Mutation Research 499 (2002) 27-52

Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network: A Proposal

Ann M. Richard[a,*] and ClarLynda R. Williams[a,b]

[a]US Environmental Protection Agency, MD-68, National Health & Environmental Effects Research Laboratories, Research Triangle Park, NC 27711 USA
[b]Environmental Protection Agency Student COOP Trainee, North Carolina Central University, Durham, NC 27707 USA

Received 8 November 2001; accepted 8 November 2001

Abstract

The ability to assess the potential genotoxicity, carcinogenicity, or other toxicity of pharmaceutical or industrial chemicals based on chemical structure information is an actively pursued and shared goal of varied academic, commercial, and government regulatory groups. These diverse interests often employ different approaches and have different criteria and use for toxicity assessments, but they share a need for unrestricted access to existing public toxicity data linked with chemical structure information. Currently, there exists no central repository of toxicity information, commercial or public, that adequately meets the data requirements for flexible analogue searching, SAR model development, or building of chemical relational databases (CRD). The Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network is being proposed as a community-supported, web-based effort to address these shared needs of the SAR and toxicology communities. The DSSTox project has the following major elements: 1) to adopt and encourage the use of a common standard file format (SDF) for public toxicity databases that includes chemical structure, text and property information, and that can easily be imported into available CRD applications; 2) to implement a distributed source approach, managed by a DSSTox Central Website, that will enable decentralized, free public access to structure-toxicity data files, and that will effectively link knowledgeable toxicity data sources with potential users of these data from other disciplines (such as chemistry, modeling, and computer science); and 3) to engage public/commercial/academic/industry groups in contributing to and expanding this community-wide, public data sharing and distribution effort. The DSSTox project's overall aims are to effect the closer association of chemical structure information with existing toxicity data, and to promote and facilitate structure-based exploration of these data within a common chemistry-based framework that spans toxicological disciplines.

Keywords: Toxicity database; Structure-searchable; SA; DSSTox; Mutagenicity; Carcinogenicity

*Corresponding author: Tel: +1-919-541-3934; fax: +1-919-541-0694.
E-mail address: richard.ann@epa.gov (A.M. Richard)

Disclaimer: This manuscript has been reviewed by the US Environmental Protection Agency and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

1. Background & DSSTox proposal

The ability to gather and explore mutagenicity, carcinogenicity and other forms of toxicity data from a chemical structure perspective is central to efforts to develop structure-activity relationship (SAR) models for toxicity prediction. Chemical structure and chemical concepts (e.g. reactive functional groups, acidity, hydrophobicity, electrophilic reactivity, free radical formation), in turn, provide a common language and framework for exploring the underlying chemical reactivity bases for diverse toxicological outcomes. Hence, chemical structure should be considered an essential identifier and scientifically useful metric for chemical toxicity databases. Many existing public toxicity databases, however, have been constructed primarily as "look-up-tables" of existing data to meet the perceived data needs of toxicologists and regulators, and thus most often do not contain chemical structures or consider potential SAR uses of the data. Instead, these databases typically utilize chemical names (usually common or commercial names) and CAS numbers (Chemical Abstract Service Registry numbers). These types of chemical identifiers are non-unique, prone to transcription and formatting errors, and devoid of any chemical information. Chemical structure as a chemical identifier, on the other hand, has universally understood meaning and scientific relevance.

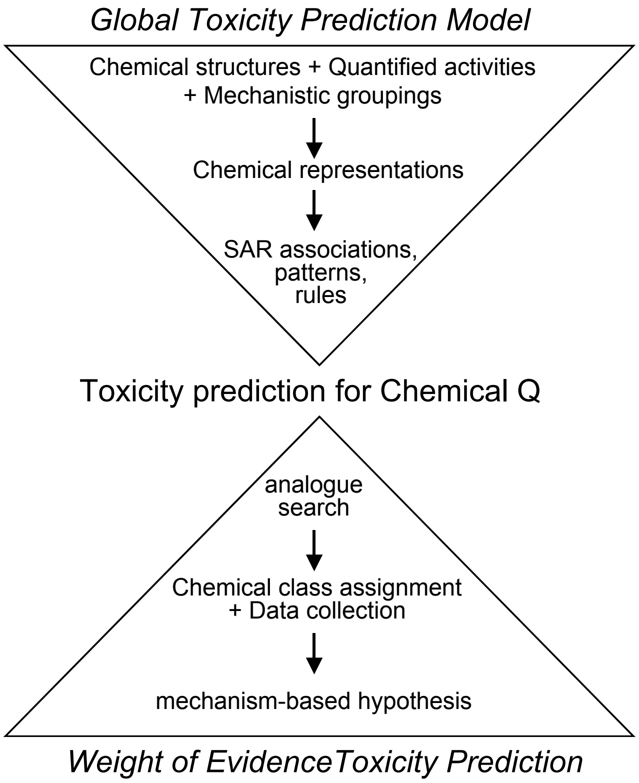

Global Toxicity Prediction Model
Chemical structures + Quantified activities
+ Mechanistic groupings
Chemical representations
SAR associations,
patterns,
rules
Toxicity prediction for Chemical Q
analogue
search
Chemical class assignment
+ Data collection
mechanism-based hypothesis
Weight of EvidenceToxicity Prediction

Fig. 1 Schematic comparison of the data-gathering activities associated with building of a global toxicity prediciton model vs. predicting the toxicity of a single chemical Q based on chemical structure.

Effective linkage of chemical toxicity data with chemical structure information can facilitate and greatly enhance data gathering and hypothesis generation in conjunction with SAR modeling efforts. These efforts generally follow one of the two paradigms represented in Fig. 1. Development of “global” SAR models for general prediction of a toxicity endpoint (e.g. rodent carcinogenicity), spanning diverse chemicals and possible activity mechanisms, requires a broad-based effort to gather toxicity data and information from wide ranging information sources. At the other end of the spectrum are efforts that expand data gathering efforts and SAR model application outward from a single chemical, to build a weight-of-evidence argument in support of a toxicity prediction. Intermediate between these two extremes, and drawing on elements of each, are efforts to develop chemical class-based or mechanism-based SARs within smaller groups of congeneric (i.e. structurally similar) chemicals. Mere knowledge of the chemical structures associated with a toxicity data base enables one to compute chemical properties that, in turn, provide the possible metrics from which SAR associations can be discerned by statistical and other means. This process of attaching structural information to existing public toxicity databases, however, can be extremely time consuming and, unless the result is publicly shared, this effort must be repeated each time a new investigator wishes to model or consider the same dataset from a structural perspective.

Beyond simply providing a listing of chemical structures, a potentially much richer ability to explore local SAR associations (i.e. within a chemical class), and to build arguments for their validity, is provided when chemical structures and toxicity data are incorporated into what is termed a "chemical relational database". For the purposes of the present article, a chemical relational database (CRD) application refers to a computer application that: 1) provides for storage of chemical records containing structures and text/data fields; 2) has some "chemical intelligence" in terms of its ability to pose general chemistry-related queries and to interpret chemical information (e.g., bonds and atoms); and 3) allows for structure/text/data relational searching across records in the database. [See Sections 2.14 and 3.3 for more detail and examples of specific CRD applications.] Relational searching refers to the coupling of diverse types of search criteria, either by sequential application (e.g., first search for chemicals containing a specified structural fragment, then search for chemicals within that group that satisfy a toxicity criteria) or using Boolean searching (e.g., search for chemicals containing a structural fragment and property value or toxicity dose range, in a single step). The power of this type of approach is apparent when considering the diverse types of information that could be included in toxicity CRDs - i.e. chemical structures, physico-chemical properties, and biological activities spanning *in vitro* and *in vivo* endpoints, multiple species, target sites, routes, and toxicological endpoints. A toxicity CRD would afford a user much greater and more flexible opportunity to explore broad-ranging aspects of chemical toxicity across multiple domains of chemical and biological information.

Currently, there exists no central repository of toxicity information, commercial or public, that adequately fulfills the broad needs for flexible and unrestricted access to public toxicity data linked with chemical structures for use in analogue and CRD searching, SAR model development, or incorporation into large corporate or in-house chemical databases. [See Section 4 for further discussion of current industry/government/commercial/public toxicity CRD database capabilities.] A major argument that could be levied in opposing the construction of such a centralized toxicity database is that the field of toxicology itself is not centralized, but rather broadly spans many domains of knowledge, biological investigation and types of information. This is clearly reflected in the current status of chemical toxicity databases, which exist in different public and private forums and in many different formats, and contain very distinct types of information. A second argument against data consolidation is that it distances the domain-specific toxicity data from the most knowledgeable sources of those data that could provide the best guidance relative to use and interpretation [1]. This, in turn, highlights a third problem: that there is currently little interaction between knowledgeable experts working in significantly different domains of toxicological study, e.g. carcinogenicity, developmental toxicity, neurotoxicity, aquatic toxicity. There is perhaps even less interaction between these distributed toxicity domain experts and the computer scientists, chemists, artificial intelligence and SAR modelers striving to understand, represent, and model diverse types of toxicological data from a chemistry standpoint.

The Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network is being proposed as a community-supported, web-based effort to address these multiple challenges and shared needs of the SAR, chemistry, computer science, and toxicology communities. The DSSTox proposal has the following three major elements:

1) **Adopt and encourage the use of a common standard file format for public toxicity databases that will include chemical structures and that can be easily imported into available chemical relational database (CRD) applications.**
Structure Data File (SDF) format (originally developed by Molecular Designs Limited, currently MDL Systems Inc.) is a public, ASCII file format that stores field-delimited structure, text and property information for any number of molecules (see also Section 2.1) [2]. SDF has already been adopted as an industry-standard import/export feature of virtually all chemical modeling and CRD applications and, thus, is ideally suited to the needs of the DSSTox project. Additionally, we are proposing to include a set of standard chemical identifier fields in all DSSTox SDF files (see Section 2.2 for details). The DSSTox SDF files that will be created for a wide variety of available public toxicological databases will then be easily convertible to data tables or importable into any commercial or private CRD application.

2) **Implement a distributed source approach that will enable decentralized, free public access to toxicity data files, and that will effectively link knowledgeable toxicity data sources with potential users and modelers of these data from other disciplines.**
The DSSTox Source refers to the person(s) or organization that compiled and currently maintains a public toxicity database for which a corresponding DSSTox SDF file has been created. Ideally, the Source would be considered the "owner" and web-based distributor of the DSSTox SDF file, would be asked to take responsibility for the file's maintenance and upgrade, and would be referenced and acknowledged in any subsequent use of that file (see Section 2.3 for further details). The DSSTox SDF file will provide summary toxicity information, as opposed to detailed descriptive text and, hence, the file may in fact be a distillation of the original toxicity database. An advantage of having the DSSTox SDF file closely associated with the Source is that a user would be encouraged to consult the Source website and original toxicity database for more complete textual descriptions, qualifications, references and guidance in the use of that toxicity data. Fig. 2 provides a schematic illustration of how a user would

retrieve DSSTox SDF files from selected distributed Sources to create a user-customized toxicity CRD that could be fully accessed, searched, reformulated, or merged with proprietary or other data.

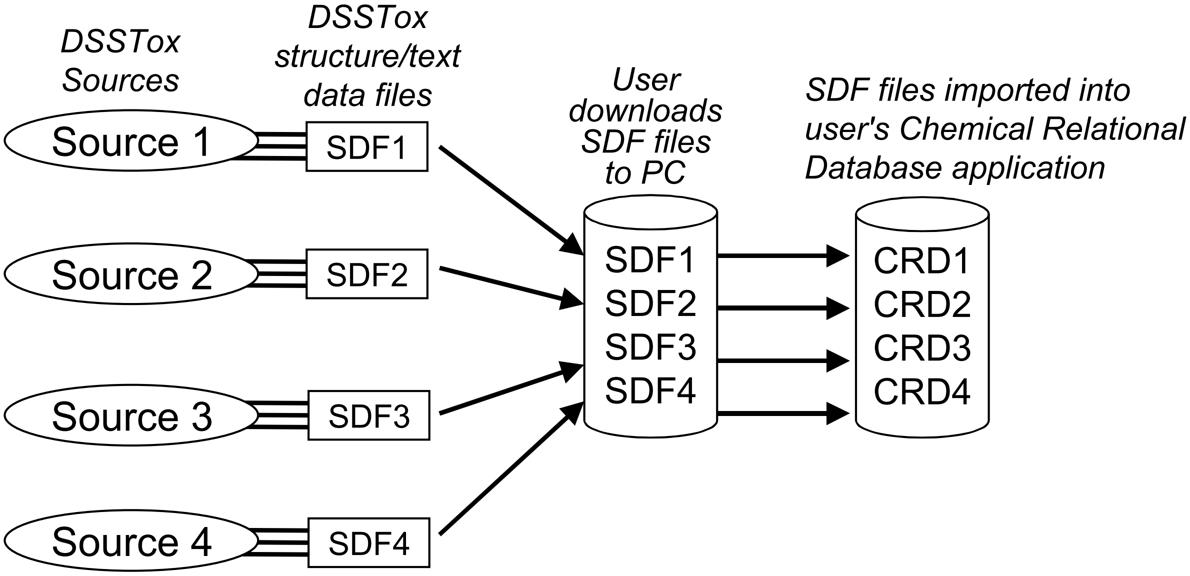

DSSTox Sources
DSSTox structure/text data files
User downloads SDF files to PC
SDF files imported into user's Chemical Relational Database application
Source 1
Source 2
Source 3
Source 4
SDF1
SDF2
SDF3
SDF4
SDF1
SDF2
SDF3
SDF4
CRD1
CRD2
CRD3
CRD4

Fig. 2 Steps to building a user-customized chemical relational database (CRD) from DSSTox Sources and SDF files.

3) **Engage public/commercial/academic/industry groups in contributing to and expanding the DSSTox public database network.**
A DSSTox Central Website (see Website Appendix, A0) will serve as the hub of the DSSTox project, providing general information on DSSTox standard file formats, field names, etc., and providing links to DSSTox Sources and SDF files, CRD vendors, and public tools and resources of general interest to the DSSTox community. Another crucial role of this website will be to connect the DSSTox user community members and to enlist their help in propagating the DSSTox recommended standards, reporting DSSTox SDF file errors to the Sources, offering enhancements to existing DSSTox SDF files, and aiding in the construction of new DSSTox SDF files. The general organization and components of this website are illustrated in Fig. 3; details of each component are provided in Section 2.10.

The remainder of this manuscript is organized into four sections. Section 2 specifies the major components of the DSSTox database network proposal. Section 3 provides details relative to construction of DSSTox SDF files for the Carcinogenic Potency Database (Website Appendix, A1) [3-6], and to viewing and searching of these files in a sample CRD application. Section 4 discusses the larger context of the DSSTox project in terms of existing toxicity database capabilities and needs, offering some perspectives of industry, government and the larger scientific community. Section 5 includes final comments pertaining to current progress on the DSSTox project, the particular relevance of the DSSTox project to the mutagenicity and carcinogenicity fields, the larger benefits of data file standardization, and our hopes for the collaborative future of this effort. Finally, a "Website Appendix" has been included at the end of this paper that will separately index and list each of the current website url addresses referred to in the text for easy future reference (i.e. Website A0, A1, A2, etc.).

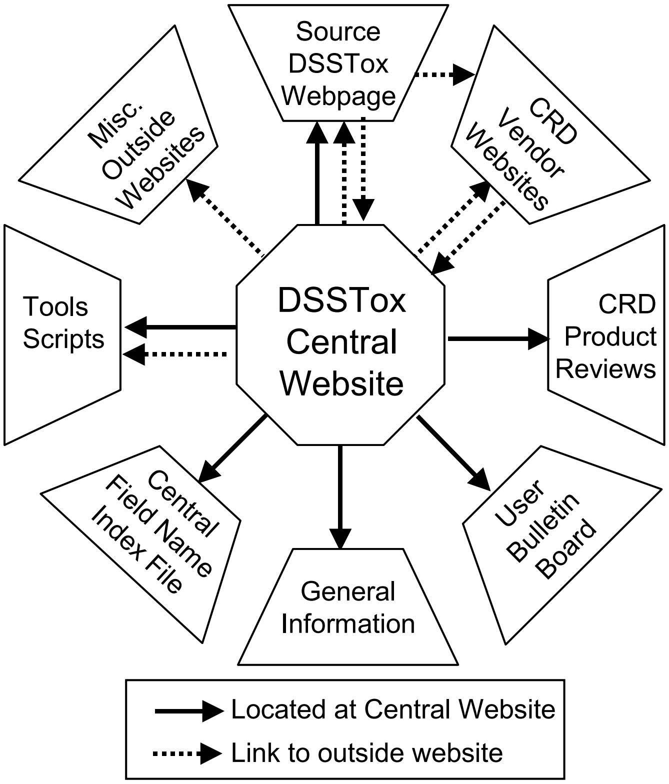


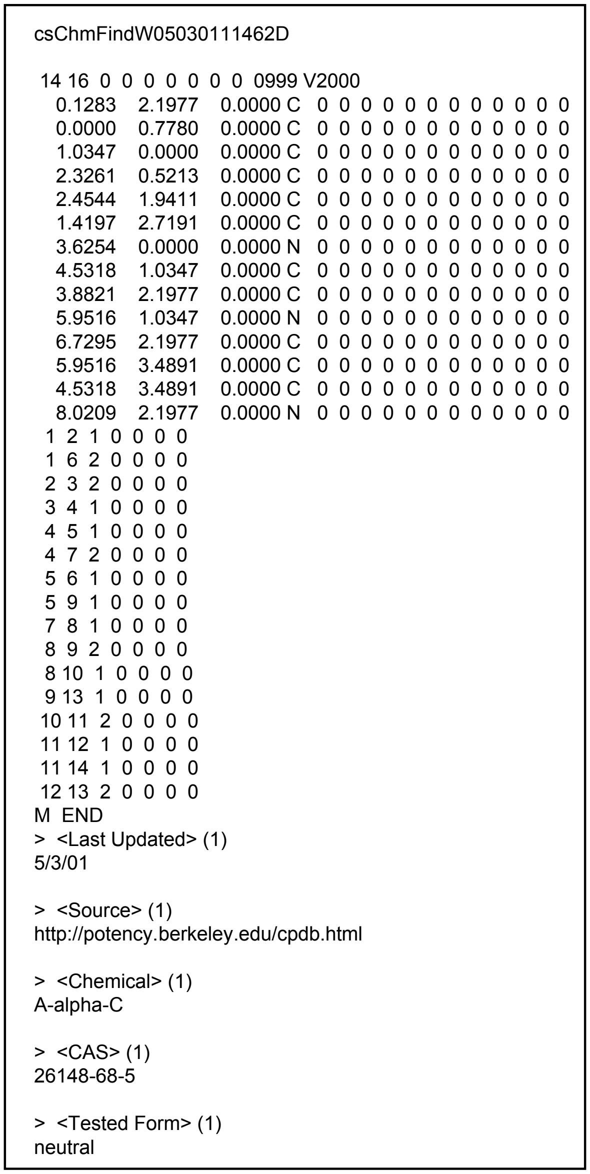
Fig. 3 Major components of the proposed DSSTox Central Website, with dashed arrows denoting interactions with outside websites (see Section 2.10 of text for further details).

2. DSSTox project specifications

2.1 SDF file

The Structure Data File format (also referred to as SDfile or SDF format) consists of a flat data-structure (i.e. no nested fields or hierarchies) that includes 2D chemical structures [2]. The number of "records" in an SDF file corresponds to the number of distinct chemicals or chemical species contained in the database. Each record contains a structure field and an unlimited number of text/data fields listed on separate lines. Each field, in turn, contains both a field name (shared by all records in the database), record identifier, and field contents (specific for the chemical record). A sample portion of an SDF file is shown in Fig. 4. SDF structures displayed in 2D representation are readily convertible to 3D structures in virtually all molecular modeling applications, enabling calculation of 3D molecular electronic and structural properties. Publically available tools for manipulating SDF files, written in the Open-Source PERL language, are also freely available on the web (see Websites A2, A3).

Although adopted as a defacto standard file format for chemistry modeling and CRD applications, we caution that there are no official industry guidelines or statements to this effect. SDF files as originally specified by MDL (Website A4) are strictly formatted [2]. However, in the course of our efforts, we have found multiple instances where CRD vendors have introduced changes to the original SDF format upon "Exporting to SDF", usually by addition of application-specific lines and information. A role of the DSSTox project will be to report such problems to users and vendors, and to recommend a "clean SDF" export option be included in CRD applications to preserve the standard and interconvertibility of the SDF format across applications.



```
csChmFindW05030111462D

 14 16  0  0  0  0  0  0  0  0999 V2000
    0.1283    2.1977    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.7780    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0  0
    1.0347    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0  0
    2.3261    0.5213    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0  0
    2.4544    1.9411    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0  0
    1.4197    2.7191    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0  0
    3.6254    0.0000    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0  0
    4.5318    1.0347    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0  0
    3.8821    2.1977    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0  0
    5.9516    1.0347    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0  0
    6.7295    2.1977    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0  0
    5.9516    3.4891    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0  0
    4.5318    3.4891    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0  0
    8.0209    2.1977    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  1  6  2  0  0  0  0
  2  3  2  0  0  0  0
  3  4  1  0  0  0  0
  4  5  1  0  0  0  0
  4  7  2  0  0  0  0
  5  6  1  0  0  0  0
  5  9  1  0  0  0  0
  7  8  1  0  0  0  0
  8  9  2  0  0  0  0
  8 10  1  0  0  0  0
  9 13  1  0  0  0  0
 10 11  2  0  0  0  0
 11 12  1  0  0  0  0
 11 14  1  0  0  0  0
 12 13  2  0  0  0  0
M  END
>  <Last Updated> (1)
5/3/01

>  <Source> (1)
http://potency.berkeley.edu/cpdb.html

>  <Chemical> (1)
A-alpha-C

>  <CAS> (1)
26148-68-5

>  <Tested Form> (1)
neutral
```

Fig. 4   Sample portion of a DSSTox SDF file for record 1 (A-alpha-C) of the Carcinogenic Potency Rodent Database (Website A1) [3-6]; the top portion of the file encodes the 2-dimensional chemical structure.

2.2 DSSTox SDF File

A DSSTox SDF file refers to an SDF file for a public toxicity database that has been standardized to conform to the DSSTox format with respect to the proposed SDF File Name convention (see Section 2.4), Field Name conventions (see Sections 2.7 and 2.8), and the inclusion of Standard Fields (see Section 2.6). Every effort will be made to insure that all DSSTox SDF files conform to strict SDF formatting such that they are importable into all major CRD applications.

Chemical structures will be represented in their neutral form whenever possible, whereas the actual "tested form" of the chemical (neutral, hydrate or salt) will be embedded in the chemical name field and CAS number and explicitly indicated in a separate chemical identifier field (see Section 2.6). Whereas most SDF records will pertain to well-defined organics with known chemical structures, many toxicity databases include records that pertain to mixtures (e.g., of isomers) or ill-defined substances (e.g, diesel fuel). These substances will not have a defined structure-field, nor will they be particularly useful in CRD searching or SAR investigations. For completeness, we have chosen to include all tested substances in the total DSSTox SDF file along with a standard field "SUBSTANCE TYPE" that will specify whether the tested substance is a “defined organic”, “mixture or ill-defined”, or “inorganic” (see Section 2.6). This field will enable a user of a CRD application to extract defined organics with known structures into a separate file in a single search step.

2.3 DSSTox Source

For present purposes, the term "Source"refers to the originator of the compiled database from which the DSSTox SDF file was created (termed the "original Source database"). An example would be the National Cancer Institute/National Toxicology Program (NCI/NTP), and associated website (A5) as the "Source" of the NTP 2-year rodent carcinogenicity bioassay database. A second example would be the Carcinogenic Potency Database (CPDB) Project [3-6], and associated website (A1), as the "Source" of the CPDB DSSTox SDF files. These examples illustrate two levels of DSSTox Sources:
1) The NTP is a primary source, i.e. the originator of all the rodent bioassay data included in the NTP database.
2) The CPDB project is a secondary source, having compiled rodent bioassay data from multiple original sources, including the NTP and a wide range of literature sources. In this case, a user would consult the DSSTox Source for further reference and guidance if the primary source of the individual data record of interest were required.

We envision the Source as the "owner" and provider of the DSSTox SDF file(s), accepting primary responsibility for maintaining these files and ensuring that they accurately reflect the contents of any corrections or enhancements to the original Source database. The Source could choose to enlist the help of the DSSTox user community in this task. Errors in chemical structures or information detected by users would be reported to the Source. In addition, users or vendors could offer enhancements (such as calculated physico-chemical properties) to the Source SDF files that could then be distributed to the larger DSSTox user community either by the Source or by the vendors, based on prior agreement and with adequate referencing.

2.4 DSSTox SDF File Name

Each DSSTox SDF file will have a unique file name that contains essential Source and contents identifier information according to the standard naming convention proposed below. The DSSTox SDF file name would be referenced in publications reporting use of that file, such that any user or interested person could trace the origin and contents of that file name through the DSSTox Central Website (see Section 2.10), the DSSTox Source Webpage (see Section 2.11), and the associated Source SDF Log file (see Section 2.5). A central index of DSSTox SDF file names and links to DSSTox Source Webpages will be found at the DSSTox Central Website. The proposed SDF file name convention is as follows:

NAMEID_V1a_00543_23mar01.sdf

NAMEID = unique Source and SDF database identifier name
There are no special guidelines except that the NAMEID would be unique to the database contents and Source, would be limited to six characters, and would be used as a general name and non-specific reference to the database and SDF file.
(examples: CPDBRO=CPDB rodent carcinogenicity database; CPDBHA=CPDB hamster carcinogenicity database, NTPCAN=NTP rodent carcinogenicity bioassay database)

V1a = Version 1, Revision a
The version number will begin with the number '1', and will increase in increments of 1 only with major additions to the database and SDF file, such as with the addition of new fields or chemicals. The revision letter will be linked to changes in the version number (starting at 'a' for each version), and will increase alphabetically (a,b,c,d, etc.) when minor corrections or modifications are made to the SDF file that do not involve addition of fields or chemical records. File modifications associated with each version update and revision beyond V1a will be documented in the corresponding Source SDF Log file (see below) of the same NAMEID and located at the DSSTox Source Webpage (see Section 2.5).

00543 = total number of records (e.g. 543)
This indicates the total number of chemicals (currently 5 digits, but can be expanded), or chemical records with associated fields, contained in the SDF file. The number will convey to a potential user the approximate size of the SDF file, and will indicate whether the file contains all records or some subset thereof. For example, a complete SDF file might contain 1300 records, of which 200 are mixtures or inorganics. A second SDF file might be created for defined organics that would be a subset of the first file and contain 1100 records. The corresponding SDF File Name in this case would change only in the total number of records.

23mar01 = ddmmmyyy = date of last SDF file revision
The date is linked to any modifications of the SDF file and conveys the date of creation or subsequent version update or revision of the database and SDF Log file.

.sdf = SDF file type extension
Although SDF is the primary generic file type that will be publically offered, there may be other application-specific files offered at the DSSTox Source Webpage. These files would contain the same basic chemical field information as the Source SDF files, but could be enhanced to include vendor-added property fields, such as octanol/water partition coefficients (i.e. logP values). Examples of other types of CRD application files include:
.cfd = Advanced Chemistry Development's ChemFolder file format (Website A6)
.cfw = CambridgeSoft's ChemOffice-ChemFinder file format (Website A7)
Another potential advantage to users in offering other types of CRD application files at the DSSTox Source Webpage is that the size of some of these files (e.g., .cfd files) may be considerably compressed from the corresponding SDF file.

2.5 DSSTox Source SDF Log File

Each DSSTox SDF file will have an associated DSSTox Source SDF Log File that will provide historical documentation for all modifications of the SDF file associated with version updates or revisions. The Log File will be in ASCII .txt format and will be located at the DSSTox Source Webpage (see Section 2.11). The naming convention of this file will mirror the convention of the DSSTox SDF File Name (see Section 2.4):

NAMEID_log_23mar01.txt

where NAMEID is shared by the corresponding SDF file, and the date indicates the latest date of modification of the SDF Log File, i.e. corresponding to the latest SDF version update or revision.

2.6 DSSTox Standard Fields

Each DSSTox SDF file will contain a common set of standard information and chemical identifier fields in addition to the Source database-specific fields. The format for allowed entries to Standard Fields will be prescribed to facilitate uniformity in across-database searching. The initial proposal is for ten Standard Fields of the following format:

STRUCTURE: 2D chemical structure in neutral form
DATE_ddmmmyy: date of creation or last version update/revision
SOURCE http://: Source url, website address
CHEMICAL: common or familiar chemical name
CAS 000000_00_0: formatted Chemical Abstracts Service number
FORMULA: empirical molecular formula
MOLWEIGHT: formula weight in atomic units
SUBSTANCE TYPE: "defined organic", "mixture or ill-defined", "inorganic"
TESTED FORM: "salt", "hydrate", "salt/hydrate", "neutral"
SMILES: linear text notation for representing 2D chemical structures

2.7 DSSTox Source Database-Specific Fields

In addition to the DSSTox Standard Fields, each DSSTox SDF file will contain fields with Source database-specific information relative to the toxicity endpoint of interest. For categorical assignments, e.g. POS or NEG, only textual field values will be allowed since character entries (e.g., +, -) are not always interpreted or transferred correctly in moving between CRD applications. For quantitative toxicity data (e.g., dose or potency values), the units of measure will be indicated in the field name. For text specification fields, such as target organ site information, the DSSTox Source Field Index File (see Section 2.8) will contain the full listing of allowed field entries. Additional fields could include calculated physico-chemical properties or other types of general purpose information. Examples include IUPAC names, alternative names or Agency-specific chemical identifiers (e.g. NTP Technical Report numbers), and calculated or vendor-supplied molecular properties (e.g. logP, pKa, vapor pressure, solubility, etc).

2.8 DSSTox Source Field Index File

Definition and description of each Field Name included in each DSSTox SDF file, including the DSSTox Standard fields, will be provided in the DSSTox Source Field Index File, located at the DSSTox Source Webpage (see Section 2.11). This information will vary in content depending on the SDF, will include description of units of measure for quantitative fields, will list allowable field entries (e.g. POS or NEG), and may contain reference or links to other Source-provided information.

2.9 DSSTox Central Field Index File

This file, located at the DSSTox Central Website (See Section 2.10), will provide a consolidated listing of all DSSTox Standard and Source SDF-specific Field names, with associated definitions and description, used in all currently available DSSTox SDF files. This file will not only consolidate the contents of all DSSTox Source Field Index Files, but will also cross reference each field name to the DSSTox SDF file in which it is contained. This feature should be particularly useful for users wishing to find common fields for performing cross-database searches, as well as for those wishing to use existing field definitions for the construction of new DSSTox SDF files.

2.10 DSSTox Central Website

This website will serve as the central hub of the DSSTox database network and will consist of the following eight sections represented schematically in Fig. 3:

1) The "DSSTox SDF Files" section will provide an indexed listing of all current DSSTox SDF file names, and links to the DSSTox Source Webpages hosting those SDF files. If located outside the DSSTox Central Website, the DSSTox Source Webpage will also link back to the DSSTox Central Website. In cases where an outside Source website does not exist or a Source chooses not to host the DSSTox Source Webpage and SDF file, this customized DSSTox Source Webpage will be accessed directly at the DSSTox Central Website.

2) The "CRD Vendor Links" section will contain links to a variety of commercial software providers of CRD applications and publicize any contributions or interactions of the vendors in supporting the DSSTox project. These CRD Vendor websites, in turn, can choose to link back to the DSSTox Central Website to provide their users with information on this resource for public toxicity information.

3) The "CRD Product Reviews" section will not endorse any particular CRD product or contain vendor-sponsored advertising or promotions. It will, however, offer some independent reviews, feature comparisons, helpful hints, and recommendations drawn from users' experiences and judgements of the strengths and limitations of the various CRD applications.

4) The "User Community" section will provide answers to frequent user questions (FAQ), and a public bulletin board to encourage free exchange of data, information, and ideas and to foster collaborations among the DSSTox user community.

5) The "General Information" section will contain detailed information on the DSSTox Project, including DSSTox file name conventions and log file specifications. It will also contain DSSTox acknowledgements, email contacts, and guidance to the DSSTox user community in providing adequate referencing and acknowledgement to the Sources of the DSSTox SDF files.

6) The "Central Field Name Index File" section will contain a posted and searchable version of the DSSTox Central Field Index File (see Section 2.9).

7) A "Tools & Scripts" section will provide links to open-source compilers (e.g. PERL - Website A3; Python, Website A8) and scripts (e.g., SDF Toolkit - Website A2) of potential use to the DSSTox community. In addition, this section will post and offer for free download open-source (PERL and Python) and application-specific scripts (e.g. based on ChemFinder's CAL scripting language) used in our initial project development and donated by the larger DSSTox user community. These could include tools to create tabular representations or otherwise manipulate SDF files, scripts for "filling" structures from one SDF file to another, and chemical property calculation modules.

8) A "Misc. Website" section will contain links to miscellaneous public resources or commercial products of possible interest to the DSSTox user community.

[Note: The DSSTox website, under development at the time of this manuscript submission, will eventually exist at the domain http://www.dsstox.net. For readers of this journal, the public launching of the website and initial DSSTox SDF offerings will be announced at www.mutationresearch.com/mutat/show/.]

2.11 DSSTox Source Webpage

Ideally, this Webpage would be customized and maintained by the Source of the database used to construct the DSSTox SDF file, and would be located at the Source website. In cases where a Source is not able or does not wish to post the DSSTox SDF, this DSSTox Source Webpage would be posted at the DSSTox Central Website, but could still be customized by the Source. At minimum, each Source Webpage would contain:

1. a link to the DSSTox Central Website;
2. the Source-specific DSSTox SDF file for public download;
3. the DSSTox Source Field Index file;
4. the DSSTox Source SDF Log file;
5. a suggested literature citation for users who publish work based on the DSSTox Source SDF files; and
6. Source contact information (e.g. "webmaster" email address) for users to report errors or to pose questions.

In addition, other application-specific DSSTox files could be offered for download from this webpage and vendor links could be provided. Further customization from the Source could include literature references or guidance in the use of the toxicity data contained in the DSSTox Source SDF file(s).

2.12 Toxicity Data

What constitutes "toxicity data" for the purposes of the DSSTox database effort will be interpreted very broadly as any experimentally determined *in vivo*, *in vitro*, or biochemical change potentially associated with a toxicological outcome or response. For the purposes of CRD searching, such data will most often consist of summary quantitative or qualitative values derived from experiment or computations. DSSTox databases will not provide toxicity information that exists in textual narrative form or that consists of details of experimentation and interpretive comments. However, a user will be encouraged to access such information at the Source. The concise summary representations of biological, chemical, or toxicity data used in DSSTox databases, in turn, will be adopted from existing Sources and scientific community standards relative to a particular set of experiments and activity measures. The DSSTox database effort will not review, reinterpret, or in any other way revise existing chemical, biological or toxicity data representations. A noteworthy exception is when errors are detected and acknowledged at the DSSTox Source, either transcription errors from the original literature sources or changes in data resulting from new experiments.

2.13 Chemical Structure-Verification

Our current efforts to create DSSTox SDF files have adopted a number of measures to verify the correctness of chemical structure information provided in original Source databases. First and foremost is ensuring the consistency of all chemical identification fields, i.e. structures, chemical names, CAS numbers, and SMILES linear text structure notation. Most of the databases that we have considered thus far have included chemical names and CAS numbers, and a few additionally have contained SMILES fields. Whenever inconsistencies have been found between any two of these fields for a given chemical record, the chemical has been flagged for further inspection and verified by independent literature or database sources. In addition, in many cases DSSTox SDF files have been populated with chemical structures by cross referencing CAS numbers or chemical names with large public sources of chemical structures (e.g., ChemFinder - Website A7, and the NCI's Structure Browser - Website A2). The ChemFinder application also processes the last control digit of the CAS number to determine if it is a valid CAS number. When SMILES are included in an original Source database, we have used SMILE conversion algorithms to populate entire databases with chemical structures, which can then be confirmed with CAS number and chemical name cross-referencing.

2.14 Chemical Relational Database (CRD) Applications

Consistent with the goal of the DSSTox project, i.e. to facilitate the widest public access to existing toxicity data linked with chemical structure, is the availability of a number of relatively low-cost, yet highly functional commercial CRD applications for the individual user. Examples include ChemFolder (Advanced Chemistry Development -ACD, Website A6), ChemFinder (ChemOffice, CambridgeSoft Inc., Website A7), and Accord for Access and Excel (Accelrys Inc., Website A9). These applications provide simple, flat data-structure (i.e. no nested fields), allow for chemical structure/data/text searching, and can be closely integrated with chemical sketching, property calculation modules, and modeling algorithms. At present, any individual can download free trial versions of these three applications off the web. More costly, server-based CRD applications include ISIS (MDL Information System Inc., Website A4) and Oracle-based (Oracle Corp., Website A10) applications. These latter applications offer higher levels of functionality in terms of field structure (i.e. nested fields), data representation (e.g. chemical reactions) and relational searching. They are also more likely to serve large corporate data management needs, such as within pharmaceutical companies, the chemical industry, and government agencies. Since the DSSTox proposal is based on the generic and public SDF format, it does not endorse, nor does it rely upon the present or future availability of any particular commercial CRD application. In the course or our initial DSSTox project development, we have made greatest use of the ChemFolder and ChemFinder PC desktop applications. ChemFinder (Cambridge Soft, ChemOffice Ultra, ver. 4) has a wide range of import/export options that were useful for DSSTox SDF file development. ChemFolder (ACD, ver. 5.0), in turn, has flexible searching options (e..g. the ability to search across separate database files) and search views (e.g. a tiling feature for viewing multiple structures simultaneously). Each CRD application differs in its range of features and has different strengths and weaknesses in relation to the DSSTox project; hence, it will be up to the individual user or organization to determine which CRD application best suits their needs.

3. DSSTox - Carcinogenic Potency Database

To provide the reader with some concrete sense for specific elements of the DSSTox proposal, we outline here details of our preliminary results in creating DSSTox files for the Carcinogenic Potency Database (Website A1) [3-6]. We also provide a few examples of the searching and viewing capabilities within a sample, representative CRD application based on the DSSTox CPDB SDF files.

3.1 CPDB - DSSTox SDF files

The Carcinogenic Potency Database (CPDB) is one of the largest and most widely used sources of public information on chronic animal cancer bioassays (Website A1) [3-6]. It includes both qualitative and quantitative information on positive and negative experiments for all substances tested under National Cancer Institute/National Toxicology Program (NCI/NTP), as well as for experiments reported in the general literature that meet a set of inclusion criteria. Although the largest proportion of the tested substances included are industrial chemicals, a number of pharmaceuticals are also present in the database. Currently, the CPDB website (A1) offers, as a free online public resource, a wide range of information and references relevant to this database. Users can access detailed experimental information for each tested substance, relating to features of experimental protocol, duration of dosing, shape of dose-response curve, author's opinion, literature citation, response, etc. Of particular interest for present purposes, users can view and download complete tables of summary toxicity information for all tested substances according to species tested. These summary tables, which were used to construct the CPDB DSSTox SDF files, include Salmonella mutagenicity values (+ and -), target organ sites for tumors, and TD_{50} values (for a given target site, the dose at which the probability of animals remaining tumorless is halved [7]; see Website A1 for further details).

Currently, the CPDB rodent carcinogenicity database (male and female, mice and rats) contains 1,354 tested substances, of which: 109 are "mixtures or ill-defined", meaning that they are mixtures of known or unspecified composition; 51 are inorganic chemicals, meaning in the broadest sense that they do not contain carbon; and the remaining 1194 are defined organics with known chemical structures. Of these defined organics, 180 were tested in a salt or hydrate form, information that is embedded in both the chemical name and CAS number, and explicitly listed in the "TESTED FORM" standard field. This information is not, however, reflected in the neutral 2D chemical structure included in the structure field of the DSSTox SDF file. In addition to the rodent carcinogenicity database, two additional databases of carcinogenicity summary data were compiled from the CPDB website (A1), one pertaining to hamsters (80 tested substances), and a second with consolidated results for all other reported species (dogs, monkeys, tree shrews, etc.) (34 tested substances). DSSTox SDF files have been created for each of these three summary toxicity databases. The DSSTox SDF File Names for the total summary carcinogenicity databases (including all tested substances) are given as follows:

Rodent CPDB: CPDBRO_V1a_01354_17sep01.sdf
Hamster CPDB: CPDBHA_V1a_00080_17sep01.sdf
Other Species CPDB: CPDBOT_V1a_00034_17sep01.sdf
The DSSTox Source Log File for the Rodent CPDB is specified as:
CPDBRO_log_17sep01.txt

3.2 CPDB - DSSTox Source Field Index Files

The DSSTox Source Field Index File contains descriptions of both Standard Fields and Source-specific Fields contained in the Source SDF files. The following is a partial representative listing of the contents of this file for the CPDB Source and rodent carcinogenicity SDF file (note, field descriptions are taken directly from Website A1):

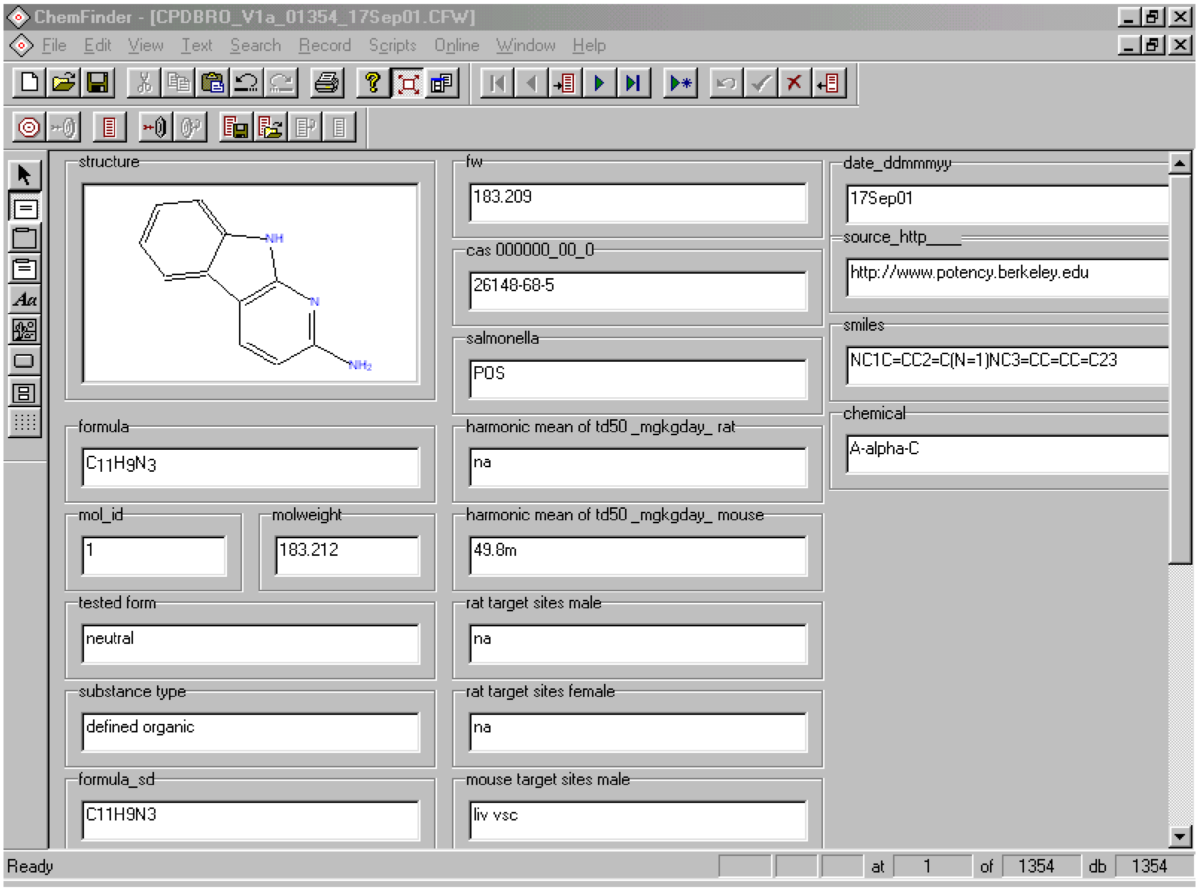
>>>
CPDB Source Field Name File

Field Name (units if applicable) (allowable values if applicable): CRD application-specific field/database-specific field/explanation
****** Units are as given in the source of the database. Allowable values are included to show possible inputs and are detailed in the explanation section. Additional information is indicated after the pound sign in each field as described below.
....
Salmonella(POS/NEG/Not Tested): All/CPDBRO, CPDBHA, CPDBOT /POS=if the chemical was evaluated as either "mutagenic" or "weakly mutagenic" by Zeiger or as "positive by the Gene-Tox Program. All other chemicals evaluated for mutagenicity by these two sources are reported as NEG.

Harmonic mean of TD_{50} (mg/kg/day Rat #h/i/m/n/s/v): All/CPDBRO/ For each positive chemical in the Gold database, results are included on carcinogenic potency (by species) and target organ (by sex-species); if there are no positive results then the word NEG appears. The classification of positivity in this summary table is based on a positive result in at least one experiment. There may additional experiments on the same chemical that are negative in the Carcinogenic Potency Database, but this is not reflected in the CPDB Summary Tables or the corresponding DSSTox SDFs. An experiment is classified as positive or negative on the basis of the author's opinion. For some chemicals the only experiments in the database for a species or a sex species group were NCI/NTP bioassays that were evaluated as inadequate, and we indicate these with an **"I"**, in the potency and target organ fields. TD_{50} is more precisely defined as that dose-rate in mg/kg body wt/day which , if administered chronically for the standard lifespan of the species, will halve the probability of remaining tumorless throughout that period. Additional information is indicated by mnemonic superscripts to the right of the TD_{50} numerical value (#) in each species: h=A mix of carcinomas of the ear duct, Zymbal's gland, oral cavity and nasal cavity

> were combined by Maltoni in his category "Head cancers", which he reports as induced by the chemical; i=Intraperitoneal or intravenous injection are the only routes of administration with positive tests for this species in the database; m=More than one positive test in the species. n=No results evaluated as positive for this species in the database are statistically significant ($p<0.1$); s=Species other than rats or mice are reported for this chemical; v=Variation is greater then ten-fold among statistically significant ($p<0.1$) TD_{50} values from different positive experiments.
>
> **Rat target sites Male (adr/cli/eso/ezy/hag/hmo/kid/lgi/liv/lun/mgl/nas/nrv/orc/ova/pan/per/pit/pre/pro/ski/smi/sto/sub/tba/tes/thy/ubl/ute/vsc):** All/CPDBRO/ Tissue codes: adr = adrenal gland; cli = clitoral gland; eso = esophagus; ezy = ear/Zymbal's gland; hag = harderian gland; hmo = hematopoietic system; kid = kidney; lgi = large intestine; liv =liver; lun = lung; mgl = mammary gland; nas = nasal cavity (includes tissues of the nose, nasal turbinates, paranasal sinuses and trachea); nrv = nervous system; orc = oral cavity (includes tissues of the mouth, oropharynx, pharynx, and larynx); ova = ovary; pan = pancreas; per = peritoneal cavity; pit = pituitary gland; pre = preputial gland; pro = prostate; ski = skin; smi = small intestine; sto = stomach; sub = subcutaneous tissue; tba = all tumor bearing animals; tes = testes; thy =thyroid gland; ubl = urinary bladder; ute = uterus; vsc = vascular system. A site is classified as a target site if the author of the published paper considered tumors to be induced by compound administration. ...
> >>

"CRD application-specific field" refers to a field that is automatically generated by a particular CRD application in addition to the normal SDF fields. [ChemFinder has two automatically generated fields - "MolWeight", "Mol-id"; ChemFolder has one automatically generated fields - "FW" (formula weight).] For more information or elaboration on any of the above biological information, the reader is referred to Website A1 and the references therein.


ChemFinder - [CPDBRO_V1a_01354_17Sep01.CFW]
File Edit View Text Search Record Scripts Online Window Help
NH
N
NH2
Ready
at 1 of 1354 db 1354

structure

formula: $C_{11}H_9N_3$

mol_id: 1

molweight: 183.212

tested form: neutral

substance type: defined organic

formula_sd: C11H9N3

fw: 183.209

cas 000000_00_0: 26148-68-5

salmonella: POS

harmonic mean of td50 _mgkgday_ rat: na

harmonic mean of td50 _mgkgday_ mouse: 49.8m

rat target sites male: na

rat target sites female: na

mouse target sites male: liv vsc

date_ddmmmyy: 17Sep01

source_http____: http://www.potency.berkeley.edu

smiles: NC1C=CC2=C(N=1)NC3=CC=CC=C23

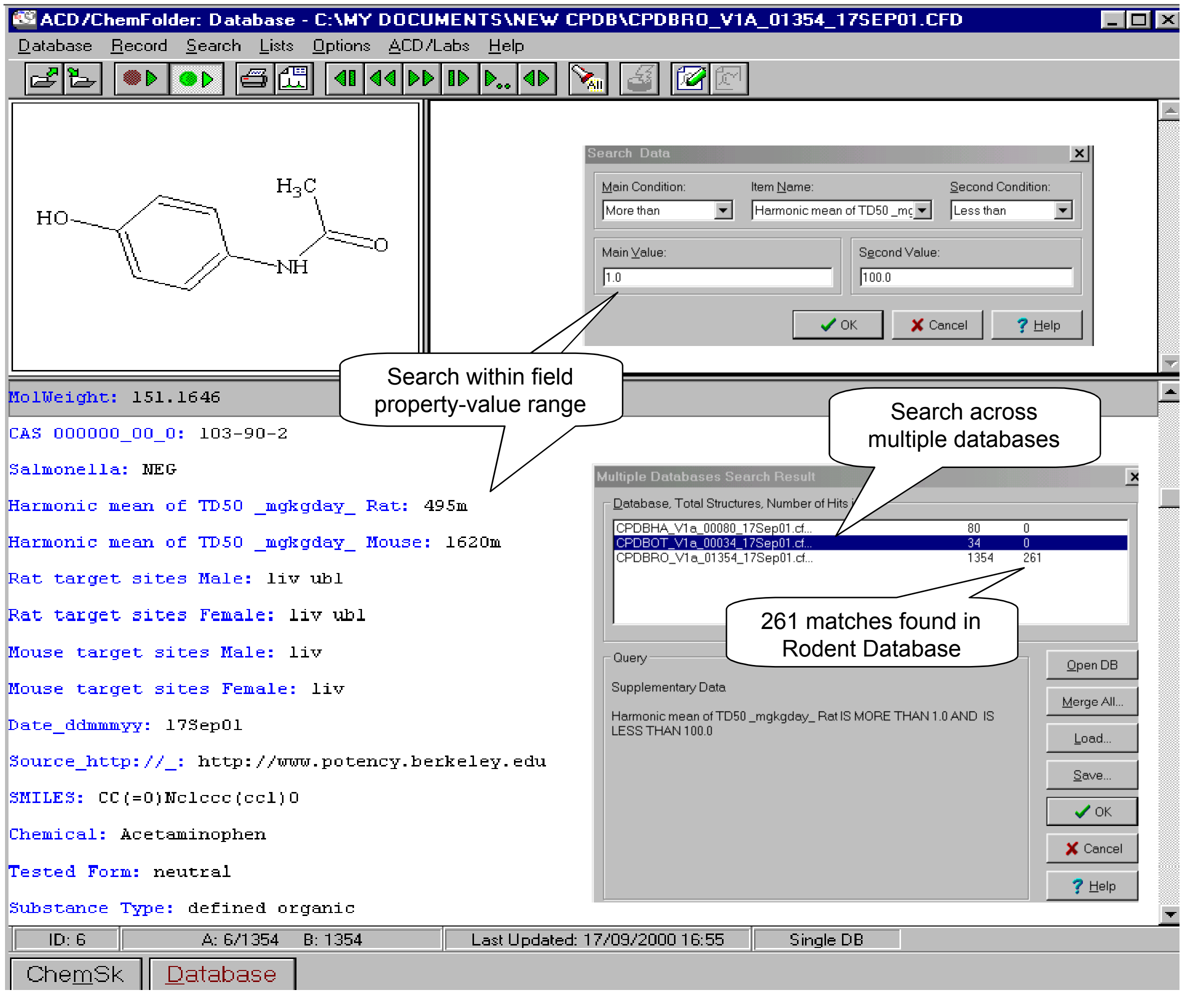
chemical: A-alpha-C

Fig. 5 ChemFinder (CambridgeSoft, Inc., ChemOffice Ultra, ver. 4) view of the first chemical record (A-alpha-C) of the Carcinogenic Potency Rodent Database showing structure and data fields (Website A1) [3-6].

3.3 CRD Application Searches and Views: DSSTox CPDBRO

Fig. 5 presents a view of the first chemical record (A-alpha-C) of the DSSTox CPDBRO database after import of the DSSTox CPDBRO SDF file into the ChemFinder CRD application (Website A7). Note that each distinct field is clearly represented by a field name and field contents, with the chemical structure automatically converted from the tabular text SDF representation (see Fig. 4) to a 2D visual representation.

Upon importing the same DSSTox CPDBRO SDF file into the ACD ChemFolder application (Website A6), we obtain a different visual representation of the field contents information in Fig. 6 (shown is record 6: acetaminophen). In addition, we have included in this view a property search window where a lower and upper limit of a property value can be specified for the search criteria. The results of the indicated search performed over multiple selected databases - in this case the three CPDB databases - is indicated in the lower embedded window in Fig. 6, showing that 261 chemical records in CPDBRO satisfy the search criteria. The ChemFolder application allows a user to subsequently view and/or save the results of this search (i.e. the 261 records) to a separate file.


ACD/ChemFolder: Database - C:\MY DOCUMENTS\NEW CPDB\CPDBRO_V1A_01354_17SEP01.CFD
Database Record Search Lists Options ACD/Labs Help
HO
H_3C
NH
O
Search within field property-value range
Search across multiple databases
261 matches found in Rodent Database
ChemSk
Database

Search Data

Main Condition: More than
Item Name: Harmonic mean of TD50 _mg
Second Condition: Less than
Main Value: 1.0
Second Value: 100.0

✓ OK ✗ Cancel ? Help

MolWeight: 151.1646

CAS 000000_00_0: 103-90-2

Salmonella: NEG

Harmonic mean of TD50 _mgkgday_ Rat: 495m

Harmonic mean of TD50 _mgkgday_ Mouse: 1620m

Rat target sites Male: liv ubl

Rat target sites Female: liv ubl

Mouse target sites Male: liv

Mouse target sites Female: liv

Date_ddmmmyy: 17Sep01

Source_http://_: http://www.potency.berkeley.edu

SMILES: CC(=O)Nc1ccc(cc1)O

Chemical: Acetaminophen

Tested Form: neutral

Substance Type: defined organic

Multiple Databases Search Result

Database, Total Structures, Number of Hits i

CPDBHA_V1a_00080_17Sep01.cf...	80	0
CPDBOT_V1a_00034_17Sep01.cf...	34	0
CPDBRO_V1a_01354_17Sep01.cf...	1354	261

Query

Supplementary Data

Harmonic mean of TD50 _mgkgday_ Rat IS MORE THAN 1.0 AND IS LESS THAN 100.0

Open DB | Merge All... | Load... | Save... | ✓ OK | ✗ Cancel | ? Help

ID: 6 | A: 6/1354 B: 1354 | Last Updated: 17/09/2000 16:55 | Single DB

Fig. 6 ChemFolder (ACD, ver. 4.55) view of the sixth chemical record (acetaminophen) of the Carcinogenic Potency Rodent Database showing a sample property range search and results (Website A1) [3-6].

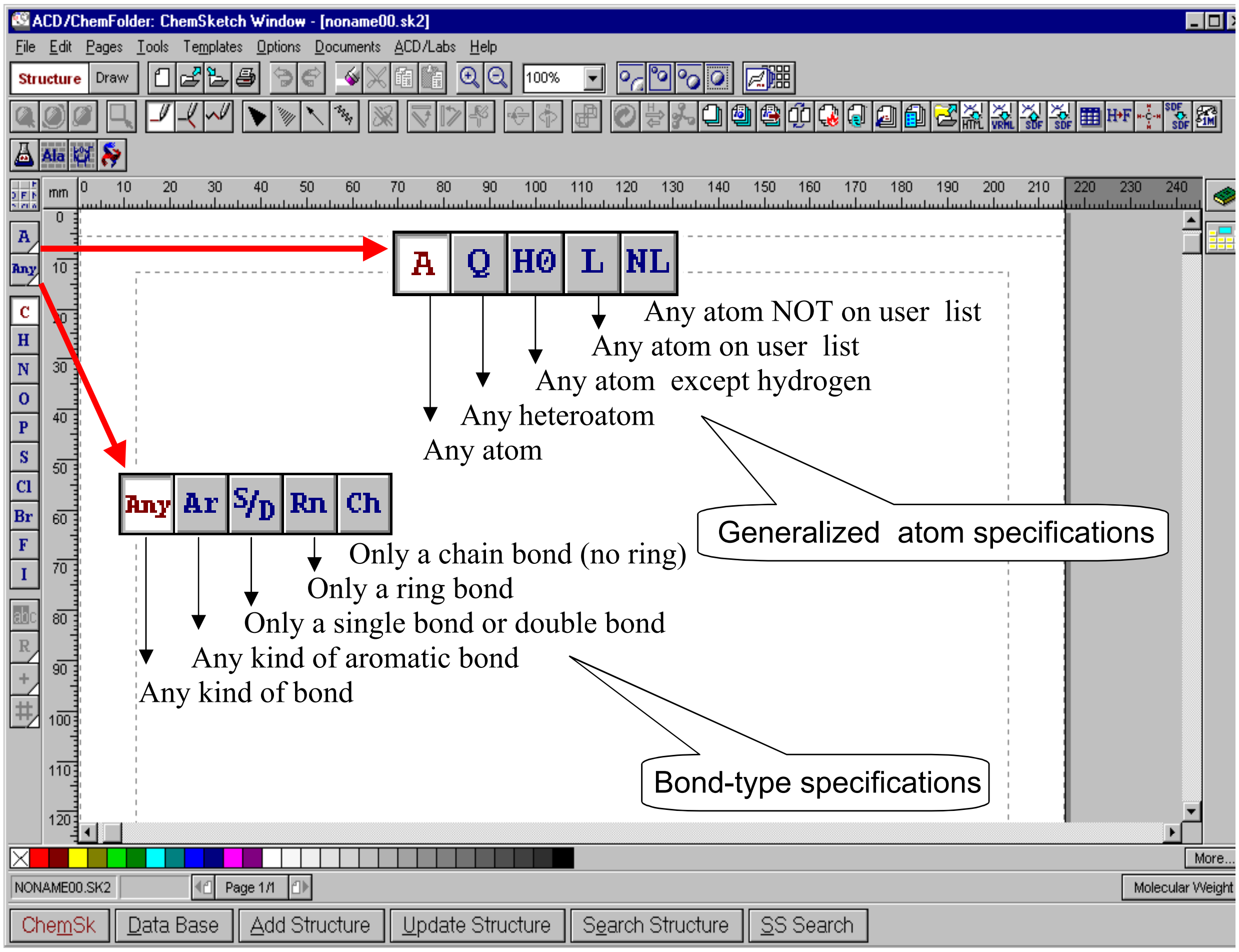
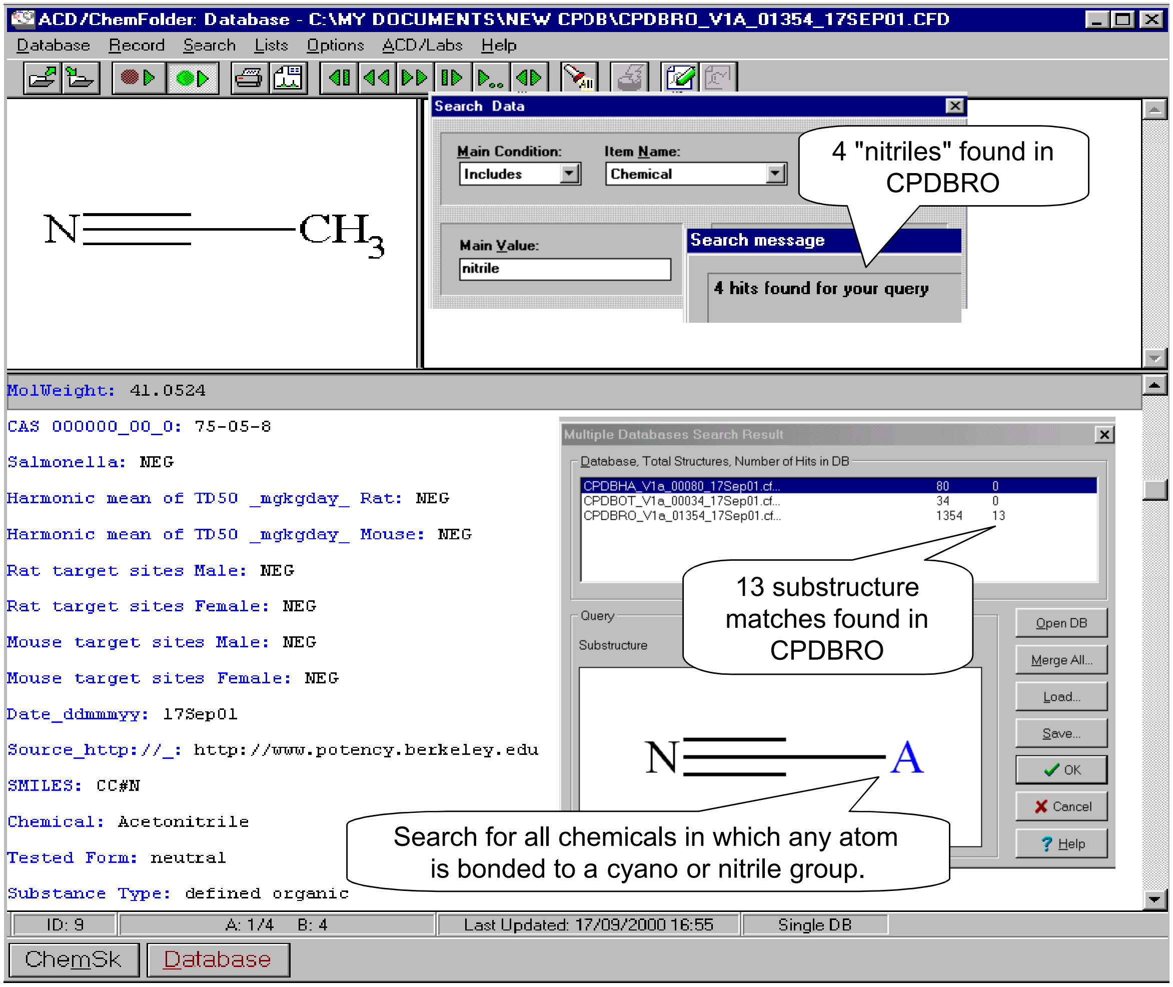

ACD/ChemFolder: ChemSketch Window - [noname00.sk2]
File Edit Pages Tools Templates Options Documents ACD/Labs Help
Structure Draw
100%
A Q HO L NL
Any atom NOT on user list
Any atom on user list
Any atom except hydrogen
Any heteroatom
Any atom
Generalized atom specifications
Any Ar S/D Rn Ch
Only a chain bond (no ring)
Only a ring bond
Only a single bond or double bond
Any kind of aromatic bond
Any kind of bond
Bond-type specifications
NONAME00.SK2
Page 1/1
More...
Molecular Weight
ChemSk Data Base Add Structure Update Structure Search Structure SS Search

Fig. 7 ChemSketch (ACD, ver. 4.56) view showing the various atom and bond specifications possible for use in structure and stubstructure-searching.

Both ChemFinder and ChemFolder have companion chemical drawing programs (i.e. ChemOffice-ChemDraw and ACD ChemSketch, respectively) that are used for drawing and inputting chemical structures or fragment search queries. In addition, both chemical drawing applications allow for generalized structure representations, e.g. to specify a fragment or chemical with a particular type of bond or atom, or to more generally specify atom criteria. Allowable options for the ACD ChemSketch application are shown in Fig. 7.

Fig. 8 displays the result of a search based on a generalized fragment query. When name searches for the terms "nitrile" and "cyano" were performed across the CPDB carcinogenicity databases, four hits were obtained for "nitrile" and no hits were obtained for "cyano". However, when the more general structural representation of this functional group is used as the search criteria, 13 hits or matches were found in the CPDBRO database. This reinforces the point made earlier that chemical names are too variable and inconsistent to serve as generalized search criteria, whereas chemical structure has the potential to more broadly and consistently capture relevant chemical information.

Finally, Fig. 9 illustrates a "tiling view" of the results of the nitrile substructure search within the ChemFolder application, providing a user with a simultaneous snapshot view of chemical structures satisfying the search criteria from Fig. 8. A user can additionally specify which field values will appear under each chemical in this tiling display. This feature could be useful for exploring structural or property-linked hypotheses within and across the DSSTox databases.



ACD/ChemFolder: Database - C:\MY DOCUMENTS\NEW CPDB\CPDBRO_V1A_01354_17SEP01.CFD

Database Record Search Lists Options ACD/Labs Help
Search Data
4 "nitriles" found in CPDBRO
13 substructure matches found in CPDBRO
Search for all chemicals in which any atom is bonded to a cyano or nitrile group.
Open DB
Merge All...
Load...
Save...
OK
Cancel
Help
ChemSk
Database

Main Condition: Includes

Item Name: Chemical

Main Value: nitrile

Search message

4 hits found for your query

MolWeight: 41.0524

CAS 000000_00_0: 75-05-8

Salmonella: NEG

Harmonic mean of TD50 _mgkgday_ Rat: NEG

Harmonic mean of TD50 _mgkgday_ Mouse: NEG

Rat target sites Male: NEG

Rat target sites Female: NEG

Mouse target sites Male: NEG

Mouse target sites Female: NEG

Date_ddmmmyy: 17Sep01

Source_http://_: http://www.potency.berkeley.edu

SMILES: CC#N

Chemical: Acetonitrile

Tested Form: neutral

Substance Type: defined organic

Multiple Databases Search Result

Database, Total Structures, Number of Hits in DB

CPDBHA_V1a_00080_17Sep01.cf...	80	0
CPDBOT_V1a_00034_17Sep01.cf...	34	0
CPDBRO_V1a_01354_17Sep01.cf...	1354	13

Query

Substructure

ID: 9 | A: 1/4 B: 4 | Last Updated: 17/09/2000 16:55 | Single DB

Fig. 8 ChemFolder (ACD, ver. 4.55) view of the ninth chemical record (acetonitrile) of the Carcinogenic Potency Rodent Database showing a sample name and substructure search and results (Website A1) [3-6].

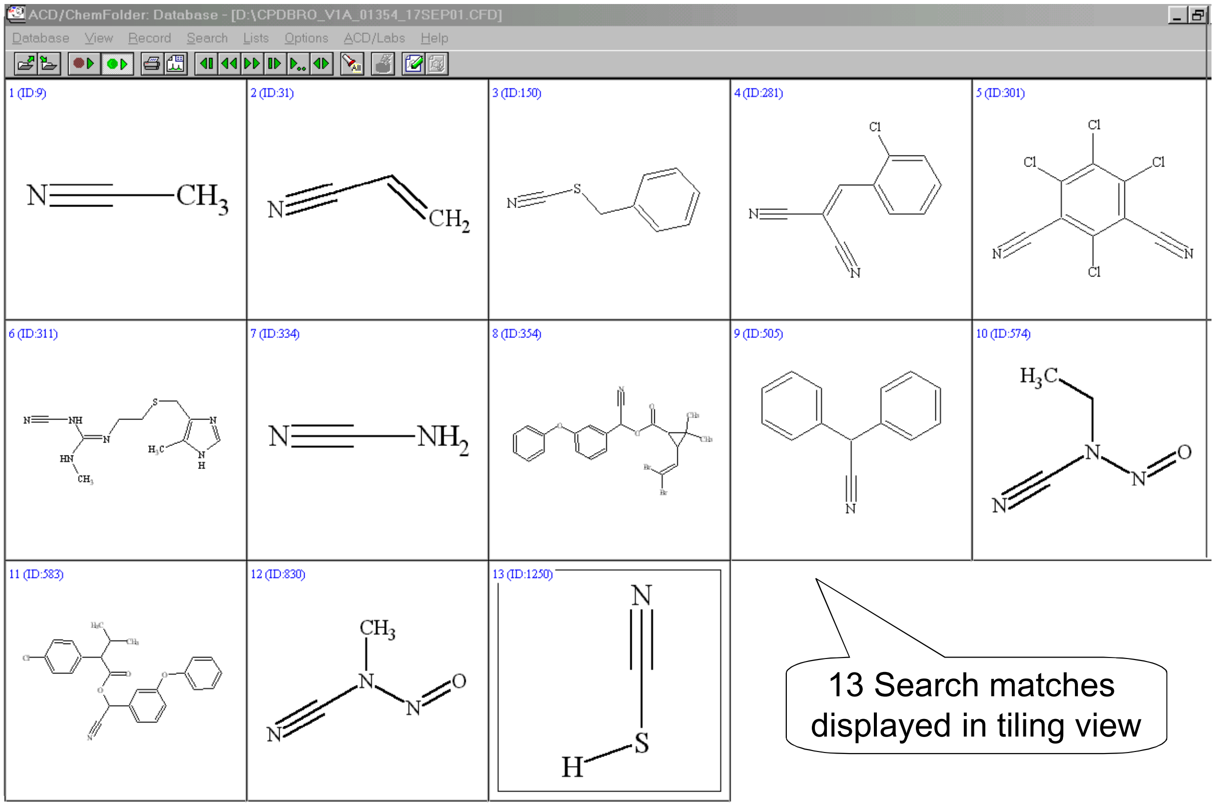
4. Other Types of Toxicity Database Efforts & Needs

In this section, we place the DSSTox project in the larger context of some current toxicity database efforts and needs spanning industry, government, and the public domain. We highlight differences between the DSSTox project and these varied efforts, and attempt to point to the potential benefits of the DSSTox project to this broader scientific community.

4.1 Some Industry Perspective

Large pharmaceutical and chemical industries have been at the forefront in their use of information technologies and data-mining tools for managing, exploring, and providing widespread corporate access to large internal libraries of chemical and biological information [8-10]. This information is considered an invaluable proprietary resource, representing huge past corporate research investments and stored corporate knowledge. A central feature of these corporate databases has been their reliance on chemical structure as a universally recognizable identifier for accessing chemical and biological information, with large, server-based ISIS (Website A4) and Oracle (Website A10) CRD applications most often forming the backbone of these efforts [10,11].

Historically, industry efforts have focused on mining such databases to create leads for the design of new chemicals with desired activities. Hence, corporate databases may contain large numbers of chemicals, but tend to be concentrated on a limited range of desired biological activities (e.g. pesticidal, pharmaceutical applications, etc.). Due to the high economic cost of undesirable and unanticipated adverse health effects, there is a burgeoning interest in the use of data mining of corporate databases for developing models for ADME-Tox (adsorption, distribution, metabolism, elimination, toxicity) prediction [10,11]. In this grand objective, toxicity screening is considered one of the most difficult prediction hurdles to overcome due to the broad range of potential targets and mechanisms spanning diverse toxicological endpoints. The ability to consolidate toxicity information on proprietary chemicals with public sources of toxicity information will allow industry to maximize the effectiveness of data mining and SAR model construction efforts [10,11]. Whereas public access to this proprietary data is not likely, extraction of knowledge contained within this data relative to toxicity may result from creative collaborative efforts [12].



13 Search matches displayed in tiling view

Fig. 9 ChemFolder (ACD, ver. 5.0) tiling view corresponding to the "nitrile" substructure search results in Figure 8 from the Carcinogenic Potency Rodent Database (Website A1) [3-6].

4.2 Some Government Perspective

In contrast to the large-corporate situation (that does not necessarily extend to the smaller pharmaceutical and chemical companies), government laboratories, academics and the public typically lag far behind industry in their management of, and access to chemical-related toxicity information. Public databases relating chemical information to toxicological activities and properties tend to be specialized and fragmented, dealing with a specific kind of test system (e.g. rodent carcinogenicity, aquatic toxicity), chemical-use category (e.g. pesticides, or hazardous air pollutants), or regulatory framework [e.g. EPA's Integrated Risk Information System (IRIS, Website A11), or EPA's

High Production Volume Chemical List (HPV, Website A12)] . When sophisticated CRD tools have been introduced for managing and accessing chemical and biological information, this has most often been to satisfy a pressing regulatory need or through the dedication and foresight of one or a few groups or individuals within specific organizations. For instance, the Office of Pollution Prevention and Toxics at the U.S. EPA has had a long history of use of local toxicity CRD databases to perform analogue searches and to develop SAR arguments to meet PreManufacture Notification (PMN) toxicity screening requirements under the Toxics Substances Control Act (TSCA) [13-16]. A second example is the recently reported development of an Oracle-based system for structure/text searching of toxicity data associated with pharmaceutical data submission requirements within the US Food & Drug Administration's Center for Drug Evaluation [11]. Neither the EPA nor the FDA have been able to migrate their internal toxicity CRDs or associated data into the public domain, in part due to presence of some confidential or proprietary information submitted by industry, and in part due to the lack of a data-sharing mechanism in place. In addition, much of the chemical toxicity information in the public domain has not been gathered and incorporated into such databases.

4.3 Web-Based Acccess to Toxicity Databases

A full accounting of currently available web-based toxicity databases, their features and limitations, is outside the scope of this report. The present DSSTox effort is taking advantage of the fact that a number of compiled toxicity databases lacking chemical structures are currently accessible on the web [17-19]. Two central resources that do, however, provide some structure-searchable access to multiple toxicity databases are worthy of mention. The ChemFinder Website (A7) (not to be confused with the ChemOffice-ChemFinder CRD application, both offered by CambridgeSoft) is a widely used public resource for retrieving chemical structures from names and CAS numbers, that also provides structure-searchable links to hundreds of publically available databases, including toxicities and many other types of biological and chemical activities (e.g. flavorings, NCI database of 3D structures, etc). The National Library of Medicine’s (NLM’s) TOXNET website (A13) similarly provides some structure and text-based searching capability across a number of public toxicity databases and is a widely used public resource. Both of these resources utilize an extensive look-up-table of chemical structures and chemical names linked to either the central toxicity database websites or to specific records of primarily textual information pertaining to chemical toxicity. However, the user is constrained in the way in which they can search and access the data, detailed characteristics of complete databases (such as numbers and types of chemicals, etc) may be unavailable or difficult to obtain, and the databases themselves can only be sampled by directed searches and often cannot be fully accessed or downloaded by the public.

4.4 Other Commercial and Public Toxicity Databases

Three main efforts to provide access to a broad range of public toxicity data in a chemical relational database format are worthy of mention. Two commercial programs, MDL's ISIS-based Toxicity Database (Website A4) and the SciVision ToxSys Database (Website A14), both offer chemical relational database searching through chemical and published toxicity data. Both, at present, are also based primarily on the extensive Registry of Toxic Effects of Chemical Substances (RTECS, Website A15) repository of published toxicity data compiled by the National Institute for Occupational Safety and Health. The major criticism of RTECS is that it includes only reports of positive chemical toxicity data and is lacking in any measure of quality review. The third effort is an industry/government cooperative project overseen by the non-profit International Life Sciences Health and Environmental Sciences Institute (ILSI, see Website A16) to develop a centralized Toxicity Structure-Activity Relational Database. None of these three toxicity CRD efforts has been described or detailed in the scientific literature, and the latter ILSI-sponsored effort remains in the early stages of development and will likely be supported by subscription fees. Hence, we will not consider details of these various efforts here except to point out two important distinctions from the present DSSTox proposal: 1) the DSSTox SDF files will be made freely and publically available; and 2) we are proposing a distributed network of standard-format public toxicity databases supported by a community-wide effort, as opposed to construction of a large application-specific, centralized database that is potentially much more difficult to construct, manage and maintain.

Additionally, we mention two prominent commercial toxicity prediction applications, TOPKAT (Website A9) and MultiCASE (Website A17). These applications provide a user with limited access to toxicity databases used in the construction and validation of their predictive models for a variety of modeled toxicity endpoints (e.g. sex/species specific rodent carcinogenicity models, mutagenicity, skin sensitization, developmental toxicity, etc).

These toxicity databases, in some cases, have been compiled from primary sources expressly for the purpose of the SAR model development by the application developers and are only available to a user with purchase of the commercial prediction program. Hence, these databases could be considered a valuable structure-linked data resource, independent of the validity or utility of the actual toxicity prediction algorithms. In their current forms, however, TOPKAT and MultiCASE both place limits on a user's open access to their internal toxicity data and do not provide a CRD searching capability; i.e. a user is limited to analogue searches constrained by the assumptions of the prediction algorithms [20,21].

4.5 Benefits of the DSSTox Approach

The DSSTox database network has the potential to complement and significantly enhance each of the above-mentioned toxicity database activities. Industry and government groups, particularly those currently accustomed to the use of CRD applications, would welcome greater access to readily importable structure-linked toxicity data files to add to existing in-house databases, particularly if those files have a clearly identifiable and knowledgeable source. On-line public data providers, such as the ChemFinder Website (A7) and TOXNET (A13), in turn, would be able to immediately expand their structure-searchable public offerings. Commercial toxicity database vendors, such as MDL's Toxicity Database (A4) and SciVision's ToxSys (A14) that have traditionally relied upon publically available toxicity databases, would have a ready source of data to expand their database contents. Furthermore, they would be challenged to add significant functionality to these public databases to justify the commercial cost. CRD vendors would find expanded interest in their products, and data-mining vendors catering to large pharmaceutical companies would be able to more easily expand their efforts into the public toxicity data arena.

A user-based toxicity CRD, assembled from DSSTox data files, could also provide an extremely valuable complement to the use of commercial toxicity prediction programs such as TOPKAT (A9) and MultiCASE (A17), whether or not a user has access to such programs. For example, numerous publications have reported extensive listings of CASE and MultiCASE biophores (i.e. structure-alerting fragments significantly associated with the toxicity of interest) for a wide variety of toxicity endpoints (see e.g. [22-25]). Users with access to a toxicity CRD, but without access to the MultiCASE or CASE programs, could make greater use of this body of published results for SAR hypothesis generation and analogue searching, or could attempt to independently scrutinize and validate these historical results. This could indeed extend to many other types of published structural-alerting features or correlates resulting from past SAR modeling efforts (see e.g. [26,27]).

5. Current Progress & Final Comments

5.1 DSSTox SDF files Under Development

DSSTox SDF files are in various stages of development for a selection of public toxicity databases spanning health and ecological endpoints. This preliminary set of databases were chosen for their accessibility and/or for the interest they have previously engendered from the SAR modeling community; the choices were not based on judgements of relative quality or importance. DSSTox SDF files for the Carcinogenic Potency Databases (Website A1 [3-6]), described in some detail in Section 3, are near completion and awaiting final review and posting at the Source (CPDB) website. Although the NTP rodent carcinogenicity database (Website 5) is largely included in the CPDB rodent carcinogenicity database, we will be creating a distinct NTP DSSTox SDF file to contain additional information and interpretation from the NTP primary source. Other DSSTox SDF files currently in development pertain to the endpoints: *Salmonella* mutagenicity (EPA/IARC Genetic Activity Profile - GAP database (Website A18) [28], EPA Gene-Tox database [29], and the NTP Salmonella database (Website A5, [30])); aquatic toxicity (EPA ECOTOX databases, Website A19); risk assessment distribution and toxicity parameters (Website A20); human behavioral neurotoxicity [31]; and estrogen receptor binding [32]. Each of these databases presents its own set of challenges to the DSSTox effort, in some cases requiring CAS number or chemical name conversion to structures, whereas in other cases availability of SMILES or an SDF file facilitates the task of creating the DSSTox SDF file considerably. Reviewing the accuracy and consistency of the chemical information is a major component of this effort, as is ensuring proper formatting of fields and providing adequate documentation in the Source Field Index files.

We are currently working to complete the development of DSSTox SDF files for the above databases, as well as to expand this effort outward to encompass a larger offering of DSSTox SDF files for public toxicity databases.

A number of toxicity databases (e.g. IRIS, Website A11; and RTECS, Website A15) are already in the public domain in some form, whereas others (e.g. pertaining to skin sensitization, developmental toxicity, biodegradation) have been published or are in private hands and could be brought into the public domain with some community involvement and assistance. DSSTox SDF file construction becomes progressively easier as common technical problems are overcome. We have written a number of program scripts to automate the process of "cleaning-up" data files or converting formats to DSSTox Standard SDF. Scripts have also been written for aiding in the verification of chemical identifier information and automatically populating SDF files with chemical structures, by cross-referencing names and CAS numbers with previous SDF files. [These scripts, written in open-source Python (Website A8) will be made available for public download from the DSSTox Central Website.] Hence, as more DSSTox SDF files are completed, the task of creating new DSSTox SDF files becomes progressively easier.

5.2 Particular Relevance of the DSSTox Project to Mutagenesis and Carcinogenesis

Relatively large numbers of chemicals have been evaluated in mutagenicity and carcinogenicity bioassays due to the regulatory emphasis and heightened level of concern associated with these endpoints. Hence, there has traditionally been greater public access to compiled databases and more SAR modeling activity devoted to the study of chemical carcinogenicity and mutagenicity than to most other types of chemical toxicity [27,33,34]). Hence, the fields of mutagenicity and carcinogenicity stand to derive significant immediate benefits from the DSSTox project in terms of improved structure-linked access to existing toxicity data. The Carcinogenic Potency Database, the NTP Rodent Carcinogenicity Bioassay Database, the EPA/IARC GAP database, and the Gene-Tox/NTP Salmonella database together represent a large portion of the data that have fueled past efforts aimed at developing structure-based models for carcinogenicity and mutagenicity prediction.

There have been a number of public challenges to the scientific community to prospectively predict chemical carcinogenicity based on chemical structure, expert knowledge and sub-chronic or in-vitro bioassay information [35-38]. These efforts, initiated by the NTP and more recently expanded by the artificial intelligence & machine-learning communities, have engaged carcinogenicity experts, computer scientists, SAR modelers, and biochemists [34,38]. A central dichotomy of the carcinogenicity prediction problem, that has traditionally plagued these efforts, is in reconciling the goal of global carcinogenicity prediction with the need to adequately capture relevant SAR prediction elements within multiple, mechanistically distinct classes of chemical carcinogens. A long advocated view has been the need to build and/or validate a global SAR prediction model based on these local mechanism-based considerations [1,13,15,21,34,38,39]. These, in turn, can be productively informed by short-term and sub-chronic bioassay results [39-43]. Revisiting the paradigm represented in Fig. 1, we stress that the ability to gather supporting information for validating any type of local SAR model or chemical toxicity prediction relies in large measure on the ability to gather relevant data on chemical and biological analogues to build a weight-of-evidence argument. A user-based CRD application built on DSSTox SDF files will significantly expand these data-gathering and argument-building capabilities among the DSSTox user community.

A final observation is that there has been a general openness of the carcinogenesis/mutagenesis community to considering the potential value and utility of SAR study. However, there are many examples in the literature of so-called SAR studies in which bioassay results are simply reported for a family or series of chemicals (sometimes as few as 3 or 4), with little or no attempt to model or interpret these results in chemical terms. Likewise, we can find an equal number of examples of published SAR studies in which little or no attempt is made to interpret the model or its predictions in a context of plausible biological mechanisms. These shortcomings arise, in large part, from the lack of interdisciplinary contact. The DSSTox project is attempting to better link these communities and to provide improved ability to place SAR results in a broader contextual framework for understanding.

5.3 Interfacing with Other Chemical Bioactivity Databases

We emphasize that one of the most important and essential elements of the DSSTox proposal is the adoption of a standard format file, SDF, with common chemical structure representation and identifier fields for public toxicity databases. This project is not creating new toxicity data, nor can it realistically attempt to tackle, at the same time, thorny issues pertaining to toxicity data representation, reproducibility, relevance, and quality. These issues rightfully reside in the domain of the toxicology experts and scientific discourse. What we are proposing is simply a way to make existing toxicity data more useful and universally accessible to toxicologists and SAR modelers alike. Within diverse fields of toxicology, data file standardization enhances our ability to explore across chemical toxicity databases from a biological or structural perspective, and to find associations that may not be anticipated or readily apparent by other means. Additionally, toxicity data file standardization, indexed by chemical structures, is a key to interfacing toxicity data with other types of standardized information relative to the effects of chemicals on biological systems. Finally, SDF file standardization of toxicity databases, now, will greatly facilitate a transition to the next generation of file standardization, whatever that may be, in the future.

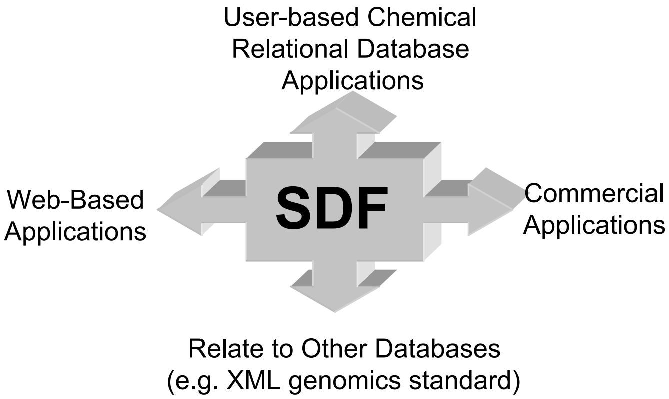


Fig. 10 Common format toxicity SDF files become widely importable into various types of database and web-based applications.

Fig. 10 summarizes four areas in which the proposed DSSTox SDF file standardization is poised to have the greatest impact. The first is in enabling any individual or corporate user to gather or create a personalized database of toxicity information; this could also mean adding data to existing CRD capabilities or creating an affordable user-based CRD searching capability. The second area of impact is in improving the coverage of current web-based structure-toxicity database search tools (e.g., ChemFinder, Website A7; and TOXNET, Website A13), where a future goal could be to create web-based search tools to specifically interface with the DSSTox project (similar to the NCI Structure Browser, Website A2). A third area of impact is in adding more public data to commercial toxicity database programs (e.g., MDL Toxicity, Website A4; and ToxSys, Website A14), and helping users derive greater benefit from published prediction algorithms and commercial toxicity prediction programs (e.g., MultiCASE, Website A17). The fourth area of impact is in interfacing with other standardized databases relative to the effects of chemicals on biological systems. For example, the file format, XML, is being proposed as a standard for storage of gene expression array results, and efforts are currently underway to create large public repositories of such data [44]. Since many experiments, in turn, collect gene expression patterns resulting from a chemical exposure, chemical structure can and should serve as a common index and metric for interfacing gene expression data to other types of chemical-activity databases [45]. The proposed DSSTox database network ideally positions toxicity databases to plug into such a chemical structure interface.

5.4 DSSTox is a Collaborative Community-Supported Effort

Progress to-date on the DSSTox project, including development of DSSTox SDF files and the DSSTox Central Website, has resulted primarily from the concentrated efforts of a few. However, the continued vitality and growth of this proposed public resource will ultimately depend upon the active support and involvement of the entire DSSTox user community who stand to derive greatest benefit from these shared efforts. A standard data file format will be propagated only if it is adopted and used by the larger toxicology and SAR community. A wider selection of DSSTox SDF files will be made available publically only if others choose to collaborate and contribute their efforts to constructing such files. The decentralized nature of the proposed DSSTox database network is not only important for maintaining close connections of toxicity data with domain experts, but is also essential for distributing responsibility for the continued growth of this project. In closing, we believe that the DSSTox database network is not just about providing structure-linked toxicity data; rather, it is a proposed construct for harnessing the shared interest and broad capabilities of a scientific community, and for catalyzing multidisciplinary interactions across the spectrum of chemical-toxicological investigation.

Acknowledgements

In keeping with the collaborative nature of the DSSTox proposal, a number of persons and organizations have provided data and technical assistance, and many others have supplied much enthusiasm and interest in helping to launch this effort. Two persons were particularly instrumental in providing inspiration for this public DSSTox effort: Thanks to Ed Thompson of Procter & Gamble for creating and demonstrating the utility of an early, Oracle-based CRD application that consolidated public gene-tox information (which became the seed for the ILSI - SAR toxicity database effort). Also special thanks to Antony Williams of Advanced Chemistry Development (ACD) for early enthusiasm and support for the project, for his generous advice and assistance in helping to create the first DSSTox SDF files, and for helping to incorporate our suggested features and improvements into the upgraded ChemFolder CRD application. Other persons that were particularly helpful in providing technical assistance were Michel Hachey and Gavin Shear, both of ACD, and Rich Talbot of CambridgeSoft Inc. Special acknowledgements to Daniel Ohuoba (2001 Shaw University High School Summer Student Mentoring Program) for SDF script development, and James Beidler (2001 EPA Summer Student Trainee - Warren Wilson College) for development of the original DSSTox Website template and SDF scripts. In addition, we thank Paul Swartz for technical assistance and helpful suggestions, and Carl Blackman, Doug Wolf, and Julian Preston for useful comments in review of this manuscript. Finally, many others across industry and government have provided data, technical assistance, and/or pledged their support to the future development of the DSSTox project. Within EPA, these include: Pauline Wagner (EPA-Office of Pesticides), Randy Brinkhuis (EPA-OPPT/Information Management Division), Michael Waters (EPA-NHEERL, Environ. Carcinogenesis Div.), Chris Russom (EPA-NHEERL/Duluth Ecology Lab, ECOTOX), and Kevin Crofton (EPA-NHEERL, Neurotox Division). Outside of EPA, these include: Lois Swirsky Gold and Thomas Slone (UC-Berkeley), Vijay Gombar and Neal Cariello (Glaxo-SmithKline), Fred Dolislager (Univ. of Tennessee - Oak Ridge), Marc Nicklaus (Nat. Cancer Inst.), Doug Bristol and Skip Eastin (Nat. Inst. of Environ. Health Sci.), Weida Tong (Nat. Ctr. for Toxicol. Research), Frank Stack (Alpha-Gamma Technol.), Dale Johnson (DD Platform/Eos Biotechnology), and Errol Zeiger. It is our hope that this list will continue expanding.

Appendix A. Websites

The following website urls were active and current at the time this manuscript writing. Since urls occasionally change over time, if a reader finds a url inactive, it is suggested that they refer to the top-level url of the company or organization to attempt to locate specific applications or documents.

A0: DSSTox Central Website (in development)
http://www.dsstox.net

A1: University of California - Berkeley, Carcinogenic Potency Project
http://potency.berkeley.edu/cpdb.html

A2: National Cancer Institute Database Browser & SDF Toolkit
http://cactus.nci.nih.gov/
http://cactus.nci.nih.gov/SDF_toolkit/

A3: O'Reilley PERL.com (Source for PERL5 compiler)
http://www.perl.com/pub/language/info/software.html

A4: MDL Systems Inc. & MDL file formats & MDL ISIS & MDL Toxicity Database
http://www.mdli.com http://www.mdli.com/cgi/dynamic/product.html?uid=$uid&key=$key&id=30
http://www.mdli.com/cgi/dynamic/product.html?uid=$uid&key=$key&id=5
http://www.mdli.com/cgi/dynamic/product.html?uid=$uid&key=$key&id=41

A5: National Institutes of Environmental Health Sciences - National Toxicology Program
http://ntp-server.niehs.nih.gov/

A6: Advanced Chemistry Development & ChemFolder CRD application
http://www.acdlabs.com
http://www.acdlabs.com/download/cfolder45.html

A7: CambridgeSoft Inc. & ChemOffice's ChemFinder CRD application
http://chemfinder.cambridgesoft.com
http://products.cambridgesoft.com/family.cfm?FID=4

A8: Open-source Python compiler
http://www.python.org/

A9: Accelrys Inc. & Accord & TOPKAT
http://www.accelrys.com/accord/
http://www.accelrys.com/offers/ci_demo/index.php
http://www.accelrys.com/products/topkat/index.html

A10: Oracle Corp.
http://www.oracle.com/

A11: EPA's Integrated Risk Information System
http://www.epa.gov/iris/

A12: EPA's High Production Volume (HPV) Challenge Program Chemical List
http://www.epa.gov/opptintr/chemrtk/hpvchmlt.htm

A13: National Library of Medicine's ToxNet
http://toxnet.nlm.nih.gov/

A14: SciVision Inc. & ToxSys
http://www.scivision.com/
http://www.scivision.com/ToxSys.html

A15: Registry of Toxic Effects of Chemical Substances (RTECS)
http://www.cdc.gov/niosh/rtecs.html

A16: International Life Sciences Institute & Structure-Activity Relationship Database
http://www.ilsi.org
http://www.ilsi.org/file/SAR.pdf

A17: MultiCASE Inc.
http://www.multicase.com/

A18: International Agency for Research on Cancer/EPA's Genetic Activity Profile Database
http://www.epa.gov/gap-db/
http://monographs.iarc.fr

A19: EPA's ECOTOX Databases
http://www.epa.gov/ecotox/
http://www.epa.gov/med/databases/fathead_minnow.html

A20: US Dept. of Energy, Oak Ridge National Lab - Risk Assessment Information System
http://risk.lsd.ornl.gov/rap_hp.shtml

References

[1] A. M. Richard, Application of artificial intelligence and computational methods to predicting toxicity, Knowledge Engineering Rev. 14 (1999) 307-317.

[2] A. Dalby, J.G. Nourse, W.D. Hounshell, A. Gushurst, D.L. Grier, B.A. Leland, J. Laufer, Description of several chemical-structure file formats used by computer-programs developed at Molecular Design Limited, J. Chem. Inf. Comput. Sci. 32 (1992) 244-255.

[3] L.S. Gold, C.B. Sawyer, R. Magaw, G.M. Backman, M. de Veciana, R. Levinson, N.K. Hooper, W.R. Havender, L. Bernstein, R. Peto, M.C. Pike, B.N. Ames, A Carcinogenic Potency Database of the standardized results of animal bioassays, Environ. Health Perspect. 58 (1984) 9-319.

[4] L.S. Gold, E. Zeiger (eds) Handbook of Carcingenic Potency and Genotoxicity Databases, Boca Raton, FL: CRC Press (1997).

[5] L.S. Gold, T.H. Slone, What do animal cancer tests tell us about human cancer risk? Overview of the Carcinogenic Potency Database, Drug Metabol. Revs. 30 (1998) 359-404.

[6] L.S. Gold, N.B. Manley, T.H. Slone, L. Rohrbach, Supplement to the Carcinogenic Potency Database (CPDB): results of animal bioassays published in the general literature in 1993 to 1994 and by the National Toxicology Program in 1995 to 1996, Environ. Health Perspect. 107 Suppl. 4 (1999) 527-600.

[7] R. Peto, M.C. Pike, L. Bernstein, L.S. Gold, B.N. Ames, The TD_{50}: A proposed general convention for the numerical description of the carcinogenic potency of chemicals in chronic-exposure animal experiments, Environ. Health Perspect. 58 (1984) 1-8.

[8] D.E. Johnson, P.E. Blower, G.J. Myatt, G.H.I. Wolfgang, Chem-tox informatics: Data mining using a medicinal chemistry building block approach, Current Opinion in Drug Discovery & Develop., 4 (2001) 92-101.

[9] C. Ahlberg, Visual exploration of HTS databases: bridging the gap between chemistry and biology, Drug Discovery Today 4 (1999) 370-376.

[10] D.E. Johnson, G.H.I. Wolfgang, Predicting human safety: screening and computational approaches, Drug Discovery Today, 5 (2000) 445-454.

[11] E.J. Matthews, R.D. Benz, J.F. Contrera, Use of toxicological information in drug design, J. Mol. Graph. Model. 18 (2000) 605-615.

[12] A.M. Richard, The optimal fragmentation principle, Reply, Drug Discovery Today: Discussion Forum 6 (2001) 235-237.

[13] C.M. Auer, J.V. Nabholz, K.P. Baetcke, Mode of action and the assessment of chemical hazards in the presence of limited data: use of structure-activity relationships (SAR) under TSCA, Section 5, Environ. Health Perspect. 87 (1990) 183-97.

[14] P.M. Wagner, J.V. Nabholz, R.J. Kent, The new chemicals process at the Environmental Protection Agency (EPA): structure-activity relationships for hazard identification and risk assessment, Toxicol. Lett. 79 (1995) 67-73.

[15] Y.T. Woo, D.Y. Lai, M.F. Argus, and J.C. Arcos, Development of structure-activity relationship rules for predicting carcinogenic potential of chemicals, Toxicol. Lett 79 (1995) 219-228.

[16] R.G. Clements, J.V. Nabholz, M.G. Zeeman, C.M. Auer, The application of structure-activity relationships (SARs) in the aquatic toxicity evaluation of discrete organic chemicals, SAR QSAR Environ. Res. 3 (1995) 203-15.

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

[18] L.L. Wright, Searching fee and non-fee toxicology information resources: an overview of selected databases, Toxicology 157 (2001) 89-110.

[19] J.D. Walker, QSARs for the world wide web - current practices, in: J.D. Walker (Ed.), Handbook on Quantitative Structure Activity Relationships (QSARs) for Pollution Prevention, Toxicity Screening, Risk Assessment and World Wide Web Applications, SETAC Press, Pensacola, FL, 2001.

[20] R. Benigni ,A.M. Richard, Quantitative structure-based modeling applied to characterization and prediction of chemical toxicity, Methods 14 (1998) 264-76.

[21] A.M. Richard, Structure-based methods for predicting mutagenicity and carcinogenicity: are we there yet?, Mutat. Res. 400 (1998) 493-507.

[22] G. Klopman, M.R. Frierson ,H.S. Rosenkranz, The structural basis of the mutagenicity of chemicals in Salmonella typhimurium: the Gene-Tox data base, Mutat Res 228 (1990) 1-50.

[23] M.H. Karol, C. Graham, R. Gealy, O.T. Macina, N. Sussman, H.S. Rosenkranz, Structure-activity relationships and computer-assisted analysis of respiratory sensitization potential, Toxicol. Lett. 86 (1996) 187-91.

[24] M. Ghanooni, D.R. Mattison, Y.P. Zhang, O.T. Macina, H.S. Rosenkranz, G. Klopman, Structural determinants associated with risk of human developmental toxicity, Am. J. Obstet. Gynecol. 176 (1997) 799-805; discussion 805-6.

[25] A.R. Cunningham, H.S. Rosenkranz, Y.P. Zhang, G. Klopman, Identification of 'genotoxic' and 'non-genotoxic' alerts for cancer in mice: the carcinogenic potency database, Mutat. Res. 398 (1998) 1-17.

[26] J. Ashby, Fundamental structural alerts to potential carcinogenicity or noncarcinogenicity, Environ Mutagen 7 (1985) 919-21.

[27] C. Hansch, Structure-activity relationships of chemical mutagens and carcinogens, Sci. Total Environ. 109-110 (1991) 17-29.

[28] M.D. Waters, H.F. Stack, N.E. Garrett, M.A. Jackson, The Genetic Activity Profile database, Environ. Health Perspect. 96 (1991) 41-45.

[29] A.E. Auletta, M. Brown, J.S. Wassom, M.C. Cimino, Current status of the Gene-Tox Program, Environ. Health Perspect. 96 (1991) 33-6.

[30] E. Zeiger, Carcinogenicity of mutagens: predictive capability of the Salmonella mutagenesis assay for rodent carcinogenicity, Cancer Res. 47 (1987) 1287-96.

[31] R.B. Dick, H. Ahlers. Chemicals in the Workplace: Incorporating Human Neurobehavioral Testing Into the Regulatory Process, Am. J. Ind. Med. 33 (1998) 439-453.

[32] R.M. Blair, H. Fang, W.S. Branham, B.S. Hass, S.L. Dial, C.L. Moland, W. Tong, L. Shi, R. Perkins, D.M. Sheehan, The estrogen receptor relative binding affinities of 188 natural and xenochemicals: structural diversity of ligands, Toxicol. Sci. 54 (2000) 138-153.

[33] R. Benigni, A. Giuliani, Quantitative structure--activity relationship (QSAR) studies of mutagens and carcinogens, Med. Res. Rev. 16 (1996) 267-84.

[34] A.M. Richard, R. Benigni., AI and SAR approaches for predicting chemical carcinogenicity: Survey and status report, SAR and QSAR in Environ. Toxicol. (2001) in press.

[35] Anonymous. (1993). Predicting chemical carcinogenesis in rodents. An International workshop. Research Triangle Park, NC, USA. National Institute of Environmental Health Sciences. 1993.

[36] D.W. Bristol, J.T. Wachsman ,A. Greenwell, The NIEHS Predictive-Toxicology Evaluation Project: Chemcarcinogenicity Bioassays, Environ Health Perspect 104S (1996) 1001-10.

[37] Srinivasan, A., King, R.D and Bristol, D.W. (1999). An assessment of submissions made to the predictive toxicology evaluation challenge. In: Proceedings of the sixteenth international joint conference on artificial intelligence (IJCAI-99). San Francisco, CA. Morgan Kaugmann, 270-275.

[38] Helma, C., Gottmann, E. and Kramer, S. (2000). Knowledge discovery and data mining in toxicology. Statis. Meth. Medical Res. 9, 1-30.

[39] Y.T. Woo, D.Y. Lai, M.F.Argus, J.C. Arcos, An integrative approach of combining mechanistically complementary short-term predictive tests as a basis for assessing the carcinogenic potential of chemicals, Environ. Carcino. & Ecotox. Revs C16 (1998) 101-122.

[40] D. Bahler, D.W. Bristol, The induction of rules for predicting chemical carcinogenesis in rodents, Proc. Int. Conf. Intell. Syst. Mol. Biol. 1 (1993) 29-37.

[41] Y. Lee, B.G. Buchanan ,H.S. Rosenkranz, Carcinogenicity Predictions for a Group of 30 Chemicals Undergoing Rodent Cancer Bioassays Based on Rules Derived from Subchronic Organ Toxicities, Environ. Health Perspect. 104S (1996) 1059-63.

[42] Woo, Y.T., Lai, D.Y., Arcos, J.C., Argus, M.F., Cimino, M.C., DeVito, S.and Keifer, L. (1997). Mechanism-based structure-activity relationship (SAR) analysis of carcinogenic potential of 30 NTP test chemicals. Environ. Carcino. & Ecotox. Revs., C15, 139-160.

[43] R. Benigni, Mouse bone marrow micronucleus assay: relationships with in vitro mutagenicity and rodent carcinogenicity, J. Toxicol. Environ. Health 45 (1995) 337-47.

[44] P. Kellam, Microarray gene expression database: progress towards an international repository of gene expression data. Genome Biology 2 (2001) reports 4011.1-4011.3.

[45] J. An, T. Nakama, Y. Kubota, A. Sarai, 3DinSight: an integrated relational database and search tool for the structure, function and properties of biomolecules, Bioinformatics 14 (1998) 188-95.