D964–D971 *Nucleic Acids Research, 2017, Vol. 45, Database issue* *Published online 29 November 2016*
doi: 10.1093/nar/gkw1077

CEBS: a comprehensive annotated database of toxicological data

Isabel A. Lea[1,*], Hui Gong[1], Anand Paleja[1], Asif Rashid[1] and Jennifer Fostel[2,*]

[1]ASRCFederal Vistronix, 430 Davis Dr, Suite 260, Morrisville, NC 27569, USA and [2]Division of the National Toxicology Program, National Institute of Environmental Health Sciences, PO Box 12233, Research Triangle Park, NC 27709, USA

Received August 08, 2016; Revised October 12, 2016; Editorial Decision October 24, 2016; Accepted November 01, 2016

ABSTRACT

The Chemical Effects in Biological Systems database (CEBS) is a comprehensive and unique toxicology resource that compiles individual and summary animal data from the National Toxicology Program (NTP) testing program and other depositors into a single electronic repository. CEBS has undergone significant updates in recent years and currently contains over 11 000 test articles (exposure agents) and over 8000 studies including all available NTP carcinogenicity, short-term toxicity and genetic toxicity studies. Study data provided to CEBS are manually curated, accessioned and subject to quality assurance review prior to release to ensure high quality. The CEBS database has two main components: data collection and data delivery. To accommodate the breadth of data produced by NTP, the CEBS data collection component is an integrated relational design that allows the flexibility to capture any type of electronic data (to date). The data delivery component of the database comprises a series of dedicated user interface tables containing pre-processed data that support each component of the user interface. The user interface has been updated to include a series of nine Guided Search tools that allow access to NTP summary and conclusion data and larger non-NTP datasets. The CEBS database can be accessed online at http://www.niehs.nih.gov/research/resources/databases/cebs/.

INTRODUCTION

The National Toxicology Program (NTP) was established by the US Department of Health and Human Services in 1978 in response to concerns about potential human health effects of environmental chemicals. The NTP provides scientific data to regulatory agencies and other health-related research groups. Chemicals studied at the NTP can be endocrine disruptors, occupational exposure mixtures, pesticides, pharmaceuticals, metals, food additives and herbal supplements; anything with the potential to impact health. The NTP conducts comprehensive testing of each substance or test article (exposure agent) in an effort to provide data for a strong scientific basis to make credible decisions that will protect public health. Testing can include evaluations of toxicity and carcinogenicity, prenatal developmental and reproductive toxicology, neurobehavioral effects, immunological effects, genetic toxicity, toxicogenomic responses, as well as chemical disposition and toxicokinetic analysis. Results and conclusions from the NTP testing program are released into the public domain as published reports or journal articles.

A great deal of toxicity information has been generated by the NTP since its inception in the 1970s. Until recently these data were made available to the public only as web-based PDF reports on an individual study basis. This made it a challenge to compare and contrast results for multiple test articles or different data endpoints for individual animals. To address this issue, the NTP designated the Chemical Effects in Biological Systems (CEBS) database as the primary repository for its data and has invested significant effort into making the data available for searching, downloading and data mining.

CEBS was developed as a public repository for toxicogenomics data by the National Center for Toxicogenomics (NCT) within the National Institute of Environmental Health Science (NIEHS). Our most recent publication in 2008 described development of CEBS to capture microarray (gene expression) and proteomics (protein expression) data (1,2) and illustrated the integration of study design parameters with toxicological assay data. The CEBS SysTox Object Model (3) and the CEBS Data Dictionary (4) were developed to promote this database model. This first version of the database permitted the CEBS user to select groups of subjects drawn from different studies, and analyze the associated microarray data. It also provided a good platform on which to build the current NTP data reposi-

*To whom correspondence should be addressed. Tel: +1 919 972 7985; Email: Isabel.lea@nih.gov
Correspondence may also be addressed to Jennifer Fostel. Email: fostel@niehs.nih.gov

tory. Since this time, CEBS has had three major goals: (i) be a repository for NTP toxicology testing data; (ii) provide a public resource for accessing, searching and reviewing all NTP toxicology data and (3) provide a public data mining resource that could be used to address toxicology related questions.

With the advent of new technologies in the field of biological science coupled with advances in database technology, access to on-line data analysis tools and large toxicological datasets is ever expanding. Many open databases and resources for toxicological information and risk assessment exist. Many of these are curated resources built on information garnered from the literature and other online resources, for example: the Comparative Toxicological Database (CTD) (5) and the Swiss Institute of Bioinformatics (SIB) (6). Some databases, including the EPA's Aggregated Computational Toxicology Resource system (ACToR) (7), PubChem (8) and Chemical Entities of Biological Interest (ChEBI) (9) act as central resources for chemical information compiled from external collections, in tandem with direct submissions or empirically generated data. Still others, Open TG-Gates (10), ArrayExpress (11) and ACuteTox (12) contain solely experimental data but with limited data types and with somewhat restricted access to metadata and study event timelines.

CEBS is unique in its role as a repository for NTP testing data and in providing access to individual animal level data in a biologically relevant framework that facilitates interpretation of data by assessment of experimental design. In this update paper we describe recent improvements and additions to CEBS. We have approached this in three ways: (i) update the back-end database design to provide the flexibility to capture as many types of data as possible; (ii) increase the data content in CEBS by capturing electronically available NTP legacy data; (iii) develop and improve search tools available on the CEBS home page to assist users in accessing individual subject (animal or plate) and NTP summary and conclusion data for genetic toxicity and carcinogenicity studies. Data for all NTP genetic toxicity, carcinogenicity, short term toxicity, and immunotoxicology studies are now accessible in the database. Large toxicogenomic reference resource datasets such as DrugMatrix data (13,14) and the Tox21 high throughput screening initiative (15–17) have also been added. CEBS does not contain all available data produced by the NTP; we are working to capture chemical disposition and toxicokinetics, and toxicogenomics data.

DATABASE DESCRIPTION

As technology and techniques continue to evolve, the generation and analysis of data has become increasingly complex. Databases designed to capture a 'standard' data structure are likely not to realize their full potential as a data repository. However, these databases are ideal for general, narrower scientific inquisitions (e.g. the Gene Expression Omnibus (GEO)). CEBS on the other hand, has been designed to capture a wide range of endpoints including various study design details, in-life observation data and qualitative and quantitative assay data for individual test subjects from *in vivo* and *in vitro* exposures (1). CEBS is a freely-available online toxicology resource that is a curated repository of empirical toxicology data (http://tools.niehs.nih.gov/cebs3/ui/).

The CEBS database has two main components: data collection and data delivery (CEBS database schema: ftp://anonftp.niehs.nih.gov/ntp-cebs/tools/Database/). The data collection component has a flexible design capable of collecting any data (to date) using the terms provided by the depositor. The data delivery component integrates data and utilizes curated synonyms, conversion rules for data units and standard normalization methods to faithfully and accurately collapse disparate depositor assay names and units into a CEBS 'standard' (defined in the CEBS Data Dictionary). The data delivery component is optimized for consistent and rapid presentation of the data to the user; the data collection component is optimized to accommodate data as it is deposited.

The CEBS database is able to capture metadata for any study design. An Investigation in CEBS is defined as a self-contained scientific enquiry. Each NTP test article is assigned to an Investigation, and each Investigation contains one or more studies. These studies encompass the comprehensive testing that the NTP conducts for each test article and may include genetic toxicity, carcinogenicity, general toxicity, toxicogenomics, and others. NTP studies are designed with multiple treatment groups, subjects, protocols, data domains (data types) and measurable effects, which are captured in CEBS along with observational data and experimental data with factors such as genetic intervention, physical interventions, and multi-factor designs. CEBS captures data at the Investigation, Study, Group, and Subject level, permitting as much or as little granularity as the depositor wishes to share. When appropriate, NTP conclusions are included at the Study level describing the potential toxicological effects of the test article under the study conditions described. The grouping of studies into a single Investigation is integral to linking multiple data types into one investigative unit for assessment of toxicological effects.

To facilitate the cross-study analysis of data, adequate study data is required to enable comparable subjects to be identified from different studies and their measured endpoints compared and understood. Figure 1 schematically illustrates the types of data in the CEBS database. Content can be largely grouped into three categories: study metadata, study data and study conclusions. Study metadata includes the experimental design with exposure groups, subjects, subject characteristics (e.g. species, age, and sex), protocols (e.g. husbandry, euthanasia, exposure, and assay) and study timeline (relative time for application of each study protocol). Study data includes all experimental data for which descriptive or measurable endpoints are collected including: in-life data (e.g. clinical observations, feed and water consumption, and body weight), post-mortem data (e.g. histopathology, clinical chemistry, hematology, organ weights, immunology, toxicogenomics, and developmental toxicity), microarray and high throughput data. Conclusions include the summary endpoints for each study. These may be NTP study conclusions (level of evidence calls), calls for individual trials in each study, descriptive and analytic statistics, activity calls and fold change data for toxicogenomics studies.

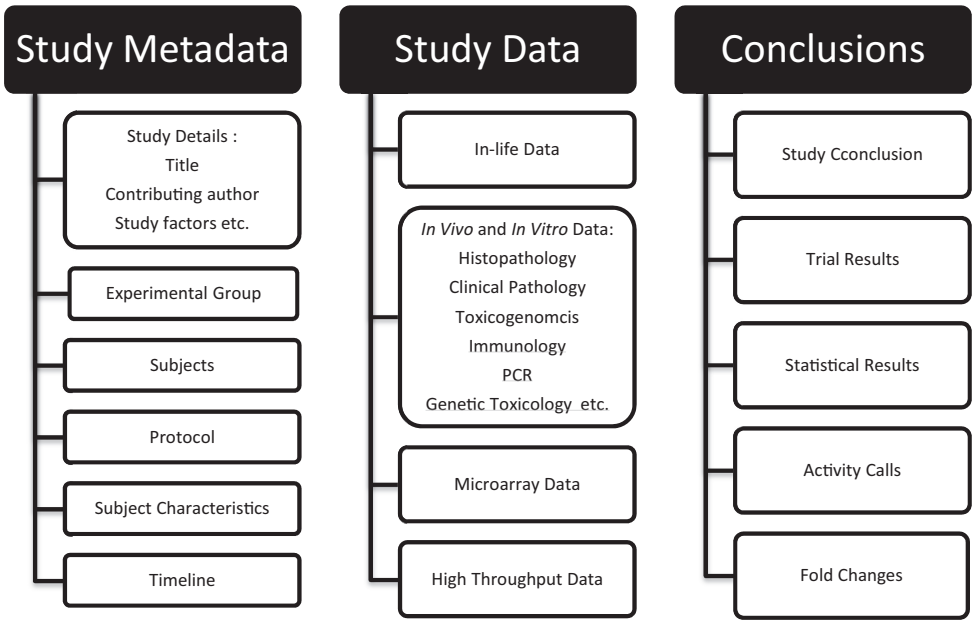

Figure 1. CEBS Data Collection Database. Data in the CEBS database are grouped into three categories: (i) Study Metadata which describe experimental design, (ii) Study Data which describe the data domains for which descriptive or measurable endpoints are collected and 3) Conclusions which describe the summary endpoints.

The organization of data in CEBS provides a fully searchable resource for which we have developed tools and features for viewing, sorting, and downloading either entire datasets or data search results. A user accessing the database can search for a single test article using various identifiers including test article name, synonym, Chemical Abstracts Services Registry Number (CASRN) or CEBS Investigation and Study Accession Numbers to access all study metadata, data and conclusions.

DATABASE DESIGN, DEVELOPMENT AND IMPLEMENTATION

Initially conceived as a microarray and proteomics database, CEBS has been redesigned and updated with an emphasis on flexible design, expansion of search tools and faster database queries to enhance the user experience. With these changes, the breadth of qualitative and quantitative information that is stored in CEBS has been greatly expanded. A description of the types of data currently stored in CEBS is provided in Supplementary Table S1.

CEBS updates were implemented by reformatting the back-end into an integrated relational database. The relational design of the data collection side of CEBS has allowed the flexibility to capture all types of data encountered to date. The data collection database is divided into three sections: Study Design, Study Execution and Study Data. The Study Design section stores high level information including the study characteristics such as study title, study type and study length. The same section also has information about the experimental treatment group, for example, sex, strain, species, dose, dose duration and exposure route. Subject information may be used to define the individual study subjects which could be animal, or cell, depending on the study type and their characteristics. The structure is flexible enough that attributes common to all subjects (e.g. sex, age) could be included at either the treatment group or subject level; this reduces the duplication of data required from the depositor. For example, if all animals in a single treatment group are female then this attribute would be a single treatment group level attribute.

The Study Execution section contains multiple tables containing information about the different protocols used in the study. Protocols which impact the subjects' in-life are collected here. There are five protocols which apply to all subject types so far encountered: Care (animal husbandry or tissue culture), Treatment (application of chemical, genetic, physical stressors or a combination of stressors or test articles), Observation (animal weights or observation), Disposition (euthanasia or cell harvest) and Preparation (specimen preparation). Assay protocols collect data on specimens after their removal from the study (e.g. tissue sections); these protocols do not impact the living subjects and are stored separately. Study events are captured along with their time components which, when linked to the Study Design and Study Data sections define a timeline that describes how, when and what happened during the study.

The Study Data section contains observation and assay data for all numeric or categorical endpoints. Simple endpoints such as body weight, food consumption or clinical chemistry measurements are assigned to one table. More complex assays such as microarray data are stored in multiple tables that allow capture of the final measured endpoint (intensity value) and values for intermediate steps such as RNA integrity values.

The database was designed with the expectation of variability in study design, study type, technology and reporting requirements and also that it should be possible to capture all data, and report legacy and current data in a similar for-

mat. The simple, flexible and robust schema developed for CEBS along with use of an appropriate data dictionary has made this possible. For example, the NTP recently began storing multi-generational reproductive and developmental toxicity parameters in CEBS yet no major updates were required to the data collection tables to assimilate these data into the database; all required changes were in the data delivery tables and in data presentation in the user interface.

The CEBS database schema is available on the CEBS website at ftp://anonftp.niehs.nih.gov/ntp-cebs/tools/Database/.

DATA CONTENT

CEBS has undergone a significant expansion in the last 5 years. Since our last update in 2008 (1), a concerted effort has been made to increase the content of the database by incorporating all electronically available NTP toxicological testing data. As of July 2016, all available NTP data for genetic toxicity, rodent long term carcinogenicity studies, rodent short term toxicity and immunotoxicology studies are accessible and searchable on the CEBS interface (8757 studies; Figure 2A). Our current efforts are focused making the Tox 21 high throughput data more easily accessible from the CEBS user interface (available from the CEBS FTP site) and capturing chemical disposition and toxicokinetics, and toxicogenomics data. As the NTP continues to generate new data, these are being made available in CEBS including new study types such as Modified One-Generation (MOG), and Reproductive Assessment by Continuous Breeding (RACB). Summary data for NTP studies, including NTP conclusions, are available in CEBS for genetic toxicity and rodent long-term carcinogenicity studies. These can be accessed, searched and downloaded using the Guided Search tools described below. Currently we are working on capturing NTP summary and conclusion information for reproductive, immunotoxicology, and high throughput screening (Tox21) studies.

The increase in content has produced a comprehensive database with a broad scope that includes not only a large collection of toxicological endpoints, but also datasets that support building predictive models (microarray and biological response endpoints in DrugMatrix), and assessment of data analysis methods and predictive models for classifying lung and liver toxicity data (Food and Drug Administration (FDA) MicroArray Quality Control (MAQC)-II data). With the availability of individual animal data and protocol information for all studies, CEBS has become a toxicological resource that supports modelling, predictive analysis and assessment of effects of time and dose on responses to experimental conditions.

Currently CEBS maintains over 11 000 test articles and 19 data types (domains in the database) although the design is extensible, so additional test articles and data domains can be added easily whenever needed. Data domains in CEBS are similar to and an extension of, the data domains defined by the Standards for Exchange of Nonclinical Data (SEND) consortium. SEND is a public effort within Clinical Data Interchange Standards Consortium (CDISC) focused on defining a format for sponsors to share preclinical data with regulators (http://www.cdisc.org/send). However, the SEND domains apply to animal studies, and it has been necessary to expand the domains in CEBS to cover all the data it stores. The CEBS domains that currently contain data are shown in Supplementary Table S1.

As a single study can contain a number of different data types, it is relevant to consider the number of studies per data type to understand the full range of data that is available in the CEBS (Figure 2B). The most frequently used data domain is genetic toxicity which occurs in 6083 studies (July 2016). The NTP has conducted genetic toxicity testing for over 30 years. The current testing program evolved from a broader initiative developed as a predictor of rodent carcinogenicity that included *in vitro* and *in vivo* assays. Both legacy and current test data are available in CEBS. After genetic toxicity, the next most prevalent data domains in CEBS are those associated with general toxicology and carcinogenicity studies: histopathology (2308 studies), organ weights (1802 studies), in-life observations (1528 studies) and gross observations (993 studies). These endpoints are collected in the majority of legacy and current NTP toxicology assessments. Other data domains are captured in more recent NTP investigations only. For example: clinical pathology (clinical chemistry (276 studies), hematology (350 studies), and urinalysis (40 studies)), immunology (97 studies), and tissue burden / biological sample analysis (57 studies). Currently there are 249 studies with reproductive data and 10 studies with developmental data. As NTP has expanded efforts to evaluate non-cancer endpoints, and to include high-throughput screening and literature analysis, we anticipate that the data domains in CEBS will expand in the coming years.

In our last publication in 2008, CEBS contained 26 microarray studies (1). Since then, we have expanded this domain considerably to include additional studies not only from NTP but other institutions also: Environmental Protection Agency (EPA), National Cancer Institute (NCI), Health and Environmental Sciences Institute (HESI), and academic and commercial laboratories. Two large datasets exist in this domain: the NTP DrugMatrix dataset (14,18) which includes results from thousands of experiments with therapeutic, industrial and environmental chemicals tested in rats or in primary rat hepatocytes, as well as the Johnson and Johnson dataset (19,20) which includes non-proprietary testing results from over 100 liver active drugs and compounds from the Johnson and Johnson Toxicogenomics program. Mining of these datasets is possible using the CEBS Guided Search tools (Look Up Gene Signatures and J&J Codelink Data) described below. All data are available for download from the CEBS FTP site (ftp://anonftp.niehs.nih.gov/ntp-cebs/datatype/).

DATA CURATION

Since 2008 we have expanded our manual content curation practices and streamlined our processes for user interface development. Results from NTP toxicity studies and other high quality publicly available studies are processed into the CEBS database in one of two ways: manually or electronically. Processing method is determined by the size of the dataset and the frequency with which data are received but

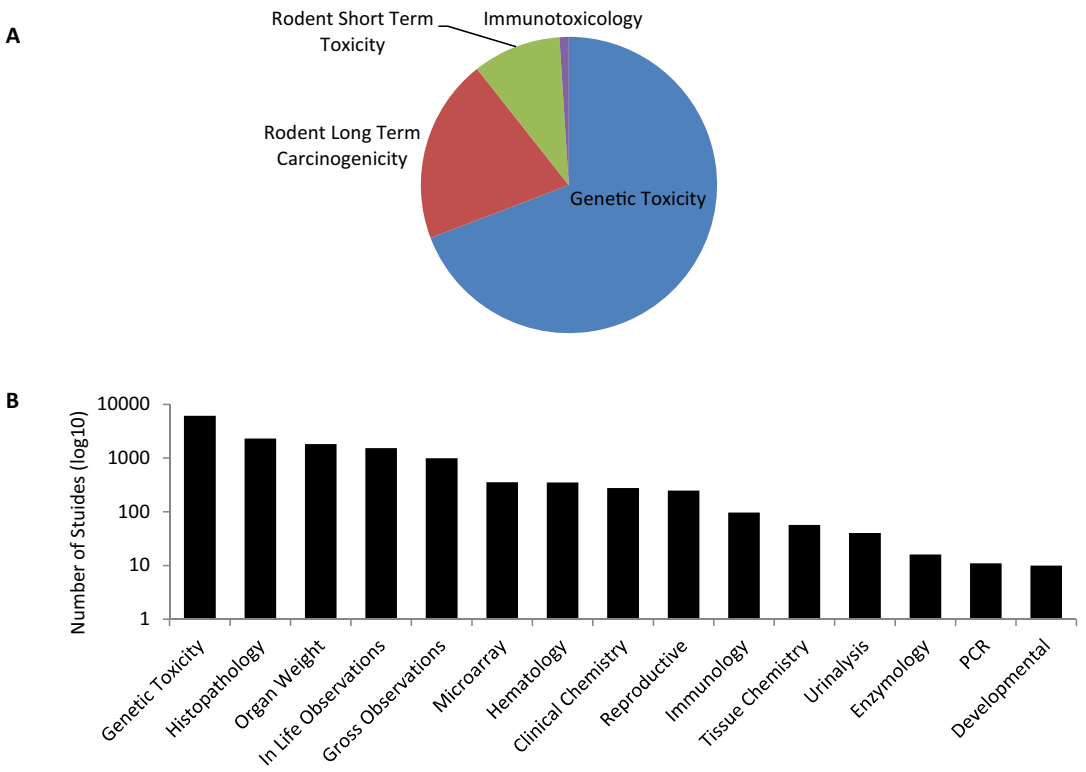

Figure 2. CEBS Content. (**A**) Number of studies and relative proportion of NTP study types in CEBS (July 2016). All NTP data currently available for these study types can be accessed from the CEBS interface. On-going studies are added as they become available. (**B**) Number of studies for each data type (NTP and non-NTP; July 2016). One study may contain multiple data types (e.g. histopathology, hematology and organ weight) and is counted in each data domain. Graph illustrates the range of data types that are available in CEBS.

in both cases data are subject to some level of manual curation.

Study data are provided to CEBS in a variety of different formats including Extensible Markup Language (XML) files, text files and Excel files. Prior to loading into the database, data curators process all data into SIFT (Simple Investigative Formatted Text) files. These are tab delimited text files with a flexible structure to support any data content.

For small datasets (e.g. a small number of studies received from a single investigator) or infrequently received data formats (e.g. data uniquely formatted for a small number of studies), data are reviewed, and processed to SIFT files manually. For large datasets with a consistent data format (e.g. NTP toxicological testing program data that are frequently deposited in CEBS and always in the same format), data are reviewed, and processed to SIFT files electronically. The processing steps employed in these cases are dependent on the format of the received dataset. In most cases, data are processed to SIFT files using Java programs written specifically for the purpose. No matter how the data are processed into SIFT, a two-step quality assurance process is performed. The first quality assurance step ensures accuracy and completeness by comparing all metadata, qualitative and quantitative assay data provided by the depositor to the SIFT file. Standardized vocabularies and ontologies are used in CEBS to control data entry and ensure data can be effectively filtered and searched. For each study, the curators review the deposited data and determine whether new terms should be added as synonyms to terms already existing in the CEBS vocabulary or whether new endpoints are indicated. When designating new controlled vocabulary terms, CEBS utilizes standards such as SEND, Ontology for Biomedical Investigations (OBI) and International Harmonization of Nomenclature and Diagnostic Criteria (INHAND) whenever possible. In the same way, all numerical values in CEBS are stored using standard database units. This ensures that all data submitted to CEBS are displayed and reviewed in a standard and comparable format.

The second quality assurance step confirms the scientific accuracy of the SIFT files by ensuring that all study components (metadata, data and study timeline) are accurately captured according to the depositor's study design. This is performed by biocuration of the SIFT and review of the data in the user interface to ensure an accurate and complete reporting of the data.

ACCESSIONING

All data in CEBS are assigned accession numbers so that each individual study and investigation is tracked. Assignment of accession numbers occurs as new studies are loaded into the database. These unique identifiers are simple to apply and constructed to be expandable. They are constructed of five parts in the format 002-02916-0004-0000-4, and have

components for study institution, depositor and study number with a checksum suffix. Users can search CEBS using an accession number to identify a specific study or dataset. CEBS accession numbers will serve as the basis for Digital Object Identifiers by the NTP.

CEBS USER INTERFACE

With the expansion in the CEBS dataset and redesign of the database, the user interface has undergone significant updates. Development of the data delivery component of CEBS underlies these changes and was designed to provide faster access to the data and to improve the end user's experience.

As the data are moved from the data collection side of CEBS to the data delivery side they are pre-processed to standardize terms, data and analysis and to improve the performance of the user interface. A series of dedicated user interface tables containing denormalized data have been designed to support each component of the user interface, and to house pre-processed data. The denormalization allows for storing redundant data and is used as a way to optimize database performance. For instance, if the user interface displays the number of studies with a particular set of attributes, then the underlying denormalized tables house both the counts (numeric) and the relevant attributes (varchar2). When a user accesses a CEBS search the data returned comes directly from the data delivery tables rather than the data collection database. These tables are shown in the CEBS schema at ftp://anonftp.niehs.nih.gov/ntp-cebs/tools/Database/.

Users who wish to access large datasets which cannot be displayed efficiently or searched in the user interface can access all data for each data domain using the CEBS FTP site (ftp://anonftp.niehs.nih.gov/ntp-cebs/datatype/). All data from the CEBS database are freely available for download in a tab delimited text format so users can employ their own analyses. These data are also accessible from the HealthData.gov website at: http://www.healthdata.gov/dataset/chemical-effects-biological-systems-cebs.

From the CEBS home page, users can search CEBS for all results for a particular test article and retrieve summary and individual subject data from all related studies. Common queries of CEBS datasets are supported with Guided Search tools that facilitate direct access and retrieval of the data of interest. Currently we have constructed nine Guided Search tools described in Supplementary Table S2. Several of these provide users with access to conclusion and summary data for chemicals tested by the NTP (Figure 3). For example, the Treatment-Related Findings Guided Search enables users to input one or more CAS numbers and view a single data table containing user specified endpoints such as NTP levels of evidence of carcinogenicity activity, site specific neoplasms associated with exposure, genetic toxicity results and conclusions, and summaries such as statistically significant neoplastic and non-neoplastic lesion incidence.

The majority of Guided Searches enable the users to define and refine search criteria for the selected dataset using a series of check boxes and pull down lists. No prior knowledge of the data or database is required to access and begin making data selections. Once search criteria are defined, a results page provides the user with a summary of the endpoint and relevant metadata. Results for any Guided Search can be downloaded and saved by the user. When a user identifies a study of interest, the metadata provided with the Guided Search enable the user to search for all data associated with the study in the Search Study section of the home page. Some tools enable users to compare their data to the NTP dataset. One example of this is the Look Up Gene Signatures Guided Search that enables users to access and mine the DrugMatrix dataset. This dataset includes clinical chemistry, hematology, histology, body and organ weight, and clinical observation data along with toxicogenomic profiles of 638 different compounds. Gene signatures for distinct phenotypes provide information about organ-specific pathologies and modes of toxicological action. To use the Look Up Gene Signature tool, users upload their own microarray data and search this dataset for similar expression profiles. The results are provided with reference to the toxicity signatures, pathology data and drug literature from DrugMatrix (13,14) and provide a way to use toxicogenomic data to perform rapid toxicological evaluations.

The underlying data for these Guided Searches are updated on a quarterly basis when the CEBS database is refreshed to add studies newly released into the public domain. The Look Up Test Article, Explore Conclusions, Organ Sites & Neoplasia, NTP Pathology Data, and Look Up BMD (Benchmark Dose) Values (21–24) Guided Searches are updated at the same time to include the new studies. The Publications Guided Search is updated as new citations are received. The DrugMatrix, J&J Codelink Data and 2012 Mouse Liver DB datasets are complete, standalone datasets for which no updates are required.

FUTURE DIRECTION AND SUMMARY

The CEBS repository has expanded rapidly in the last 5 years in response to the data deposition needs of the NTP. A 370-fold increase in the number of curated studies in CEBS has been achieved by incorporating NTP legacy and current toxicological testing data. In the coming years, CEBS will continue to capture individual animal experimental data from NTP and other sources. To accommodate this expansion the data collection part of CEBS has been re-designed to an integrated relational database. This has ensured the current system is flexible enough to capture disparate study types and was implemented with a design approach that will safeguard longevity and facilitate growth.

The expanded CEBS dataset has prompted updates to the data delivery part of CEBS. Creation of dedicated series of denormalized tables has provided faster access to the data and enabled the development of Guided Search tools that make the CEBS datasets more accessible to users with no prior knowledge of their contents. One of the limitations of the current CEBS interface is the technology platform on which it is built. CEBS was built as an Adobe Flex application that requires users to install and run the Adobe Flash plugin. As the database currently has a user community that access the data through the web interface, the platform on which the interface runs is of critical importance. The most common challenges for users of the system are

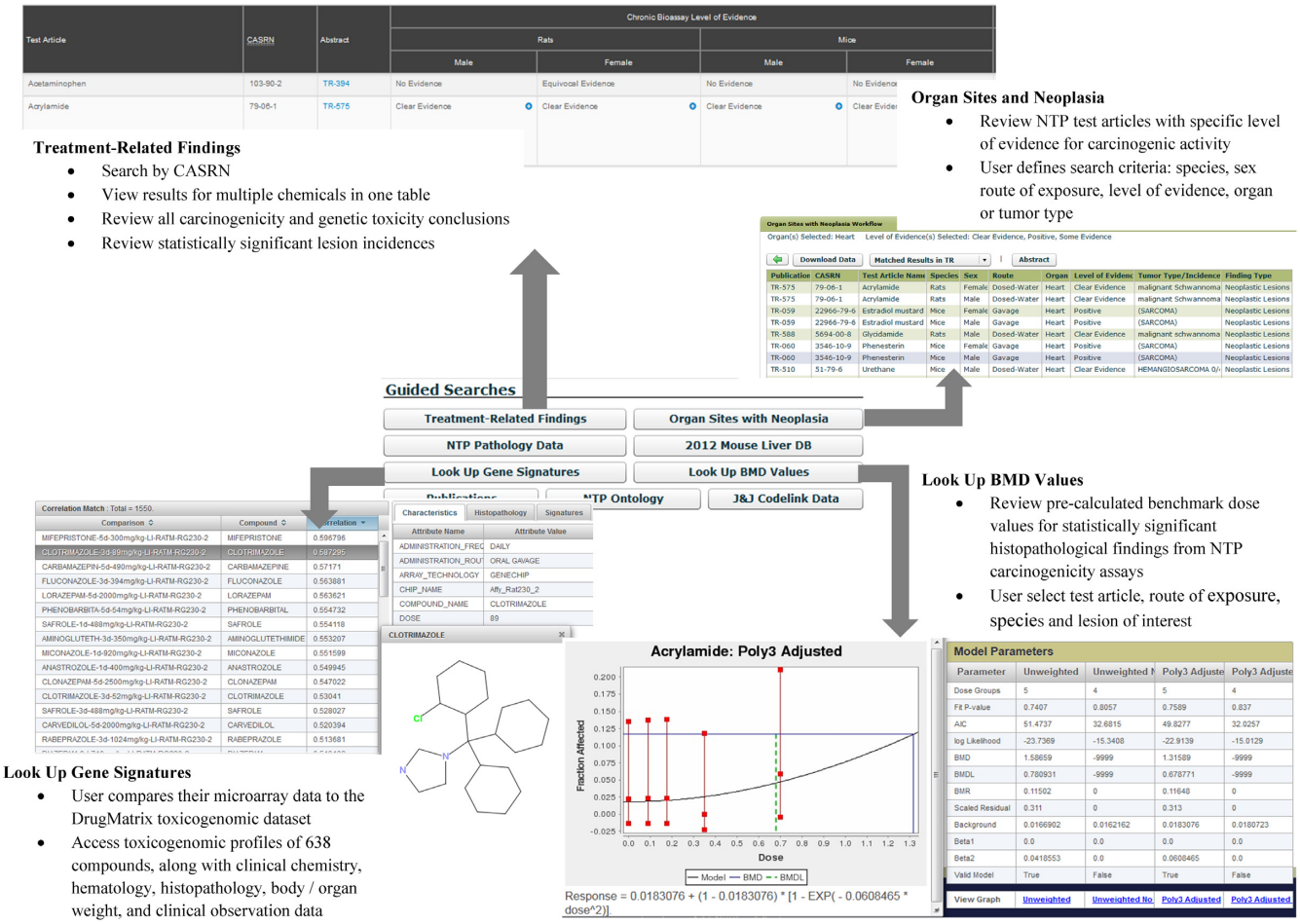

Figure 3. Guided Search Functionality in CEBS. Guided searched accessible from the CEBS home page allow users to access and search NTP summary and conclusion data (Treatment-Related Findings and Organ Sites and Neoplasia), review pre-calculated benchmark dose values for NTP carcinogenicity studies (Look Up BMD Values) and access DrugMatrix toxicogenomic profiles (Look Up Gene Signatures).

related to the limitations of the Adobe Flex platform used by the NIEHS. To address this issue, we are beginning a new effort to remove the dependency on Flex by updating the interface to an HTML platform, permitting a responsive design that works across all user devices. In addition, the size of searches that can be performed in real time is currently limited by the response time of the infrastructure. We are therefore working to install new servers and move the data delivery component of CEBS from Oracle to MongoDB and Elasticsearch.

Our goal for development of CEBS has been to develop a toxicological resource that provides access to an accurate, stable dataset from a user friendly platform. To promote this goal we have put emphasis on our data curation and quality assurance practices and expanded the functionality of our user interface. Currently an effort is under way to integrate calculated endpoints such as BMD analysis of neoplastic and non-neoplastic lesions, as well as statistically significant changes in endpoints such as body weight, organ weight or immunotoxicology responses into the Treatment Related Findings Guided Search. With these changes, the tool will concurrently provide many conclusion and summary values for many test articles and will promote their simultaneous evaluation and the possibility of new correlations and findings. Another initiative is to design a simple Guided Search tool that will allow users to access Tox21 high throughput screening data (15,25). The challenges inherent in display of large datasets can limit their usefulness. Our goal is to develop tools for the Tox21 data that will enable user access to the level of detail required for their analysis in a simple and straightforward format.

The expansion of CEBS described in this paper now places the database in a favorable position to start integrating with other database resources. As any single database cannot contain all relevant information, the next step in development of CEBS as toxicology resource is to incorporate links to external resources with toxicological data such as PubChem (8), ACToR (5) or CTD (7). This effort will advance the utility of CEBS and increase its value as a toxicology resource of use to environmental health scientists.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank all members of the CEBS database team who have made contributions to building and populating the database over the years. We would also like to acknowledge the many NTP scientists that have provided feedback that have helped improve the functionality and utility of CEBS.

FUNDING

ASRC Federal Vistronix under National Institute of Environmental Health Sciences (NIEHS) contracts 'Support for Toxicological Data for the National Toxicology Program (NTP)' [HHSN31620120054W and HHSN273201000063U GS06F0629Z]. Funding for open access charge: National Institute of Environmental Health Sciences.
Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. Waters,M., Boorman,G., Bushel,P., Cunningham,M., Irwin,R., Merrick,A., Olden,K., Paules,R., Selkirk,J., Stasiewicz,S. *et al.* (2003) Systems toxicology and the Chemical Effects in Biological Systems (CEBS) knowledge base. *EHP Toxicogenomics*, **111**, 15–28.
3. Xirasagar,S., Gustafson,S.F., Huang,C.C., Pan,Q., Fostel,J., Boyer,P., Merrick,B.S., Tomer,K.B., Chan,D.D., Yost 3rd,K.J. *et al.* (2006) Chemical effects in biological systems (CEBS) object model for toxicology data, SysTox-OM: design and application., *Bioinformatics* **22**, 874–882.
4. Fostel,J., Choi,D., Zwickl,C., Morrison,N., Rashid,A., Hasan,A., Bao,W., Richard,A., Tong,W., Bushel,P.R. *et al.* (2005) Chemical effects in biological systems–data dictionary (CEBS-DD): a compendium of terms for the capture and integration of biological study design description, conventional phenotypes, and 'omics data. *Toxicol. Sci.*, **88**, 585–601.
5. Davis,A.P., Grondin,C.J., Lennon-Hopkins,K., Saraceni-Richards,C., Sciaky,D., King,B.L., Wiegers,T.C. and Mattingly,C.J. (2015) The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic Acids Res.*, **43**, D914–D920.
6. Members, S.I.B.S.I.o.B. (2016) The SIB Swiss Institute of Bioinformatics' resources: focus on curated databases. *Nucleic Acids Res.*, **44**, D27–D37.
7. Judson,R.S., Martin,M.T., Egeghy,P., Gangwal,S., Reif,D.M., Kothiya,P., Wolf,M., Cathey,T., Transue,T., Smith,D. *et al.* (2012). Aggregating data for computational toxicology applications: the U.S. Environmental Protection Agency (EPA) Aggregated Computational Toxicology Resource (ACToR) System. *Int. J. Mol. Sci.*, **13**, 1805–1831.
8. Kim,S., Thiessen,P.A., Bolton,E.E., Chen,J., Fu,G., Gindulyte,A., Han,L., He,J., He,S., Shoemaker,B.A. *et al.* (2016). PubChem Substance and Compound databases. *Nucleic Acids Res.*, **44**, D1202–D1213.
9. Hastings,J., Owen,G., Dekker,A., Ennis,M., Kale,N., Muthukrishnan,V., Turner,S., Swainston,N., Mendes,P. and Steinbeck,C. (2016) ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.*, **44**, D1214–1219.
10. Igarashi,Y., Nakatsu,N., Yamashita,T., Ono,A., Ohno,Y., Urushidani,T. and Yamada,H. (2015) Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Res.*, **43**, D921–D927.
11. Kolesnikov,N., Hastings,E., Keays,M., Melnichuk,O., Tang,Y.A., Williams,E., Dylag,M., Kurbatova,N., Brandizi,M., Burdett,T. *et al.* (2015) ArrayExpress update–simplifying data submissions. *Nucleic Acids Res.*, **43**, D1113–D1116.
12. Kinsner-Ovaskainen,A., Rzepka,R., Rudowski,R., Coecke,S., Cole,T. and Prieto,P. (2009) Acutoxbase, an innovative database for in vitro acute toxicity studies. *Toxicol. In Vitro*, **23**, 476–485.
13. Ganter,B., Tugendreich,S., Pearson,C.I., Ayanoglu,E., Baumhueter,S., Bostian,K.A., Brady,L., Browne,L.J., Calvin,J.T., Day,G.J. *et al.* (2005) Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. *J. Biotechnol.*, **119**, 219–244.
14. Natsoulis,G., El Ghaoui,L., Lanckriet,G.R., Tolley,A.M., Leroy,F., Dunlea,S., Eynon,B.P., Pearson,C.I., Tugendreich,S. and Jarnagin,K. (2005) Classification of a large microarray data set: algorithm comparison and analysis of drug signatures. *Genome Res.*, **15**, 724–736.
15. Hsieh,J.H., Sedykh,A., Huang,R., Xia,M. and Tice,R.R. (2015) A data analysis pipeline accounting for artifacts in Tox21 quantitative high-throughput screening assays. *J. Biomol. Screen*, **20**, 887–897.
16. Chen,S., Hsieh,J.H., Huang,R., Sakamuru,S., Hsin,L.Y., Xia,M., Shockley,K.R., Auerbach,S., Kanaya,N., Lu,H. *et al.* (2015) Cell-based high-throughput screening for aromatase inhibitors in the Tox21 10K library. *Toxicol. Sci.*, **147**, 446–457.
17. Behl,M., Hsieh,J.H., Shafer,T.J., Mundy,W.R., Rice,J.R., Boyd,W.A., Freedman,J.H., Hunter,E.S. 3rd, Jarema,K.A., Padilla,S. *et al.* (2015) Use of alternative assays to identify and prioritize organophosphorus flame retardants for potential developmental and neurotoxicity. *Neurotoxicol. Teratol.*, **52**, 181–193.
18. Ganter,B., Snyder,R.D., Halbert,D.N. and Lee,M.D. (2006) Toxicogenomics in drug discovery and development: mechanistic analysis of compound/class-dependent effects using the DrugMatrix database. *Pharmacogenomics*, **7**, 1025–1044.
19. McMillian,M., Nie,A.Y., Parker,J.B., Leone,A., Bryant,S., Kemmerer,M., Herlich,J., Liu,Y., Yieh,L., Bittner,A. *et al.* (2004) A gene expression signature for oxidant stress/reactive metabolites in rat liver. *Biochem. Pharmacol.*, **68**, 2249–2261.
20. McMillian,M., Nie,A.Y., Parker,J.B., Leone,A., Kemmerer,M., Bryant,S., Herlich,J., Yieh,L., Bittner,A., Liu,X. *et al.* (2004) Inverse gene expression patterns for macrophage activating hepatotoxicants and peroxisome proliferators in rat liver. *Biochem. Pharmacol.*, **67**, 2141–2165.
21. U.S. EPA. (2005) Guidelines for carcinogen risk assessment. *Federal Register*, **70**, 177650–18717.
22. Crump,K. (2002) Critical issues in benchmark calculations from continuous data. *Crit. Rev. Toxicol.*, **32**, 133–153.
23. Crump,K.S. (1995) Calculation of benchmark doses from continuous data. *Risk Anal.*, **15**, 79–89.
24. Crump,K.S. (1984) A new method for determining allowable daily intakes. *Fundam. Appl. Toxicol.*, **4**, 854–871.
25. Attene-Ramos,M.S., Miller,N., Huang,R., Michael,S., Itkin,M., Kavlock,R.J., Austin,C.P., Shinn,P., Simeonov,A., Tice,R.R. *et al.* (2013) The Tox21 robotic platform for the assessment of environmental chemicals–from vision to reality. *Drug Discov. Today*, **18**, 716–723.