ACToR – Aggregated Computational Toxicology Resource

Richard Judson[1,3], Ann Richard[1], David Dix[1], Keith Houck[1], Fathi Elloumi[1], Matthew Martin[1], Tommy Cathey[2], Thomas R. Transue[2], Richard Spencer[2], Maritja Wolf[2]

(1) National Center for Computational Toxicology, U.S. Environmental Protection Agency, 109 T.W. Alexander Drive, Research Triangle Park NC 27711

(2) Lockheed Martin, A Contractor to the U.S. Environmental Protection Agency, 109 T.W. Alexander Drive, Research Triangle Park NC 27711

(3) Corresponding Author: judson.richard@epa.gov, 919-541-3085

ABSTRACT

ACToR (Aggregated Computational Toxicology Resource) is a database and set of software applications that bring into one central location many types and sources of data on environmental chemicals. Currently, the ACToR chemical database contains information on chemical structure, *in vitro* bioassays and *in vivo* toxicology assays derived from more than 150 sources including the U.S. Environmental Protection Agency (EPA), Centers for Disease Control (CDC), U.S. Food & Drug Administration (FDA), National Institutes of Health (NIH), state agencies, corresponding government agencies in Canada, Europe and Japan, universities, the World Health Organization (WHO) and non-governmental organizations (NGOs). At the EPA National Center for Computational Toxicology, ACToR helps manage large data sets being used in a high throughput environmental chemical screening and prioritization program called ToxCast™.

INTRODUCTION

Computational Toxicology is an emerging field that aims to use modern computational and molecular biology techniques to understand and predict chemical toxicity. A particular area where this approach is being applied is in chemical screening and prioritization. In the U.S., there are an estimated 30,000 unique chemicals in wide commercial use (>1 ton/yr)(Muir and Howard, 2006), and only a relatively small subset of these have been sufficiently well characterized for their potential to cause human or ecological toxicity to support regulatory action. This "data gap" is well documented (EPA, 1998; Allanou *et al.*, 1999; Birnbaum *et al.*, 2003; Guth *et al.*, 2005; Applegate and Baer, 2006; Krewski *et al.*, 2007). The standard approach to determine a chemical's toxicity profile involves performing *in vivo* studies on rodents and other species, and can take 2-3 years and cost millions of dollars per chemical. Clearly, this strategy is neither practical nor viable for evaluating tens of thousands of chemicals; hence, the large inventories of existing chemicals for which little or no test data are available. An alternative approach is to attempt to cover much larger regions of chemical space by employing more efficient *in vitro* methods. One strategy applies relatively inexpensive and rapid high throughput screening (HTS) assays to a large set of chemicals, followed by the use of these results to prioritize a much smaller subset of chemicals for more detailed analysis. The "prioritization score" for a chemical would be based on derived signatures, or patterns extracted from the HTS data, which are predictive of particular effects or modes of chemical toxicity. Chemicals of known toxicity comprise the reference or training set that is used to develop and validate predictive signatures. HTS assays that yield data for the predictive signatures would then be run on chemicals of unknown toxicity (the test chemicals), and a prioritization score for those chemicals would be produced. The U.S. EPA has made a significant investment in this approach through the recent launch of the ToxCast^{TM} research program (Dix *et al.*, 2007). ToxCast is screening hundreds, and eventually thousands of environmental chemicals using hundreds of HTS assays towards the two goals of developing predictive toxicity signatures, and using these signatures to prioritize chemicals for further testing. In this EPA context, the term "environmental chemicals" refers primarily to industrial chemicals

and pesticides used or produced in large enough quantities to pose significant potential for human or ecological exposure.

There are multiple computational aspects to this approach. First, some of the screening assays themselves may be computational (*in silico*). Second, a robust database and data analysis infrastructure are required to manage the large data volumes produced by a large-scale HTS program. Third, one needs high quality *in vivo* toxicology data on as large and diverse group of chemicals as possible in order to develop and validate the predictive signatures. Currently, such toxicity data are available from a number of sources, but these data are widely dispersed and often not sufficiently annotated or fully accessible for computational use.

To support the EPA's ToxCast screening and prioritization effort, as well as other EPA programs, we are developing a system called ACToR, for Aggregated Computational Toxicology Resource. ACToR is a set of linked databases and software applications that bring together many types and sources of data on environmental chemicals into one central location. Currently, the ACToR chemical and assay databases contain information on chemical structure, *in vitro* bioassays and *in vivo* toxicology assays derived from more than 150 sources including the EPA, CDC, FDA, NIH, state agencies, corresponding government agencies in Canada, Europe and Japan, universities, the World Health Organization and NGOs. An important set of data collections come from the DSSTox project (Distributed Structure-Searchable Toxicity)(Richard and Williams, 2002) at the EPA which produces curated collections of chemical structures with corresponding assay data. The design of ACToR has followed that of the NIH PubChem Project in many respects, but has been generalized to allow for the broader types of data that are of interest to toxicologists and environmental regulators. The current ACToR web interface is also designed to meet the needs of scientists focused on the study of chemical toxicity.

This paper briefly outlines the design of the ACToR database and the types of data it contains, and will illustrate its utility in the context of developing training and validation data sets for chemical screening and prioritization projects.

MATERIALS AND METHODS

Organization of the Database: The current version of ACToR is focused mainly on capturing information on chemicals and assays of chemical-biological effects. Plans are underway to extend this to capture relevant genomic and biological pathway information. The organizing principles for the design of the chemical/assay system are largely derived from the PubChem project, which is capturing chemical structure and HTS information on millions of chemicals in its role as the main data repository for the NIH Molecular Libraries Roadmap (Austin *et al.*, 2004). The main organizing principle of PubChem centers on the three main types of data that are catalogued: substances indexed by substance identifier (SID), compounds (i.e., chemical structures, indexed by compound identifier (CID), and bioassays indexed by assay identifier (AID). A PubChem substance is a single chemical entity submitted by one data source and often corresponds to the physical substance on which some experiment was performed. A compound is a generic chemical entity that corresponds to a unique chemical structure. Since a substance is defined as being both data source and experiment-specific, many substances (SIDs) may map to a single compound (CID). A bioassay, indexed by AID, represents a specific type of test data associated with one or more substances.

In ACToR, these ideas are generalized somewhat, although the model is close enough such that all data from PubChem can be easily loaded into ACToR. In ACToR, a substance is similarly defined as a unique chemical from a single “data collection” (see below) and is minimally characterized by a data collection-specific SID and a chemical name. Most often, the substance will also have synonyms, a CAS (Chemical Abstracts Service) registry number (CASRN) and multiple other parameters. A compound always has an associated chemical structure and a data collection-specific CID, in addition to optional parameters derived directly from chemical structure, such as SMILES and InChI

representations and a molecular weight. Note that since ACToR is in essence a "super-aggregator", pulling in large external data collections such as PubChem, it also stores the source-labeled CIDs from each independent collection (e.g., PubChem CID, DSSTox CID). The data collection-specific SIDs and CIDs are called SOURCE_NAME_SID and SOURCE_NAME_CID and are alphanumeric strings of the form PUBCHEM_1234 or SIDS_2345. Additionally, ACToR internally uses sequentially generated unique numeric SIDs and CIDs.

Data on chemicals across data collections are aggregated using the concept of a generic chemical, which for this purpose takes the place of the compound in PubChem. The vast majority of chemicals in PubChem have defined structures, while many environmental chemicals are complex, and often undefined, mixtures. However, most environmental chemicals, along with their related toxicity data are indexed by a more discriminating CAS registry number (CASRN) rather than by chemical structure. Because of this, ACToR aggregates information based on CASRN. A generic chemical is defined by a CASRN, a preferred name, an ACToR CID and a unique generic chemical identifier or GCID. All data on all substances sharing a particular CASRN are attached to the corresponding generic chemical. An advantage of using CASRN is that different numbers will be assigned to a pure substance versus a mixture of isomeric forms or a mixture of unrelated compounds. All of these cases, however, may share a common compound PubChem CID and representative structure. Disadvantages of using CASRN include the fact that they are not always available or unique for a given substance (e.g. CASRN can be retired and replaced), they do not typically distinguish to the level of compound purity grade (e.g., analytical vs. technical grade), and they are tied to a non-public registry system (Chemical Abstracts Service (CAS) SciFinder). Nonetheless, CASRN are sufficiently general to serve as the basis for aggregation. Because only a small fraction of PubChem substance records contain a CASRN, we perform a second level of aggregation and pull in all PubChem substances that share the structure or PubChem CID associated with a particular GCID. Currently, the two main sources of chemical structure data in ACToR are EPA DSSTox and PubChem. Because DSSTox structures are quality reviewed, hand curated, and reconciled with chemical name and

CASRN, they are always preferred over structures automatically generated and provided by disparate sources in PubChem. Figure 1 illustrates the basic relationships between substance, compounds, generic chemicals and assays, which are described next.

In ACToR, an assay is a collection of data values associated with a set of substances and can be represented in a rectangular matrix. An assay is associated with an AID, a name, a category, and one or more phenotypes. Examples of assay categories are listed in Table 1 and reflect our focus on chemical toxicity and its origin in detailed molecular biological interactions. As one can see, the concept of an assay as implemented in ACToR is purposely broad so as to capture any information potentially relevant to understanding toxicity and evaluating risk for environmental chemicals. An assay also can have one or more components, which correspond to the columns of the rectangular data matrix. Each component is defined by an assay component identifier (ACID), the corresponding AID, a name, a description, units (when applicable) and a data type (FLOAT, INTEGER, CATEGORICAL, TEXT, BOOLEAN, URL). The actual data values are called assay results and are linked to the assay, the assay component and the original data-collection-specific substance.

Because ACToR is intended to support hazard identification and risk assessment, assays can be labeled by a series of “phenotypes” for which they contain information. The set of phenotypes implemented in ACToR span both traditional toxicology study areas: General Chemical Hazard, Acute Toxicity, Subchronic Toxicity, Chronic Toxicity, Carcinogenicity, Developmental Toxicity, Reproductive Toxicity, Neurotoxicity, Developmental Neurotoxicity, Immunotoxicity, Dermal Toxicity, Respiratory Toxicity, Genotoxicity and Ecotoxicity.

Data Sources: ACToR is importing data from a large number of public sources (currently >150), which are referred to as data collections. A data collection will usually include a set of substances and may have corresponding compounds (chemical structures) and one or more assays. The largest source of data currently in ACToR in terms of substances and assay data points is PubChem, which is itself a compilation of multiple

data sources (57 of which have data included in ACToR). Most assay data in PubChem comes from HTS assays run by the Molecular Libraries Screening Centers Network (MLSCN)(Austin *et al.*, 2004) on compounds from the Molecular Libraries Small Molecule Repository (MLSMR). However, the vast majority of chemicals in PubChem have no assay data and come from collections of molecular structures from chemical manufacturer catalogs (e.g., SIGMA) or virtual screening libraries (e.g., ZINC).

The balance of the data collections within ACToR pertain more specifically to environmental chemicals. These collections are from the US EPA, CDC, FDA, NIH, equivalent agencies in Europe, Japan and Canada, the World Health Organization, universities and several states and NGOs. Some of these specific sources are described below.

To be included in ACToR, a data collection must meet several criteria. First, it has to be publicly available with no restrictions on redistribution. An important goal of the ACToR project is to create a widely usable, freely distributable, open source system. Any conclusions drawn from these data should be subject to independent confirmation, which is made possible by this open source data model. Second, the collection should contain information on environmentally-relevant chemicals. We have not to this point included a number of data sets focused exclusively on pharmaceutical compounds, although toxicological information on these compounds is potentially informative. Third, if a source consists of a web-accessible database, we require an index of the chemicals in the database in order to link that web resource back into ACToR. Several web databases that allow local searching by name or CASRN do not provide full access or a list of the chemicals needed for indexing; hence, these are not currently included in the database. However, there are a number of important data collections without publicly available indexes, so the ACToR user interface provides URL links to allow the user to search these databases on a chemical-by-chemical basis based on CASRN and/or name. TOXNET and its component databases are the main data collections in this category. In addition to compiling data from other databases, selected tabular information from the primary toxicology literature is also being captured in ACToR.

Software Aspects: The ACToR database is implemented using MySQL. Software to preprocess and load data is written in Perl and the web interfaces are written in Java. The use of 100% open-source software will allow the entire system to be easily distributed to other interested groups.

Search and Browsing: The current version of the database allows one to browse by data collection or assay, and to search chemicals by name and structure using a chemical drawing applet and standard chemical similarity algorithms.

RESULTS

Table 2 gives summary statistics on the current composition of the database. As already mentioned, the vast majority of substances come from PubChem, although the overlap of that set with chemicals of environmental interest is relatively small. ACToR contains all substances, compounds and assay results from PubChem, but the table only gives counts for chemicals that can be indexed by CAS registry number, which yields just over 500,000 unique or generic chemicals.

To illustrate the utility of ACToR, we show how the aggregated data can be used to evaluate sets of chemicals for use in developing and validating toxicology signatures for a screening and prioritization approach. This approach is more fully explored elsewhere(Judson *et al.*, In preparation). Our focus is on environmental chemicals having sufficiently high production and use volumes such that there is potential for human and ecological exposure. The sets of chemicals used are summarized in Table 3. Because of overlaps between these lists, the current total number of generic chemicals considered is 11,139. This exact number will fluctuate over time because the chemicals included in the lists periodically change due to altered use-patterns, introduction of new chemicals, and discontinuation of use of others. On average, each of these chemical substances has information derived from 2-3 sources, although some of chemicals are found in a dozen or more data collections. Of these chemicals, 7,512 have an associated chemical structure

and CID assigned. Of the chemicals without a structure, many are mixtures or complex substances (e.g., mica, milk, mink oil and molasses, all of which are pesticide inert ingredients).

The primary *in vivo* toxicology assays (either tabulated or not) are those derived from National Toxicology Program (NTP) studies and from ToxRefDB. The majority of data currently in the ToxRefDB database, a component of the larger ACToR system, contains summary results of primary toxicology studies submitted to the EPA on pesticide active ingredients (Martin *et al.*, 2007). Typically these data have been extracted from EPA Office of Pesticide Programs (OPP) evaluations of studies based on EPA Office of Prevention, Pesticides and Toxic Substances (OPPTS) harmonized test guidelines (http://www.epa.gov/opptsfrs/home/guidelin.htm). ToxRefDB captures details of study design and dose series data from the areas of histopathology, clinical chemistry, hematology, gross anatomy, pathology (neoplastic and non-neoplastic), urinalysis and mortality. Data is aggregated at the level of animal treatment group (dose and time). Summary data from ToxRefDB is entered into the ACToR assay tables. NTP primary tabular data is also being entered into ToxRefDB for chemicals not covered by OPPTS sources. The DSSTox program has indexed all of the studies in the NTP database by chemical and study type, and this index is in ACToR.

The secondary *in vivo* toxicology study data is derived from Risk-Based Concentrations (RBC); WHO Classifications of Pesticide Hazards; the Cancer Potency Database (CPDB) (Richard *et al.*, 2006); the EPA Fat Head Minnow Database (Russom *et al.*, 2007); the FDA Maximum Daily Dose Database (Matthews *et al.*, 2004); and IRIS (Integrated Risk Information System) (Richard *et al.*, 2007) . With the exceptions of RBC and the WHO pesticide data, the data for these sets are taken from the DSSTox database. Web site URLs for all of the data sources used for this analysis are given in Table 4.

The category of *in vivo* toxicology (summary calls) describes sources where experts have reviewed the toxicology literature and have made a definitive statement about a particular chemical and endpoint, for instances labeling a chemical as a proven

human carcinogen. Data sources for this category are California EPA Determination of Cancer and Developmental Risks (Proposition 65); CERCLA Priority List of Hazardous Substances; the FDA Everything Added to Food in the United States List (EAFUS); Health Canada Priority Substance Lists; EPA OPP Inert (other) Pesticide Ingredients categories; NTP 11th Report on Carcinogens (RoC); and the Disinfection By-products Database (Woo *et al.*, 2007).

Data sources included under Toxicity Summary Reports on the Web are Cancer Potency Database; National Toxicology Program (NTP) reports (Burch *et al.*, 2007); IRIS; EPA HPV Information System; CDC Agency for Toxic Substances and Disease Registry (ATSDR); Center for the Evaluation of Risks to Human Reproduction (CERHR); DrugBank; EPA Pesticide Fact Sheets; EXTOXNET Pesticide Information Profiles; INCHEM Concise International Chemical Assessment Documents (CICAD); INCHEM Environmental Health Criteria Monographs (EHC); INCHEM International Agency for Research on Cancer (IARC); ITER TERA Risk Assessments; Ministry of Health Labor and Welfare (Japan) Risk Assessments; NTP 11th Report on Carcinogens (RoC); and OECD Screening Information Data Sets (SIDS) for High Volume Chemicals; ESIS (European chemical Substances Information System) and its subsets ESIS HPV, ESIS LPV (low production volume); ESIS PBT (Persistent Bioaccumulating Toxins), and ESIS ORATS (Online European Risk Assessment Tracking System). Note that some sources of data are included in multiple assay categories.

Table 5 summarizes the amount of toxicology information that we have currently captured in ACToR, selected by assay categories for the set of 11,139 environmental chemicals being analyzed. About half of these chemicals have some publicly available toxicology data within the sets of information we have currently compiled. Primary *in vivo* toxicology data is available for 1,447 chemicals (13%), and secondary *in vivo* toxicology data is available for a total of 1,405 chemicals (13%). A total of 5,205 chemicals (47%) have one or more summary *in vivo* toxicity calls or determinations, which are derived by experts who have curated data from the primary scientific literature. Finally 5,244 chemicals (47%) have one or more summary text reports on chemical

toxicity available on the web. However, many of these (especially from the ESIS LPV list) simply state that no hazard or toxicology information is available for that chemical. We emphasize once again that these are conservative numbers as there are still large collections of data yet to be compiled and loaded into the database. The bottom line, though, is that there is little detailed in vivo toxicology information for the majority of these environmental chemicals.

The EPA ToxCast program is a major driver of the development of the ACToR system. The goal of ToxCast is to develop and test methods for chemical screening and prioritization by linking the results of *in vitro* assays to *in vivo* toxicity data (Dix *et al.*, 2007). The most rigorous chemical toxicity testing data are derived from whole animal human health guideline studies, many of which are being captured in ToxRefDB and ACToR, initially for a set of 308 unique chemicals that are being used in Phase I of ToxCast. OPP-required guideline studies and NTP studies are the primary source of the *in vivo* data that will be used in ToxCast. A secondary source is the data from IRIS assessments. Table 6 shows the number of chemicals in the ToxCast Phase I set of 308 that have data from OPP guideline studies, from NTP or from IRIS for a set of key areas including acute toxicity, subchronic toxicity, chronic toxicity, carcinogenicity, developmental toxicity, reproductive toxicity, immunotoxicity, neurotoxicity and genotoxicity. As one can see, there is currently good coverage for many of these areas, but several will require searches through other ACToR-catalogued data collections in order to build the complete analysis data set.

DISCUSSION

This paper briefly describes ACToR (Aggregated Computational Toxicology Resource), which is a set of linked databases and analysis tools that aggregate a large number of data sets of relevance to environmental chemicals and toxicology. The utility of the system was illustrated with an example showing the amount of data available from multiple sources that can be used for developing training and validation sets for high-throughput chemical screening and prioritization efforts.

ACToR is not alone in its goal of aggregating large sets of chemical structure and assay data. PubChem is the largest effort currently available, with information on more than 10 M unique chemical compounds. PubChem currently focuses on aggregating data from *in vitro* HTS assays as the primary data repository for the MLSCN. PubChem allows more generalized types of assay data to be submitted and displayed, but their query engine is not tailored to the types of custom toxicology-based queries needed for our purposes. However, their underlying data model maps easily into the ACToR application and serves as a useful model for our internal data organization. This has allowed us to import all of the PubChem data and easily integrate it with other data sources. Another important comparison is with TOXNET which is a collection of multiple data sources covering many aspects of chemical toxicity. TOXNET has a common search engine that allows the user to easily find data from multiple sources. However, it is a closed system which does not allow a user to pull together datasets that are useful for computational purposes. One unique aspect of the ACToR system is that it is pulling together the data from PubChem (focused on chemical structure and HTS *in vitro* assay data) and TOXNET (focused on *in vivo* toxicology data) and combining it in a way that it can be used for computational analysis. We are in the process of extracting selected tabular data from TOXNET to include directly into the ACToR database.

eChemPortal is an Organization for Economic Co-operation and Development (OECD) effort very similar to ACToR. It is aggregating information on HPVs and pesticides among others. eChemPortal currently contains links to 7 large database systems, some of which contain what in ACToR are multiple individual databases (e.g., INCHEM contains 11 individual databases). Unlike eChemPortal, which provides links to web pages for the component databases, ACToR extracts tabular data from the individual sources and makes it searchable in an aggregated fashion. A system called Vitic is being developed as a collaboration between IUCLID and a number of pharmaceutical companies with the goal of being an international toxicology information center (Judson *et al.*, 2005). Finally, the European substances Information System or ESIS provides links to a number of databases including EPA HPV, IUCLID and EINECS (European INventory of Existing Commercial chemical Substances). The CEBS

(Chemical Effects in Biological Systems) project at the NIEHS is constructing a multi-domain information repository to hold the detailed results and summaries of *in vivo* and *in vitro* toxicology experiments (Waters *et al.*, 2003).

We have made use of several reviews of the toxicology data landscape to select data collections to be included in ACToR. Yang et al. have recently published two such reviews (Yang *et al.*, 2006a; Yang *et al.*, 2006b). In 2001 and 2002, a pair of review collections were published surveying the landscape of toxicity data available on the internet (Brinkhuis, 2001; Poore *et al.*, 2001; Felsot, 2002; Junghans *et al.*, 2002; Patterson *et al.*, 2002; Polifka and Faustman, 2002; Russom, 2002; Winter, 2002; Wolfgang and Johnson, 2002; Young, 2002; Richard and Williams, 2003).

We foresee several key uses for ACToR. One is the derivation of training and validation data sets for ToxCast and other chemical screening and prioritization efforts. A second is to serve as a unique resource for researchers developing fully computational models linking chemical structure with *in vitro* and *in vivo* assays. Third, this large structure-searchable database can be a valuable resource for reviewers within the EPA and other regulatory agencies who are examining new chemicals submitted for marketing approval. Reviewers can use the system to search for structural analogs of the novel compounds and, if available, easily locate potentially informative *in vitro* bioassay and *in vivo* toxicology data on related compounds. This can, in turn, inform their decisions on the novel chemicals under review.

ACToR is a rapidly evolving system. Future developments will involve bringing in additional sources of information; extraction of tabular data from on-line text documents linked to chemicals; addition of more curated chemical structures; and the construction of a more flexible query and data export interface. Additionally, this system will allow the construction of workflow processes for prioritization of data capture, quality control and chemical prioritization scoring. The system is currently being used at the EPA to support the ToxCast chemical screening and prioritization program. We are working towards public release of the system in 2008.

Assay Category	Description	Examples
Physicochemical	Physical and chemical properties (in vitro and/or in silico)	Molecular weight, logP, boiling point
Biochemical	Biochemical (non-cell-based) (in vitro and/or in silico)	Enzyme inhibition or receptor binding constants
Genomics	Gene expression values or signatures	Result of *in vitro* or *in vivo* microarray analysis
Cellular	Cell-based assay	Cell culture cytotoxicity
Tissue	Tissue slice assays	Tissue slice cytotoxicity
In vivo toxicology (tabular primary)	Tabulated results from primary animal-based studies of chemical effect	Clinical chemistry, histopathology, developmental and reproductive assays
In vivo toxicology (study listing primary)	Primary studies are available but have not been tabulated	Clinical chemistry, histopathology, developmental and reproductive assays
In vivo toxicology (tabular secondary)	Tabulated data from secondary sources of *in vivo* toxicology studies	Clinical chemistry, histopathology, developmental and reproductive assays
In vivo toxicology (summary calls)	Derived summary determinations of risk	Chemicals determined to pose a defined risk of human cancer
In vivo toxicology (summary report via URL)	Links to text reports on the web for which specific data values are not directly accessible in tabular form	Reports from EPA Integrated Risk Information System (IRIS), National Toxicology Program (NTP)
Regulatory	Listings of chemicals that fall under specific environmental laws or government mandates	U.S. Toxic Substances Control Act (TSCA)
Chemical Category	Listing of structural or use categories, often intended for prioritization efforts	Phthalate

Table 1: Categories of assays in ACToR

Data Collections	232
Source-specific Substances	964,083
Compounds (chemical structures)	404,196
Generic Chemicals	504,871
Generic Chemicals with Structure	390,379
Assays	1,592
Assay Components	10,733
Assay Results	6,118,231

Table 2: Summary statistics for the ACToR database. Assay Results are individual data points for a single substance and a single assay component. The numbers only include substances having CAS registry numbers. A much large number of substances, compounds and assay results are included from PubChem, but are not currently indexed as generic chemicals.

Data Collection	**Chemicals**
HPV- High Production Volume chemicals produced or imported in quantities >1M lb/year http://www.epa.gov/hpv/pubs/update/hpvchmlt.htm	2810
IUR- Inventory Update Rule, chemicals produced or imported in quantities >10,000 lb/year (2002 list) (also referred to as MPVs or medium Production volume chemicals) http://www.epa.gov/oppt/iur/tools/data/2002-comp-chem-records.htm	5375
Pesticide active ingredients including anti-microbials and food-use pesticides http://www.epa.gov/pesticides/factsheets/registration.htm	3476
Pesticide inert ingredients (or “other ingredients”) http://www.epa.gov/opprd001/inerts/lists.html	3850
TRI- Toxic Release Inventory provides reports of toxic chemical releases and other waste management activities http://www.epa.gov/tri/	577
Drinking water chemical contaminants, disinfection byproducts, and chemical contaminant candidates http://www.epa.gov/safewater/ccl/index.html	120
EDC - Draft list of chemicals considered for endocrine disruption screening http://www.epa.gov/endo/pubs/edspoverview/index.htm	73

Table 3: Sources of lists of environmental chemicals

Data Source	**URL**
CDC Agency for Toxic Substances and Disease Registry (ATSDR)	http://www.atsdr.cdc.gov/toxfaq.html
California EPA Proposition 65	http://www.oehha.ca.gov/prop65/prop65_list/Newlist.html
Cancer Potency Database (DSSTox)	http://potency.berkeley.edu http://www.epa.gov/ncct/dsstox/sdf_cpdbas.html
Chemical Abstracts Service (CAS) SciFinder,	http://www.cas.org/
Center for the Evaluation of Risks to Human Reproduction (CERHR)	http://cerhr.niehs.nih.gov/chemicals/index.html
CERCLA Priority List of Hazardous Substances	http://www.atsdr.cdc.gov/cercla/05list.html
eChemPortal	http://webnet3.oecd.org/echemportal/ParticipatingDb.aspx
DrugBank	http://redpoll.pharmacy.ualberta.ca/drugbank
EPA Disinfection By-products Database (DSSTox)	http://www.epa.gov/ncct/dsstox/sdf_dbpcan.html
EPA Fathead Minnow Database (DSSTox)	http://www.epa.gov/ncct/dsstox/sdf_epafhm.html
EPA HPV Challenge Program	http://www.epa.gov/hpv/
EPA HPV Information System	http://www.epa.gov/ncct/dsstox/sdf_hpvcsi.html
EPA Integrated Risk Assessment System (IRIS)	http://www.epa.gov/iris http://www.epa.gov/ncct/dsstox/sdf_iristr.html
EPA Pesticide Fact Sheets (Conventional Chemicals)	http://www.epa.gov/opprd001/factsheets
EPA Office of Pesticides (OPP) Inert (other) Pesticide	http://www.epa.gov/opprd001/inerts/lists.html

Ingredients	
EPA Risk Based Concentrations (RBC)	http://www.epa.gov/reg3hwmd/risk/human/index.htm
EPA ToxCast Program	http://www.epa.gov/comptox/toxcast/
European substances Information System (ESIS)	http://ecb.jrc.it/esis
EXTOXNET Pesticide Information Profiles	http://extoxnet.orst.edu
FDA Everything Added to Food in the United States	http://vm.cfsan.fda.gov/~dms/eafus.html
FDA Maximum Daily Dose Database	http://www.epa.gov/ncct/dsstox/sdf_fdamdd.html
Health Canada Priority Substance Lists	http://www.hc-sc.gc.ca/ewh-semt/contaminants/existsub/categor/_result_substance/index_e.html
INCHEM Concise International Chemical Assessment Documents	http://www.inchem.org/pages/cicads.html
INCHEM Environmental Health Criteria Monographs	http://www.inchem.org/pages/ehc.html
INCHEM International Agency for Research on Cancer (IARC)	http://www.inchem.org/pages/iarc.html
ITER TERA Risk Assessments	http://www.tera.org
Ministry of Health Labor and Welfare (Japan) Risk Assessments	http://wwwdb.mhlw.go.jp/ginc/html/db1.html
Molecular Libraries Small Molecule Repository (MLSMR)	http://mlsmr.glpg.com/MLSMR_HomePage/project.html
National Toxicology Program (NTP)	http://ntp.niehs.nih.gov/ http://www.epa.gov/ncct/dsstox/sdf_ntpbsi.html

NIH Molecular Libraries Roadmap	(http://nihroadmap.nih.gov/molecularlibraries/)
NTP 11th Report on Carcinogens (RoC)	http://ntp.niehs.nih.gov/ntpweb/index.cfm?objectid=035E5806-F735-FE81-FF769DFE5509AF0A
OECD Screening Information Data Sets (SIDS) for High Volume Chemicals	http://www.chem.unep.ch/irptc/sids/OECDSIDS/indexcasnumb.htm
PubChem	http://pubchem.ncbi.nlm.nih.gov
TOXNET	http://toxnet.nlm.nih.gov
WHO Classifications of Pesticide Hazard	http://www.inchem.org/documents/pds/pdsother/class.pdf

Table 4: URLs for sources of data described in this article

Assays	0	1	2	3	4	5	>5
In vivo toxicology (study listing primary)	9692	686	341	85	42	51	242
In vivo toxicology (tabular secondary)	9734	861	240	163	91	43	7
In vivo toxicology (summary calls)	5934	2776	950	419	258	184	618
In vivo toxicology (summary report via URL)	5895	2706	1121	573	344	203	297

Table 5: (Number of chemicals) x (number of assays) in ACToR for the 11,139 environmental chemicals being analyzed.

Phenotype	ToxRefDB	IRIS	NTP	Total	% Coverage
Acute Toxicology	0	126	0	126	41
Subchronic Toxicology	235	0	0	235	77
Chronic Toxicology	274	126	38	291	95
Carcinogenicity	263	126	38	285	93
Developmental Toxicology	274	126	4	290	95
Reproductive Toxicology	251	126	0	277	91
Immuno- Toxicology	0	126	1	126	41
Genotoxicity	0	126	60	147	78
Neuro- Toxicology	0	126	0	126	41

Table 6: Number of the ToxCast Phase I 308 chemicals for which data is captured in ACToR from primary guideline studies or from IRIS assessments for key areas of toxicology.

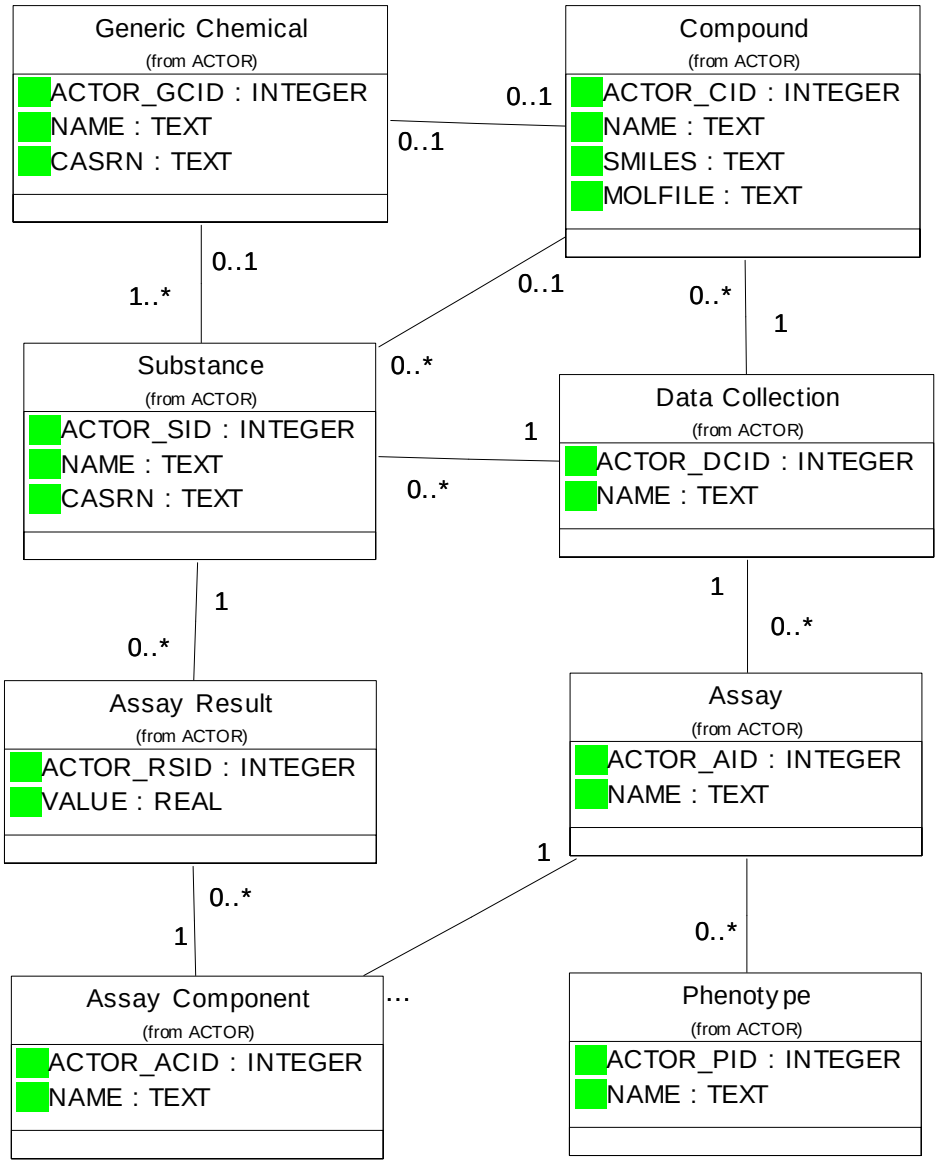

Figure 1 – Schematic, entity-relationship (ER) diagram of the ACToR database schema showing key relationships between substances, compounds, generic chemicals, data collections and assays. The annotations on the connecting lines the range of number of entities for each relationship. For instance, a substance can have 0 or 1 generic chemicals (indicated by 0..1), while a generic chemical can have 1 to many substances (indicated by 1..*). The zero/one-to-many relationships are implemented through separate tables (not shown). The actual schema contains 44 tables.

References

Allanou, R., Hansen, B., and van det Bilt, Y. (1999). Public Availability of Data on EU High Production Volume Chemicals.

Applegate, J., and Baer, K. (2006). Strategies for Closing the Data Gap.

Austin, C. P., Brady, L. S., Insel, T. R., and Collins, F. S. (2004). NIH Molecular Libraries Initiative. *Science* **306**, 1138-1139.

Birnbaum, L. S., Staskal, D. F., and Diliberto, J. J. (2003). Health effects of polybrominated dibenzo-p-dioxins (PBDDs) and dibenzofurans (PBDFs). *Environ Int* **29**, 855-860.

Brinkhuis, R. P. (2001). Toxicology information from US government agencies. *Toxicology* **157**, 25-49.

Burch, J., Eastin, W. C., Bowden, B., Wolf, M. A., and Richard, A. M. (2007). DSSTox National Toxicology Program Bioassay On-line Database Structure-Index Locator File: SDF File and Documentation.

Dix, D. J., Houck, K. A., Martin, M. T., Richard, A. M., Setzer, R. W., and Kavlock, R. J. (2007). The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol Sci* **95**, 5-12.

EPA (1998). Chemical Hazard Data Availability Study. Office of Pollution Prevention and Toxics.

Felsot, A. S. (2002). WEB resources for pesticide toxicology, environmental chemistry, and policy: a utilitarian perspective. *Toxicology* **173**, 153-166.

Guth, J., Denison, R., and Saas, J. (2005). Background Paper for Reform No. 5 of the Louisville Charter for Safer Chemicals: Require Comprehensive Safety Data for All Chemicals.

Judson, P. N., Cooke, P. A., Doerrer, N. G., Greene, N., Hanzlik, R. P., Hardy, C., Hartmann, A., Hinchliffe, D., Holder, J., Muller, L., Steger-Hartmann, T., Rothfuss, A., Smith, M., Thomas, K., Vessey, J. D., and Zeiger, E. (2005). Towards the creation of an international toxicology information centre. *Toxicology* **213**, 117-128.

Judson, R., Dix, D. J., Houck, K., Richard, A. M., Martin, M. T., Kavlock, R. J., Dellarce, V., Holderman, T., Tan, S., Carpenter, T., and Smith, E. (In preparation). The Toxicity Data Landscape for Environmental Chemicals.

Junghans, T. B., Sevin, I. F., Ionin, B., and Seifried, H. (2002). Cancer information resources: digital and online sources. *Toxicology* **173**, 13-34.

Krewski, D., D Acosta, J., Anderson, M., Anderson, H., III, J. B., Boekelheide, K., Brent, R., Charnley, G., Cheung, V., Green, S., Kelsey, K., Kervliet, N., Li, A., McCray, L., Meyer, O., Patterson, D. R., Pennie, W., Scala, R., Solomon, G., Stephens, M., J Yager, J., and Zeize, L. (2007). *Toxicity Testing in the Twenty-first Century: A Vision and a Strategy* National Academies Press, Washington D.C.

Martin, M. T., Houck, K. A., McLaurin, K., Richard, A. M., and Dix, D. J. (2007). Linking Regulatory Toxicological Information on Environmental Chemicals with High-Throughput Screening (HTS) and Genomic Data. *The Toxicologist CD- An official Journal of the Society of Toxicology* **96**, 219-220.

Matthews, E. J., Kruhlak, N. L., Weaver, J. L., Benz, R. D., and Contrera, J. F. (2004). Assessment of the health effects of chemicals in humans: II. Construction of an

adverse effects database for QSAR modeling. *Curr Drug Discov Technol* **1**, 243-254.

Muir, D. C., and Howard, P. H. (2006). Are there other persistent organic pollutants? A challenge for environmental chemists. *Environ Sci Technol* **40**, 7157-7166.

Patterson, J., Hakkinen, P. J., and Wullenweber, A. E. (2002). Human health risk assessment: selected Internet and world wide web resources. *Toxicology* **173**, 123-143.

Polifka, J. E., and Faustman, E. M. (2002). Developmental toxicity: web resources for evaluating risk in humans. *Toxicology* **173**, 35-65.

Poore, L. M., King, G., and Stefanik, K. (2001). Toxicology information resources at the Environmental Protection Agency. *Toxicology* **157**, 11-23.

Richard, A., and Williams, C. (2003). Public Sources of Mutagenicity and Carcinogenicity Data: Use in Structure-Activity Relationship Models. In *QSARS of Mutagens and Carcinogens* (R. Benigni, Ed.), pp. 145-173. CRC Press, New York.

Richard, A. M., Gold, L. S., and Nicklaus, M. C. (2006). Chemical structure indexing of toxicity data on the internet: moving toward a flat world. *Curr Opin Drug Discov Devel* **9**, 314-325.

Richard, A. M., M.A., W., and J., B. (2007). DSSTox EPA Integrated Risk Information System (IRIS) Toxicity Review Data: SDF File and Documentation.

Richard, A. M., and Williams, C. R. (2002). Distributed structure-searchable toxicity (DSSTox) public database network: a proposal. *Mutat Res* **499**, 27-52.

Russom, C. L. (2002). Mining environmental toxicology information: web resources. *Toxicology* **173**, 75-88.

Russom, C. L., Williams, C. R., Stewart, T. W., Swank, A. E., and Richard, A. M. (2007). DSSTox EPA Fathead Minnow Acute Toxicity Database (EPAFHM): SDF Files and Documentation.

Waters, M., Boorman, G., Bushel, P., Cunningham, M., Irwin, R., Merrick, A., Olden, K., Paules, R., Selkirk, J., Stasiewicz, S., Weis, B., Van Houten, B., Walker, N., and Tennant, R. (2003). Systems toxicology and the Chemical Effects in Biological Systems (CEBS) knowledge base. *EHP Toxicogenomics* **111**, 15-28.

Winter, C. K. (2002). Electronic information resources for food toxicology. *Toxicology* **173**, 89-96.

Wolfgang, G. H., and Johnson, D. E. (2002). Web resources for drug toxicity. *Toxicology* **173**, 67-74.

Woo, Y. T., Williams, C. R., Fields, N., and Richard, A. M. (2007). DSSTox EPA Water Disinfection By-Products with Carcinogenicity Estimates Database (DBPCAN): SDF Files and Documentation.

Yang, C., Benz, R. D., and Cheeseman, M. A. (2006a). Landscape of current toxicity databases and database standards. *Curr Opin Drug Discov Devel* **9**, 124-133.

Yang, C., Richard, A. M., and Cross, K. P. (2006b). The art of data mining the minefields of toxicity databases to link chemistry to biology. *Curr. Comput.-Aided Drug Dis.* **2**, 135-150.

Young, R. R. (2002). Genetic toxicology: web resources. *Toxicology* **173**, 103-121.