EADB: An Estrogenic Activity Database for Assessing Potential Endocrine Activity

Jie Shen,* Lei Xu,† Hong Fang,‡ Ann M. Richard,§ Jeffrey D. Bray,¶ Richard S. Judson,§ Guangxu Zhou,* Thomas J. Colatsky,†† Jason L. Aungst,¶¶ Christina Teng,††† Steve C. Harris,* Weigong Ge,* Susie Y. Dai,# Zhenqiang Su,* Abigail C. Jacobs,** Wafa Harrouk,†† Roger Perkins,* Weida Tong,* and Huixiao Hong*,1

*Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, Arkansas 72079; †School of Materials Science and Engineering, Chang’an University, Nan Er Huan Zhong Duan, Xi’an City 710064, China; ‡Office of Scientific Coordination, National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, Arkansas 72079; §National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina 27711; ¶ Division of Reproductive and Urological Products and †Division of Drug Safety Research, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland 20993; ††Division of Food Contact Notifications, Office of Food Additive Safety, Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, College Park, Maryland 20740; ‡‡Division of National Toxicology Program, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, North Carolina 27709; Office of the Texas State Chemist and Veterinary Pathobiology, Texas A&M University, College Station, Texas 77843; **Office of New Drugs and †††Division of Nonprescription Clinical Evaluation, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland 20993

1To whom correspondence should be addressed at Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079. Fax: (870) 543-7854. E-mail: huixiao.hong@fda.hhs.gov

Received May 20, 2013; accepted July 16, 2013

Endocrine-active chemicals can potentially have adverse effects on both humans and wildlife. They can interfere with the body’s endocrine system through direct or indirect interactions with many protein targets. Estrogen receptors (ERs) are one of the major targets, and many endocrine disruptors are estrogenic and affect the normal estrogen signaling pathways. However, ERs can also serve as therapeutic targets for various medical conditions, such as menopausal symptoms, osteoporosis, and ER-positive breast cancer. Because of the decades-long interest in the safety and therapeutic utility of estrogenic chemicals, a large number of chemicals have been assayed for estrogenic activity, but these data exist in various sources and different formats that restrict the ability of regulatory and industry scientists to utilize them fully for assessing risk-benefit. To address this issue, we have developed an Estrogenic Activity Database (EADB; http://www.fda.gov/ScienceResearch/BioinformaticsTools/EstrogenicActivityDatabaseEADB/default.htm) and made it freely available to the public. EADB contains 18,114 estrogen activity data points collected for 8212 chemicals tested in 1284 binding, reporter gene, cell proliferation, and in vivo assays in 11 different species. The chemicals cover a broad chemical structure space and the data span a wide range of activities. A set of tools allow users to access EADB and evaluate potential endocrine activity of chemicals. As a case study, a classification model was developed using EADB for predicting ER binding of chemicals.

Disclaimer: The findings and conclusions in this article have not been formally disseminated by the U.S. Food and Drug Administration, the U.S. Environmental Protection Agency, and National Institutes of Health and should not be construed to represent any agency determination or policy.

Key Words: endocrine disruptor; estrogen receptor; estrogenic activity; database.

Endocrine-active chemicals and endocrine disruptors (EDs) have been the subject of intense scientific discussions over the past two decades because of their potential to interfere with hormone (endocrine) systems in both humans and wildlife (De Coster and van Larebeke, 2012; Zoeller et al., 2012). In 1996, the U.S. Congress passed two laws, the Food Quality Protection Act of 1996 (FQPA 1996) and the Safe Drinking Water Act Amendments of 1996 (SDWA Amendments 1996). Pursuant to the two acts, the U.S. Environmental Protection Agency (EPA) launched the Endocrine Disruptor Screening Program (EDSP) to evaluate chemicals for possible effects on the endocrine system in humans and wildlife (EPA, 1998; Willett et al., 2011).

The endocrine system is composed of glands that produce and secrete hormones and their corresponding receptors and metabolizing and synthesizing enzymes (i.e., steroidogenic), (Luu-The and Labrie, 2010; Miller, 2002; Nelson and Bulun, 2001; Wilson, 2009), as well as the proteins that compete for EDs in serum (Hong et al., 2012). EDs can mimic the effect of endogenous hormones and exert a harmful effect by causing inappropriate responses or can block the interaction of hormones with endogenous receptors, resulting in adverse effects on developmental, reproductive, neuroendocrine, and immune systems (Daston et al., 2003). Binding to hormone receptors is one class of molecular initiating events that can
lead to disruption of the endocrine function. The tier 1 assays of EPA's EDSP were developed to identify chemicals with the potential to interact with major endocrine pathways, and include tests for direct interactions with the estrogen receptor (ER) and androgen receptor (Willett et al., 2011). ER is arguably the most important receptor and has been the subject of extensive study (Shanle and Xu, 2011; Watson et al., 2010).

ER belongs to the nuclear receptor superfamily and is widely expressed in various tissues within the body. In humans, there are two major subtypes of ER, ER-α and ER-β, which have a high degree of structural homology (Hall and McDonnell, 2005). EDs can bind to ER and interfere with normal estrogen signaling through genomic and nongenomic pathways. In the genomic pathway, estrogenic molecules such as 17β-estradiol (E2) bind to ER to form ER-ligand complexes, which then create ER-ER dimers that recruit cofactors and undergo a substantial conformational change induced by ligand binding, leading to the mature transcription factor. The ER-ER transcription factor complex can then directly bind to DNA or create transcription factors that bind to DNA, such as specificity protein 1 and activator protein 1 (Safe and Kim, 2008; Wormke et al., 2003). In the nongenomic pathway, estrogenic molecules bind to membrane-bound ERs and subsequently interfere with the membrane ER-mediated activation of second-messenger and protein kinase signaling (Yager and Davidson, 2006). Each of these mechanisms involving ER binding can lead to downstream changes that have the potential to disrupt the endocrine system depending on dose and situation.

In the last two decades, a large number of chemicals have been assayed for estrogenic activity by government agencies and academic research groups (Blair et al., 2000; Brannham et al., 2002; Shen et al., 2010, 2012). In addition, new chemical entities have been synthesized to target ER for the treatment of various diseases (Komm and Chines, 2012; Minutolo et al., 2011; Silverman, 2010). To enable and optimize the use of the data generated by these studies, we developed the Estrogenic Activity Database (EADB) to provide both scientific and regulatory communities a comprehensive and up-to-date resource for evaluating potential endocrine activity of chemicals. EADB was developed in a Java Web Start application with ORACLE database. We also implemented EADB in the Instant JChem (http://www.chemaxon.com/), which facilitates easy browsing, querying, and exporting functions. The database incorporates the most extensive collection of chemicals with publicly available estrogenic activity data obtained from in vitro and in vivo assays. The chemicals contained in EADB are from diverse sources, including drugs, pesticides, industrial chemicals, consumer product chemicals, and new chemical entities.

The estrogenic activity data curated in EADB are converted to standardized representations for comparability, demonstrate a high degree of concordance for the large majority of chemicals for which multiple study results are available, and cover a wide range of activity types and values. The large chemical space coverage and standardized representations of estrogenic activity data make EADB a useful resource and tool for assessing potential estrogenic activity of chemicals.

EADB is publicly available from http://www.fda.gov/ScienceResearch/BioinformaticsTools/EstrogenicActivityData baseEADB/default.htm. It provides the scientific community a free resource to search estrogenic activity data for chemicals of interest and to develop predictive models for assessing potential estrogenic activity of chemicals for which no estrogenic activity data are available. As a case study to demonstrate the utility of the database, a classification model for predicting ER binding of chemicals was developed using EADB.

The U.S. Food and Drug Administration (FDA) is in the process of consolidating information on substances into an agency-wide Substance Registration System (SRS). The SRS assigns a Unique Ingredient Identifier (UNII) that is used in product listing to unambiguously identify a substance. A future version of EADB will contain the UNII, which will allow this essential resource to be integrated into the regulatory framework across the FDA and outside the FDA.

MATERIALS AND METHODS

Data sources and curation. Data curated for EADB were extracted from three sources: the published literature, the FDA’s Endocrine Disruptor Knowledge Base (EDKB), and other publicly available data sets. Figure 1 illustrates the EADB data sources together with the workflows to curate the data. The published literature (up through 30 June 2012) covered several research areas, primarily toxicology, environmental science, and medicinal chemistry. Estrogenic in vivo data (uterotrophic assays) were transferred from EDKB. Public data sets from the National Toxicology Program and Ministry of Economy, Trade and Industry (METI) of Japan were also incorporated into EADB. The estrogenic data and associated metadata from the raw sources were separated into four categories: assay descriptions, activity data, references, and chemicals. As shown in Figure 1, each category contained multiple-related data entries. The assays were further categorized into four types: binding assays, reporter gene assays, cell proliferation assays, and in vivo assays. For binding assays, ER subtypes and ER protein domains (full-length or ligand-binding domain) used in the assays were recorded in EADB. The estrogenic activity data for each compound in each assay were recorded separately and associated with the assay description, literature or database reference, and chemical structure using the internal identification numbers. For compounds in the Chemical database of European Molecular Biology Laboratory (ChEMBL) (https://www. ebi.ac.uk/chembl/; Gaulton et al., 2012), the structures, chemical abstracts service (CAS) numbers, and chemical names (including synonyms) were retrieved directly from ChEMBL. For all other compounds, the structures were drawn and named using Marvin Sketch (http://www.chemaxon.com/).

Briefly, all the data were sourced from accessible references and curated manually. Chemical structures, assay descriptions, and activity data were parsed into different data formats and linking between different types of data was generated at the same time. The data were manually checked. Obvious typing errors were corrected immediately. Suspected errors were carefully corrected by reading the original source.

Data standardization. The chemical structure of each compound in EADB was processed and de aromatized by JChem Standardizer (http://www. chemaxon.com/). The estrogenic activity data were standardized as described herein. In each assay, if E2 (a potent and active endogenous estrogen) was tested, then the estrogenic activity data of the other tested compounds was normalized to the relative activity (RA) data of E2. For example, the binding
affinity data were normalized to relative binding affinity (RBA) data for the binding assays in which E2 was assayed. In the same way, the activity data were normalized to RA data for the reporter gene assays and cell proliferation assays when E2 was tested. Both RBA and RA of E2 were set to 100 for normalization. After normalization, base-10 logarithmic transformation was applied to the RA data, resulting in the endpoints logRBA and logRA in EADB. Except for the unit-less RA data, activity data were standardized to nanomolar concentrations. Some of the activity data in the original references are qualitative descriptions rather than quantitative values, such as “not active,” “no binding,” or “weak binding.” We used −10,000 to represent “not active” and −5000 to represent “weak binding” in EADB, in order to facilitate numeric search on the activity data.

Data schemes and database implementation. All the data in EADB comprise two hierarchical data tree relational representations (Fig. 1 and Supplementary fig. S1). The first data tree named DATA is a joint table containing the data from tables of estrogenic activity data, assay descriptions, and references. This data tree corresponds to the Biological Activity Interface for EADB (Fig. 1). The compound table is a subtree with a one-to-many relationship with the root (Supplementary fig. S1). The second data tree named MOLECULES contains the molecular structures and other chemical identification information. This data tree corresponds to the Chemical Interface for EADB (Fig. 1). It contains three subtrees: the estrogenic activity data, CAS, and synonym tables (Supplementary fig. S1). Two interfaces (form views) enable browsing, querying, and other functions in EADB. The Java Web Start application consists of a client front-end and a database storing all the data. The Instant JChem version uses the local database to store and manage the data trees.

Data sets used in chemical structure space comparison. To examine the chemical structure space coverage of EADB, three data sets from DrugBank, FDA’s UNII from SRS, and the DSSTox TOX21S inventory were used to compare with EADB. DrugBank contains 6516 FDA-approved and experimental drugs. The chemical structures were downloaded from http://www.drugbank.ca/downloads. The UNII from SRS were generated based on molecular structures and/or descriptive information (http://fdasis.nlm.nih.gov/srs/srs.jsp). The UNII list contains 26,733 unique CAS numbers. The chemical structures were generated from the CAS numbers using an in-house program and output to a structure-data file (SDF). The TOX21S data set contains 8193 chemicals undergoing testing under the cross-federal agency Tox21 high-throughput screening (HTS) program (http://epa.gov/ncct/Tox21/; Tice et al., 2013). The chemical structures were downloaded from the EPA DSSTox Web site (http://www.epa.gov/ncct/dsstox/sdf_tox21s.html).

Calculation of Mold2 molecular descriptors. Chemical structure space can be described using molecular descriptors. The molecular descriptors used in this study were generated using Mold2 (http://www.fda.gov/ScienceResearch/BioinformaticsTools/Mold2/default.htm), a free software tool developed at the FDA and demonstrated to be reliable for numerically describing chemical structures (Hong et al., 2008). Specifically, 777 Mold2 descriptors were separately generated for the chemicals in each of the four SDF files that have molecular descriptions of chemicals from EADB, DrugBank, UNII, and TOX21S. The
calculated Mold² descriptors were output in text files for subsequent chemical structure space coverage analysis.

**Comparison of chemical structure spaces.** Principal component analysis (PCA) was applied on the four data sets. Prior to PCA, the molecular descriptors were filtered using Shannon entropy (Shannon, 1948). More specifically, for each of the four data sets, Shannon entropy was first calculated for each of the 777 Mold² molecular descriptors. Then, the molecular descriptors were sorted based on their Shannon entropy values, and the top 300 molecular descriptors were retained. Thereafter, the 300 molecular descriptors were scaled into values between 0 and 1. Finally, PCA was applied to the data sets represented by the scaled values, and the first 3 principal components (PCs) were used to compare the chemical structure spaces.

The chemical structure space of a data set is defined by a cuboid in the three-dimensional (3D) space of the first 3 PC. In addition to visual comparison, chemical structure spaces of two data sets i and j are compared quantitatively using two measures: chemical coverage (CCij) and chemical distribution similarity (CDSij). CCij is the ratio of chemicals covered by both chemical structure spaces (within both cuboids) to total of chemicals in an individual data set. It is used to measure how many chemicals in a data set are covered by the other data set. To measure the similarity of distributions of chemicals between two data sets, each of the covered two cuboids is divided into 216 subspaces (6 × 6 × 6 in the 3D space) first. Then, number of chemicals for each of the 216 subspaces is counted for the two data sets separately. The CDSij is measured using the correlation coefficient between the two vectors of 216 chemical counts.

**Analysis of data concordance.** When a chemical has multiple data points from the same type of assays, e.g., ER binding assays, the data concordance within the assay type for the chemical is defined and calculated by using Equation 1 as follows:

$$\text{Concordance} = \frac{d_{\text{active}} - d_{\text{inactive}}}{d_{\text{active}} + d_{\text{inactive}}} \times 100\%$$

Where $d_{\text{active}}$ represents number of active data points and $d_{\text{inactive}}$ indicates number of inactive data points. The concordance values of chemicals with multiple data points from the same type of assays are given in Supplementary Table S2.

When analyzing data concordance between two related but different types of assays, e.g., ER binding assays versus reporter gene assays, the overall concordance is defined and calculated using Equation 2 as follows:

$$\text{Concordance}_{\text{overall}} = \frac{c_{\text{active}} + c_{\text{inactive}} + c_{\text{discordant}}}{c_{\text{total}}} \times 100\%$$

Where $c_{\text{active}}$ is the number of chemicals with all active data points from the two types of assays, $c_{\text{inactive}}$ is the number of chemicals with all inactive data points, $c_{\text{discordant}}$ is the number of chemicals with discordant data points within the same assay type (i.e., some are active and the rest are inactive) for both types of assays, and $c_{\text{total}}$ is the number of chemicals assayed in both assay types. The concordances for the different activity sets within an assay type (all active, all inactive, and discordant data) in each type of assays are defined and calculated by using the following Equations 3–5:

$$\text{Concordance}_{\text{active}} = \frac{c_{\text{active}}}{n_{\text{active}}} \times 100\%,$$

$$\text{Concordance}_{\text{inactive}} = \frac{c_{\text{inactive}}}{n_{\text{inactive}}} \times 100\%,$$

$$\text{Concordance}_{\text{discordant}} = \frac{c_{\text{discordant}}}{n_{\text{total}}} \times 100\%,$$

Where $n_{\text{active}}$, $n_{\text{inactive}}$, and $n_{\text{total}}$ are numbers of chemicals with all active, all inactive, and discordant data from the same type of assays, respectively. $c_{\text{active}}$, $c_{\text{inactive}}$, and $c_{\text{discordant}}$ are the same as defined in Equation 2.

**Development of classification model for predicting ER binding activity of chemicals.** As a case study to demonstrate the utility of EADB in predictive toxicology, we developed a model for predicting ER binding activity. After removing 103 chemicals with discordant ER binding activity data (i.e., concordance calculated from Equation 1 is < 100%), EADB had 4719 ER binders (chemicals with concordant positive results in all the tested binding assays) and 675 nonbinders (chemicals with concordant negative results in all the tested binding assays). The molecular structures of these 5394 chemicals were exported from EADB in an SDF file that was used for calculating the 777 Mold² (Hong et al., 2008) molecular descriptors. Thereafter, the descriptors with constant value across all 5394 chemicals were removed. The values for each of the remaining 633 Mold² descriptors were then scaled to values between 0 and 1. The supervised machine learning methodology, decision forest (DF; Tong et al., 2003), was used to build the ER binding activity prediction model based on the scaled Mold² descriptors. To assess the performance of the DF model, fivefold cross-validation was conducted as shown in Figure 2. In one cross-validation step, the 5394 chemicals were randomly split into five equal portions. Four of the five portions were used to train a DF model, which was then used to predict ER binding activity for the remaining portion. This process was repeated sequentially so that each of the five portions was left out once as the testing set. The prediction results were then averaged to provide the estimate of model performance. The fivefold cross-validation was repeated 10 times using different random divisions of the 5394 chemicals. Accuracy, sensitivity, specificity, balanced accuracy, and Mathews’ correlation coefficient (MCC, defined and calculated in Equation 6) were calculated and reported for each of the 10 cross-validations.

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FN) \cdot (TN + FP)}}$$

TP, TN, FP, and FN indicate numbers of true positives, true negatives, false positives and false negatives, respectively.

**Statistical analysis.** All statistical analyses, including two-tailed t-test, PCA, and box-plot, as well as Shannon entropy calculation and scaling, were conducted using packages in R 2.15.1 (http://www.r-project.org/).

**RESULTS**

A comprehensive set of estrogenic activity data from a variety of data sources was assembled and curated (Fig. 1), with the primary data source being the published literature. We systematically searched the literature published before 30 June 2012 by using Web of Knowledge with keywords of “estrogen receptor” or “estrogenic.” In total, 14,873 data were curated from the literature. Estrogenic activity data from 444 papers published in 21 journals were loaded into EADB (Supplementary Table S1 lists the detail of the publications and the corresponding summaries of the data curated). The second major data source were reports and databases in the public domain, including 667 estrogenic activity data from the Interagency Coordinating Committee on the Validation of
Alternative Methods report “Current Status of Test Methods for Detecting Endocrine Disruptors: In Vitro Estrogen Receptor Binding Assays” (http://iccvam.niehs.nih.gov/docs/endo_docs/final1002/erbdbrd/ERBd034504.pdf) and 938 data from Risk Assessment of Endocrine Disrupters, METI, Japan (http://www.meti.go.jp/english/report/data/g020205ae.html). The 1640 in vivo uterotrophic assay activity data (1604 of mouse and 36 of rat) from FDA’s EDKB were included.

In total, EADB contains 18,114 estrogenic activity data for 8212 molecules tested in 1284 assays (binding assays, reporter gene assays, cell proliferation assays, and in vivo assays). Table 1 presents the statistics on chemicals, data, assays, and references in EADB.

The same type of assays often measured and reported various endpoints in different units. Table 2 summarizes the endpoints and their corresponding data curated in EADB. The endpoints, units, and transformation methods are recorded in the database to retain all the original information pertaining to the data stored in EADB.

Species used in the assays and the procedures of the assays are important for assessing potential estrogenic activity of chemicals, and these types of metadata have also been entered into EADB. Table 3 lists the 11 species verifiably used in the assays curated into EADB. We had difficulties confirming the species used in the assays based on the original publications for 12% (2059) of the data and consequently marked those as “unknown” in EADB.

**User Interfaces**

EADB provides different user interfaces (Fig. 3) to accommodate users with different knowledge backgrounds or different purposes. The biological data focused interface (Fig. 3A) stresses examining chemical structures with a specific estrogenic activity. The chemical structure focused interface (Fig. 3B) stresses exploring estrogenic activity data for specific chemicals. The primary component of EADB is the activity table located in the right of the window. It displays the database content and the querying results. The searching and chemical structure displaying panels are in the

---

**TABLE 1**

<table>
<thead>
<tr>
<th>Assay type</th>
<th>Chemicals</th>
<th>Data</th>
<th>Assays</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binding</td>
<td>5494</td>
<td>10,853</td>
<td>751</td>
<td>377</td>
</tr>
<tr>
<td>Reporter gene</td>
<td>1371</td>
<td>2633</td>
<td>234</td>
<td>80</td>
</tr>
<tr>
<td>Cell proliferation</td>
<td>1540</td>
<td>3039</td>
<td>297</td>
<td>107</td>
</tr>
<tr>
<td>In vivo</td>
<td>1351</td>
<td>1640</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>8212</td>
<td>18,165</td>
<td>1284</td>
<td>447</td>
</tr>
</tbody>
</table>
left of the window. They provide the structure searching and data filtering functions (Fig. 3A). By clicking the “Individual Compound” button, the molecule interface would popup, and all the information, including the molecular identifications, properties, and experimental data, will be shown clearly in this window (Fig. 3B).

The Instant JChem version of EADB provides similar interfaces as shown in Supplementary figure S2. Each interface contains three functional windows. The project window shows the organization of database components in EADB, including data trees and interfaces for using the database. The query window provides users an easy and visually oriented way to build complex queries through logical operations. The main window displays the database content of search results in a manner expected for different purposes. The project window and query window are the same for both interfaces. The only difference between the two interfaces comes from the main windows. Supplementary figure S2A gives a screenshot of the chemical structure focused interface, whereas Supplementary figure S2B shows the biological data focused interface implemented in EADB. The chemical structure focused interface displays molecular structure and related information such as name, physicochemical properties, CAS, synonyms, as well as links to PubChem (http://pubchem.ncbi.nlm.nih.gov/) and ChemSpider (http://www.chemspider.com/) in the left panel (Supplementary fig. S2A). The right panel of the main window is a joint dynamic table that lists estrogenic activity data related to the compound displayed in the left panel, with each row describing one datum and the columns representing different types of data such as estrogenic activity data, assay descriptions, and literature references. The biological data focused interface shows all the activity data in a dynamic table along with related assay, reference, and chemical structure displayed in the right side (Supplementary fig. S2B).

Table 4 summarizes the database functions implemented in EADB. Detailed instructions on using the database and the functions implemented are given in the EADB users’ manual (Supplementary data).

### Chemical Space Coverage

The utility of the database for assessing estrogenic potential of chemicals largely depends on structural similarity between

<table>
<thead>
<tr>
<th>Endpoint</th>
<th>Description</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binding assay</td>
<td>Log transfer of RBA compared with E2</td>
<td>10,853</td>
</tr>
<tr>
<td>logRBA</td>
<td>50% inhibition concentration</td>
<td>8478</td>
</tr>
<tr>
<td>IC&lt;sub&gt;50&lt;/sub&gt;</td>
<td>50% effective concentration</td>
<td>1128</td>
</tr>
<tr>
<td>EC&lt;sub&gt;50&lt;/sub&gt;</td>
<td>Binding affinity</td>
<td>63</td>
</tr>
<tr>
<td>Ki</td>
<td>Dissociation constant</td>
<td>728</td>
</tr>
<tr>
<td>Kd</td>
<td>Association constant</td>
<td>19</td>
</tr>
<tr>
<td>Ka</td>
<td>Inhibition</td>
<td>41</td>
</tr>
<tr>
<td>INH</td>
<td></td>
<td>396</td>
</tr>
<tr>
<td>Reporter gene</td>
<td>Log transfer of RA compared with E2</td>
<td>2633</td>
</tr>
<tr>
<td>logRA</td>
<td>Effective concentration equal to 10% of E2</td>
<td>1212</td>
</tr>
<tr>
<td>logRA100</td>
<td>Log transfer of effective concentration equal to 10% of E2</td>
<td>518</td>
</tr>
<tr>
<td>REC10</td>
<td>Log transfer of relative proliferative effect compared with E2</td>
<td>13</td>
</tr>
<tr>
<td>EC&lt;sub&gt;50&lt;/sub&gt;</td>
<td>50% effective concentration</td>
<td>13</td>
</tr>
<tr>
<td>IC&lt;sub&gt;50&lt;/sub&gt;</td>
<td>50% inhibition concentration</td>
<td>60</td>
</tr>
<tr>
<td>Ki</td>
<td>Antagonist activity</td>
<td>496</td>
</tr>
<tr>
<td>INH</td>
<td>Inhibition percentage</td>
<td>25</td>
</tr>
<tr>
<td>Agonism</td>
<td>Agonistic estrogenic activities</td>
<td>167</td>
</tr>
<tr>
<td>Agonism</td>
<td>Antiestrogenic antagonistic activities</td>
<td>71</td>
</tr>
<tr>
<td>Cell proliferation</td>
<td>Log transfer of relative proliferative effect compared with E2</td>
<td>57</td>
</tr>
<tr>
<td>logRPE</td>
<td>Log transfer of relative proliferative potency compared with E2</td>
<td>90</td>
</tr>
<tr>
<td>logRPP</td>
<td>Log transfer of relative efficiency compared with E2</td>
<td>207</td>
</tr>
<tr>
<td>logRE</td>
<td>50% effective concentration</td>
<td>18</td>
</tr>
<tr>
<td>EC&lt;sub&gt;50&lt;/sub&gt;</td>
<td>50% inhibition concentration</td>
<td>1649</td>
</tr>
<tr>
<td>IC&lt;sub&gt;50&lt;/sub&gt;</td>
<td>50% effective dose</td>
<td>44</td>
</tr>
<tr>
<td>GI&lt;sub&gt;50&lt;/sub&gt;</td>
<td>Concentration of drug that reduces cell growth by 50%</td>
<td>403</td>
</tr>
<tr>
<td>IC&lt;sub&gt;30&lt;/sub&gt;</td>
<td>30% inhibition concentration</td>
<td>10</td>
</tr>
<tr>
<td>Ki</td>
<td>Constant of cytotoxicity</td>
<td>22</td>
</tr>
<tr>
<td>INH</td>
<td>Inhibition percentage</td>
<td>480</td>
</tr>
<tr>
<td>Agonism</td>
<td>Agonistic activity</td>
<td>7</td>
</tr>
<tr>
<td>Antagonism</td>
<td>Antagonistic activity</td>
<td>52</td>
</tr>
<tr>
<td>In vivo</td>
<td></td>
<td>1640</td>
</tr>
<tr>
<td>logRP</td>
<td>Log transferred relative potency compared with E2</td>
<td>1640</td>
</tr>
<tr>
<td></td>
<td>Human</td>
<td>Rat</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td><strong>Binding</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemical</td>
<td>2458</td>
<td>1619</td>
</tr>
<tr>
<td>Data</td>
<td>5531</td>
<td>2241</td>
</tr>
<tr>
<td>Assay</td>
<td>430</td>
<td>107</td>
</tr>
<tr>
<td>Reference</td>
<td>195</td>
<td>66</td>
</tr>
<tr>
<td><strong>Reporter gene</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemical</td>
<td>1225</td>
<td>32</td>
</tr>
<tr>
<td>Data</td>
<td>2164</td>
<td>32</td>
</tr>
<tr>
<td>Assay</td>
<td>186</td>
<td>2</td>
</tr>
<tr>
<td>Reference</td>
<td>66</td>
<td>2</td>
</tr>
<tr>
<td><strong>Cell proliferation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemical</td>
<td>1464</td>
<td>1</td>
</tr>
<tr>
<td>Data</td>
<td>2942</td>
<td>1</td>
</tr>
<tr>
<td>Assay</td>
<td>292</td>
<td>1</td>
</tr>
<tr>
<td>Reference</td>
<td>107</td>
<td>1</td>
</tr>
<tr>
<td><strong>In vivo</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemical</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>Data</td>
<td>0</td>
<td>36</td>
</tr>
<tr>
<td>Assay</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Reference</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemical</td>
<td>4468</td>
<td>1630</td>
</tr>
<tr>
<td>Data</td>
<td>10,637</td>
<td>2310</td>
</tr>
<tr>
<td>Assay</td>
<td>908</td>
<td>111</td>
</tr>
<tr>
<td>Reference</td>
<td>298</td>
<td>68</td>
</tr>
</tbody>
</table>

The chemical being considered and chemicals selected for comparison. To assess the applicability of EADB in the evaluation of drug product safety in terms of estrogenic potential, we compared chemical structure space in EADB with drugs in DrugBank (Knox et al., 2011), a database containing the most known marketed drug products. The $CC_{ij}$ for the chemical structure spaces of EADB and DrugBank (Fig. 4A) were 99.68 and 99.75%, respectively, indicating 99.68% chemicals in DrugBank are covered by the chemical space of EADB and 99.75% chemicals in EADB are covered by the chemical space of DrugBank. The $CDS_{ij}$ between DrugBank and EADB was 0.793, indicating the distributions of chemicals in chemical structure spaces of DrugBank and EADB were comparable.

SRS is a compilation of the substances used in drugs, biologics, foods, and medical devices regulated by the FDA. For evaluating utility of EADB in assessment of estrogenic potential for all the FDA-regulated products, comparative analysis of chemical structure spaces between the UNII list in SRS and the chemicals in EADB was conducted. We observed that the chemical structure space of EADB is similar to the chemical structure space of UNII list in SRS (Fig. 4B) with a high $CC_{ij}$ value for EADB (100%) and a slightly lower $CC_{ij}$ value for UNII list (97.12%). Moreover, the $CDS_{ij}$ between UNII list and EADB in the covered space was 0.802.

As a multiple agency collaborative project, Tox21 (Kavlock et al., 2009) aims to develop, validate, and translate innovative, HTS chemical testing methods to characterize key interactions in cellular pathways for toxicological evaluation of a wide range of environmental and commercial chemicals that are regulated by and of interest to the EPA, the National Institutes of Environmental Health Sciences, the National Institutes of Health (NIH), National Chemical Genomics Center (NCGC), and the FDA. TOX21S lists the unique chemical inventory currently undergoing HTS testing in Tox21. The Tox21 library includes approximately one-third marketed drugs (NCGC) with the remaining two thirds comprising a broad diversity of environmental chemicals of concern for potential exposure or toxicity. The Tox21 inventory could be considered to be broadly representative of the chemical structure space needed for toxicological evaluation across EPA, NIH, and FDA programs. To further assess the applicability of EADB for safety evaluation of potentially estrogenic chemicals, we compared the chemical structure spaces between EADB and TOX21S. The $CC_{ij}$ values were 99.29 and 99.70% for TOX21S and EADB, respectively, indicating the chemical structure spaces (Fig. 4C) of the two data sets are similar. Furthermore, the distributions of chemicals in the covered chemical structure space are similar, and the $CDS_{ij}$ was 0.644. After excluding 997 drug compounds from TOX21S, the chemical structure space of the environmental chemicals remained high coverage with EADB (Fig. 4D) with slightly lower $CC_{ij}$ values of 99.26 and 99.68% for TOX21S and EADB, respectively, and a slight lower $CDS_{ij}$ of 0.603.

**Estrogenic Activity Ranges**

To ensure EADB contains a suitable set of chemicals with a sufficiently wide estrogenic activity range for assessing estrogenic potential, we analyzed the distribution of estrogenic activity data in EADB. The results plotted in Figure 5 demonstrate a wide estrogenic activity range, including inactive chemicals.
Fig. 3. Snapshots of EADB interfaces. Both the biological data focused interface (A) and chemical structure focused interface (B) consist of the panels of molecular structure and assay data. The query and filtering functions are implemented in the biological data focused interface. The chemical structure focused interface can be opened by clicking the “Show” Individual Compound at the top of the biological data focused interface.
EADB contains four types of assays and each of them has many data endpoints (Table 2). Among the 10,818 data from binding assays, 8471 were obtained from binding assays with the reference compound, E2, assayed in the same experiments. Out of these 8471 binding data, 799 showed no binding activity or very weak binding affinity (the left panel of Fig. 5A). The distribution of the logRBA values of the remaining 7672 are plotted in the right panel of Figure 5A. In EADB, (17β)-3-aminoestra-1,3,5(10)-trien-17-ol, a synthesized E2 analog (Wiese et al., 1997), is the most potent ER binder, with a logRBA value of 3.876, whereas desethylatrazine is the least potent ER binder, with a logRBA value of −4.7. Thus, EADB contains ER binding affinity data that span a wide range of more than eight orders of magnitude. Moreover, the distribution of logRBA values plotted in the right panel of Figure 5A is not sparse. The binding activity distributions of other endpoints with > 100 data points also show wide binding affinity ranges (Supplementary fig. S3).

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Browsing</td>
<td>The database or searching results can be browsed easily in different ways.</td>
</tr>
<tr>
<td>Searching</td>
<td>Searching can be carried out on structure (substructure search, super structure search, similarity search, full search, R-group search, and exclusion search) or on data, including numerical data (various estrogenic activity data) and text data (assay descriptions and literature references), as well as logical combinations of multiple searching operations.</td>
</tr>
<tr>
<td>Updating</td>
<td>The database can be updated through adding new chemicals or estrogenic activity data and editing the structures or data whenever errors are found.</td>
</tr>
<tr>
<td>Exporting</td>
<td>Structures and data can be exported in various formats</td>
</tr>
</tbody>
</table>

**TABLE 4**

**Database Functions Implemented in EADB**

![FIG. 4.](image)

Comparisons of EADB chemical structure space with chemical structure spaces of DrugBank (A), UNII (B), TOX21S (C), and the environmental chemicals in TOX21S (D; after excluding 997 drug compounds). PCA was conducted on the five sets of chemicals that are described by the chemical descriptors calculated using Mold. The first 3 PCs were used to represent chemical structure spaces for each of the five data sets. The pairwise comparisons of chemical structure spaces between EADB and the four data sets were performed using scatter plots of the three PCs. Color codes: red for EADB; blue for DrugBank; cyan for UNII; green for TOX21S and black for TOX21S after excluding drug compounds.
For the reporter gene assays, 1211 data were from experiments in which the reference compound E2 was also tested. Among those 1211 data, 105 showed no activity in the reporter gene assays (the left-most bar in Fig. 5B). The distribution of the logRA values for the remaining 1106 are plotted in the right panel of Figure 5B. The largest logRA is 2.841 for (1S,4R,5R)-4-(2-fluoro-4-hydroxyphenyl)-2,2,6-trimethyl-3-oxabicyclo[3.3.1]non-6-en-1-ol and the smallest logRA is −5.375 for 1-chloro-2-[2,2-dichloro-1-(4-chlorophenyl)ethenyl]benzene. The majority (> 78%) of logRA values are between −3 and 3, spanning more than six orders of magnitude (the right panel of Fig. 5B). Similar to the binding data, the reporter gene activity data in EADB are not sparse. The reporter gene activity distributions of other endpoints with >100 data are similar to that for logRA values (Supplementary fig. S3).

The cell proliferation assays covered a range of kinds of experiments and corresponding diverse sets of endpoints (Table 2). Some assays tested antiproliferation activity of compounds. For the endpoints (e.g., IC$_{50}$, ED$_{50}$, GI$_{50}$) of those assays, the original data were recorded in EADB without normalization and transformation. The most prevalent type of cell proliferation activity data in EADB are IC$_{50}$ values (concentration of testing chemical that reduces cell growth by 50%). The 1512 IC$_{50}$ values from cell proliferation assays are between 100 µM and 0.1 nM, spanning a wide range of more than six orders of magnitude. There are 134 inactive data for endpoint IC$_{50}$ (the left-most bar in Fig. 5C) for which an IC$_{50}$ value cannot be detected or extrapolated. The distribution of IC$_{50}$ values plotted in Figure 5C indicates that the cell proliferation activity data are not sparse. The cell proliferation activity distributions of other endpoints with >100 data are similar to that for IC$_{50}$ data (Supplementary fig. S5).

The in vivo assay data were generated from two different experiments. The data have been normalized to the endpoint logRP, which is the base-10 logarithm of relative potency compared with E2 (Table 2). Of the 1640 in vivo data, a very high proportion, 1455, are inactive. The remaining 185 span the logRP values range from −4 to 4, covering eight orders of magnitude (Fig. 5D).
Data Concordance

We found that the majority of the chemicals in EADB are concordant (Supplementary table S3) within-assay types. Very few chemicals are discordant: 3.3, 13.2, 14.7 and 4.8% for binding, reporter gene, cell proliferation, and in vivo assays, respectively (Fig. 6A). An interesting observation, but consistent with expectation, is that the more complex the biological endpoint in an in vitro assay, the higher the probability of discordance (three left-most bars in Fig. 6A).

For chemicals with discordant data, the distributions of chemicals are plotted in Figure 6B. Most of these chemicals have very low concordance, i.e., the number of active data is equal or close to number of inactive data.

We also analyzed concordance among the different types of assays. We first identified the chemicals tested using two types of assays. The concordance between the two compared types of assays was then analyzed using Equations 2–5.

There were 667 chemicals tested in both binding and reporter gene assays. Their overall concordance was 73.0% (Supplementary table S4). Further, concordance between these assays for active chemicals was much higher than for inactive chemicals and for chemicals with discordant data (i.e., partially active). Interestingly, most of the chemicals that did not show activity in any binding assay (101/110) were active in reporter gene assays.

Activity data from both binding and cell proliferation assays were available for 768 chemicals in EADB. The overall concordance between binding data and cell proliferation data was 83.5% (Supplementary table S5). However, the concordance for chemicals with all active data was a higher 96.2 and 86.2% for binding and cell proliferation, respectively. It should be noted, however, that the high concordances in this case are largely due to the very high “all active” rates for both types of assays. Similar to the concordance between binding and reporter gene assays, the concordance between binding and cell proliferation assays for chemicals having all inactive data was very low, again heavily influenced by the overall very low rates of “all inactive” chemicals for both binding and cell proliferation, some 1–2%.

EADB contains only 145 chemicals with activity data from both reporter gene and cell proliferation assays. Their overall...
concordance was 79.3% (Supplementary table S6). Once again, high concordance for chemicals with all active data, 87.4 and 88.8% for reporter gene and cell proliferation, respectively, was observed, whereas concordance for chemicals having all inactive data was much lower, again, largely due to the much higher incidences of “all active” versus “all inactive” in the two assay groups.

Analysis of data concordance between in vivo and in vitro assays was conducted, and the results are summarized in Supplementary table S7 (between binding and in vivo), Supplementary table S8 (between reporter gene and in vivo), and Supplementary table S9 (between cell proliferation and in vivo). Note that the total number of overlapping chemicals being compared in each case is significantly smaller than for the in vitro to in vitro comparisons in Supplementary tables S4–S6. Also noteworthy is the more balanced distribution of “all active” versus “all inactive” for the in vivo assay group in each case. As expected, the concordance between in vivo and in vitro assays was much lower than those between in vitro assays. Furthermore, chemicals that show estrogenic activity in an in vivo assay most likely exert estrogenic activity in an in vitro assay, whereas a large portion of chemicals active in in vitro did not show estrogenic activity in in vivo.

Concordance analyses demonstrated that the estrogenic activity data in EADB are generally concordant both within the same type of assays and between different types of assays, indicating the usefulness and reliability of EADB for safety assessment related to estrogenic potential of chemicals. In summary, the within-assay type concordance is higher than the cross-assay type concordance. Moreover, the concordance between in vivo and in vitro assays is lower than the concordance between the in vitro assays.

**Prediction of ER Binding Activity**

As a case study to demonstrate the utility of EADB, a DF model was developed for predicting ER binding activity. Ten repetitions of fivefold cross-validations (Fig. 2) were performed to estimate the predictive performance of the DF model, and the results are given in Figure 7. The mean accuracy, sensitivity, specificity, balanced accuracy, and MCC were 93.84 (SD = 0.25%), 98.03 (SD = 0.21%), 64.53 (SD = 2.51%), 81.35 (SD = 1.29%), and 69.66% (SD = 1.50%), respectively.

**DISCUSSION**

EADB is a rich data source for research and regulatory scientists to use to assess a chemical’s potential for estrogenicity. Endocrine disruption in both humans and wildlife is a priority concern for environmental sciences, particularly where a no effect level of exposure may be nonexistent. Given that the different modes of estrogen action figure prominently in assessing such potential and the continuing need to assess a vast and growing number of industrial chemicals for potential estrogenicity, the EADB fills an important need. FDA regulates therapeutic compounds that may contain ER agonists, partial

![FIG. 7. Performance of the 10 iterations of fivefold cross-validation. Accuracy was plotted in diamonds, sensitivity in up-triangles, specificity in down-triangles, MCC in circles, and balanced accuracy in stars.](image-url)
agonists, or antagonists, as well as medical devices, cosmetics, veterinary medicine products, and foods and food packaging that may contain estrogenic compounds.

EADB’s value is best realized for screening and prediction. Chemical structure and similarity search capabilities provide a simple means of comparing an untested chemical structure with the body of data for tested chemical structures. More valuable still is the use of the data to supervise the training of predictive models to estimate estrogenic activity solely based on chemical structure. Preferably, training set chemicals are selected to span the chemical structure space and activity range of untested chemicals on which the model will be used. Care should be taken to exclude false-positive or false-negative data points. In addition, a reasonable balance between active and inactive chemicals is desirable. Such models are valuable to industry and regulatory authorities alike to screen for potential estrogenic activity and to inform decisions as to the need for additional testing. According to EPA’s EDSP21 work plan (EPA, 2011), EPA will use computational or in silico models and molecular-based in vitro HTS assays to prioritize and screen chemicals to determine their potential to interact with endocrine systems. As a case study, the DF model we developed for predicting ER binding activity based on EADB is more accurate in cross-validation than earlier models (Hong et al., 2011). However, the performance of the model is very stable with a very small SD, 0.25%, for the 10 iterations of fivefold cross-validation. The lower specificity (64.5%) compared with the sensitivity (98.0%) is likely influenced by the very unbalanced nature of the data set (87.5% ER binders and only 12.5% ER nonbinders), reflecting less structural information about inactives compared with actives. The specificity could be expected to be improved by adding more inactive chemicals to the training set.

EADB provides an open public resource to quickly estimate the potential estrogenic activity of a new chemical entity before any testing has begun. Safety evaluation is an important part of the FDA’s mission, with risk assessment a key part of the evaluation. For cases of inadvertent exposure in regulated products where extensive testing in animals and humans is not routinely conducted, EADB may provide sufficient evidence that estrogenic activity is unlikely. For drugs, risk assessment usually takes a number of factors into consideration, such as the indication, patient population, route of exposure, duration, and the safety margins calculated from nonclinical findings at exposures relative to the expected clinical exposure. EADB would permit rapid assessment to determine if a testing for endocrine activity should be conducted earlier in development to mitigate the risk, or if any additional studies would even be needed beyond those normally conducted.

Two publicly available chemical databases, PubChem developed by the NIH (Wang et al., 2009) and ChEMBL developed by European Bioinformatics Institute (Gaulton et al., 2012), provide comprehensive and well-organized biological databases for a large number of chemicals. Two large toxicity-specific data resources, ACToR (Aggregated Computational Toxicology Resource, http://actor.epa.gov) developed by EPA (EPA, 2012) and TOXNET (Toxicology Data Network, http://toxnet.nlm.nih.gov) developed by NIH’s National Library of Medicine (Wexler, 2001), provide free access and easy searching across publicly available data for evaluating the potential risks of chemicals to human health and the environment (the full ACToR database is available for downloading, whereas TOXNET is only searchable online). However, these resources lack domain-specific knowledge for endocrine-active chemicals. Comparative Toxicogenomics Database (CTD, http://ctdbase.org) provides information about interactions between environmental chemicals and gene products and their relationships to diseases (Davis et al., 2013). It contains information on > 600,000 chemical-gene interactions, including thousands of chemical-ER interactions. Different from EADB, CTD is a biology-oriented database and focuses on the chemical-gene associations but not specific assay data. The FDA’s EDKB is an endocrine activity-specific knowledge base. It was developed to serve as a free resource for scientists to foster development of predictive computational toxicology models and to reduce dependency on slow and expensive animal experiments (Ding et al., 2010). EDKB provides domain-specific knowledge and estrogenic activity data, along with data for other types of endocrine-related endpoints, and has been frequently used by scientists for > 10 years. However, the EDKB has not been recently updated. To enhance the knowledge base, EDKB is now undergoing redevelopment to incorporate up-to-date and comprehensive sources of data related to all aspects of endocrine activity. EADB will be one of the databases incorporated into the new version of EDKB. EADB contains 5700 new chemicals and 15,000 new estrogenic activity data that are not included in EDKB. EADB only collected estrogenic activity data, whereas EDKB has other types of data related to endocrine activity such as androgenic activity data. Therefore, users are suggested to use EADB when interested in estrogenic activity and EDKB for other types of endocrine activity.

Using EADB, we observed that the percentage of discordant chemicals increases as the complexity of the biological endpoint increases across the in vitro assays. Binding assays are biochemical in nature and generally much simpler than the other two types of in vitro assays that involve cellular processes and functions. In spite of the fact that different binding assays use ERs extracted from different species or different ER domains, or that the assays might have different experimental procedures, they all directly measure binding affinity of a chemical with ERs and, thus, could easily provide similar results for the same chemical. The underlying mechanisms of reporter gene and cell proliferation assays are more complex than binding assays, and thus are likely to be more variable within the assay type and sensitive to differing experimental protocols. Protocol variation might explain our observation that the percentages of chemicals with discordant estrogenic
activity data within reporter gene and cell proliferation assays are higher than within binding assays. There are only two uterotropic models in EADB, and most chemicals tested in one of these in vivo assays were not tested in the other. The low percentage (4.8%) of discordant chemicals in in vivo assays is likely influenced by the limited amount of data. Alternatively, this could mean that only potent and efficacious compounds were tested to confirm the in vitro findings.

Interstingly, the concordance distributions of chemicals with discordant data are quite uneven (Fig. 6B). More than 50% of chemicals with discordant data have the lowest concordance of zero by Equation 1; such chemicals have an equal number of active and inactive assay results. Closer examination of the experiments for discordant chemicals could delineate causes of the differing experimental outcomes. There are also discordant chemicals that may be selective ER modulators, i.e., active in some types of cells but not others, that possibly differs in how they affect cofactor recruitment. These discordant chemicals could be useful in the forward validation step. However, until such time as these discordances are better understood, we suggest that the chemicals with very low concordance among different assays should be removed from a training data set prior to developing predictive models, as we did in developing the DF model for predicting ER binding activity.

A key use of in vitro ER assays is to screen compounds for their ability to interact with ER, and then prioritize these chemicals for more rigorous in vivo testing. An in vitro assay needs to be very sensitive for this purpose in that it should have a minimal number of false negatives. Data in Supplementary tables S7–S9 show that the binding and transactivation assays have high but not perfect sensitivity. There are a few chemicals that are active in some but not all in vitro assays, which then show activity in the in vivo uterotropic assay. The properties and experimental results for these chemicals should be examined further. However, the main point is that no single in vitro assay is perfect, so one would want to deploy multiple in vitro assays in the initial screening step to minimize false negatives. For example, Supplementary table S9 shows that the cell proliferation assays included in EADB poorly identify chemicals that will be in vivo positive. Hence, such assays would not be appropriate for a screening and prioritization effort.

Both potency and efficacy data are important for evaluating endocrine activity of chemicals. Currently, EADB has rich potency data. However, the amount of efficacy data is relatively small. The effort to curate more comprehensive efficacy data is ongoing. The future version of EADB is expected to include more efficacy data.

In summary, EADB is the most comprehensive public database of chemicals assayed for estrogenic activities. It contains carefully curated estrogenic activity data extracted from a wide array of public and literature sources for >8000 chemicals. With the powerful database functions implemented in EADB, users can easily browse, query, and export the data. Where multiple data are available for a given chemical, the data curated in EADB display a high degree of concordance in activity calls. Additionally, the results span a wide range of estrogenic activity potency. These characteristics make EADB a valuable resource and a convenient tool for assessing potential estrogenic activity of chemicals and for developing predictive models, as demonstrated by high accuracy of the DF model developed based on the database. EADB is openly available to the public, it can be supplemented, corrected, and updated, and it is easily accessed and used by scientists.

SUPPLEMENTARY DATA

Supplementary data are available online at http://toxsci.oxfordjournals.org/.

FUNDING

Research Participation Program at the National Center for Toxicological Research (J.S.) administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Food and Drug Administration.

ACKNOWLEDGMENTS

The authors acknowledge and thank Dr. Lawrence N. Callahan (Substance Registration System, Division of Scientific Computing and Medical Information, U.S. Food and Drug Administration) for his critical evaluation of an earlier draft of this article and for his editing and constructive comments. All high-performance computations were performed using the Blue Meadow in Food and Drug Administration Scientific Computing Lab. The authors declare that there are no conflicts of interest.

REFERENCES


