Although annotation will focus on genes and proteins with associated toxicologic data, an inclusive set of sequence data will be stored locally in CTD to *a*) maximize the value of comparative sequence analyses that may be performed using integrated computational tools, *b*) prevent exclusion of sequences with potential toxicological significance, *c*) allow querying of annotated features, and *d*) provide integration with data from other sources. Subsets of nucleotide sequences will be acquired from the National Center for Biotechnology Information (NCBI; *http://www.ncbi.nlm.nih.gov*). CTD will store all nucleotide reference sequences for human (*Homo sapiens*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), and fruitfly (*Drosophila melanogaster*), thereby providing a nonredundant set of sequences for these particular species (Pruitt and Maglott 2001). All nucleotide sequences for other vertebrates and invertebrates will be loaded from GenBank (*http://www.ncbi.nlm.nih.gov/Sitemap/index.html* GenBank; Benson et al. 2002). Protein sequences for the corresponding organisms will be acquired from SWISS-PROT (*http://ca.expasy.org/sprot/*), which provides a comprehensive, annotated, and nonredundant protein sequence data set (O'Donovan et al. 2002). Direct submissions of sequence data to CTD will not be accepted to avoid duplication of information loaded from GenBank and SWISS-PROT. Information will be updated from these databases frequently to ensure that CTD remains current and comprehensive.

During phase I, references associated with genes and proteins will be identified from GenBank and SWISS-PROT sequence records and the NCBI literature database PubMed (*http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed*). Candidate associations between genes, proteins, and toxic agents will be identified using queries to search the titles, abstracts, and Medical Subject Headings (MeSH) of references (Lipscomb 2000; Young 2002) included in CTD. For queries of genes and proteins, nomenclature inconsistencies will be accounted for initially by including synonyms identified in public biologic databases also addressing this issue, such as Locus Link (*http://www.ncbi.nlm.nih.gov/LocusLink/*) and the Mouse Genome Informatics databases (*http://www.informatics.jax.org/*). Queries for toxic agents will be constructed using a hierarchical vocabulary that will enhance MeSH's Chemicals Index and Chemicals and Drugs category by supplementing it with chemical information from the U.S. EPA, the U.S. Fish and Wildlife Service, and the National Toxicology Program. Criteria for queries will be established in collaboration with investigators from other NIEHS MFBS centers and other investigators from the scientific community with expertise in molecular biology and toxicology. All associations between data sets in CTD will be labeled "not reviewed" until a curator has confirmed their accuracy.

During phase II, we will evaluate and integrate analysis tools for sequence similarity searches (e.g., WU-BLAST) (Altschul et al. 1990), multiple alignments (e.g., ClustalW) (Thompson et al. 1994), and phylogenetic analysis (e.g., PHYLIP) (Felsenstein 1993). Currently, many web sites offer BLAST capabilities against statically defined data sets that include sequences from specific organisms, groups of organisms, or databases. These data sets are often either too inclusive, resulting in an overabundance of "hits," or exclude organisms of interest. By storing sequences and related data locally in a relational database, it will be possible for users to define customized data sets. This capability will permit highly focused sequence analysis, such as restricting BLAST searches to a specific combination of taxa. In addition, large-scale automated sequence analysis will be possible.

During phase III, we will develop a World Wide Web (WWW) interface for CTD that will include user registration and comment forms, basic and advanced query options to access data for sequences, references, and toxic agents, and a platform for analyzing sequences. At the completion of phase III, CTD will be made accessible to collaborators and participating members of the community to evaluate its functionality and test the system. On the basis of feedback from the scientific community, we will then work with MFBS center investigators in subsequent years to continue the data curation process and prioritize the inclusion of additional data sets such as expressed sequence tags, single nucleotide polymorphisms, and data from microarray experiments.

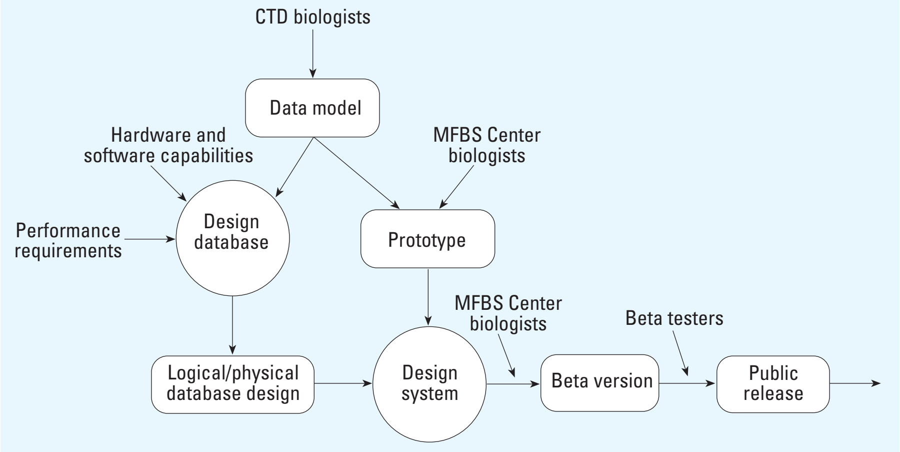

Figure 1. Software development life cycle. The CTD system will be implemented in stages. A data model was designed prior to developing functional specifications and a prototype system. Biologists will evaluate content and functionality throughout the development life cycle.

Implementation. CTD is being designed using a data-driven approach in which the data model is developed prior to specifying system functions (Figure 1). This approach will *a*) promote reusability of data, *b*) establish a consistent set of names and definitions for data, *c*) determine what functions the system will support, and *d*) provide a concise overview of the system's scope (Simsion 1994). CTD will be implemented in an Oracle relational database. The current data model includes 40 entities with well-documented definitions, including text descriptions of all entities and attributes, data types, constraint definitions, and representative values. CTD will include a curation tool and WWW user interface. Oracle Forms Developer will be used to develop the first generation of the curation tool, which will be used to annotate and modify data. This tool is tightly integrated with the Oracle database and provides client-side validation, reusable components, and rapid prototyping capability. The WWW interface will be developed using the Python programming language.

World Wide Web interface. The CTD WWW interface will combine the familiar paradigms of NCBI and Mouse Genome Informatics databases. Simple and advanced query forms will be available to retrieve information about genes, including nucleotide and protein sequences, as well as references, toxic agents, reagents, and taxonomy. Each of these major categories will have a resource page providing a description of associated data and links to resources with supplemental information. Data will be highly integrated within CTD and with external databases.

Community involvement. MDIBL is committed to involving the scientific community in the development of CTD. To this end, we are formally collaborating with investigators at each of the NIEHS MFBS centers; hosting conferences to evaluate the progress and strategic plan of CTD; attending national meetings to promote awareness of and participation in CTD development; and planning online mechanisms for feedback and data submissions. From its inception, CTD has benefited from significant community support. In April 2000, 45 biologists and bioinformatics experts attended a conference at MDIBL (MDIBL 2000) to address the application of bioinformatics in toxicology research. Discussions at this meeting formulated the initial plan for a toxicogenomics database and were the foundation for the NIEHS-phased innovation grant application that now funds CTD. In May 2002 MDIBL hosted a workshop (MDIBL 2002) to promote dialog about genomic databases in the scientific community and to seek feedback about the progress of CTD. Because of the success and utility of these meetings, another conference is planned for 2004.

Community Invitation

To ensure that CTD is a valuable resource for the scientific community, we invite participation in its development. Specific challenges for which we encourage feedback include addressing nomenclature inconsistencies, clustering sequence data from diverse species, and determining the role of microarray data in CTD. Defining strategies to meet these challenges will have broad implications for molecular biologists and toxicologists.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410.

Aller SG, Lombardo ID, Bhanot S, Forrest JN, Jr. 1999. Cloning, characterization, and functional expression of a CNP receptor regulating CFTR in the shark rectal gland. Am J Physiol 276:C442–C449.

Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, et al. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science 297:1301–1310.

Ballatori N, Villalobos AR. 2002. Defining the molecular and cellular basis of toxicity using comparative models. Toxicol Appl Pharmacol 183:207–220.

Baxevanis AD. 2002. The Molecular Biology Database Collection: 2002 update. Nucleic Acids Res 30:1–12.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL. 2002. GenBank. Nucleic Acids Res 30:17–20.

Blake JA, Richardson JE, Bult CJ, Kadin JA, Eppig JT. 2002. The Mouse Genome Database (MGD): the model organism database for the laboratory mouse. Nucleic Acids Res 30:113–115.

Cai SY, Wang W, Ballatori N, Boyer JL. 2001. Bile salt export pump is highly conserved during vertebrate evolution and its expression is inhibited by PFIC type II mutations. Am J Physiol Gastrointest Liver Physiol 281:G316–322.

Cai SY, Wang W, Soroka CJ, Ballatori N, Boyer JL. 2002. An evolutionarily ancient Oatp: insights into conserved functional domains of these proteins. Am J Physiol Gastrointest Liver Physiol 282:G702–G710.

Dooley K, Zon LI. 2000. Zebrafish: a model system for the study of human disease. Curr Opin Genet Dev 10:252–256.

Felsenstein J. 1993. PHYLIP Phylogeny Inference Package 3.5. Seattle, WA:The University of Washington.

Gamba G, Saltzberg SN, Lombardi M, Miyanoshita A, Lytton J, Hediger MA, et al. 1993. Primary structure and functional expression of a cDNA encoding the thiazide-sensitive, electroneutral sodium-chloride cotransporter. Proc Natl Acad Sci USA 90:2749–2753.

Hahn M. 2002. Aryl hydrocarbon receptors: diversity and evolution. Chem Biol Interact 141:131–160.

Koonin EV, Aravind L, Kondrashov AS. 2000. The impact of comparative genomics on our understanding of evolution. Cell 101:573–576.

Lipscomb CE. 2000. Medical Subject Headings (MeSH). Bull Med Libr Assoc 88:265–266.

MDIBL (Mount Desert Island Biological Laboratory). 2002. Conference on Bioinformatics of Genes and ESTs Relevant to Membrane Cellular Toxicology, 28–29 April 2000, Salsbury Cove, ME.

MDIBL (Mount Desert Island Biological Laboratory). 2002. Conference on Community Participation in Genomic Databases, 3–5 May 2002, Salsbury Cove, ME.

O'Donovan C, Martin MJ, Gattiker A, Gasteiger E, Bairoch A, Apweiler R. 2002. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. Brief Bioinform 3:275–284.

Poland A, Knutson JC. 1982. 2,3,7,8-Tetrachlorodibenzo-*p*-dioxin and related halogenated aromatic hydrocarbons: examination of the mechanism of toxicity. Annu Rev Pharmacol Toxicol 22:517–554.

Pruitt KD, Maglott, DR. 2001. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res 29:137–140.

Russom CL. 2002. Mining environmental toxicology information: web resources. Toxicology 173:75–88.

Schmidt JV, Bradfield CA. 1996. Ah receptor signaling pathways. Annu Rev Cell Dev Biol 12:55–89.

Simsion G. 1994. Data Modeling Essentials: Analysis, Design, and Innovation. London:International Thomson Computer Press.

Sprague J, Doerry E, Douglas S, Westerfield M. 2001. The Zebrafish Information Network (ZFIN): a resource for genetic, genomic and developmental research. Nucleic Acids Res 29:87–90.

Thomas RS, Penn SG, Holden K, Bradfield CA, Rank DR. 2002. Sequence variation and phylogenetic history of the mouse *Ahr* gene. Pharmacogenetics 12:151–163.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680.

U.S. EPA. New Chemicals Program. Washington, DC:U.S. Environmental Protection Agency. Available: http://www.epa.gov/opptintr/newchems/invntory.htm [accessed 27 January 2003].

Waters M, Boorman G, Bushel P, Cunningham M, Irwin R, Merrick A, et al. 2003. Systems Toxicology and the Chemical Effects in Biological Systems (CEBS) Knowledge Base. Environ Health Perspect 111:811–824 (2003).

Wexler P. 2001. TOXNET: an evolving web resource for toxicology and environmental health information. Toxicology 157:3–10.

Whelan S, Lio P, Goldman N. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends Genet 17:262–272.

Wittbrodt J, Shima A, Schartl M. 2002. Medaka—a model organism from the Far East. Nat Rev Genet 3:53–64.

Xu JC, Lytle C, Zhu TT, Payne JA, Benz E, Forbush B. 1994. Molecular cloning and functional expression of the bumetanide-sensitive Na-K-Cl cotransporter. Proc Natl Acad Sci USA 91:2201–2205.

Young RR. 2002. Genetic toxicology: web resources. Toxicology 173:103–121.