A novel approach: chemical relational databases, and the role of the ISSCAN database on assessing chemical carcinogenicity

Romualdo Benigni[a], Cecilia Bossa[a], Ann M. Richard[b] and Chihae Yang[c]

[a] *Dipartimento di Ambiente e Connessa Prevenzione Primaria, Istituto Superiore di Sanità, Rome, Italy*
[b] *National Center for Computational Toxicology, US Environmental Protection Agency, Research Triangle Park, North Carolina, USA*
[c] *LeadScope Inc., Columbus, Ohio, USA*

Summary. Mutagenicity and carcinogenicity databases are crucial resources for toxicologists and regulators involved in chemicals risk assessment. Until recently, existing public toxicity databases have been constructed primarily as "look-up-tables" of existing data, and most often did not contain chemical structures. Concepts and technologies originated from the structure-activity relationships science have provided powerful tools to create new types of databases, where the effective linkage of chemical toxicity with chemical structure can facilitate and greatly enhance data gathering and hypothesis generation, by permitting: a) exploration across both chemical and biological domains; and b) structure-searchability through the data. This paper reviews the main public databases, together with the progress in the field of chemical relational databases, and presents the ISSCAN database on experimental chemical carcinogens.

Key words: database, mutagenicity, carcinogenicity, chemical structure.

Riassunto *(Un approccio innovativo: i database chimico relazionali e il ruolo del database ISSCAN per la valutazione della cancerogenesi chimica)*. Basi di dati di cancerogenesi e mutagenesi sono essenziali per la stima del rischio chimico. Finora queste si presentavano essenzialmente come tavole statiche, ma i progressi nel campo delle relazioni struttura-attività hanno permesso di creare nuove tipologie dove l'unione del dato tossicologico con la struttura chimica permette di legare ricerche in ambiti chimico e biologico, e di esplorare i dati dal punto di vista strutturale. Questo articolo presenta le principali basi di dati pubbliche assieme agli sviluppi delle nuove banche dati chimico relazionali, e illustra la banca dati ISSCAN sui cancerogeni chimici.

Parole chiave: basi di dati, mutagenesi, cancerogenesi, struttura chimica.

INTRODUCTION

Currently, the public has access to a variety of databases containing mutagenicity and carcinogenicity data. These resources are crucial for the toxicologists and regulators involved in the risk assessment of chemicals, which necessitate access to all the relevant literature, and capability to search across toxicity databases using both biological and chemical criteria. In this field, rapid progress has taken place both in terms of initiatives and technological innovation. In particular, public Internet resources to support biological and toxicological activity evaluation of chemicals have expanded greatly and are ushering in a new era of public information access and data mining in support of toxicity assessment.

In the context of the recent dramatic changes in regulations and regulatory needs worldwide, the progress in toxicological databases, and in database technology is particularly timely and provides an absolutely *sine qua non* tool for the regulatory implementations. As a matter of fact, increasing demands and expectations are being placed on predictive toxicology in support of the new European REACH legislation and other pieces of legislation worldwide [1], and the need emerges for more structured organization and harnessing of legacy toxicity data, and maximal utilization of these data [2]. Until now, the assessment of chemical risk in the European Union (EU) has been largely based on traditional toxicology. However legislative, societal and practical realities (too many chemicals, too few resources) have created new inducements and opportunities to encourage use and acceptance of "alternative" approaches, which can reduce substantially the need for experimental toxicological testing.

Address for correspondence: Romualdo Benigni, Dipartimento di Ambiente e Connessa Prevenzione Primaria, Istituto Superiore di Sanità, Viale Regina Elena 299, Rome, Italy. E-mail: romualdo.benigni@iss.it.

In 2003, the European Commission (EC) adopted a legislative proposal for a new chemical management system called REACH (Registration, Evaluation and Authorisation of Chemicals). Article 13(1) of the legal text of the draft REACH regulation states that [3]: "Information on intrinsic properties of substances may be generated by means other than tests, in particular through the use of qualitative or quantitative structure-activity relationship models or from information from structurally related substances (grouping or read-across), provided that the conditions set out in Annex XI are met".

REACH is expected to introduce a dramatic change in the present EU regulatory schemes [4]. It will provide a basis for the use of structure-activity relationships models, together with other "non-testing" approaches, for predicting the environmental and toxicological properties of chemicals, in the interests of time-effectiveness, cost-effectiveness and animal welfare. According to an assessment carried out by the European Chemicals Bureau (ECB), the *in vivo* mutagenicity studies, shortly followed by carcinogenicity, are posing high demand for test-related recourses [5, 6].

In particular, the science of the relationships between chemical structure and the biological activity of molecules is expected to play a new role and support three distinct activities: category formation, "read-across", and (Quantitative) Structure-Activity Relationships ((Q)SAR). A chemical category is a group of chemicals whose physicochemical and human health and/or environmental toxicological properties are likely to be similar or follow a regular pattern as a result of structural similarity. If this similarity is recognized with sufficient evidence, all the chemicals in the category can be considered (and regulated) in the same way. Another approach to fill data gaps is read-across. In the read-across approach, endpoint information (*e.g.*, carcinogenicity) for one chemical is used to predict the same endpoint for another chemical, which is considered to be "similar" in some way (usually on the basis of structural similarity). Regarding the third approach, the scientific foundation of (Q)SAR models lies in physical organic chemistry, where features of a chemical and its properties are used to estimate chemical behaviour and activity solely from the knowledge of chemical structure. (Q)SAR modeling has been widely used in pharmacology, toxicology and physical chemistry [7], and its capabilities and limitations are relatively well understood [8-10]. Regarding the use of (Q)SAR, a recent project supported by the European Chemicals Bureau (ECB) surveyed the models for mutagenicity and carcinogenicity in the public domain: the results are summarized in [4] and [11].

The extensive use of estimation techniques such as (Q)SARs, read-across and grouping of chemicals, where appropriate and in a suitably constrained context, has the potential to effect huge reductions in use of animals for modeled toxicity endpoints. At the same time, all these approaches need to be fed by adequate amounts of good quality data and databases.

DATABASES OF CHEMICAL MUTAGENS AND CARCINOGENS IN THE PUBLIC DOMAIN

Among the sources of freely available data pertaining to toxicity on chemical substances, one of the principal resources is the TOXNET database of the National Library of Medicine (NLM) (http://toxnet.nlm.nih.gov/). TOXNET is a cluster of different databases, collecting information on toxicology, hazardous chemicals, environmental health, and toxic releases. From the website, it is possible to search across and within the databases by several identifiers, such as chemical name, CAS (Chemical Abstract Service) number, molecular formula, classification code, locator code, and structure or substructure (with the CHEMID PLUS protocol). Among the TOXNET databases, the Chemical Carcinogenesis Research Information System (CCRIS) and the GENE-TOX databases deal specifically with mutagenicity and carcinogenicity data.

CCRIS contains over 8000 chemical records with animal carcinogenicity, mutagenicity, tumor promotion, and tumor inhibition test results provided by the National Cancer Institute (NCI). Test results have been reviewed by experts and all the records are written in a standardized textual format.

GENE-TOX was developed by the US Environmental Protection Agency (USEPA) and contains genetic toxicology (mutagenicity) test data, resulting from expert peer review of the open scientific literature, on over 3000 chemicals. The GENE-TOX program was established within EPA to select assay systems for evaluation, review data in the scientific literature, and recommend proper testing protocols and evaluation procedures for these systems.

Another repository of experimental carcinogenicity data available on the web is the Carcinogenic Potency Database (CPDB) (http://potency.berkeley.edu/cpdb.html). This database collects the results from over 6000 chronic, long-term animal cancer bioassays on over 1500 chemicals published in the general literature through 1997 and by the National Cancer Institute/National Toxicology Program through 1998. CPDB is organized alphabetically by chemical name. All experiments of a chemical are listed under the name of the test agent; for each experiment, information is included on test animals, features of experimental protocol, and carcinogenicity results in detail, including literature citation. CPDB is downloadable in pdf, xls or txt formats, and searchable by chemical name, CAS number, or author. Most recently, chemical-specific summary data pages have been provided on the CPDB website to make these data more accessible through chemical or structure searching (see, *e.g.*, the result of a search on acetaldehyde: http://potency.berkeley.edu/chempages/ACETALDEHYDE.html).

The US National Toxicology Program (NTP) makes available on the web (http://ntp.niehs.nih.gov/) data from more than 500 long-term toxicology and carcinogenesis bioassays collected by the NTP and its predecessor, the National Cancer Institute's Carcinogenesis

Testing Program, and organized in a database at the National Institute of Environmental Health Sciences (NIEHS). These data can be accessed as technical reports; the user can browse them directly or make text searches (by chemical name or CAS number, for example), or download the reports in pdf format. In addition, detailed experimental study data, to the level of individual animal observations, are housed in an Oracle NTP on-line database, with limited searchable access to detailed data on thousands of experiments provided to the public on the NTP website.

To enhance their structure-searchability and use in modeling applications, both the CPDB and the on-line NTP database have been "chemically-indexed" by the USEPA's National Center for Computational Toxicology DSSTox (Distributed Structure-Searchable Toxicity) database project (www.epa.gov/ncct/dsstox/), which emphasizes quality procedures for accurate and consistent chemical structure annotation of toxicological experiments. Chemical structures and summary mutagenicity and carcinogenicity data have been published for the entire CPDB inventory (www.epa.gov/ncct/dsstox/sdf_cpdbas.html; recently updated), along with the URL address locating the specific chemical data webpage on the CPDB website provided for each indexed chemical substance. Chemical structures and indicators of data availability (1 = yes, 0 = no) have also been provided for the entire chemical inventory of the online NTP database, for each of the 4 main NTP study areas (Developmental, Immunological, Genetox, and Chronic Cancer Bioassays) (see below for more information on the DSSTox project).

From the International Agency for Research on Cancer (IARC) website it is possible to access the *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans* (www-cie.iarc.fr/). In these documents, independent assessments by international experts of the carcinogenic risks to humans posed by a variety of agents, mixtures and exposures, are published. The Monographs are searchable by key word, CAS number, synonym or chemical name.

Recently, a very useful tool that is expanding access to a wide range of toxicological databases, as well as other public biological activity databases available on the web has been created by the National Center for Biotechnology Information (NCBI) through the PubChem project (http://pubchem.ncbi.nlm.nih.gov). PubChem is a public information system (tightly integrated into the cluster of biological and literature databases hosted at NCBI, such as PubMed http://www.ncbi.nih.gov/entrez/query.fcgi) that links chemical identifiers (such as chemical name, CAS number and chemical structures) to biological activity knowledge of substances. It should be remarked that PubChem is not an independently curated database, but rather a user-depositor system that aggregates standardized data from many sources, providing a tool to interrogate databases in the public domain in the US (including both toxicological and biomedical ones). The PubChem interfaces provide extensive query capabilities on textual and numeric information, as well as a comprehensive set of structure-based query methodologies. PubChem was originally created to house all the bioassay data of the NIH Molecular Libraries Initiative Screening program, whose goal is to process hundreds of thousands of chemicals through up to several thousands of high-throughput bioassay screens, using chemistry to probe biology at the fundamental cellular and protein receptor level (http://nihroadmap.nih.gov/molecularlibraries/). PubChem has expanded, however, as a user-depositor public data repository, housing large amounts of public bioassay data, including the NLM TOXNET and USEPA DSSTox inventories. PubChem has also significantly expanded its tools and capabilities for analyzing chemicals across bioactivity space, through summary activity assignments (active or inactive, or a binned range of activities).

Recent reviews [12-14] surveyed the current status of public toxicity databases in terms of their diverse content and structure, and provide a useful complement to the information summarized above.

NEW NEEDS AND NEW TOOLS: CHEMICAL RELATIONAL DATABASES

Until recently, many existing public toxicity databases have been constructed primarily as "look-up-tables" of existing data, and most often did not contain chemical structures. These databases typically utilize chemical names (usually common or commercial names) and CAS numbers which are non-unique and commercially registered and, therefore, unsuitable for a unique, public identifier. In addition, often the organization of the data follows that of the literature on paper, and does not lend easily itself to informatics implementation.

Recently, concepts and computer techniques that originated from the structure-activity relationships science have provided powerful tools to create new types of databases, where the ability to retrieve data is strongly improved both in qualitative and quantitative terms. In fact, whereas the indexing (identifier) elements in traditional databases, such as names and CAS numbers, are non-unique, prone to errors and devoid of intrinsic information, chemical structure as a chemical identifier has universally understood meaning and scientific relevance. Chemical structure and chemical concepts (*e.g.*, reactive functional groups, acidity, hydrophobicity, electrophilic reactivity, free radical formation) provide a common language and framework for exploring the similarity among chemicals and 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. Effective linkage of chemical toxicity data with chemical structure information can facilitate and greatly enhance data gathering and hypothesis generation in conjunction with (Q)SAR modeling efforts [15].

Thus, a crucial point is that of collecting and standardizing portions of the existent knowledge in a way that allows: a) exploration across both chemical and biological domains; and b) structure-searchability through the data. These characteristics may be gained when chemical structures and toxicity data are incorporated into what is termed a Chemical Relational Database (CRD). CRD is a special type of relational database whose main informational unit is a chemical structure and whose fields are attributes or data associated with that chemical structure.

In order to be accessed with a CRD application, the information has to be stored in specialized file formats. Among them, Structure Data File (SDF) format has become as the most widely used public standard for exchange of structure/data information on chemicals. SDF files are simple text files that adhere to a strict format for representing multiple chemical structure records and associated data fields. Each record in the file is composed of a "structure" section where the 2D or 3D structure of the molecule is represented as MOLfile format, and a second section composed of numerical or text data fields *(Figure 1)*. Hence, SDF files are very versatile: they can accommodate many types of data, are easily edited and manipulated by programming scripts, and could be easily ported to other types of standard formats, such as the mark-up languages, XML and CML (for further information on issues related to chemical annotation [4, 15]).

SEARCHING CAPABILITY OF CRD DATABASES

Even though simple, useful searches can be performed with widely available informatics tools, such as XLS (Excel-readable) files of chemicals with annotated toxicity and/or properties, where it is possible, *e.g.*, to retrieve substances within a predefined range of toxicity values. However, coding the chemical structure in the SDF file allows one to perform remarkably more complex searches by using specialized CRD software programs: most commercially available CRD applications provide substructure and functional group search features, different algorithms for searching compounds chemically similar to query ones (similarity search), and text and data field search functions (for information on commercial and public software applications, see the DSSTox website: www.epa.gov/ncct/dsstox/SDFViewerBrowserCRDs.html).

When the SDF file is imported into a CRD application, it is possible to do structure/text/data relational searching across records in the database. All these operations are collectively termed "data mining" [14]. In *Figure 2*, as an example the substructure searching results using aniline as query structure are depicted. The result of the search consists of all chemicals in a database (*i.e.*, an SDF file) containing aniline as basic substructure. In this way, it is possible to identify subsets of chemicals according to any structural query (*i.e.*, functional group, or molecular substructure).

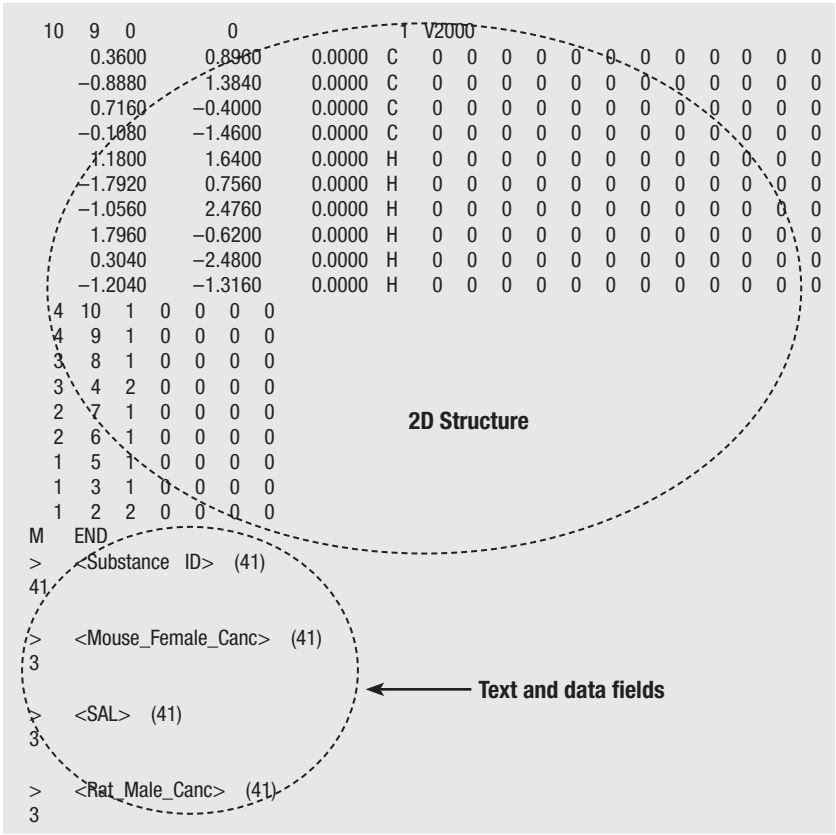

Fig. 1 | *A sample SDF file, containing both structural (top) and data (bottom) information.*

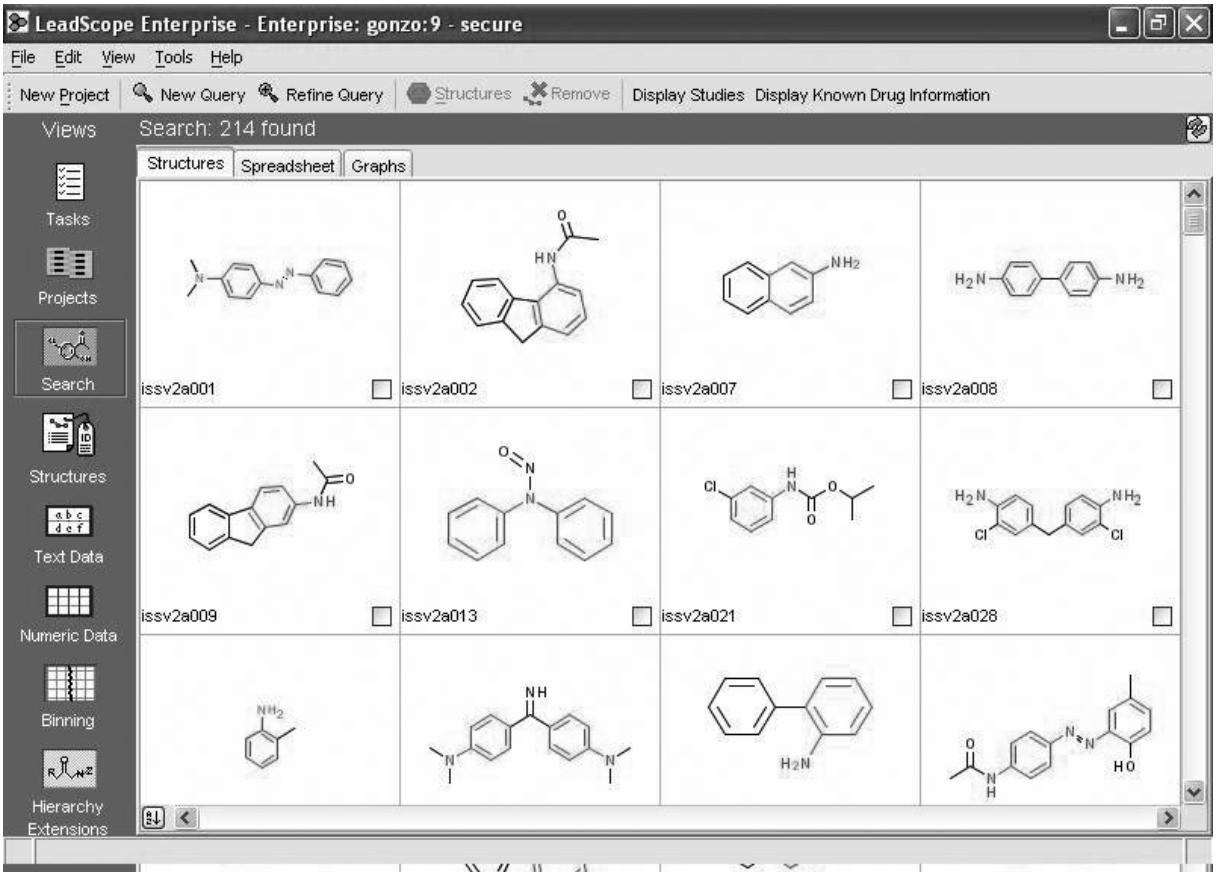

Fig. 2 | *Example of substructure searching in a database of diverse chemicals. All the chemicals including aniline as a substructure are highlighted. The search was performed with the program LeadScope (LeadScope Inc., Ohio).*

Another very useful feature with the addition of visual analytic tools is the possibility of characterizing a database by its component functional groups or chemical classes. An example of this capability is presented in *Figure 3*, by applying a CRD application to the SDF file. The figure shows that the chemicals in the database are divided into chemical classes, and the frequency in each class is given. In addition, it is possible to add colors to each class bar, pointing visually to the abundance in each class of the chemicals active and inactive for some selected property (*e.g.*, carcinogenicity). Visualization of the retrieved data makes easier and more immediate the understanding of the results of the query.

The above data mining capabilities can be expanded to perform more complex searches, by formulating queries where specific combinations of structures, data and text (*i.e.*, "chemical profiles") are searched for in the database at the same time.

Another crucial operation that can be performed on structural databases is that of calculating chemical similarity between pairs of chemicals [16]. Based on the structural motifs in common to two chemicals, the degree of similarity can be quantified on, *e.g.*, a 0 to 1 scale, and the resulting similarity value can be used as supporting evidence in the process of identifying categories of similar chemicals.

A more sophisticated use of data mining approaches allowed by modern CRD applications is the identification of one or more common structural patterns among groups of chemicals with similar characteristics or profiles (*e.g.*, toxicity). Such patterns, when identified, can be used as predictive models to estimate the toxicity of other chemicals, with similar structural patterns [14].

THE DSSTOX DATABASE PROJECT

In view of the powerful opportunities provided by the CRD technology, a major problem is that of transforming the available databases according to the new standards. A considerable progress is represented by PubChem that allows the user to browse through the US public databases individually and collectively according to structural criteria. However, even though this design permits a user to explore and download all or portions of the available information, there is no quality review of the structural inventory of PubChem in relation to bioassay data, which come from a large number of user-depositors

or sources with various levels of quality review applied to their data; hence, it is largely a "user-beware" public resource. New initiatives are now being developed to address this concern in the world of toxicity data. An example of project designed to provide the user with self-contained data files that can be readily incorporated into CRD and used freely is the Distributed Structure-Searchable Toxicity (DSSTox) Database Network, which is a project of the USEPA (www.epa.gov/comptox/).

A primary objective of the DSSTox website (www.epa.gov/ncct/dsstox) is to serve as a central community forum for publishing standard-format, structure-annotated chemical toxicity data files for open-access, public use, and for use in CRD applications. DSSTox efforts include the careful quality annotation of chemical structures, standardization and documentation of toxicity data in collaboration with toxicity data experts, and open public access to toxicity databases.

In the initial phase, data files were not structure-searchable on the DSSTox web site itself, but the data files could be downloaded in their entirety and freely used. Since September 2007, a DSSTox structure-browser offered on the DSSTox website allows structure/substructure/similarity-searching through all DSSTox data file content, and can be additionally accessed from off-site collaborators (*e.g.*, CPDB, EPA IRIS, NTP) for website searching through either local content (*e.g.*, just the content of the originator's website) or broader searching through the DSSTox inventory and, soon to be added, providing external links to PubChem.

At present, the DSSTox data file cluster includes six separate databases: CPDBAS – Carcinogenic Potency Project Summary Tables (Source, LS Gold, Carcinogenic Potency Project, UC Berkeley); DBPCAN – EPA Disinfection By-products Carcinogenicity Estimates Database (Source, YT Woo, USEPA, Office of Pollution Prevention & Toxics); EPAFHM – EPA Fathead Minnow Acute Toxicity Database (Source, C. Russom, USEPA, Mid-Continental Ecology Division-Duluth); NCTRER – FDA NCTR Estrogen Receptor Binding Database (Source, Weida Tong and Hong Fang, National Center for Toxicological Research, Jefferson, Arkansas); FDAMDD – FDA Maximum Recommended Daily Dose (Source, Edwin Matthews and R. Daniel Benz, US FDA, Rockville, MD), and the newest data file, IRISTR (Source, USEPA's Integrated Risk Information System Toxicity Reviews), which includes 34 toxicity-related content fields. Additionally, the DSSTox file inventory includes 2 structure-locator files, HPVCSI (USEPA's High Production Volume Challenge Program) and NTPBSI

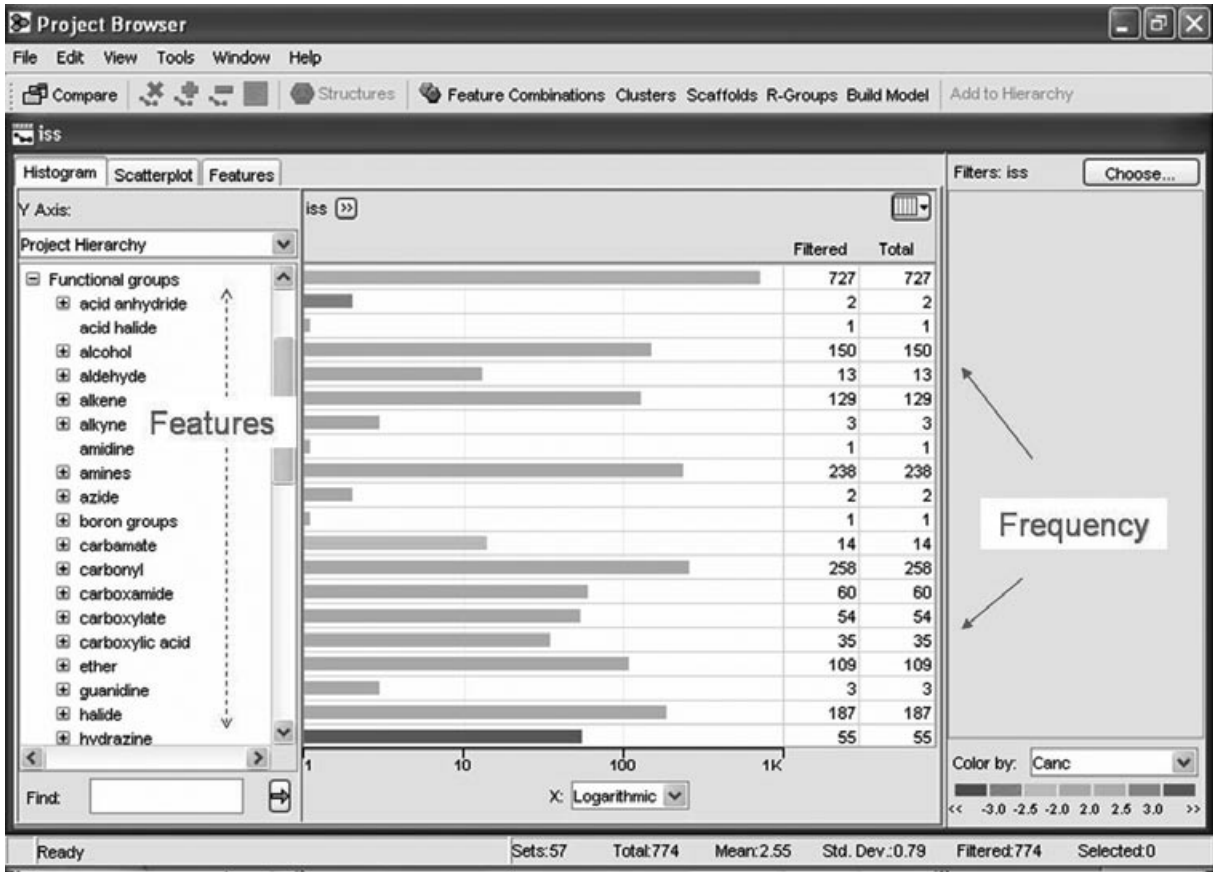

Fig. 3 | *Example of classification of the chemicals in a database by chemical classes. The analysis was performed with the program LeadScope (LeadScope Inc., Ohio).*

(National Toxicology Program Bioassay) containing URL addresses to chemical-specific data pages, and 2 structure-index files containing only a chemical structure listing, NTPHTS (National Toxicology Program High-Throughput Screening) and TOXCST (EPA's National Center for Computational Toxicology ToxCast testing program).

Each DSSTox database is published as a separate and distinct module that adheres to standard conventions in SDF data file format, file names, chemical structure fields, and minimum documentation requirements. Together with the SDF file, the DSSTox provides an MS Excel-readable file (.xls) (reporting the non-structural data), and an Acrobat-readable file (.pdf) which displays the traditional graphical representation of the chemicals. In addition, the DSSTox website provides a detailed guide on the use of files, and a rich documentation on the entire subject of databases and related concepts [12, 17]. The collected DSSTox published inventory contains over six thousand unique chemical substances relevant to toxicology and can be merged for structure-searching, or ported into CRD applications.

THE ISSCAN DATABASE ON CHEMICAL CARCINOGENS

As pointed out above, currently the public has access to a variety of toxicity databases; however, these publicly available data may not be immediately suitable for use. One general issue is that of data quality, both from a chemical and biological perspective. Beyond its most obvious meaning (data "must" be of good quality, otherwise any inference based on them is simply devoid of any value), there are more subtle problems linked to this issue. For example, for each chemical the CCRIS (as well as the CPDB) reports all the available experimental results. There are cases where more than one experiment, with contradictory results, exist for a given chemical. There are also cases where the experimental protocols differ to a large extent. In all these cases, the database user has to employ her/his expert judgement to make an activity assignment. Together with the data issue and linked to it, is that of the data standardization, which can become extremely critical for some more formalized applications, such as QSAR analyses [9]. These approaches need highly summarized representations of the activity of the chemicals (*i.e.*, a unique number for the potency of the active compounds; a dichotomous classification into actives/inactives). But the large public databases often do not meet these modeling requirements. One example is the NTP on-line database that includes high-level detail on animal bioassays and genetic toxicity experiments for several thousands of chemicals, respectively, but which does not provide ready access to data for the entire chemical study inventory, relational access to particular slices of the data, or aggregate summarizations of the data according to the requirements of QSAR modeling.

To alleviate the above problems, at the Istituto Superiore di Sanità (ISS) a new database on chemical carcinogens called ISSCAN: "Chemical carcinogens: structures and experimental data" has been built. The data can be freely downloaded from the ISS website: www.iss.it/ampp/dati/cont.php?id=233&lang=1&tipo=7 or from the DSSTox site: www.epa.gov/ncct/dsstox/sdf_isscan_external.html.

The ISSCAN database contains information on chemical compounds tested with the long-term carcinogenicity bioassay on rodents (rat, mouse). The specific characteristics of the ISSCAN database in respect to other databases should be emphasized. First, the ISSCAN initiative is aimed at providing the scientific and regulatory community with carcinogenicity calls that have been re-checked, in order to ensure the quality of the data. The data were cross-checked on different sources of information available; contradictions were solved going back to the original papers, and results based on insufficient protocols were not included. Second, the biological data (carcinogenicity and Salmonella mutagenicity) were coded in numerical terms that can be used directly for QSAR analyses. This aspect of being QSAR-ready eliminates the intermediate passage of data transformation that often is problematic for the QSAR practitioner without specific toxicological expertise.

The general structure of the database is inspired by that of the DSSTox. The ISSCAN database is composed of standard chemical data fields, such as 2D structure, chemical name and synonyms, CAS registry number, molecular weight, chemical formula and SMILES notation, together with biological data fields: carcinogenic potency in rat and mouse, mutagenicity in *Salmonella typhimurium* (Ames test), carcinogenicity results in the four experimental groups most commonly used for the cancer bioassay, carcinogenicity results from the NTP experimentation (when available), overall carcinogenicity, together with the source of carcinogenicity data. *Figure 4* displays the information reported by ISSCAN for a representative chemical.

From the website it is possible to download four different files:

1) an SDF file containing chemical structures together with chemical and biological data;
2) a PDF file with a detailed explanation and guidance of use;
3) a PDF file with 2D chemical structures of the substances;
4) an XLS file of the data.

At present, the second updated version of ISSCAN is available, including 890 chemicals tested for rodent carcinogenicity (the main primary sources of data are the NTP, CPDB, CCRIS, and IARC repositories). It is our plan to accomplish the evaluation of the remaining chemicals by the year 2008. Since the SDF file cannot be read by users without specialized software applications, it is also our plan to make available on our website a tool suitable for simple analyses.

It should be emphasized that this type of project (ISSCAN) is not in opposition to other databases (*e.g.*, CCRIS, CPDB) that follow the philosophy of reporting vast amounts of data at different hierarchical levels, also including contradictory evidence when existing. In contrast and complementary to these efforts, the ISSCAN initiative is aimed at providing the end-user with information that is revised and re-organized for a specific aim, whereas the above databases have the important role of keeping track of all the available information. Even when the knowledge contribution of portions of such databases looks very minor (*e.g.*, data from experiments with few animals and old protocols), this – in a different context – may turn out to be very useful for, *e.g.*, planning further studies.

CONCLUSIONS

The key to a rapid progress in the field of chemical toxicity databases exploitation is that of combining information technology with the chemical structure as identifier of the molecules. This permits an enormous range of operations (*e.g.*, retrieving chemicals or chemical classes, describing the content of databases, finding similar chemicals, crossing biological and chemical interrogations, etc.) that other more classical databases cannot allow. In the foreseeable future, this trend will become even more pervasive: a clear demonstration of this trend is the creation by NCBI of the chemically-interrogable PubChem database fully integrated with the traditional, textual PubMed (http://www.ncbi.nlm.nih.gov/sites/entrez) repository of biomedical information. At the same time, there is a proliferation of new tools aimed at fully exploiting the possibilities afforded by CRDs. Together with private companies, public bodies have entered the arena. We will quote only two examples.

The European Chemicals Bureau (ECB) has developed (through IdeaConsult Ltd.) ToxTree. This is a freely available application from the ECB website (http://ecb.jrc.it/qsar) able to estimate different types of toxic hazards by applying structural rules.

Another tool is the (Q)SAR Application ToolBox, developed under the umbrella of the OECD, for which a pilot version is currently under development. The application will be made publicly available during the first half of 2008 (www.oecd.org/env/existingchemicals/qsar). This application links a number of existing tools as well as a library of existing (Q)SAR models and will allow a user to:

a) make estimations for single chemicals, and receive the results of all the (Q)SAR estimates for all the models covering the appropriate domain, for the relevant endpoints that the user wishes to estimate;
b) receive summary information on the validation results of the model according to the OECD validation principles so that the user can decide for which regulatory purpose the estimate can be used; (Q)SAR models would be incorporated into the toolbox as they come forward from member countries with the information on their validation according to the OECD principles;
c) receive a list of analogues, together with their (Q)SAR estimates; and
d) receive estimates for metabolite activation/detoxification information. The Toolbox will link a number of public domain tools, and make them available to the user according to a flexible workflow.

It should be emphasized that the use of CRD is a key element for both of the QSAR Applications, the ECB ToxTree and the OECD (Q)SAR ToolBox.

To further expand the reach of databases, new challenges have to be addressed. All the databases will need a maintenance system in order to permit the integration of additional data to their historical body, and to possibly modify the technical implementation as soon as technology improves. Another crucial challenge is that of integrating different databases, and different levels of information (*i.e.*, overall study result, tests with certain experimental conditions, dose level group results, results from the individual test subject level) in the same database [13]. The public ToxML data schema for toxicology study areas, implemented by Leadscope for several US Food and Drug Administration (FDA) data sets, is a prominent example of an effort addressing this need. In addition, database standardization and CRD-accessibility will be a high priority, and the collection of new data has to planned accordingly at the same moment the experimental design is laid out. An example of this new trend is ToxCast [18], the recently established EPA program to perform high-throughput analyses of cellular and whole-

Formula	C15H13NO
FW	223.2699
Substance ID	2
Mouse_Female_Canc	ND
SAL	3
Rat_Male_Canc	ND
TD50_Rat	NP
TD50_Mouse	ND
Rat_Female_Canc	1
Canc	1
MolWeight	223.28
Mouse_Male_Canc	ND
Mouse_Male_NTP	ND
ChemName	4-Acetylaminofluorene
Rat_Male_NTP	ND
Reference	CPDB
SMILE	O=C(Nc3c2c1ccccc1Cc2ccc3)C
Rat_Female_NTP	ND
CAS	28322-02-3
Mouse_Female_NTP	ND
Synonyms	4-AAF; 4-acetamidofluorene; n-4-fluorenylacetamide; n-9H-fluoren-4-ylacetamide; n-fluoren-4-ylacetamide.

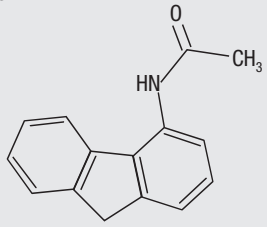

Fig. 4 | *Example of the information reported by the ISSCAN database on chemical carcinogens for a representative chemical.*

animal toxicity on a chosen set of compounds: the standardization of data and CRD-accessibility will be a necessary requirement in order to fully exploit the value of these data (for more information, see: www.epa.gov/ncct/toxcast/).

Acknowledgements

This work was partially granted by the EU FP6 Contract n. 037017 OSIRIS "Optimized strategies for risk assessment of industrial chemicals through Integration of non-test and test information"

Disclaimer

This manuscript does not necessarily reflect the views and policies of the USEPA, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

Submitted on invitation.
Accepted on 16 December 2007.

References

1. Organisation for Economic Co-operation and Development. *Report on the Regulatory Uses and Applications in OECD Member Countries of (Quantitative) Structure-Activity Relationship ((Q)SAR) Models in the Assessment of New and Existing Chemicals.* 58. 2006. OECD Series on Testing and Assessment. Paris: OECD; 2006. (ENV Monograph No. 58).
2. Richard AM. Future of predictive toxicology. An expanded view of "chemical toxicity", future of toxicology perspective. *Chem Res Toxicol* 2006;19:1257-62.
3. Commission of the European Communities. *Proposal concerning the registration, evaluation, authorisation and restriction of chemicals (REACH)*. (COM(2003)644Final). Bruxelles: EU; 2003.
4. Benigni R, Netzeva TI, Benfenati E, Bossa C, Franke R, Helma C, Hulzebos E, Marchant C, Richard A, Woo Y-T, Yang C. The expanding role of predictive toxicology: an update on the (Q)SAR models for mutagens and carcinogens. *J Environ Sci Health* C 2007;25:53-97.
5. Pedersen F, de Brujin J, Munn SJ, and Van Leeuwen, K. *Assessment of additional testing needs under REACH. Effects of (Q)SARs, risk based testing and voluntary industry initiatives.* Ispra: Joint Research Centre; 2003. (JRC report EUR 20863 EN).
6. Van der Jagt K, Munn SJ, Torslov J, de Brujin J. *Alternative approaches can reduce the use of test animals under REACH. Addendum to the Report "Assessment of addtional testing needs under REACH. Effects of (Q)SARs, risk based testing and voluntary industry initiatives".* Ispra: Joint Research Centre; 2004. (JRC Report EUR 21405 EN).
7. Hansch C, Leo A. *Exploring QSAR. 1. Fundamentals and applications in chemistry and biology.* Washington DC: American Chemical Society; 1995.
8. Hansch C, Hoekman D, Leo A, Weininger D, Selassie CD. Chem-bioinformatics: comparative QSAR at the interface between chemistry and biology. *Chem Rev* 2002;102:783-812.
9. Franke R, Gruska A. General introduction to QSAR. In: Benigni R (Ed.). *Quantitative structure-activity relationhsip (QSAR) models of mutagens and carcinogens.* Boca Raton: CRC Press; 2003. p. 1-40.
10. Benigni R. Structure-activity relationship studies of chemical mutagens and carcinogens: mechanistic investigations and prediction approaches. *Chem Rev* 2005;105:1767-800.
11. Benigni R, Bossa C, Netzeva TI, Worth AP. *Collection and evaluation of (Q)SAR models for mutagenicity and carcinogenicity. Office for the Official Publications of the European Communities. EUR - Scientific and Technical Research Series.* Luxenbourg; 2007. (EUR 22772 EN). Available from: http://ecb.jrc.it/documents/QSAR/EUR_22772_EN.pdf; last visited 21/11/2007.
12. Richard AM, Williams CR. Public sources of mutagenicity and carcinogenicity data: use in structure-activity relationship models. In: Benigni R (Ed.). *Quantitative Structure-Activity Relationship (QSAR) models of mutagens and carcinogens.* Boca Raton: CRC Press; 2003. p. 145-74.
13. Yang C, Benz RD, Cheeseman MA. Landscape of current toxicity databases and database standards. *Curr Opinion Drug Discov Develop* 2006;9:124-33.
14. Yang C, Richard AM, Cross KP. The art of data mining the minefields of toxicity databases to link chemistry to biology. *Curr Comput Aid Drug Des* 2006;2:135-50.
15. Richard AM, Gold LS, Nicklaus MC. Chemical structure indexing of toxicity data on the Internet: moving toward a flat world. *Curr Opinion Drug Discov Develop* 2006;9:314-25.
16. Gallegos Saliner A. Mini-review on chemical similarity and prediction of toxicity. *Curr Comput Aid Drug Des* 2006;2:105-22.
17. Richard AM. DSSTox web site launch: improving public access to databases for building structure-toxicity prediction models. *Preclinica* 2004;2:103-8.
18. Dix DJ, Houck KA, Martin MT, Richard AM, Setzer MW, Kavlock RJ. The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol Sci* 2007;95:5-12.