Structure-activity models of chemical carcinogens: state of the art, and new directions

Romualdo Benigni and Cecilia Bossa
Dipartimento di Tecnologie e Salute, Istituto Superiore di Sanità, Rome, Italy

Summary. Chemical carcinogenicity has been the target of numerous attempts to create predictive models alternative to the animal ones, ranging from short-term biological assays (*e.g.* mutagenicity tests) to theoretical models. Among the theoretical models, the application of the science of structure-activity relationships (SAR) has earned special prominence. The qualitative approach to SAR has lead to the identification of a large number of reactive chemical substructures that are both mutagenic and carcinogenic. More sophisticated developments are the quantitative structure-activity relationship (QSAR) models, that link the physical chemical or structural properties of the molecules to the toxicological endpoints. Both approaches provide strong support to the process of risk assessment of the chemicals, especially in the phase of priority setting. Among the areas potentially able to contribute to further developments of (Q)SAR, the novel chemical relational databases are presented and discussed. The freely downloadable ISSCAN database on chemical carcinogens is presented.

Key words: structure-activity relationships, QSAR, carcinogen, mutagen, database.

Riassunto *(Relazioni struttura-attività di agenti carcinogenici: stato dell'arte e prospettive)*. Il campo della cancerogenesi chimica ha visto una lunga serie di tentativi di creare modelli alternativi a quelli animali, dai saggi a breve termine di mutagenesi ai modelli teorici. Un posto particolare tra quelli teorici hanno i modelli basati sulle relazioni struttura-attività. Nella sua versione qualitativa, tale approccio ha portato alla identificazione di una varietà di sottostrutture, o gruppi funzionali legati all'induzione di mutazioni e/o cancro. Approcci più sofisticati sono le relazioni quantitative struttura-attività, che legano le proprietà tossicologiche delle molecole alle loro proprietà chimico fisiche o strutturali. Sia l'approccio qualitativo che quello quantitativo forniscono un valido contributo alla stima del rischio delle sostanze chimiche, soprattutto nella fase dell'identificazione di priorità. Tra le aree di sviluppo di tali approcci, si illustrano le novità rappresentate dalle basi di dati chimico-relazionali, e si presenta la base di dati sui cancerogeni chimici ISSCAN, liberamente disponibile sul sito Internet dell'Istituto Superiore di Sanità.

Parole chiave: relazioni struttura-attività, QSAR, cancerogeni, mutageni, banche dati.

INTRODUCTION

Toxicology faces the parallel tasks of performing safety evaluations that support the development of new chemicals before human exposures are permitted, and assessing the potential hazard posed by exposures to chemicals that lack safety evaluations. The accelerating pace of chemical discovery and synthesis has heightened such a need for efficient prioritization and toxicity screening methods. Chemical carcinogenicity has been the target of numerous attempts to create alternative predictive models, ranging from short-term biological assays (*e.g.* mutagenicity tests) to theoretical models. Among the theoretical models, the application of structure-activity relationships (SAR) concepts has earned special prominence. SAR uses the wide and well established body of knowledge of chemistry to rationalize the chemical-life interactions and to "domesticate" the chemicals, to make them safer and more efficient [1-3]. In addition, as a result of recent policy developments in different countries it is expected that the use of SAR for regulatory purposes will sharply increase. For example, the European Union adopted a legislative proposal for a new chemical management system called REACH (registration, evaluation and authorisation of chemicals), which is intended to harmonise the information requirements applied to new and existing chemicals [4]. REACH provides for the use of SAR for predicting the environmental and toxicological properties of chemicals, in the interests of time-effectiveness, cost-effectiveness and animal welfare. The development of SARs for human health endpoints will also contribute to meeting the needs of the seventh amendment to the EU cosmetics directive [5]. According to a recent assessment by the European Chemicals Bureau (ECB), approximately 3.9 million additional vertebrate test animals could be used as a consequence of the implementation of REACH, if alternative methods are not accepted by regulatory authorities and adopted by industry [6]. However, a considerable reduction in animal

Indirizzo per la corrispondenza (Address for correspondence): Romualdo Benigni, Dipartimento di Tecnologie e Salute, Istituto Superiore di Sanità, Viale Regina Elena 299, 00161 Rome, Italy. E-mail: rbenigni@iss.it.

use could be obtained if alternatives were applied more extensively: acceptance of theoretical models based on SAR and related techniques would lead to a saving of 1.3 to 1.9 million test animals. Another example comes from the ongoing program in Canada on the existing chemicals. Each of the more than 23 000 substances on the domestic substances list, an inventory of chemicals and biological agents that were in commerce in Canada between January 1984 and December 1986, must be "categorized" by September, 2006. The purpose of categorization is to determine which substances on the list may have the greatest potential for exposure to the population, or are persistent or bioaccumulative and "inherently toxic" to human beings or non-human organisms. Given the paucity of experimental data available for the large number of substances in the list, this task necessitates increased reliance on alternative methods, in at least the early stages, to set priorities for additional testing or for further assessment. More information on the approaches used by the Canadian Authorities can be found at the following web sites: www.hc-sc.gc.ca/ewh-semt/contaminants/existsub/index_e.html and www.ec.gc.ca/substances/ese.

CHEMICAL REACTIVITY AND MECHANISMS OF ACTION

The overlap between mutagenicity and carcinogenicity is quite large, since the first step of the carcinogenic process often consists of one or more mutations (somatic mutation theory of cancer). A unifying approach to the rationalization of this area has been that chemical mutagens and carcinogens act by attacking the DNA, based on their electrophilicity *per se* or after metabolic transformation [7]. This concept, together with experimental observation, has led to the identification of several chemical functional groups and substructures (structural alerts, SAs) that can cause both mutation and cancer; these include carbonium ions (alkyl-, aryl-, benzylic-), nitrenium ions, epoxides and oxonium ions, aldehydes, polarized double bonds (α,β-unsaturated carbonyls or carboxylates), peroxides, free radicals and acylating intermediates [8, 9]. The identification of the SAs has been a very important scientific advancement, since it has provided means to design safer compounds by avoiding the known SAs. In addition, the SAs for mutagenicity and carcinogenicity have been incorporated into expert systems for predicting toxicological effects of chemicals (*e.g.* DEREK, OncoLogic) [10]. However, some classes of carcinogens (called epigenetic carcinogens) do not act through genotoxic mechanisms; for these classes, the progress in the identification of SAs has been much slower [11].

QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIPS

Whereas the SAs define the potential for the chemicals to be carcinogenic or mutagenic, the actual modulation of this potential depends on a series of factors (*e.g.*, molecular weight, physical state, solubility, degree of chemical reactivity, etc.) which vary within each individual class of compounds. In fact, chemicals sharing the same SA (*e.g.* aromatic amine) can behave in different ways: some are active (to different extents) and some are inactive. Such modulating factors can be approximated by minor, context-dependent SAs (*e.g.*, the substituents in the different positions of the skeleton of a chemical with a primary SA.

However, the qualitative approach based on the recognition of SAs has clear limitations; for example, it is not possible to predict the activity of substructures which are not in the list of known SAs, or which have a combination of SAs. A powerful generalization is provided by the quantitative structure-activity relationship (QSAR) analysis, which is based on a limited number of physical chemical properties with general relevance, and produces a mathematical model of the chemical determinants of the biological activity. The physical chemical properties of interest for the biological activity of the chemicals are hydrophobic, electronic and steric effects [3, 12].

QSARs have been generated for a number of individual chemical classes (including aromatic amines, nitroarenes, quinolines, triazenes, polycyclic aromatic hydrocarbons, lactones) [13, 14]. The majority are relative to *in vitro* mutagenicity; however, a number of QSAR models for the animal carcinogenicity exist as well. A great aspect of the QSARs for the individual chemical classes, as performed according to the classical Hansch approach, is that they point to the physical chemical determinants of the biological activity of the compounds; thus they have considerably contributed to the understanding of the mechanisms of chemical mutagenicity and carcinogenicity.

The other important goal of the use of QSAR analyses is the risk assessment of chemicals: once formulated, the QSARs can be employed for estimating the activity of other chemicals not tested experimentally. It should be remarked that, before using the QSAR models for prediction purposes, their validity should be assessed very carefully. Whereas several statistical approaches and criteria (*e.g.*, leave-one-out, leave-many-out, shuffling, leverage, etc.) [15] can be used to assess the internal statistical goodness of a QSAR model, the most stringent criterion is to challenge the model to predict the activity of a number of chemicals (external test set) of the same chemical class, not used for the generation of the model itself. Obviously, the activity of the chemicals in the test set has to be known. In our hands, carefully generated, mechanistically based QSAR models have shown a very high predictive ability (more than 90% correct predictions of external data sets) [16, 17].

GENERAL PREDICTION MODELS FOR NONCONGENERIC DATA SETS

Unfortunately, each QSAR model is specific for one individual chemical class and the database of experimental results is not enough populated of representative

chemicals to provide a basis for modeling the carcinogenicity or mutagenicity of each chemical class of interest; in addition, the chemicals of interest change with the time: *e.g.* new drugs and new dyes are put in commerce continuously, so that similar chemicals, already tested for carcinogenicity and from which to derive QSAR models, are not available. The response to such a situation is represented by a series of attempts to develop SARs and QSARs for noncongeneric sets of chemicals, *i.e.* "general" prediction models in order to hopefully cope with the thousands of chemicals present in the environment. Several approaches have been made commercially available as well [10, 14, 18-20]. It should be remarked that these general prediction models are quite different in nature [10, 14, 18-20] from the classical QSARs; they range from human experts panels that inspect the chemical structures "by eye" and make inferences based on the activity of similar, previously tested chemicals, to computerized expert systems, to quantitative approaches devised specifically for sets of noncongeneric chemicals. On the contrary the classical QSAR approaches are specific for congeneric chemicals, *i.e.* sets of chemicals belonging to the same chemical class, and acting through the same mechanism of action.

The general prediction systems usually do not contribute much new mechanistic knowledge, but have to be judged for their ability to provide correct predictions of untested chemicals. A stringent criterion is prospective prediction: the predictions are performed on compounds whose experimental results were not existing, or not known when the model was built.

To evaluate the validity of the prediction approaches for noncongeneric chemicals, important prospective prediction exercises were performed in the past decade under the aegis of the US National Toxicology Program (NTP) [21, 22]. The exercises invited the modeling community to submit predictions on the toxicity of chemicals that were in the process of being bioassayed by the NTP; at the end of the experimentation, the actual results were compared with the theoretical predictions. Because of their unbiased character, they represent the most important source of information on the ability of the prediction systems to deal with "real" chemicals. Details and summaries on the NTP prediction exercises are presented in [14].

A NTP comparative exercise regarded the prediction of the mutagenicity in *Salmonella typhimurium* of 100 chemicals of different chemical classes. Comparable performance was attained by: a) a human expert (John Ashby) that inspected the chemical structures and made his predictions based on the presence of SAs in the structures; and b) two computerized systems (Topkat, Case) (74–76% overall correct prediction) [23].

Another NTP comparative exercise regarded the prediction of the rodent carcinogenicity of 44 compounds of different chemical classes [21]. A detailed analysis of the results is in [24]. For approaches that relied solely on information derived from the chemical structure, the overall accuracy was in the range 50-65%, whereas the human experts Tennant and Ashby [21] – that combined the inspection of SAs with the knowledge of short term mutagenicity tests or subchronic bioassay results – attained 75% accuracy.

A second comparative exercise on the prediction of rodent carcinogenicity of 30 chemicals in the progress of being bioassayed was performed by NTP [22]; an analysis of the results of this exercise is in [25]. The highest overall accuracy in this second exercise was 65%, and was attained by human experts (the OncoLogic team) that inspected the chemical structures and reasoned in terms of chemical analogy with known carcinogens and noncarcinogens. The experimental Syrian hamster embryo cells transformation assay had a similar overall performance.

Together with the above comparative prediction exercises, a number of other external validation exercises have challenged the most popular commercial software programs (DEREK, Multicase, Topkat). A common finding of these studies is that the performance in prediction varied very much according to the subset of chemicals to be predicted: this indicates that the level of uncertainty of the predictions is quite high, and is not known in advance [14, 26].

VALUE OF THE PREDICTION APPROACHES

Overall, it can be concluded that predictions for the individual chemicals cannot be taken at face value and cannot replace the experiments, when necessary. Their main role is to complement the information of different nature and from different sources. At the same time, the structure-activity-based methods can have a great role, *e.g.*, in priority setting of large numbers of chemicals. This has been brilliantly demonstrated by the selection process for chemicals to be bioassayed by the NTP: this selection was operated by human experts largely based on the recognition of SAs. As a matter of fact, the proportion of carcinogens among the "suspect" chemicals was almost ten times higher than that relative to the chemicals selected only on production/exposure considerations [27]. Another positive evidence is that the presently available knowledge permits the identification of the genotoxic, transpecies-transex-carcinogens more efficiently than that of the nongenotoxic ones [28]; the former are likely to be the most harmful for the human health.

It should be emphasized that the different (Q)SAR approaches vary largely in terms of reliability. First are the mechanistically-based QSARs for individual chemical classes; as shown by our experience reported above, they can have a very high predictive ability. These methods apply to congeneric sets of compounds. Among the "general" approaches for noncongeneric chemicals, the best results have been obtained by panels of human experts, that inspect the chemical structures and apply subjective criteria of "similarity" with known carcinogens and noncarcinogens. On the other hand the commercially available "all purpose" prediction software systems can be a useful support for the expert judgment as well, provided that they offer "transparent" predictions and not only black box responses. Transparency includes declaring the SAs

and the rules used for formulating the predictions, as well as the list of chemicals – with known activity – similar to the query chemical (possibly together with a similarity measure).

BUILDING IMPROVED (Q)SAR

In the recent years, an aspect that has earned special attention is that of the generation of databases to be used for the development of (Q)SAR models for mutagenicity and carcinogenicity models. In fact, the generation of better predictions is strictly related to the accessibility of high quality experimental data for the scientific community. This is crucial, for example, for retrieving enough representative chemicals of individual chemical classes, in order to generate QSARs. This is also crucial to the process of human expert judgement, which necessitates: a) access to all the relevant literature; and, b) capability to search across toxicity databases using both biological and chemical criteria [9].

In this field, a rapid progress has taken place, both in terms of initiatives and of technological innovation.

DATABASES OF CHEMICAL MUTAGENS AND CARCINOGENS

Among the sources of freely available data 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 web site, it is possible to search across and within the databases by several identifiers, such as Chemical Name, CAS Registry 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 format. GENE-TOX 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 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 in the web is the Carcinogenic Potency Database (CPDB) (http://potency.berkeley.edu/cpdb.html). This database collects the results from 6153 chronic, long-term animal cancer tests on 1485 chemicals, that have been 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.

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 tests, 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). To access the data, stored 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. From the International Agency for Research on Cancer (IARC) Web site it is possible to access the IARC Monographs on the Evaluation of Carcinogenic Risks to Humans (http://monographs.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.

Another source of carcinogenicity data is the Survey of Compounds Which Have Been Tested for Carcinogenic Activity (CD-ROM Version 4.0, GMA Industries Inc.). This collection contains data extracted from experimental carcinogenicity research results, published in the literature between 2000 and 1934 (and earlier) for over 10 000 chemicals. Principal chemical identifiers for searching through these data are the chemical name and CAS number.

Recently, a very useful tool to browse through the different toxicology 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, 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. The PubChem interfaces provide extensive query capabilities on textual and numeric information, as well as a comprehensive set of structure-based query methodologies.

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 do not contain chemical structures or consider potential SAR uses of the data. These databases typically utilize chemical names (usu-

```
  10   9   0            0                    1   02000
        0.3600        0.8960        0.0000   C    0    0    0    0    0    0    0    0    0    0    0    0
       -0.8880        1.3840        0.0000   C    0    0    0    0    0    0    0    0    0    0    0    0
        0.7160       -0.4000        0.0000   C    0    0    0    0    0    0    0    0    0    0    0    0
       -0.1080       -1.4600        0.0000   C    0    0    0    0    0    0    0    0    0    0    0    0
        1.1800        1.6400        0.0000   H    0    0    0    0    0    0    0    0    0    0    0    0
       -1.7920        0.7560        0.0000   H    0    0    0    0    0    0    0    0    0    0    0    0
       -1.0560        2.4760        0.0000   H    0    0    0    0    0    0    0    0    0    0    0    0
        1.7960       -0.6200        0.0000   H    0    0    0    0    0    0    0    0    0    0    0    0
        0.3040       -2.4800        0.0000   H    0    0    0    0    0    0    0    0    0    0    0    0
       -1.2040       -1.3760        0.0000   H    0    0    0    0    0    0    0    0    0    0    0    0
  4   10   1   0   0   0   0
  4    9   1   0   0   0   0
  9    8   1   0   0   0   0
  3    4   2   0   0   0   0
  2    7   1   0   0   0   0
  2    6   1   0   0   0   0
  1    5   1   0   0   0   0
  1    3   1   0   0   0   0
  1    2   2   0   0   0   0
M    END
>    <Substance ID>  (41)
41

>    <Mouse_Female_Canc>  (41)
3

>    <SAL>  (41)
3

>    <Rat_Male_Canc>  (41)
3
```

2D Structure

Text and data fields

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

ally common or commercial names) and Chemical Abstract Service (CAS) registry numbers. These types of chemical identifiers are non-unique, prone to transcription and formatting errors, and devoid of any chemical information. On the contrary, 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 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 [29].

Thus, a crucial point is that of collecting and standardizing portions of the existent knowledge, in a way that allow: 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. Most commercially available CRD applications provide substructures and functional groups search features, different algorithms for searching compounds chemically similar to query ones (similarity search), and text and data field search functions.

In order to be accessed with a CRD application, the information have 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. SDFiles 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 by a first part, where the 2D or 3D structure of the molecule is represented as MOLfile format, and a second part with numerical or text data field *(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.

THE DSSTOX DATABASE CLUSTER

A most remarkable example of database designed according to the novel criteria is the distributed Structure-Searchable Toxicity (DSSTox) Database

Network, which is a project of the US Environmental Protection Agency's Computational Toxicology Program (www.epa.gov/comptox). A primary objective of the DSSTox web site (www.epa.gov/nheerl/dsstox) is to serve as a central community forum for publishing standard-format, structure-annotated chemical toxicity data files for open-access, public use. In this initial phase, data files cannot be structure-searched on the DSSTox web site itself, but the data files can be downloaded in their entirety and freely used.

At present, the DSSTox cluster includes five separate databases: CPDBAS – Carcinogenic Potency Project Summary Tables (Source, L.S. Gold, Carcinogenic Potency Project, UC Berkeley); DBPCAN – EPA Disinfection By-products Carcinogenicity Estimates Database (Source, Y.T. Woo, US EPA, Office of Pollution Prevention & Toxics); EPAFHM – EPA Fathead Minnow Acute Toxicity Database (Source, C. Russom, US EPA, 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).

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 web site provides a detailed guide on the use of files, and a rich documentation on the entire subject of databases and related concepts [30, 31].

THE ISSCAN DATABASE ON CHEMICAL CARCINOGENS

To provide the experts with a decision support tool, and to provide the scientific community with good quality data for modeling purposes, at the Istituto Superiore di Sanità a CRD database on chemical carcinogens was built. The database is called ISSCAN: "Chemical carcinogens: structures and experimental data", and is located on the web site of the Institute. The data can be freely downloaded from the address: http://progetti.iss.it/ampp/hhhh/hhhh.php?id=233,

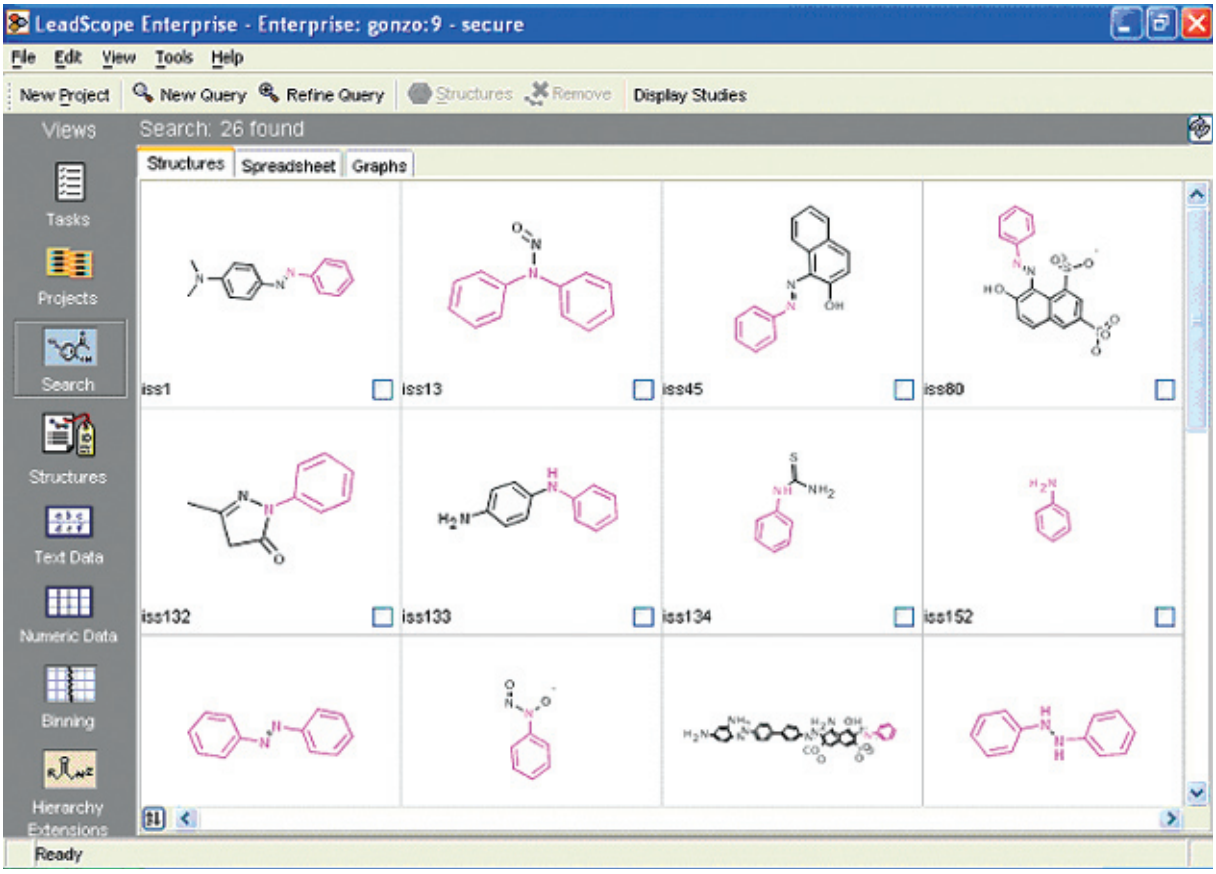

Fig. 2 | *Example of substructure searching with Leadscope. The chemicals including aniline as a substructure are pointed out by the search capability.*

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 data were cross-checked on different sources of information available.

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.

From the web site 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.

The specific characteristics of the ISSCAN database in respect to other databases should be emphasized. The primary goal was twofold: a) providing the scientific and regulatory community with carcinogenicity calls that have been re-checked one by one, in order to ensure the quality of the data; b) coding the biological data (carcinogenicity and Salmonella mutagenicity) 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.

MORE ON THE SEARCHING CAPABILITY OF CRD DATABASES

Together with simple searches on the XLS (Excel-readable) file, more complex searches can be performed on the SDF file by using specialized CRD software programs, such as *e.g.*, Leadscope (Leadscope Inc., Columbus, OH) and ChemFolder (ACD/ChemFolder, Toronto).

When the SDFile is imported into a CRD application, it is possible to do structure/text/data relational searching across records in the database. In *Figure 2*, as an example, the substructure searching results by the program Leadscope, using aniline as a query structure, are depicted. The result of the search consists of all

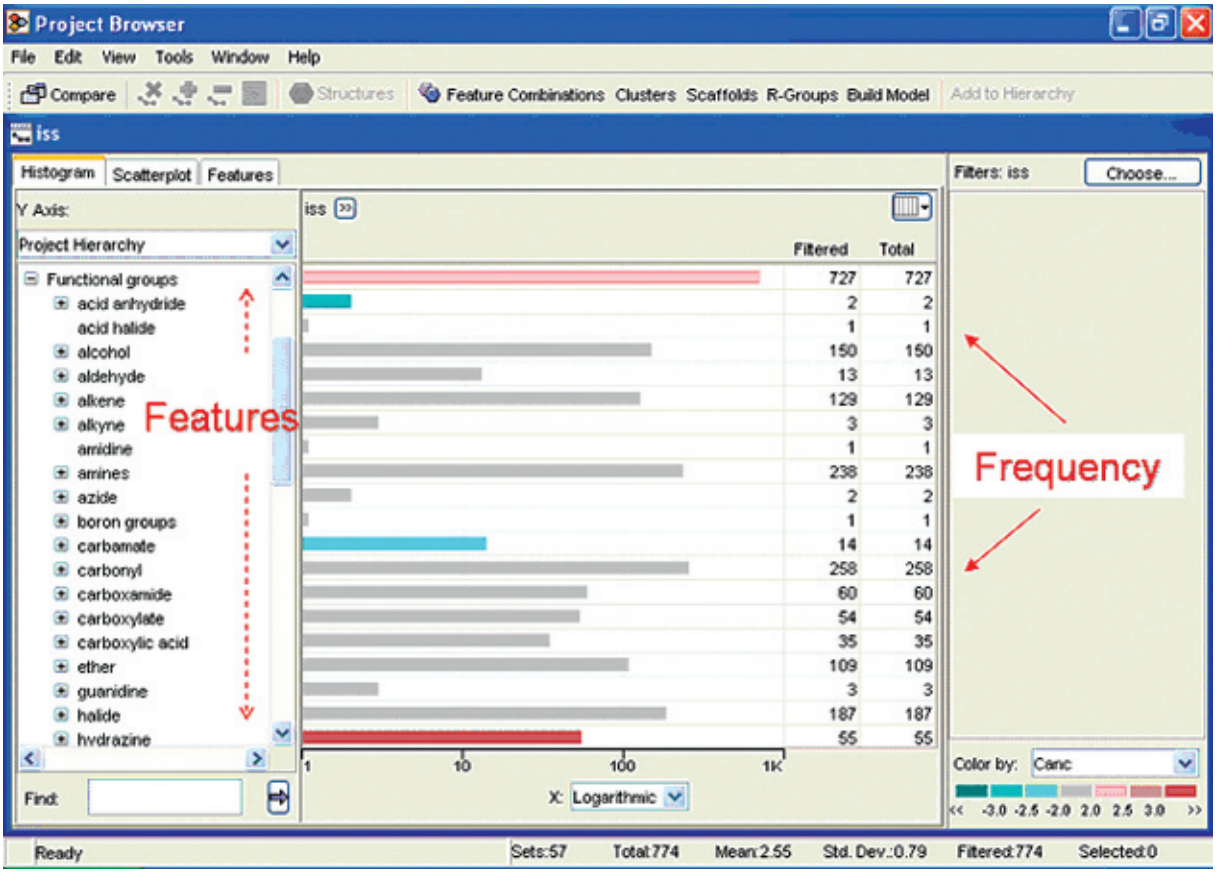

Fig. 3 | *Example of classification of the chemicals in a database by chemical classes. The classification capability of Leadscope has been used.*

chemicals in the database containing aniline as basic substructure.

Another very useful feature for (Q)SAR purposes is the possibility of searching through the database by functional group or chemical classes. An example of this capability is presented in *Figure 3*, using Leadscope as a CRD application to read the SDFile. The figure shows that the chemicals in the database are divided into chemical classes, and the frequency in each class is given. Moreover, 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 (here carcinogenicity).

CONCLUSIONS

The evidence summarized in this work points to both limitations and potentiality of the (Q)SAR approaches. At the level of large numbers of chemicals (*e.g.*, priority setting), the careful use of (Q)SAR provides a very effective support; it also contributes to the mechanistic understanding of chemical – life interactions. Within this context, there is much space for further research and improvements. In this paper, we have illustrated the potentiality coming from the recent developments in databases. We conclude with the reflections made by Rainer Franke regarding the role of (Q)SAR in drug research: "As the drug discovery process is of a very complex nature, effective drug design requires an entire spectrum of techniques in which QSAR methods still play an important role. (…) The real power of (Q)SAR methods is to extract and synthesize information from data to obtain hypotheses that can be put to experimental test. No dramatic overnight discoveries of wonder drug will result, but an increase in the chance of success due to indications of promising directions is a realistic expectation" [12]. The above concept applies to the use of (Q)SAR in predictive toxicology as well.

Submitted on invitation.
Accepted on 10 April 2006.

References

1. Ariens EJ. Domestication of chemistry by design of safer chemicals: structure-activity relationships. *Drug Metab Rev* 1984;15: 425-504.
2. Hansch C. Quantitative structure-activity relationships and the Unnamed science. *Acc Chem Res* 1993;26:147-53.
3. Hansch C, Leo A. Exploring QSAR. 1. *Fundamentals and applications in chemistry and biology.* Washington, DC: American Chemical Society; 1995.
4. European Commission. *Proposal concerning the registration, evaluation, authorization and restriction of chemicals (REACH).* COM(2003)644Final. 2003.
5. European Parliament and of the Council. Directive 2003/15/EC of the European Parliament and of the Council of 27 February 2003 amending Council Directive 76/768/EEC on the approximation of the laws of the Member States relating to cosmetic products. *Off J Eur Union* 2003;L66:26-35.
6. Van der Jagt K, Munn S, 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: European Commission Joint Research Centre; 2004. (JRC Report EUR 21405 EN).
7. Miller JA, Miller EC. Ultimate chemical carcinogens as reactive mutagenic electrophiles. In: Hiatt HH, Watson JD, Winsten JA (Ed.). *Origins of human cancer.* Cold Spring Harbor: Cold Spring Harbor Laboratory; 1977. p. 605-28.
8. Ashby J. Fundamental structural alerts to potential carcinogenicity or noncarcinogenicity. *Environ Mutagen* 1995;7:919-21.
9. Woo YT, Lai DY, McLain JL, Ko Manibusan M, Dellarco V. Use of mechanism-based structure-activity relationships analysis in carcinogenic potential ranking for drinking water disinfection by-products. *Environ Health Perspect* 2002;110(Suppl. 1):75-87.
10. Benigni R, Richard AM. Quantitative structure-based modeling applied to characterization and prediction of chemical toxicity. *Methods* 1998;14:264-76.
11. Woo YT. Mechanisms of action of chemical carcinogens, and their role in Structure-Activity Relationships (SAR) analysis and risk assessment. In: Benigni R (Ed.). *Quantitative Structure-Activity Relationship (QSAR) models of mutagens and carcinogens.* Boca Raton: CRC Press; 2003. p. 41-80.
12. 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.
13. Hansch C. Structure-activity relationships of chemical mutagens and carcinogens. *Sci Tot Environ* 1991;109/110:17-29.
14. Benigni R. Structure-activity relationship studies of chemical mutagens and carcinogens: mechanistic investigations and prediction approaches. *Chem Revs* 2005;105:1767-800.
15. Eriksson L, Jaworska JS, Worth AP, Cronin MTD, McDowell RM, Gramatica P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ Health Perspect* 2003;111: 1361-75.
16. Benigni R, Andreoli C, Conti L, Tafani P, Cotta-Ramusino M, Carere A, Crebelli R. Quantitave structure-activity relationship models correctly predict the toxic and aneuploidizing properties of six halogenated methanes in *Aspergillus nidulans*. *Mutagenesis* 1993;8:301-5.
17. Benigni R, Conti L, Crebelli R, Rodomonte A, Vari' MR. Simple and α,β-unsaturated aldehydes: correct prediction of genotoxic activity through structure-activity relationship models. *Environ Mol Mutagen* 2005;46(4):268-80.
18. Dearden JC, Barratt MD, Benigni R, Bristol DW, Combes RD, Cronin MTD *et al.* The development and validation of expert systems for predicting toxicity. *Alternative to Laboratory Animals (ATLA)* 1997;25:223-52.
19. Richard AM, Benigni R. AI and SAR approaches for predicting chemical carcinogenicity: survey and status report. *SAR QSAR Environ Res* 2002;13:1-19.
20. Cronin MTD, Dearden JC. QSAR in toxicology. 4. Prediction of non-lethal mammalian toxicological end points, and expert systems for toxicity prediction. *Quant Struct - Act Relat* 1995;14: 518-23.
21. Tennant RW, Spalding J, Stasiewicz S, Ashby J. Prediction of the outcome of rodent carcinogenicity bioassays currently being conducted on 44 chemicals by the National Toxicology Program. *Mutagenesis* 1990;5:3-14.
22. Bristol DW, Wachsman JT, Greenwell A. The NIEHS predictive-Toxicology Evaluation Project: chemcarcinogenicity biossay. *Environ Health Perspect* 1996;104(Suppl. 5):1001-10.

23. Zeiger E, Ashby J, Bakale G, Enslein K, Klopman G, Rosenkranz HS. Prediction of Salmonella mutagenicity. *Mutagenesis* 1996; 11:474-84.
24. Benigni R. The first US National Toxicology Program exercise on the prediction of rodent carcinogenicity: definitive results. *Mutat Res* 1997;387:35-45.
25. Benigni R, Zito R. The second National Toxicology Program comparative exercise on the prediction of rodent carcinogenicity: definitive results. *Mutat Res Revs* 2004;566:49-63.
26. Crettaz P, Benigni R. Prediction of rodent carcinogenicity of 60 pesticides by the DEREKfW Expert System. *J Chem Inf Model* 2005;45:1864-73.
27. Fung VA, Barrett JC, Huff J. The carcinogenesis biossay in perspective: application in identifying human cancer hazards. *Environ Health Perspect* 1995;103(7-8):680-3.
28. Benigni R, Zito R. Designing safer drugs: (Q)SAR-based identification of mutagens and carcinogens. *Curr Top Med Chem* 2003;3:1289-300.
29. Richard AM, Williams CR. Distributed structure-searchable toxicity (DSSTox) public database network: a proposal. *Mutat Res* 2002;499:27-52.
30. 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.
31. Richard AM. DSSTox web site launch: improving public access to databases for building structure-toxicity prediction models. *Preclinica* 2004;2:103-8.