New perspectives in toxicological information management, and the role of ISSTOX databases in assessing chemical mutagenicity and carcinogenicity

Romualdo Benigni*, Chiara Laura Battistelli, Cecilia Bossa, Olga Tcheremenskaia and Pierre Crettaz†

Istituto Superiore di Sanità, Environment and Health Department, 00161 Rome, Italy; *Federal Office of Public Health, Consumers Protection Unit, CH-3097 Bern, Switzerland

Received on January 10, 2013; revised on January 29, 2013; accepted on January 30, 2013

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 necessitates access to all the relevant literature, and the capability to search across toxicity databases using both biological and chemical criteria. Towards the larger goal of screening chemicals for a wide range of toxicity endpoints of potential interest, publicly available resources across a large spectrum of biological and chemical data space must be effectively harnessed with current and evolving information technologies (i.e. systematised, integrated and mined), if long-term screening and prediction objectives are to be achieved. A key to rapid progress in the field of chemical toxicity databases 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. This article describes the progress in the technology of toxicity databases, including the concepts of Chemical Relational Database and Toxicological Standardized Controlled Vocabularies (Ontology). Then it describes the ISSTOX cluster of toxicological databases at the Istituto Superiore di Sanità. It consists of freely available databases characterised by the use of modern information technologies and by curation of the quality of the biological data. Finally, this article provides examples of analyses and results made possible by ISSTOX.

Databases on chemical toxicological information

‘A toxicity database is a valued set of electronic information (i.e., data) that can be related to the toxicity of substances of which the information is accessible by computer, organised by software, and utilised for safety and risk analysis of chemicals, product discovery and development, or academic research for settings in the biomedical and toxicological sciences’ (10). Complementary to the above is the definition of most important uses of toxicity databases, i.e.

- Data-mining information. Data mining is a technique that provides analysis of large databases to identify unsuspected relationships or hidden patterns in a data set that can be used to predict future behaviour;
• ‘Read-across’ of various sources (also regulators) and endpoints. In the read-across approach, the endpoint information for one or a few chemicals is used to predict the same endpoint for another chemical, considered to be ‘similar’ in some way (usually on the basis of structural similarity) (7,11);
• Building a computational model [source for (Q)SAR data sets].

Many authors have pointed to a ‘lack of data’ problem with current databases (12–15) because of lack or fragmentation of data (most data do not exist in electronic or exchangeable formats), lack of standardisation dictionary and controlled vocabularies, and poor accessibility. Among other issues, the need for the development and adoption of shared public standards for toxicity databases should be emphasised. These standards should include controlled vocabulary and hierarchical data relationships or ontologies (i.e. using the same terminology to describe the same things and incorporating the layered relationships of different terms to one another), should be derived from close working knowledge of the toxicity study domain (e.g. carcinogenicity, mutagenicity, developmental toxicity and neurotoxicity) and should be inclusive of chemical structure.

Table I displays a selection of the publicly available carcinogenicity and genetic toxicology/mutagenicity databases, including a short description and the main features and use. The list is representative—but not exhaustive—of databases in the public domain.

From a structural point of view, until lately, toxicity databases have been constructed as ‘look-up tables’ of existing data and did not contain chemical structure. Typically, the use of chemical name or CAS number, as identifier, is non-unique and prone to errors. In the new types of databases instead, defined as Chemical Relational Database (CRD), the main informational unit is a chemical structure and the fields are data (e.g. toxicity) associated with that chemical structure. Therefore, using the chemical structure as a chemical identifier has a universally understood meaning and scientific relevance for chemical toxicity databases: effective linkage of chemical toxicity data with chemical structure information can facilitate and greatly enhance data gathering and hypothesis generation.

The use of CRD permits the exploration across both chemical and biological domains and structure-search ability through the data. In order to be accessed with a CRD application, the information has to be stored in specialised file formats, the most widely used for exchange of structure/data information on chemicals is Structure Data File (SDF) format. When the SDF file is imported into a CRD application, it is possible to do structure/text/data relational searching (‘data mining’) across records in the database, as follows:

• Searching results using a basic substructure or functional group as query structure;
• Classification of the chemicals by chemical classes (frequency in each class can be given);
• Formulating query with specific operation of structure–data–text (chemical profiles);
• Calculating chemical similarity between pairs of chemicals;
• Identification of one or more common structural patterns among groups of chemicals with similar characteristics or profiles of toxicity; these patterns can be used as predictive models for estimate the toxicity of chemicals with similar structural patterns.

In addition, SDF files are very versatile. They can accommodate many types of data, are easily edited and manipulated by programming scripts and can be easily ported to other standard formats, such as the mark-up languages XML and CML (4) (see also http://www.epa.gov/ncct/dsstox/SDFViewerBrowserCRDs.html).

Whereas the current technology provides a wide range of possibilities, a major problem that remains is that of collecting good-quality data into databases according to these new standards.

The ISSTOX toxicity databases

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 perspectives. 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 Chemical Carcinogenesis Research Information System (CCRIS) database [as well as the Carcinogenic Potency Database (CPDB) database] reports all available carcinogenicity bioassay results. However, there are cases where contradictory results exist for a given chemical. In other cases, the experimental protocols applied to the same chemical in different experiments differ to a large extent. In all these cases, the database user has to employ her/his expert judgement to make an activity assignment. Other problems often encountered in databases are that the same experiment/result is listed more than once, under different headings. All these emphasise the need for critically selecting data to be included in the databases.

Together with the data issue and linked to it is that of the data standardisation, which can become extremely critical for some more formalised applications, such as QSAR analyses. These approaches need highly summarised representations of the activity of the chemicals (i.e. a unique number for the carcinogenic potency of the active compounds and a dichotomous classification into active/inactive chemicals). But the large public databases often do not meet these modelling requirements. One example is the NTP online database that includes high-level detail on animal bioassays and genetic toxicity experiments for several thousands of chemicals, 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 summarisations of the data according to the requirements of QSAR modelling.

Recognising the above problems, the ISS has started a project to generate new databases on chemical toxicity. The project is called ISSTOX, and at present includes five databases, each relative to a different toxicological endpoint. Part of the project (ISSMIC and ISSBIOC databases) has been developed in collaboration with the Swiss Federal Office of Public Health. The databases are (i) long-term carcinogenicity bioassay in rodents (rat and mouse) (ISSSCAN), (ii) in vitro mutagenesis in Salmonella typhimurium (Ames test) (ISSSTY), (iii) in vivo mutagenesis (micronucleus test) (ISSMIC), (iv) in vitro cell transformation (ISSCTA) and (v) carcinogenicity and
### Table I. Databases on carcinogenicity and mutagenicity in the public domain

<table>
<thead>
<tr>
<th>Database/link</th>
<th>Description</th>
<th>Main features and use</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACToR</td>
<td>Aggregated Computational Toxicology Resource (ACToR) by the US EPA National Center for Computational Toxicology (NCCT); contains data on environmental chemicals and toxicology (including chemical structure, physico-chemical values, in vitro and in vivo toxicity assays) for over 500,000 compounds derived from &gt;1000 sources data. ToxRefDB, ToxCastDB and ExpoCastDB are included. Data searchable by name, CAS number or structure name</td>
<td>Cluster of environmental chemicals and toxicological databases. Planned to become a fully relational database</td>
</tr>
<tr>
<td>ATSDRToxProfiles</td>
<td>Agency for Toxic Substances and Disease Registry (ATSDR) includes information on &gt;1800 chemicals, hazardous substances and mixtures found at National Priorities List (NPL) sites. These hazardous substances are ranked based on frequency of occurrence at NPL sites, toxicity and potential for human exposure. Data searchable by CAS number, Substance Name, Synonym or Trade name</td>
<td>Toxicity experimental data</td>
</tr>
<tr>
<td>CPDB</td>
<td>Carcinogenic Potency Database (CPDB) by University of California (Berkeley); contains results of 6400 chronic, long-term animal cancer tests on 1547 chemicals. Data downloadable in easy-to-use format (pdf, xls and txt), and searchable by chemical name, CAS number or author. ‘Chemically-indexed’ in the DSSTox database</td>
<td>Toxicity experimental data</td>
</tr>
<tr>
<td>Danish (Q)SAR database</td>
<td>Danish EPA; estimates (predictions not experimental data) from over 70 (Q)SAR models and health effects for &gt;160,000 chemicals</td>
<td>(Q)SAR database</td>
</tr>
<tr>
<td>DSSTox</td>
<td>Distributed Structure-Searchable Toxicity Database (DSSTox) Network; downloadable, structure-searchable, standardised chemical structure files associated with toxicity data. Emphasise quality procedures for accurate and consistent chemical structure annotation of toxicological experiments</td>
<td>Cluster of toxicity databases, (Q)SAR-ready and Chemical Relational Database</td>
</tr>
<tr>
<td>ECHA C&amp;L Inventory database</td>
<td>C&amp;L Inventory database by ECHA contains classification and labelling information on substances (30,000 dosiers), received from manufacturers and importers. It also includes the list of harmonised classifications. Data searchable by CAS number, EC number, name, tonnage band (min. and max.) and registration fields (such as registrant, country, type, etc.)</td>
<td>Regulatory classification and labelling database</td>
</tr>
<tr>
<td>ESIS</td>
<td>European Chemical Substances Information system (ESIS) organised by European Chemicals Bureau (ECB) of European Commission’s Joint Research Centre (JRC). Provide access to a collection of European regulatory inventories and data associated to them on 2500 chemicals related to risk and safety. It includes EINECS (European Inventory of Existing Commercial Chemical Substances), ELINCS (European List of Notified Chemical Substances), BPD (Biocidal Products Directive) active substances, PBT (Persistent, Bioaccumulative and Toxic) or vPvB (very Persistent and very Bioaccumulative), CLP/GHS (Classification, Labelling and Packaging of substances and mixtures), HPVCs (High Production Volume Chemicals) and LPVCs (Low Production Volume Chemicals), IUCLID Chemical Data Sheets and others. Data searchable by name, CAS number or molecular formula</td>
<td>Toxicity experimental data</td>
</tr>
<tr>
<td>EXTOXNET PIP</td>
<td>Extension Toxicology Network (EXTOXNET) Pesticide Information Profiles (PIP) database by University of California, Davis, Oregon State University, Michigan State University, Cornell University and the University of Idaho. It includes toxicological and ecological effects, environmental fate, physical properties and exposure guidelines. Data searchable by name, CAS number or molecular formula</td>
<td>Risk assessment</td>
</tr>
<tr>
<td>HERA</td>
<td>Human and Environmental Risk Assessments (HERA): toxicity and risk data on ingredients of household cleaning products. Data searchable by name and CAS number</td>
<td>Risk assessment</td>
</tr>
<tr>
<td>IARC</td>
<td>International Agency for Research on Cancer (IARC); monographs on the evaluation of carcinogenic risks to humans of &gt;900 chemicals. Data searchable by key word, CAS number, synonym or chemical name</td>
<td>Risk assessment</td>
</tr>
</tbody>
</table>
### Table I. Continued

<table>
<thead>
<tr>
<th>Database/link</th>
<th>Description</th>
<th>Main features and use</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPCS INCHEM <a href="http://www.inchem.org/">http://www.inchem.org/</a></td>
<td>Public, searchable, peer-reviewed chemical safety-related publications and database records from international bodies, including Concise International Chemical Assessment Document (CICADS), International Chemical Safety Cards (ICSCs), Pesticide Data Sheets (PDSs), Screening Information Data Set (SID) for High Production Volume Chemicals. Chemicals included are contaminants in the environment and food. Data searchable by key word, CAS number, synonym or chemical name</td>
<td>Risk assessment</td>
</tr>
<tr>
<td>ISSTOX Chemical Toxicity Databases <a href="http://www.iss.it/ampp/dati/cont.php?id=233&amp;lang=1&amp;tipo=7">http://www.iss.it/ampp/dati/cont.php?id=233&amp;lang=1&amp;tipo=7</a></td>
<td>Chemical Toxicity databases from Istituto Superiore di Sanità (ISS), Italy, designed to be usable for SARs studies for toxicity prediction: (i) long-term carcinogenicity bioassay on rodents (rat and mouse) (ISSCAN), (ii) in vitro S. typhimurium mutagenesis (Ames test) (ISSSTM), (iii) in vivo mutagenesis (micronucleus test) (ISSMIM), (iv) Cell Transformation Assays (ISSCTA) and (v) Mutagenicity and Carcinogenicity of Biocides (ISSBIOC)</td>
<td>Cluster of toxicity databases, (Q)SAR-ready, Chemical Relational Database</td>
</tr>
<tr>
<td>JRC (Q)SAR Model Inventory <a href="http://iehp.jrc.ec.europa.eu/our_databases/jrc-qsar-inventory">http://iehp.jrc.ec.europa.eu/our_databases/jrc-qsar-inventory</a></td>
<td>The JRC QSAR Model Inventory is an inventory of information on the validity of (Q)SAR models by JRC. The database help to identify valid (Q)SARs (e.g. for the purposes of REACH). The QSAR Model Reporting Format (QMRF) is a harmonised template structured according to the OECD principles for the validation of (Q)SAR models (QMRF) is a harmonised template structured according to the OECD principles for the validation of (Q)SAR models</td>
<td>(Q)SAR database</td>
</tr>
<tr>
<td>NTP <a href="http://ntp.niehs.nih.gov/">http://ntp.niehs.nih.gov/</a></td>
<td>US NIH/NIEHS National Toxicology Program (NTP); contains testing status and information of agents registered in the USA of public health interest (&gt;500 2-year studies and &gt;2000 genetic toxicity studies), accessioned as technical reports; searchable by chemical name or CAS number, downloadable the reports in pdf form. 'Chemically-indexed' in the DSSTox database</td>
<td>Primary toxicological data, including historical controls</td>
</tr>
<tr>
<td>PAN pesticide <a href="http://www.pesticideinfo.org/">http://www.pesticideinfo.org/</a></td>
<td>Pesticide Action Network (PAN) North America; contains toxicity data (human acute toxicity, human carcinogenicity, reproductive/developmental toxicity, endocrine disruption, neurotoxicity and ecotoxicity) and regulatory information on &gt;6500 pesticides. Searchable by CAS number, chemical name, Trade name or chemical category</td>
<td>Toxicity experimental data</td>
</tr>
<tr>
<td>PubChem <a href="http://pubchem.ncbi.nlm.nih.gov/">http://pubchem.ncbi.nlm.nih.gov/</a> and <a href="http://www.ncbi.nlm.nih.gov/pcassay">http://www.ncbi.nlm.nih.gov/pcassay</a></td>
<td>National Center for Biotechnology Information (NCBI); it is not an independent database, but a depositor system with standardised data from toxicological/biological database sources and from literature. Contains chemical structure-annotated data submissions, with summary bioassay data</td>
<td>Biological activity database, including toxicity. Links to other databases</td>
</tr>
<tr>
<td>TOXNET <a href="http://toxnet.nlm.nih.gov/">http://toxnet.nlm.nih.gov/</a></td>
<td>TOXicology Data NETwork by US National Library of Medicine (NLM); databases on toxicity, hazardous chemicals, environmental health and toxic releases. It included Chemical Carcinogenesis Research Information System (CCRIS), Genetic Toxicology Data Bank (GENETOX), Integrated Risk Information System (IRIS), International Toxicity Estimates for Risk (ITER) and Household products database. Searchable within and across the databases by chemical name, CAS number, molecular formula, classification code, locator code, and structure or substructure</td>
<td>Cluster of toxicity databases</td>
</tr>
<tr>
<td>TSCATS <a href="http://www.syrres.com/what-we-do/product.aspx?id=136">http://www.syrres.com/what-we-do/product.aspx?id=136</a></td>
<td>The TSCA Test Submissions (TSCATS) database is an online index to unpublished, non-confidential studies and technical reports submitted by industry to US EPA. It contains chemical testing results and adverse effects of chemicals on health and ecological systems. Search by CAS number, formula or chemical name</td>
<td>Toxicity experimental data</td>
</tr>
</tbody>
</table>

mutagenicity results for biocides and plant protection products (or pesticides) (ISSBIOC).


The structure of the databases is inspired by that of the Distributed Structure-Searchable Toxicity (DSTox) Network of the US Environmental Protection Agency (EPA): [http://www.epa.gov/nct/dstox/](http://www.epa.gov/nct/dstox/). Similar to the DSTox spirit, the ISSTOX project wants to contribute to the free diffusion of scientific data in a standardised, easy to read format. From the ISSTOX website, for each database it is possible to download (together with a guidance for use) three different files: (i) an SDF file containing chemical structures together with chemical and biological data, (ii) a PDF file with 2D chemical structures of the substances and (iii) an XLS file of the data.

Whereas the structure and organisation of the databases permit an easy interrogation and use with technologically advanced procedures, particular attention has been given to the
quality of the biological data selected for inclusion. The critical selection and review of data have followed different criteria, depending on the type of data and related issues.

The ISSCAN database (version v4n, \(n = 1150\)) contains information on chemical compounds tested with the long-term carcinogenicity bioassay on rodents (rat and mouse). 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 (retrieved from CPDB), 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. The main primary sources of data are the NTP, CPDB, CCRIS and IARC repositories. When available, summary mutagenicity in *S. typhimurium* (Ames test) information was included. The data were cross-checked on the different sources of information available; contradictions were solved going back to the original articles, and results based on insufficient protocols were not included.

The ISSMIC database (version v4b, \(n = 566\)) contains results relative to the *in vivo* micronucleus mutagenicity assay. Among the *in vivo* mutagenicity assays, the micronucleus is the test with the largest database. Standard chemical data fields were included together with biological data fields: *in vivo* micronucleus test outcomes in bone marrow cells, peripheral blood cells and splenocytes in four experimental groups (mouse, rat, male and female); *in vivo* micronucleus test outcomes in genetically modified mice (in bone marrow and peripheral blood cells); details such as strain, route of administration and toxicity biomarker (target cell toxicity). Sources of data are Toxnet, NTP, Leadscope FDA CRADA Toxicity Database (http://www.leadscope.com/). As for ISSCAN, the data were cross-checked on the different sources of information available; contradictions were solved going back to the original articles, and results based on insufficient protocols were not included. Particular attention was given to the definition of negative results. Very often—especially in old literature—negative calls are reported without any evidence that the chemicals actually interacted with the target cells (e.g. bone marrow cells). In the preparation of ISSMIC, the target cells toxicity was given the highest priority as demonstration of interaction with target cells, and for each experiment the information was included in the database. Thus, ISSMIC results are stratified into the categories of (i) positives (\(n = 190\)), (ii) borderline results (\(n = 48\)), (iii) negatives with clear evidence of cell exposure (\(n = 97\)) and (iv) inconclusives (\(n = 231\)). Inconclusive results include (i) chemicals that neither induce micronuclei nor target cells toxicity and (ii) chemicals that do not induce micronuclei and for which target cell toxicity information is not available in the literature (16).

The ISSSTY database on *S. typhimurium* (Ames test) mutagenicity (version v1a, \(n = 7367\)) is relative to the *in vitro* assay considered as the most reliable indicator of DNA reactive chemicals (17). Its very large database (several thousands of chemicals, often with many repetitions) posed new problems to the curation of data, and did not permit the scrutiny of the individual results; thus, a different strategy was adopted. The data were downloaded automatically from the CCRIS database in the Toxnet website. For chemicals with repeated experiments, the number of reported positive and negative studies in each strain were counted, and the call was assigned when at least 60% of results were concordant. An overall outcome was assigned as well. A chemical was considered to be positive when at least one positive study (with or without metabolic activation) was reported (regardless of the strain). The decision on the negative overall calls was more complicated: because different *Salmonella* strains are specific for different types of genetic damage, it was required that negative results were reported for at least two strains with different specificities. Thus, a negative is defined when two conditions are fulfilled: (i) no positive or equivocal results are present in any strain and (ii) negative outcomes exist for at least one strain among TA1535 or TA100 or TA97 (with and without metabolic activation) and at least one strain among TA1538 or TA98 or TA1537 (with and without metabolic activation).

The ISSCTA database (version v1a, \(n = 370\)) provides data on chemicals tested with the *in vitro* cell transformation assay. The data refer to results obtained with different experimental systems: (i) Syrian hamster embryo cells, \(pH = 6.7\), (ii) Syrian hamster embryo cells, \(pH \geq 7\), (iii) a cell line derived from embryos of BALB/c mice, (iv) a cell line derived from embryos of C3H mice. (v) Bhas 42 cell line (established from the BALB/c 3T3 cells through the transfection with a plasmid pBR322 containing Ha-MuSV-DNA, clone H1 [v-Ha-ras]) (18). The selection of results presented no major problems because it was possible to build on two previous, curated compilations of data (19,20).

The ISSBIOC database (version v2a, \(n = 299\)) contains information on 110 biocides with a product-type number according to Annex II of the EU Commission Regulation (EC) 1451/2007 and further 189 chemicals from other sources. The toxicological information consists of rodent bioassay and *in vitro* mutagenicity in *S. typhimurium* (Ames test). Standard chemical data fields together with biological data are included: carcinogenicity results in the four experimental groups most commonly used for the cancer bioassay, overall carcinogenicity, overall mutagenicity *in vitro* in *S. typhimurium* (Ames test) from the ISSSTY_v1a database, the source of carcinogenicity data. The main primary sources of carcinogenicity data are Aggregated Computational Toxicology Resource (ACToR), the Communication and Information Resource Centre for Administrations (CIRCA), CPDB, Fluoride Action Network (FAN) Pesticide Project, IARC database, ISSSCAN_v4a, NTP, Toxnet database (CCRIS from the cluster of toxicological databases Toxnet). Information on the product type (PT) according to the Biocides Directive 98/8/EC (concerning the placing of biocidal products on the market) or the specific type of use for the chemical (e.g. insecticide, herbicide, fungicide, acaricide, etc.) are included.

### Exploiting ISSTOX

The collection of data, their critical review in terms of quality of the biological results, together with the implementation into organised databases that exploit the modern CRD technology, permit a wide range of operations and analyses that would be impossible or very difficult otherwise. Several examples can be quoted.

#### Support for QSAR models

A characterising feature of the CRD databases is that the chemical structures are coded in such a way as to be directly
usable for SARs studies. The use of experimental data for SAR studies amplifies their informative value (e.g. through toxicity predictions) and contributes to the reduction and replacement of animal experimentation. In our laboratory, the ISSCAN database has been used to build QSARs for the mutagenicity and carcinogenicity of the aromatic amines. Because of its environmental and industrial importance, the aromatic amines are the single chemical class most studied for their ability to induce mutations and cancer. The large database of mutagenicity and carcinogenicity results has been analysed with QSAR approaches, leading to models that permit the predictions of mutagenic and carcinogenic potencies, as well as the discrimination between mutagenic/non-mutagenic and carcinogenic/non-carcinogenic amines (21-23).

Within a more general perspective, the ISSCAN database has been used within a systematic survey of the value and predictivity of QSAR models available in the public domain for mutagenicity and carcinogenicity. A range of good-quality QSARs have been selected from the open literature, and then challenged to predict real external test sets (i.e. chemicals never considered by the authors that have developed the models). Such external test sets were retrieved from ISSCAN, and predictions were computed with the selected models. It appears that the QSARs for potency (applicable only to toxic chemicals) generated predictions 30-70% correct, whereas the QSARs for discriminating between active and inactive chemicals were 70-100% correct in their external predictions: thus, the latter can be used with good reliability for applicative purposes (24). More recently, ISSCAN has been used to assess the performance of the OncoLogic expert system for predicting chemical carcinogenicity. A number of compounds tested with the long-term carcinogenicity bioassay on rodents were extracted from the ISSCAN database and then subjected to OncoLogic. The concordance between OncoLogic predictions and the bioassay results was remarkably high (86% concordance in the subset of chemicals to which Oncologic can be applied) (25).

Even though the availability of data in a QSAR-ready format drastically facilitates QSAR modelling, it should be emphasised that mechanistic aspects and QSAR know-how are of central importance when generating and assessing the predictivity of a QSAR model. These aspects tend to be neglected with the advent of the increased number of computer programs that can produce hundreds of parameters per compound automatically. Clearly, such a situation calls for automatic procedures to arrive at QSAR models, and this in turn leads to an overestimation of the value of the statistical assessment. Predictions obtained in this way can be very misleading. As Corwin Hansch pointed out, ‘there is no substitute for extensive experience … in physical organic chemistry and QSAR’ (5). This holds true for the development of a QSAR model, its validation and its application to predict the behaviour of new substances. Another aspect to be considered is that because the role of QSAR as a regulatory tool is undoubtedly going to expand in the near future, models and tools that ensure transparency, consistency and acceptability of the estimation methods will be needed. For this reason, the Organization for Economic Cooperation and Development (OECD) has drafted principles for QSAR validation that provide the best currently available harmonised framework to assist in judging the suitability of approaches for regulatory use (26).

Generating structural alerts for predictive toxicology

A very active field in predictive toxicology is the research on structural alerts (SA) for toxic effects, i.e. particular reactive groups or substructures in the molecule responsible for the toxic effects. The recognition of the SAs is still at the basis of the mechanistic science of toxicology and provides powerful means of intervention to generate safer chemicals. Knowledge on action mechanisms, as exemplified by the SAs, is routinely used in QSARs assessment in a regulatory context, and is at the basis of popular commercial and non-commercial software systems. Fundamental work in this field has been performed by John Ashby following the electrophilicity theory of carcinogenicity of the Millers (27). In more recent times, the mechanistic knowledge on chemical carcinogens has been implemented into computerised expert systems that permit faster and more flexible assessment of chemicals. In our laboratory, ISSCAN has been used as a basis for developing an updated compilation of SAs for carcinogenicity and bacterial mutagenicity (9,28). The updated SAs list has been implemented into computerised rules for the expert system Toxtree. Toxtree (http://ecb.jrc.ec.europa.eu/qsar/qsar-tools/index.php?c=TOXTREE) is an open-source, freely available software application that places chemicals into categories and predicts various kinds of toxic effect by applying various decision tree approaches. Toxtree was developed by IdeaConsult Ltd (Sofia, Bulgaria) under the terms of an ex-European Chemicals Bureau (ECB) contract (29). Recently, the carcinogenicity/mutagenicity rules have been implemented into the OECD (Q)SAR Toolbox as well (see below for more details on the Toolbox).

Another study, SAs for the induction of micronuclei in vivo (rodents) has been compiled. In vivo genotoxicity studies—shortly followed by carcinogenicity—are posing high demand for test-related rescourses in terms of animal lives and resources. Among them, the micronucleus test in rodents is the most widely used as a follow up to positive in vivo mutagenicity results: therefore, the development and extensive use of estimation techniques based on the concept of SAs might have a huge saving potential for this endpoint. To this aim, the ISSMIC database on in vivo micronucleus test results was used as a basis for compiling a new set of SAs, thus making available a tool for preliminary screening of potential in vivo mutagens (16,30).

Analysis of risk assessment strategies

A field that gets great benefits from the use of well-structured toxicological databases is that of strategies for the risk assessment of chemicals. The CRD databases not only provide the necessary amount of experimental evidence relative to the different test methods but also permit an efficient comparison and combination of test results by unequivocally indexing the chemicals through their structure. Thus, different hypotheses and schemes can be easily analysed and compared. As a matter of fact, after several decades of investigations, the need for tools able to predict chemical carcinogens in shorter times, and at a lower cost in terms of animal lives and money is still a research priority. Still unsolved issues in the present testing strategies and regulations are (i) lack of alternative assays able to identify non-genotoxic carcinogens, (ii) exaggerated rate of misleading (‘false’) positive results of the in vitro mammalian cells—based mutagenicity short-term tests and (iii) extremely low sensitivity of in vivo mutagenicity short-term tests. As a consequence, the present strategies for identifying carcinogens
without the use of the rodent bioassay are not an efficient filter for protecting human health. In our laboratory, the various ISSTOX databases were combined with other public and commercial databases and several analyses were performed. It was concluded that the combination of *Salmonella* and SAs for the DNA-reactive carcinogens and *in vitro* cell transformation assays for non-genotoxic carcinogens permits the identification of a very large proportion of carcinogens without the use of, and in much shorter time than with animal experiments (18,31). Another conclusion was that the *in vivo* mutagenicity assays, in their present form, do not have added value, and rather impair the prediction ability of the *in vitro* tests alone (16).

**Support for predictive toxicology applications**

As repositories of data, the databases are invaluable tools for the human experts involved in risk assessment. When implemented into modern computing tools, they support tasks much more complex and sophisticated than simple interrogations. In particular, they can be integrated within applications that, through the combination of vast amounts of organised data and predictive tools, permit a variety of operations. A typical example is the OECD (Q)SAR Application Toolbox; this is a standalone, free software application intended to be used by governments, the chemical industry and the toxicologists to fill gaps in toxicity data needed for assessing the hazards posed by chemicals (http://www.oecd.org/document/54/0,3343,en_2649_34379_42923638_1_1_1,00.html). Documents on the use of the QSAR Toolbox and of the underlying rationale are available at: http://www.oecd.org/document/54/0,3343,en_2649_34379_42923638_1_1_1,00.html#Guidance_Documents_and_Training_Materials_for_Using_the_Toolbox.

The Toolbox contains databases with results from experimental studies, a library of QSAR models, tools to estimate missing experimental values by read-across (i.e. extrapolating results from tested chemicals to untested chemicals), tools to define categories of chemicals to be regulated or considered in a similar way from the point of view of risk and tools to estimate missing experimental values by trend analysis (i.e. interpolating or extrapolating from a trend in results for tested chemicals to untested chemicals within a category).

The Toolbox is able to quickly evaluate chemicals for common mechanisms or modes of action as well as for common toxicological behaviour or consistent trends among results related to regulatory endpoints. This is accomplished through a large collection of ‘profilers’ (e.g. lists of SAs) for the various toxicological endpoints. The profiling step is followed by the formation of a category of chemicals ‘similar’ to the query one. A chemical category is a group of chemicals with physical-chemical and toxicological properties that are likely to be similar or follow a regular pattern because of their similar chemical structure. Practically, the formation of a category is performed by retrieving in the Toolbox databases chemicals with a SAs pattern similar to that of the query chemical. Using this category approach, not every chemical needs to be tested for every endpoint because data gaps may be filled by read-across from a tested chemical to an untested chemical, trend analysis (interpolation or extrapolation) or related QSAR methods. The category approach used in the Toolbox enables robust hazard assessment through mechanistic comparisons without testing.

As the definition of a chemical category for data gap filling is the critical step in the workflow of the Toolbox, the selection and use of the profilers and databases are crucial. The Toolbox provides several options to assist the user in defining and refining the category definition. The current version of the software includes the ISSTOX databases, as well as several profilers generated and validated by the ISS on the basis of the ISSTOX data. Thus, the use of CRD databases is a key element for the Toolbox application. **Towards a toxicological ontology**

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, often heterogeneous and annotated to different terminology, 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). Database standardisation and CRD accessibility are a high priority, and the collection of new data should preferably be planned accordingly at the same moment the experimental design is laid out.

A critical step in designing databases and their full exploitation is the creation of a standardised controlled vocabulary (ontology). From its original philosophical roots, the word and idea of ontology has been adopted by the sciences as a formal representation of a set of concepts within a knowledge domain and the relationships between those concepts. It is used to reason about the properties of that domain and may be used to define the domain. Thus, an ontology is a ‘formal, explicit specification of a shared conceptualisation’. An ontology provides a shared controlled vocabulary, enriched by relationship between terms, which can be used to model a domain—i.e. the type of objects and/or concepts that exist, and their properties and relations. The controlled vocabulary is a collection of preferred terms that are used to assist in more precise retrieval of content. Controlled vocabulary terms can be used for categorising content, building labelling systems and creating style guides and database schema. A well-known example of controlled vocabulary is a taxonomy. Ontology-based informatics approaches provide very powerful tools to organise and retrieve information: Ontologies are used in artificial intelligence, the Semantic Web, software engineering, biomedical informatics, library science and information architecture as a form of knowledge representation about the world or some part of it (32).

The application of ontology in predictive toxicology is a relatively recent endeavour in which investigators are developing or applying ontologies towards the solution of toxicology problems. Use of ontology allows data to be organised and combined with metadata and vocabularies for study terms and experimental protocols. Communication between toxicology resources using shared ontology supports toxicology data integration in a more reliable way, and to then be able to construct more complex queries over the data such as linking data with biological processes (33). To implement the ontologies, the DL species of the Web Ontology Language (OWL DL) (http://www.w3.org/TR/owl-features/), supported by the Protégé OWL editor (http://protege.stanford.edu/) is the most commonly employed technical language.

Recently, OECD and European Chemical Agency (ECHA) have started a project aimed at developing toxicological
ontologies, with the Istituto Superiore di Sanità as main contractor. This project will facilitate data gap filling for the predictive toxicology approach through a better integration and standardisation of experimental data from databases within the OECD (Q)SAR Toolbox, including ISSSTOX databases. The resulting ontology will ensure efficient potential further extensions of the databases, and will permit the harmonisation of exchange of toxicological data and predictions between the Toolbox, user input and the International Uniform Chemical Information Database (IUCLID), which is a system for managing data on intrinsic and hazard properties of chemical substances and registration within REACH (http://iuclid.eu/). An ambitious long-term goal is to create the basis for a systematic Toxicological Ontology, which—at present—is lacking. A first phase of the project was concluded in 2012 and has generated ontologies for three toxicological endpoints: carcinogenicity, repeated dose toxicity and reproductive/developmental toxicity. A new phase of the ontology project is in progress and focuses on skin and respiratory sensitisation, skin irritation and corrosion, eye irritation and corrosion (34).

Conclusions

The need for speeding up the toxicological assessment of chemicals and of using less animals and less-expensive tools has strongly stimulated the development of predictive toxicology and of structure-based approaches (35). To make the predictive toxicology methods more effective, the field needs access to the largest possible knowledge base of previous toxicology findings, which need to be interoperable and comparable. As a matter of fact, access to high-quality experimental data sources is extremely important for the successful satisfaction of regulatory demands in risk assessment as required by, e.g., REACH legislation. The use of well-structured databases could significantly improve storage, exchange and use of information from experiments already carried out is essential to avoid unnecessary repetition of experiments and to support increased access to data and associated metadata including experimental protocols and scientific concepts. The ISSSTOX databases, as shown by the examples of applications reported, fit into this general perspective. In conclusion, the development of open, public and computable toxicological datasets based on standardised controlled vocabularies will support data management, model building, integrated analysis, validation and reporting, including regulatory reporting and alternative testing submission requirements, leading to new scientific advances in a mechanistically based predictive toxicology.

Funding


References


