D786–D792 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 9 September 2008*
doi:10.1093/nar/gkn580

Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical–gene–disease networks

Allan Peter Davis, Cynthia G. Murphy, Cynthia A. Saraceni-Richards, Michael C. Rosenstein, Thomas C. Wiegers and Carolyn J. Mattingly*

Department of Bioinformatics, The Mount Desert Island Biological Laboratory, Salisbury Cove, ME 04672, USA

Received June 7, 2008; Revised August 26, 2008; Accepted August 27, 2008

ABSTRACT

The Comparative Toxicogenomics Database (CTD) is a curated database that promotes understanding about the effects of environmental chemicals on human health. Biocurators at CTD manually curate chemical–gene interactions, chemical–disease relationships and gene–disease relationships from the literature. This strategy allows data to be integrated to construct chemical–gene–disease networks. CTD is unique in numerous respects: curation focuses on environmental chemicals; interactions are manually curated; interactions are constructed using controlled vocabularies and hierarchies; additional gene attributes (such as Gene Ontology, taxonomy and KEGG pathways) are integrated; data can be viewed from the perspective of a chemical, gene or disease; results and batch queries can be downloaded and saved; and most importantly, CTD acts as both a knowledgebase (by reporting data) and a discovery tool (by generating novel inferences). Over 116 000 interactions between 3900 chemicals and 13 300 genes have been curated from 270 species, and 5900 gene–disease and 2500 chemical–disease direct relationships have been captured. By integrating these data, 350 000 gene–disease relationships and 77 000 chemical–disease relationships can be inferred. This wealth of chemical–gene–disease information yields testable hypotheses for understanding the effects of environmental chemicals on human health. CTD is freely available at http://ctd.mdibl.org.

INTRODUCTION

Environmental agents are postulated to play a critical role in the etiology of many human diseases (1–4), and chemicals are an important component of the environment. To understand the impact of environmental chemicals on human health, we have developed the Comparative Toxicogenomics Database (CTD; http://ctd.mdibl.org) as a unique tool to provide connections between chemicals, genes/proteins and diseases that may not otherwise be apparent, and to provide the basis for testable hypotheses about the mechanisms underlying the etiology of environmental diseases (5–7).

Several valuable chemical, gene and disease databases currently exist. Each one has its advantages. Many public chemical databases, such as PharmGKB (8), DrugBank (9), ChemBank (10) and STITCH (11) focus on drugs and other small molecules, providing an invaluable resource for therapeutic research. There are several microarray resources that provide varying degrees of data for chemicals, genes and diseases. Chemical Effects in Biological Systems (CEBS) (12) is a public repository and tool for chemically relevant microarray, proteomics, clinical chemistry, hematology and histopathology data. ArrayExpress (13) and Gene Expression Omnibus (GEO) (14) are public repositories for microarray data. Although the latter contain chemically relevant data, these data are not their expressed priority. ArrayTrack (15) is an installable application and database for managing and analyzing microarray data. Currently, only users at the US Food and Drug Administration (FDA) may submit their data; however, non-FDA users have access to ArrayTrack functionality. ChEBI (16) is an excellent dictionary for chemical entities, but outsources its information on the biology of those chemicals to other databases via external links. PubChem (14) is a repository of chemical substance information, compound structures and biological activities of small molecules, but does not integrate that data with official gene symbols or disease information. OMIM (17) and HGMD (18), two of the most commonly cited disease databases, annotate genetic diseases, but do not provide any associated chemical information. Some gene databases, such as GeneCards (19) and PubGene (20), have recently included gene–chemical

*To whom correspondence should be addressed. Tel: +1 207 288 3605; Fax: +1 207 288 2130; Email: cmattin@mdibl.org

associations, but those relationships are established via text-mining algorithms and are not reviewed or validated by professional biocurators. KEGG (21) and Reactome (22) map chemicals, genes and (in the case of KEGG) disease information to pathways, but the pathways and interactions are generically applied to orthologous proteins and all species, and it is not always clear which reference supports which pathway relationship. CTD is distinct from these databases in three ways: (i) it focuses on environmental chemicals; (ii) it integrates curated and imported data, allowing users to explore connections between chemicals, genes, and diseases; and (iii) it functions not only as a repository for information, but also as a resource for generating novel hypotheses about environmental diseases and chemical actions.

ENVIRONMENTAL CHEMICALS AND DISEASE

It is becoming well established that environmental agents influence chronic disease susceptibility (23). There are numerous types of environmental agents, including infectious agents (bacteria, viruses and parasites), diet, radiation and chemicals. One way that chemicals might influence diseases is by interacting with genes and proteins. Environmental chemicals can affect genes in multiple, nonexclusive ways, including mutagenesis (24), altered methylation (3), physical interaction (25) and influencing gene expression or protein function. Conversely, naturally occurring genetic polymorphisms may affect chemical susceptibility and result in increased disease predisposition (26). To help understand the complex effects of the environment on human health, CTD focuses its manual curation effort on environmental chemicals (e.g. arsenic, heavy metals and dioxins), how those chemicals interact with genes or proteins in different species and how they relate to human diseases.

MANUAL CURATION

CTD biocurators capture three types of core data from the literature: chemical–gene (and protein) interactions, chemical–disease relationships and gene–disease relationships. These data are curated in a structured format using controlled vocabularies and are integrated to establish a triad of chemicals, genes and diseases (Figure 1a, Table 1).

A major strength of CTD is that these core data are manually curated from the literature by professional biocurators (27), ensuring accuracy. CTD does use text mining to triage the literature, but each reference (abstract or full-text) is read by a biocurator to identify interactions and relationships, and all curated data is supported by its source citation. Some databases rely solely on text mining and report interactions based on co-occurrences of a chemical and gene in a document. However, this method has several limitations: co-occurrence of terms does not always imply a valid chemical–gene interaction; chemical names and gene symbols are challenging to text mine accurately because of their many synonyms and correspondence with common words (e.g. 'lead', 'find', 'up', 'for', 'a'); and to date, text-mining tools have not accommodated types of molecular interactions. The manual curation approach at CTD allows biocurators to validate every interaction and relationship, ensure that the correct

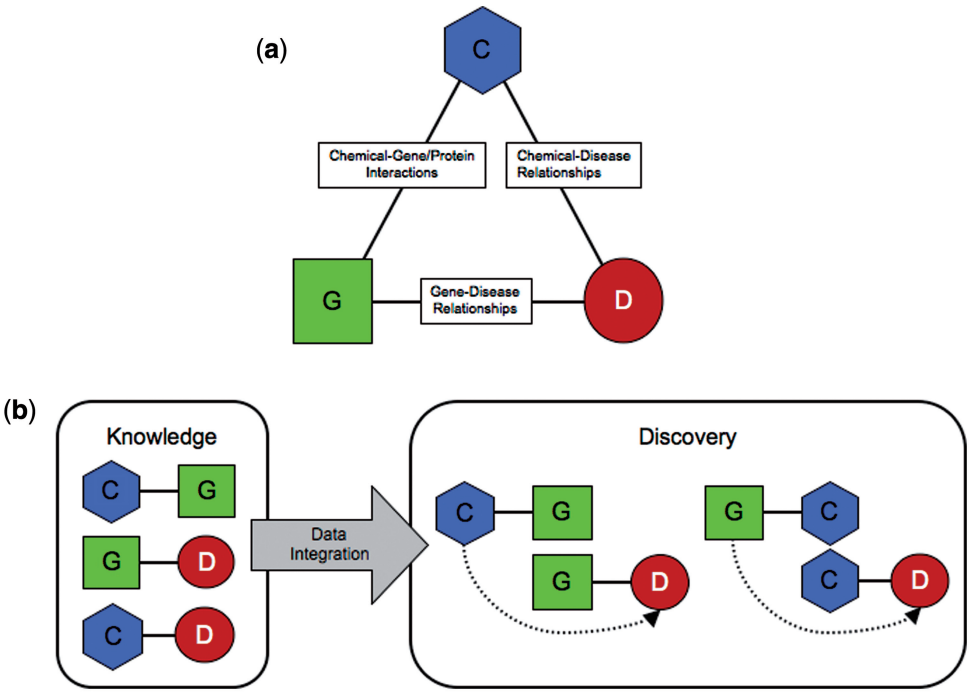

Figure 1. CTD curation and integration paradigm. (**a**) Biocurators capture three types of data from the literature for chemicals (C), genes (G) and diseases (D): C–G interactions, G–D relationships and C–D relationships. These three relationships generate a chemical–gene–disease triad. (**b**) The integration of these three data sets enables users to infer novel connections between chemicals–diseases and genes–diseases (dashed arrow).

chemical name and gene symbol is used, and generate detailed descriptions of the types of interaction. Data are uploaded to the database monthly.

The use of controlled vocabularies provides numerous advantages: the curation process is streamlined, different biocurators capture data in a consistent manner, users retrieve data reproducibly and quality control is feasible. The hierarchical structure of chemical and disease vocabularies in CTD also enable users to query data using general (e.g. heterocyclic compounds) or specific (e.g. 2-hydroxytacrine) terms. The following vocabularies are used by biocurators for curation and are integrated in the database to facilitate querying:

(1) *Chemicals*. The CTD chemical vocabulary was derived from a modified subset of the chemicals and supplementary concepts in the 'Drugs and Chemicals' category of Medical Subject Headings (MeSH) from the National Library of Medicine (28).
(2) *Chemical qualifiers*. A chemical in a curated interaction can be qualified as an 'analog' or 'metabolite' to describe a chemical derivative (e.g. arsenic disulfide analog or benzo[a]pyrene metabolite).
(3) *Genes*. CTD uses official gene symbols and names from the National Center for Biotechnology Information's (NCBI) Entrez-Gene database (14).
(4) *Gene qualifiers*. The form of a gene can be specified with one of 15 gene qualifiers (e.g. DNA, promoter and mRNA, protein).
(5) *Actions*. CTD curators developed a hierarchical vocabulary of 50 diverse terms (e.g. binding, phosphorylation, activity) to describe specific molecular interactions between a chemical and gene. A complete list of action codes with their definitions is available via the Help link for interactions on the CTD glossary page or query pages. Chemical–gene–disease relationships are qualified as molecular marker/mechanism or therapeutic.
(6) *Diseases*. The CTD disease vocabulary comprises terms from the disease subset of MeSH (28) and OMIM (17). OMIM terms were mapped to a single equivalent term from MeSH whenever possible (e.g. OMIM's Lung Cancer maps directly to MeSH's Lung Neoplasm). OMIM diseases were mapped to multiple MeSH terms when a single equivalent term was not available (e.g. OMIM's Chronic Myeloproliferative Disorder with Eosinophilia was mapped to MeSH's Myeloproliferative Disorders and Eosinophilia terms). This mapping enables users to retrieve data for specific diseases (e.g. Papillon-Lefevre Syndrome and Cafe-au-Lait Spots) or related groups of diseases (e.g. skin diseases).
(7) *Organisms*. The CTD organism vocabulary consists of the Eumetazoa portion (vertebrates and invertebrates) of the NCBI Taxonomy database (14).
(8) *References*. All curated data are derived from literature in PubMed and are associated with a unique PubMed identifier (14).

Table 1. CTD curated data status

Curated data	Number[a]
Chemical–gene interactions	116 067
Chemicals	3971
Genes	13 300
Direct gene–disease relationships	5925
Inferred gene–disease relationships	349 300
Direct chemical–disease relationships	2569
Inferred chemical–disease relationships	76 970

[a]As of July 2008.

EXTERNAL DATA INTEGRATION

Community-accepted controlled vocabularies and identification numbers allow integration with other databases that use the same terms. CTD enhances its core data pages (Chemical, Gene and Disease) with links to the following external resources (data are updated monthly):

(1) *CTD Chemical pages*. Enhanced with chemical structures from ChemIDPlus (29); chemical reports from Chemical Carcinogenesis Research Information System, GENE-TOX and Hazardous Substances Data Bank (30); and links to DrugBank (9), MeSH (28) and TOXLINE (30) for online literature searching.
(2) *CTD Gene pages*. Enhanced with Gene Ontology (GO) annotations (31), KEGG pathways (21), nucleotide and amino acid sequences from UniProt (32), DDBJ (33), EMBL (34) and GenBank (35); links to NCBI Entrez-Gene pages (14); and microarray reports from the EDGE database (36). Protein sequence pages (associated with a Gene page) are, in turn, additionally linked to records from GenPept (14), InterPro (37), PRINTS (38), PROSITE (39), ProDom (40), SMART (41), Pfam (42) and, when appropriate, to a species-specific gene page in FlyBase (43) or ZFIN (44).
(3) *CTD Disease pages*. Enhanced with KEGG pathways (21), and definitions and synonyms from MeSH (28) and OMIM (17).

TURNING KNOWLEDGE INTO DISCOVERIES

A powerful feature of CTD is the integration of curated chemical, gene and disease core data from the literature (knowledge) to generate new, putative discoveries (Figure 1b, Table 1). For example, if chemical A interacts with gene B (via a curated chemical–gene interaction) and independently gene B is associated with disease C (via a curated gene–disease relationship), then it may be inferred or hypothesized that chemical A has a relationship with disease C (inferred via gene B). This integration provides possible chemical–gene–disease connections that may not otherwise be apparent.

The molecular basis of most environmental diseases is still not clear. CTD can act as a discovery tool to generate testable hypotheses about the mechanisms underlying the etiology of environmental diseases. This approach was

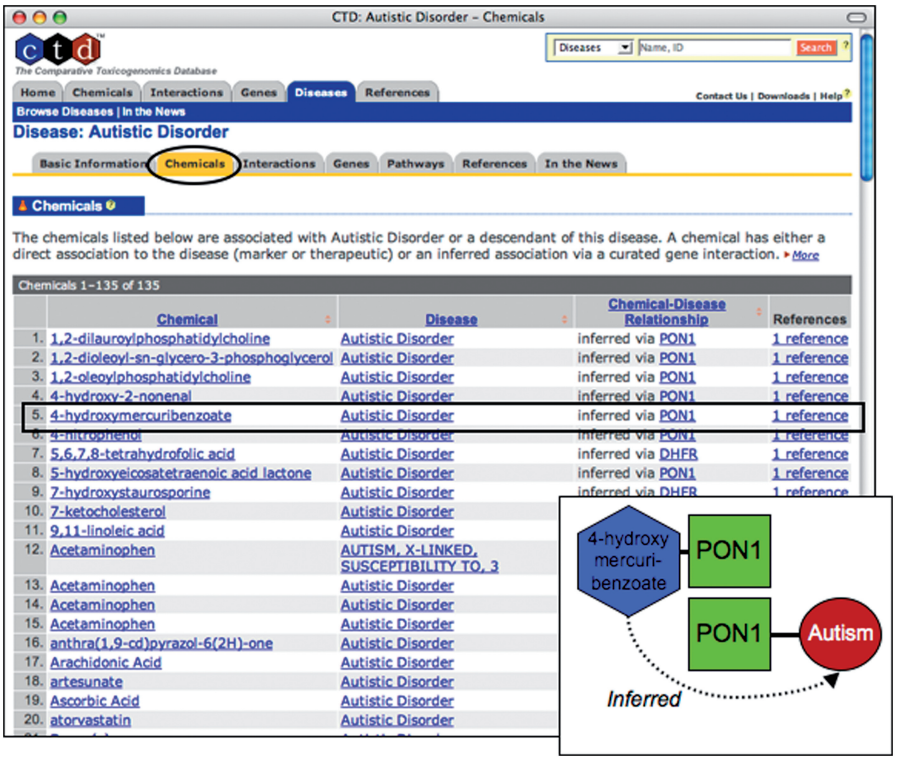

Figure 2. Discovering putative chemical–gene–disease networks for autism. The Chemical tab (red circle) on the CTD Disease page for Autistic Disorder displays chemicals (e.g. 4-hydroxymercuribenzoate) with an inferred connection to the disease based upon interaction with a gene already known to be associated with autism (e.g. PON1), allowing a putative, novel, chemical–gene–disease network to be proposed (insert). The cited reference (red box) will take the user to a page that provides a link to the curated interactions between 4-hydroxymercuribenzoate and PON1 as well as the reference describing the PON1–Autistic Disorder relationship.

recently supported by analyzing the CTD arsenic data set, wherein CTD correctly predicted types of diseases that may be associated with arsenic exposure and set of genes that may be involved in modulating arsenic-related diseases, such as lung cancer and diabetes. A similar analysis can be applied to any environmental chemical or disease. For example, chemical–gene networks connecting environmental exposure to autism can be discovered by going to the CTD disease page for autism (Autistic Disorder) and clicking on the 'Chemicals' tabs (Figure 2). Here, users will find a list of chemicals that interact with genes known to be associated with autism, generating a hypothetical chemical–gene–disease network for the disease. These predictions of environmental exposure can now be addressed and tested in the laboratory.

Since the chemical–gene–disease triad connects all nodes with curated edges (Figure 1a), a user can explore CTD by their chemical, gene or disease of interest and discover a novel connection to any of the other two nodes. This makes CTD a valuable discovery tool for any laboratory studying chemistry, genetics or human health.

QUERYING CTD

Users have several options for querying data in CTD. A keyword search box appears on every CTD page (Figure 3). This box contains a pick-list to allow queries of chemicals, genes, diseases, GO terms, organisms or references. Keywords may include terms, symbols or accession IDs, and Boolean operators are supported.

Users can also perform detailed searches using Gene, Interaction and Reference Query pages. Many terms associated with curated and imported data can be used as search parameters. For example, queries may include GO annotations, KEGG pathways, chemical classes, types of chemical interactions, associated diseases and organisms to ask questions, such as: polychlorinated biphenyls affect the activity of which transcription factors? What proteins involved in limb development interact with a heavy metal? What chemicals downregulate members of the glycine metabolism pathway? This detailed querying option allows users to find data beyond the limits of a specific chemical, gene or disease term, and start to analyze data from the perspective of broader biological concepts and systems.

A new batch query tool allows users to download data associated with lists of chemicals, genes or diseases. Users choose the type of results to retrieve, which include curated chemical–gene interactions, curated chemical or gene associations, disease relationships, GO associations or pathway associations. This feature provides important biological insights into groups of chemicals, genes or diseases such as: what is the predominant molecular function associated with the 863 genes that interact with paraquat, or what disease is most commonly associated with this list of heavy metals? Batch query results can be downloaded

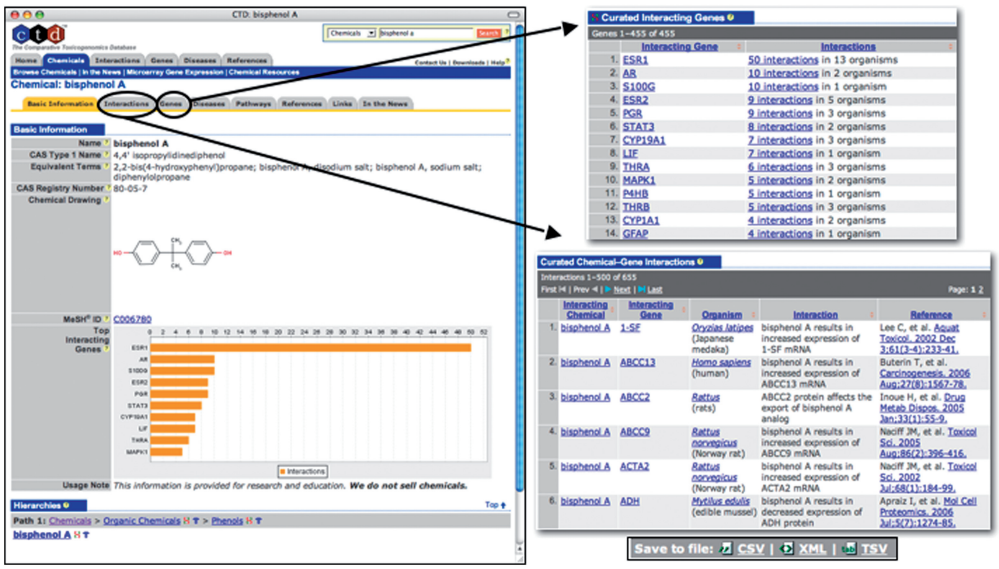

Figure 3. CTD Chemical page for bisphenol A. CTD uses tabs to organize data. The Basic Information tab provides names, synonyms, identification numbers, structures, a graph of the top 10 interacting genes and the chemical hierarchy. The Gene tab shows all 455 genes that interact with bisphenol A, the number of curated interactions between the chemical and gene, and the number of unique organisms for which an interaction has been curated. The Interactions tab displays the 655 detailed molecular interactions between bisphenol A and those 455 genes. The table is divided into columns for Interacting Chemical, Interacting Gene, Organism, Interaction and Reference. Columns can be sorted by clicking on their headers. All chemicals, genes, interactions, references and diseases (data not shown) are hyperlinked to their respective CTD pages, allowing the user to navigate integrated data. Complete data pages can be downloaded and saved as CSV, XML or TSV files onto the user's desktop by using the 'Save to File' function at the bottom of the page. The keyword search box is seen in the upper right corner of the bisphenol A Chemical page.

in CSV (comma-separated values), TSV (tab-separated values) or XML (extensible markup language) format.

VIEWING DATA

The curation and integration paradigm in CTD allows users to explore data from the perspective of a chemical, gene or disease. All chemical, gene and disease terms are hyperlinked to respective detail pages, which organize associated data on tabbed pages (Figure 3). The data for each tabbed page is presented in a table, which can be sorted by columns or downloaded in CSV, TSV or XML format. The location and content of tabbed data pages in CTD are described:

(1) *Basic information*. Chemical, Gene and Disease pages. Lists the official names, symbols, synonyms, chemical structures (where applicable), accession identifiers (IDs), definitions and links to external resources. Chemical and Disease pages also show the position of the term in the corresponding hierarchical vocabulary, allowing users to adjust the specificity of the data they are viewing. Chemical and Gene pages display a graph of the top 10 interacting genes or chemicals, respectively.
(2) *Interactions*. Chemical, Gene and Disease pages. Lists curated chemical–gene interactions.
(3) *Genes*. Chemical and Disease pages. Lists the interacting genes for chemicals. Lists associated genes and type of association (direct or inferred) for diseases.
(4) *Chemicals*. Gene and Disease pages. Lists the interacting chemicals for genes. Lists associated chemicals and type of association (direct or inferred) for diseases.
(5) *Diseases*. Chemical and Gene pages. Lists the associated disease and type of association (direct or inferred).
(6) *Pathways*. Chemical, Gene and Disease pages. Lists and provides links to associated KEGG pathways. On Chemical pages, KEGG associations are inferred via interacting genes.
(7) *GO*. Gene pages. Lists the GO annotations and their sources.
(8) *Sequences*. Gene pages. Lists the protein and nucleic acid sequences for species within the Eumetazoa portion (vertebrates and invertebrates) of the NCBI Taxonomy database (14).
(9) *References*. Chemical, Gene and Disease pages. Lists pertinent literature, and indicates which papers have been manually curated. Cites the chemicals, genes and diseases from each reference.
(10) *In the news*. Chemical and Disease pages. Lists relevant, current articles from the mainstream media.
(11) *Links*. Chemical pages. Provides links to other chemical databases.

SUMMARY AND FUTURE DIRECTIONS

CTD is a unique scientific resource that promotes understanding about the effects of environmental chemicals on human health. It provides chemical–gene interactions, chemical–disease relationships and gene–disease relationships that are manually curated by biocurators using controlled vocabularies. By integrating these core data, CTD

functions as a discovery tool for identifying connections between chemicals, genes and diseases not otherwise apparent in other biological resources, and for generating testable hypotheses about the mechanisms underlying the etiology of environmental diseases.

Future development of CTD will aim to further expand the depth of its curated data and enhance the data query and visualization capabilities. Specifically, text-mining tools will be incorporated to increase the efficiency of manual data curation, the number of databases to which CTD is reciprocally linked will increase to improve integration of relevant databases in the public domain and additional query and data visualization strategies will be explored to introduce graphical representations of complex data relationships (e.g. chemical–gene–protein interactions and chemical–gene–disease relationships). CTD will continue to be publicly available and the community is encouraged to contact us with comments and suggestions so that we may continue to enhance its value.

FUNDING

This work is supported by the National Institute of Environmental Health Sciences (ES014065) and the INBRE program of the National Center for Research Resources (RR016463) of the National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Brody,J.G., Moysich,K.B., Humblet,O., Attfield,K.R., Beehler,G.P. and Rudel,R.A. (2007) Environmental pollutants and breast cancer: epidemiologic studies. *Cancer*, **109**, 2667–2711.
2. Clavel,J. (2007) Progress in the epidemiological understanding of gene-environment interactions in major diseases: cancer. *C. R. Biol.*, **330**, 306–317.
3. Dolinoy,D.C. and Jirtle,R.L. (2008) Environmental epigenomics in human health and disease. *Environ. Mol. Mutagen*, **49**, 4–8.
4. Schwartz,D. and Collins,F. (2007) Medicine. Environmental biology and human disease. *Science*, **316**, 695–696.
5. Mattingly,C.J., Colby,G.T., Forrest,J.N. and Boyer,J.L. (2003) The comparative toxicogenomics database (CTD). *Environ. Health Perspect.*, **111**, 793–795.
6. Mattingly,C.J., Rosenstein,M.C., Colby,G.T., Forrest,J.N. Jr and Boyer,J.L. (2006) The Comparative toxicogenomics database (CTD): a resource for comparative toxicological studies. *J. Exp. Zoolog. A Comp. Exp. Biol.*, **305**, 689–692.
7. Mattingly,C.J., Rosenstein,M.C., Davis,A.P., Colby,G.T., Forrest,J.N. Jr and Boyer,J.L. (2006) The comparative toxicogenomics database: a cross-species resource for building chemical-gene interaction networks. *Toxicol. Sci.*, **92**, 587–595.
8. Klein,T.E., Chang,J.T., Cho,M.K., Easton,K.L., Fergerson,R., Hewett,M., Lin,Z., Liu,Y., Liu,S., Oliver,D.E. *et al.* (2001) Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics research network and knowledge base. *Pharmacogenomics J.*, **1**, 167–170.
9. Wishart,D.S., Knox,C., Guo,A.C., Cheng,D., Shrivastava,S., Tzur,D., Gautam,B. and Hassanali,M. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, **36**, D901–D906.
10. Seiler,K.P., George,G.A., Happ,M.P., Bodycombe,N.E., Carrinski,H.A., Norton,S., Brudz,S., Sullivan,J.P., Muhlich,J., Serrano,M. *et al.* (2008) ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.*, **36**, D351–D359.
11. Kuhn,M., von Mering,C., Campillos,M., Jensen,L.J. and Bork,P. (2008) STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.*, **36**, D684–D688.
12. Waters,M., Stasiewicz,S., Merrick,B.A., Tomer,K., Bushel,P., Paules,R., Stegman,N., Nehls,G., Yost,K.J., Johnson,C.H. *et al.* (2008) CEBS–Chemical effects in biological systems: a public data repository integrating study design and toxicity data with microarray and proteomics data. *Nucleic Acids Res.*, **36**, D892–D900.
13. Parkinson,H., Kapushesky,M., Shojatalab,M., Abeygunawardena,N., Coulson,R., Farne,A., Holloway,E., Kolesnykov,N., Lilja,P., Lukk,M. *et al.* (2007) ArrayExpress–a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.*, **35**, D747–D750.
14. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al.* (2008) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **36**, D13–D21.
15. Tong,W., Cao,X., Harris,S., Sun,H., Fang,H., Fuscoe,J., Harris,A., Hong,H., Xie,Q., Perkins,R. *et al.* (2003) ArrayTrack–supporting toxicogenomic research at the U.S. food and drug administration national center for toxicological research. *Environ. Health Perspect.*, **111**, 1819–1826.
16. Degtyarenko,K., de Matos,P., Ennis,M., Hastings,J., Zbinden,M., McNaught,A., Alcantara,R., Darsow,M., Guedj,M. and Ashburner,M. (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, **36**, D344–D350.
17. McKusick,V.A. (2007) Mendelian inheritance in man and its online version, OMIM. *Am. J. Hum. Genet.*, **80**, 588–604.
18. Stenson,P.D., Ball,E.V., Mort,M., Phillips,A.D., Shiel,J.A., Thomas,N.S., Abeysinghe,S., Krawczak,M. and Cooper,D.N. (2003) Human gene mutation database (HGMD): 2003 update. *Hum. Mutat.*, **21**, 577–581.
19. Safran,M., Chalifa-Caspi,V., Shmueli,O., Olender,T., Lapidot,M., Rosen,N., Shmoish,M., Peter,Y., Glusman,G., Feldmesser,E. *et al.* (2003) Human gene-centric databases at the Weizmann institute of science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res.*, **31**, 142–146.
20. Jenssen,T.K., Laegreid,A., Komorowski,J. and Hovig,E. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.*, **28**, 21–28.
21. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
22. Vastrik,I., D'Eustachio,P., Schmidt,E., Joshi-Tope,G., Gopinath,G., Croft,D., de Bono,B., Gillespie,M., Jassal,B., Lewis,S. *et al.* (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.*, **8**, R39.
23. Olden,K. and Wilson,S. (2000) Environmental health and genomics: visions and implications. *Nat. Rev. Genet.*, **1**, 149–153.
24. Dixon,K. and Kopras,E. (2004) Genetic alterations and DNA repair in human carcinogenesis. *Semin. Cancer Biol.*, **14**, 441–448.
25. Delaney,J.C. and Essigmann,J.M. (2008) Biological properties of single chemical-DNA adducts: a twenty year perspective. *Chem. Res. Toxicol.*, **21**, 232–252.
26. Gonzalez,F.J. and Gelboin,H.V. (1993) Role of human cytochrome P-450s in risk assessment and susceptibility to environmentally based disease. *J. Toxicol. Environ. Health*, **40**, 289–308.
27. Salimi,N. and Vita,R. (2006) The biocurator: connecting and enhancing scientific data. *PLoS Comput. Biol.*, **2**, e125.
28. Sewell,W. (1964) Medical Subject Headings in Medlars. *Bull. Med. Libr. Assoc.*, **52**, 164–170.
29. Tomasulo,P. (2002) ChemIDplus-super source for chemical and drug information. *Med. Ref. Serv. Q*, **21**, 53–59.
30. Wexler,P. (2004) The U.S. national library of medicine's toxicology and environment health information program. *Toxicology*, **198**, 161–168.
31. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.*, **25**, 25–29.
32. The Uniprot Consortium. (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190–D195.

33. Sugawara,H., Ogasawara,O., Okubo,K., Gojobori,T. and Tateno,Y. (2008) DDBJ with new system and face. *Nucleic Acids Res.*, **36**, D22–D24.
34. Kanz,C., Aldebert,P., Althorpe,N., Baker,W., Baldwin,A., Bates,K., Browne,P., van den Broek,A., Castro,M., Cochrane,G. *et al.* (2005) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **33**, D29–D33.
35. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2008) GenBank. *Nucleic Acids Res.*, **36**, D25–D30.
36. Hayes,K.R., Vollrath,A.L., Zastrow,G.M., McMillan,B.J., Craven,M., Jovanovich,S., Rank,D.R., Penn,S., Walisser,J.A., Reddy,J.K. *et al.* (2005) EDGE: a centralized resource for the comparison, analysis, and distribution of toxicogenomic information. *Mol. Pharmacol.*, **67**, 1360–1368.
37. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Buillard,V., Cerutti,L., Copley,R. *et al.* (2007) New developments in the InterPro database. *Nucleic Acids Res.*, **35**, D224–D228.
38. Attwood,T.K., Bradley,P., Flower,D.R., Gaulton,A., Maudling,N., Mitchell,A.L., Moulton,G., Nordle,A., Paine,K., Taylor,P. *et al.* (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.*, **31**, 400–402.
39. Hulo,N., Bairoch,A., Bulliard,V., Cerutti,L., Cuche,B.A., de Castro,E., Lachaize,C., Langendijk-Genevaux,P.S. and Sigrist,C.J. (2008) The 20 years of PROSITE. *Nucleic Acids Res.*, **36**, D245–D249.
40. Servant,F., Bru,C., Carrere,S., Courcelle,E., Gouzy,J., Peyruc,D. and Kahn,D. (2002) ProDom: automated clustering of homologous domains. *Brief Bioinform.*, **3**, 246–251.
41. Letunic,I., Copley,R.R., Schmidt,S., Ciccarelli,F.D., Doerks,T., Schultz,J., Ponting,C.P. and Bork,P. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res.*, **32**, D142–D144.
42. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
43. Wilson,R.J., Goodman,J.L. and Strelets,V.B. (2008) FlyBase: integration and improvements to query tools. *Nucleic Acids Res.*, **36**, D588–D593.
44. Sprague,J., Bayraktaroglu,L., Bradford,Y., Conlin,T., Dunn,N., Fashena,D., Frazer,K., Haendel,M., Howe,D.G., Knight,J. *et al.* (2008) The Zebrafish information network: the zebrafish model organism database provides expanded support for genotypes and phenotypes. *Nucleic Acids Res.*, **36**, D768–D772.