Reprinted with permission from CRC Press:
QSARs of Mutagens and Carcinogens, Ed. R. Benigni (2003)

Chapter 5: Public Sources of Mutagenicity and Carcinogenicity Data: Use in Structure-Activity Relationship Models

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

[a]U. S. Environmental Protection Agency, Mail Drop B143-09
National Health and Environmental Effects Research Laboratory
Research Triangle Park, NC 27711 USA

[b]U.S. Environmental Protection Agency Student COOP Trainee
North Carolina Central University
Durham, NC 27707

CONTENTS

Revised 14 Feb 03

5.1 INTRODUCTION

Publicly supported compilations of mutagenicity and carcinogenicity data are available for a significant number and variety of environmental and industrial chemicals and, to a lesser extent, pharmaceutical chemicals. These datasets represent tremendous past investment in *in vivo* and *in vitro* chemical toxicity testing, primarily driven by government regulatory concerns. These datasets also are the historical informational basis from which virtually all past structure-activity relationship (SAR) models of mutagenic and carcinogenic activity have been derived, and mechanism-based SAR inferences pertaining to these endpoints have been gleaned. It follows that the nature, representation and availability of these data exert a governing influence on the success of derived SAR models. Less appreciated, however, is the role that SAR modeling, itself, can play in assessing data quality, consistency, and completeness. Furthermore, SAR modeling can offer objective means for assessing information content as a function of how these data are pooled, classified, or otherwise interpreted by toxicologists and regulators. In this sense, existing representations of mutagenicity and carcinogenicity data constitute the working interface between toxicologists and SAR modelers.

Schematically illustrated in Figure 5.1 are two generic categories of SAR modeling activities with different data requirements. The top half of the figure represents SAR global model development for a broad toxicity endpoint of interest, such as rodent carcinogenicity or *Salmonella* mutagenicity. In this case, biological activity data are gathered for as wide a range of chemical structures as possible.

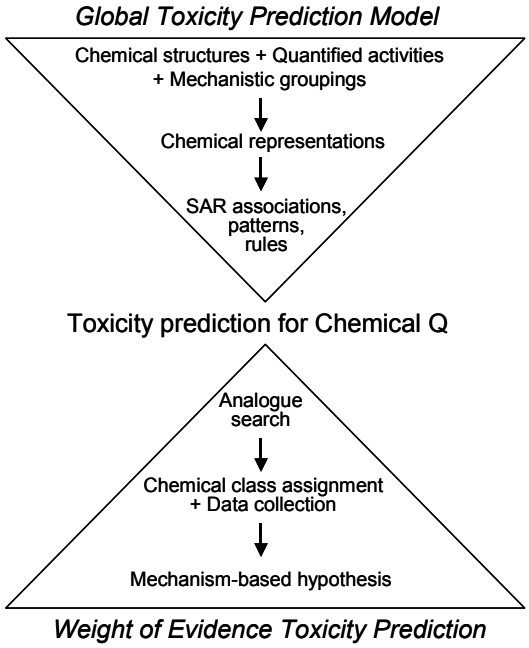

FIGURE 5.1. Schematic illustrating different types of data gathering for SAR model development and toxicity prediction.

Automated algorithms are then employed to extract rules, statistical associations, patterns, etc. that can be applied to toxicity prediction of new chemicals. This type of modeling activity is knowledge-based and exploratory in nature, and has the potential for generating *a priori* SAR hypotheses for subsets and subclasses of the larger dataset. Artificial intelligence (AI) and statistical approaches that fall under this category of SAR modeling activity, and issues associated with their application to modeling of rodent carcinogenicity, have been reviewed[1,2] and are discussed elsewhere in this volume.

A second type of SAR modeling activity, represented in the lower half of Figure 5.1, refers to the process of data gathering towards the goal of toxicity prediction for a single chemical or chemical class of interest. Preexisting SAR models, from commercial sources or previous model studies, can be used to generate SAR predictions for a chemical or class of interest. An example of this approach is illustrated in the study of Moudgal et al.[3] in which the TOPKAT carcinogenicity prediction module[4] was applied to predicting potential carcinogenicity for a series of 244 small organic chemicals detected as water disinfection by-products. Increased confidence in an individual toxicity prediction of this sort is gained from surveying the original training database for examples of structurally similar chemicals with a common basis for activity. In addition, or alternatively, one could perform analog searches of existing data to build a mechanism-based rationale for an SAR prediction of a chemical or class of chemicals. Analogs imply structurally or biofunctionally similar compounds, where the definition of similarity is informed by expert judgment and chemical knowledge. A mechanistic SAR analog approach to prediction is described in Chapter 2 of this volume and illustrated in a study by Woo et al.[5], in which the same water disinfection by-product chemicals as considered in the Moudgal et al. study[3] were evaluated and ranked for potential carcinogenicity.

The first part of this chapter considers issues pertaining to the nature, representation, and availability of mutagenicity and carcinogenicity data as they relate to SAR modeling and prediction problems. Prominent sources of publicly available mutagenicity and carcinogenicity data are listed, along with indication of the availability of chemical structure linkages and complete database access that have the potential to greatly facilitate SAR modeling efforts. An essential consideration in the use of these datasets for SAR modeling, which is discussed in some detail for rodent carcinogenicity, is the degree to which these data represent objective, quantitative experimental measures of a biological endpoint or biochemical event. Alternatively, it is important to know to what extent expert judgment and consensus have been brought to bear on interpreting and classifying an experimental result, as well as the aim of this classification. The discussion considers how the representation and nature of modeled biological data strongly influence the resulting characteristics and success of SAR models. Examples from the literature are used to illustrate how SAR models, in turn, can themselves generate insight into issues of mechanistic complexity and biological relevance of a particular toxicity endpoint representation.

For the purposes of this discussion, we focus primarily on a uniquely large and varied body of work associated with application of the CASE/M-CASE SAR technology[6-8] to global modeling of mutagenicity and carcinogenicity. In particular, we are interested in those studies in which generic data representation and database

issues have been explicitly considered and explored. This focus does not represent endorsement of this SAR prediction technology over any others, nor does it forgive the challenges of modeling non-congeneric data for complex biological endpoints. These issues have been considered in some depth by others[2,9-14].

A basic tenet of SAR study is that the quality of model predictions is highly dependent upon the training set, or knowledge-base, used to derive the SAR models. Returning to Figure 5.1, we conclude that broad access to quality data is essential for building global SAR prediction models, and for validating individual predictions of these and more focused models using analog searches. The last section of this chapter briefly considers some new technologies and initiatives aimed at promoting greater structure-linked access to public toxicity databases for facilitating SAR exploration and model development. This includes a survey of relational database initiatives and data-mining applications pertaining to carcinogenicity and mutagenicity endpoints.

5.2 PUBLIC SOURCES OF CARCINOGENICITY AND MUTAGENICITY DATA

5.2.1 Online Resources

A number of literature reviews offer listings and descriptions of publicly accessible online and digital resources containing chemical mutagenicity and carcinogenicity data. The interested reader should consult these reviews for more detailed description of websites and their contents. Brinkhuis[15] provides an extensive survey of US government public websites, offering information on many types of chemical toxicity, including mutagenicity and carcinogenicity. Richard et al.[16] survey online toxicity databases with particular emphasis on those providing linkages to chemical structure information. In addition, an issue of the journal Toxicology (published by Elsevier Science) is devoted entirely to review of online digital information and tools, with articles organized according to toxicology discipline and/or regulatory application[17]. In that issue, Young[18] broadly surveys genetic toxicology resources and includes discussion of the TOXNET databases of the National Library of Medicine (NLM), as well as the CHEMID PLUS protocol, which enables structure seachability across and within these databases. Also in that issue, Junghans et al.[19] survey a wide range of cancer information resources, including the International Agency for Research on Cancer (IARC) monographs, TOXNET resources, the Berkeley Carcinogenic Potency Database (CPDB) maintained by L. S. Gold, and the National Cancer Institute/ National Toxicology Program (NCI/NTP) rodent bioassay and genetic toxicity databases administered by the National Institutes for Environmental Health Sciences (NIEHS). Table 5.1 provides a listing and description of websites that are the most prominent public sources of chemical mutagenicity and carcinogenicity information.

5.2.2 Chemical Structures Availability

Although it would seem that abundant public information pertaining to chemical carcinogenicity and mutagenicity is available for SAR model development,

TABLE 5.1
Selected Online Public Resources for Carcinogenicity and Mutagenicity Data for Use in SAR Modeling

Website URL [a]	Sponsor/Database	Mutagenicity/ STT [b]	Cancer Bioassay	Structures?/ Searchable?[c]	Downloadable?[d]	Description
http://ntp-server.niehs.nih.gov/	National Cancer Institute /National Toxicology Program (NCI/NTP)	SAL, MLA	Mouse, rat	Yes/No	No	Technical reports of mutagenesis and long-term rodent bioassays and summary results for over 500 chemical substances; two-dimensional and three-dimensional structures available.
http://toxnet.nlm.nih.gov/	National Library of Medicine (NLM)/TOXNET	SAL	Misc.	Yes/Yes	Yes, without structures	TOXNET site maintains multiple toxicity databases searchable by text and structure; NLM site offers full ftp download of database textual content; without structures.
http://toxnet.nlm.nih.gov/	Environmental Protection Agency (EPA)/ Gene-Tox	Misc.	—	Yes/Yes	Yes, without structures	Genetic toxicity info on more than 3000 chemicals for variety of assay systems abstracted from the literature and reviewed.
http://toxnet.nlm.nih.gov/	Chemical Carcinogenesis Research Information System (CCRIS)	Misc.	Misc.	Yes/Yes	No	Summary records abstracted from the literature on carcinogenicity, tumor promotion and inhibition, and mutagenicity on over 8000 chemicals; with references.

TABLE 5.1 (*Continued)*
Selected Online Public Resources for Carcinogenicity and Mutagenicity Data for Use in SAR Modeling

Website URL [a]	Sponsor/Database	Mutagenicity/ STT [b]	Cancer Bioassay	Structures?/ Searchable?[c]	Downloadable?[d]	Description
http://toxnet.nlm.nih.gov/ or http://www.epa.gov/iris/	EPA/ Integrated Risk Information System (IRIS)	Misc.	Misc.	Yes/Yes	No	EPA summary analysis of available toxicity data in support of human health risk assessment for over 500 chemicals; mostly textual content.
http://toxnet.nlm.nih.gov/ or http://www.mdli.com/products/ toxicity.html	Nat. Inst. for Occupational Safety & Health/ Registry of Toxic Effects of Chemical Substances (RTECS)	Misc.	Misc.	Yes/Yes	No	Literature-abstracted acute and chronic toxicity data for over 70,000 chemicals; structure-searchable database maintained and commercially available through MDL, Inc; older version accessible through TOXNET.
http://potency.berkeley.edu/ cpdb.html	Univ. of California – Berkeley/Carcinogenic Potency Database (CPDB) Project	SAL	Mouse, rat, hamster misc.	No/No	Yes, without structures	Chronic animal cancer bioassay results with TD_{50} potencies for over 1300 chemicals abstracted from literature sources and the NCI/NTP testing program, data reviewed and managed by L.S. Gold.
http://www.epa.gov/gap-db/	EPA/ Genetic Activity Profiles (GAP)	Misc.	Links to IARC reviews	Yes/No	Yes	Genetic toxicity information for over 600 chemicals tested in a wide range of STTs, abstracted from the literature, graphical profiles and tabular listings.

TABLE 5.1 (*Continued*)
Selected Online Public Resources for Carcinogenicity and Mutagenicity Data for Use in SAR Modeling

Website URL [a]	Sponsor/Database	Mutagenicity/ STT [b]	Cancer Bioassay	Structures?/ Searchable?[c]	Downloadable?[d]	Description
http://monographs.iarc.fr	World Health Organization (WHO)/ International Agency for Research on Cancer (IARC)	Misc.	Misc.	No/No	No	Published authoritative monographs on carcinogenic hazards to humans posed by more than 800 agents, authored by expert working groups; textual content.
http://cactus.nci.nih.gov/	National Cancer Institute (NCI)/ Structure Database Browser	Tumor inhibition cell line	No	Yes/Yes	Yes, with Structures	Two-dimensional structure and relational searching through NCI Development Therapeutics Program (DTP) Human Tumor Cell Line Screen database for over 37,000 chemicals, full data accessibility, three-dimensional structures available.
http://www.chemfinder.com	CambridgeSoft/ ChemFinder	Misc.	Misc.	Yes/Yes	No	Two-dimensional structure-searchable queries with links to over 300 online public databases, some of which contain mutagenicity or carcinogenicity data.

[a] Website urls were active and current at the time of submission of this review; if a url becomes inactive, we suggest referring to the top-level url of the company or organization to relocate specific information.
[b] Database contains mutagenicity and/or short-term test (STT) information related to the carcinogenic process; SAL=Ames *Salmonella typhimurium* assay, MLA=mouse lymphoma assay.
[c] Database contains chemical structure information (two-dimensional and/or three-dimensional); database is searchable online by chemical structure.
[d] Entire database contents (as opposed to individual chemical results) can be downloaded from website without cost, with or without chemical structures.

this information has, for the most part, not been organized or made available for distribution with the needs of SAR practitioners in mind. The most glaring deficiency is the absence of chemical structure information in many online, public databases, with toxicity information most commonly indexed and searchable only by Chemical Abstracts Number (CAS) or chemical name. Even in cases where chemical structure information is currently provided and online toxicity data records are searchable by chemical structure or substructure (see Table 5.1 for examples), tabular listings of toxicology endpoint data linked to chemical structure cannot be downloaded in full (the single current exception is the NCI Structure Browser for accessing the Human Tumor Cell Line Screen database). Hence, SAR practitioners relying on public online sources for carcinogenicity or mutagenicity data have had to expend considerable effort to extract summary toxicity results and add chemical structure information to databases prior to undertaking modeling. Because no forum for public sharing is in place, in most cases this process is repeated with each new investigator undertaking to model the same dataset. Commercial toxicity prediction, database, and data-mining applications have addressed this need to some extent by providing structure-linked versions of public toxicity databases that include carcinogenicity and mutagenicity data (see Table 5.2). However, these programs are costly and inaccessible to many, do not survey all public datasets of possible interest, and do not in all cases provide unrestricted access to the toxicity data contained within. Some public initiatives aimed at improving this situation will be discussed in Section 5.5.

5.3 TOXICITY DATA REPRESENTATIONS: CARCINOGENICITY

5.3.1 Nature of Existing Data

Structure-activity relationship practitioners generally rely upon whatever description and quantification of the toxicity endpoint of concern is represented within public databases and do not typically undertake review of individual toxicity experiments or activity assignments; rather, this data representation is presumed to reflect the best judgment of toxicology domain experts as to biological relevance. It is essential to recognize, however, that the nature of such endpoint quantification and activity assignments can profoundly impact resulting SAR models. Of particular value to SAR modelers are downloadable tabular compilations of mutagenicity or carcinogenicity data that provide objective and standard comparative measures of a well-defined activity for a broad diversity of chemical structures. A number of important data quality considerations in this regard should be noted. Were data generated under strict experimental protocols overseen by the same laboratory or organization (e.g., NCI/NTP rodent carcinogenicity bioassay results)? Is the database a bibliographic compilation of literature results reported from many laboratories, such as RTECS, CPDB, CCRIS, EPA Gene-Tox (see Table 5.1 for descriptions)? If so, were the results abstracted from the literature with no external review (e.g., RTECS, CCRIS), or were the results reviewed and interpreted by experts in the field (e.g., CPDB, Gene-Tox)? Does the database contain only examples of compounds and results that demonstrate some positive toxicity (e.g., RTECS) or does the

TABLE 5.2
Commercial Toxicity Prediction, Database and Data-mining Applications That Contain Mutagenicity and Carcinogenicity Databases Compiled from Public Sources

Website URL [a]	Company/ Application	Type	Public Data Sources [b]	Structure Searchable?[c]	Description
http://www.mdli.com/products/toxicity.html	MDL Inc./ Toxicity	Relational bibliographic database	RTECS, Misc.	Yes	Oracle-based system runs thru MDL/ISIS Host, extends data records from RTECS to 150,000+ chemicals, toxicity data abstracted from the published literature; with references.
http://www.scivision.com/ToxSys.html or http://www.scivison.com/QSARIS.html	SciVision/ ToxSys and QSARIS	Relational bibliographic database and QSAR development tools	RTECS, Misc.	Yes	Desktop application, originally built from RTECS records, enhanced with records from other public databases, 230,000+ chemicals, endocrine disruptors, etc. QSARIS contains property calculation and statistical analysis tools for facilitating construction of QSAR/SAR models; linked to ToxSys database.
http://www.multicase.com/	MultiCASE, Inc./ M-CASE, CASE	SAR toxicity prediction	NCI/NTP, EPA/Gene-Tox, CPDB	No	Contains 10 rodent (rat/mouse) carcinogenicity SAR models: four species/gender models for NCI/NTP, rat and mouse summary models for NCI/NTP and CPDB, and overall rodent models for NCI/NTP and CPDB. Contains three models for summary Ames SAL mutagenicity data from NTP and EPA/ Gene-Tox; database exploration allowed only within constraints of prediction algorithm.

TABLE 5.2 (*Continued)*
Commercial Toxicity Prediction, Database and Data-mining Applications That Contain Mutagenicity and Carcinogenicity Databases Compiled from Public Sources

Website URL [a]	Company/ Application	Type	Public Data Sources [b]	Structure Searchable[c]	Description
http://www.accelrys.com/products/topkat/	Accelyrs/ TOPKAT	SAR toxicity prediction	FDA-CDER (NCI/NTP, CPDB, NCI, FDA, IARC, EPA); SA; misc. sources	Yes	Contains 8 species/gender (rat/mouse/male/female) multisite vs. single site models, and one weight-of-evidence carcinogenicity SAR discriminant model, all based on FDA-CDER classification of published data. Contains 10 chemical-class-specific discriminant models for summary SAL mutagenicity data from various sources.
http://www.leadscope.com/	LeadScope, inc./ ToxScope	Data-mining, SAR development	RTECS, CPDB, NCI/NTP	Yes	Provides interactive data exploration and filtering by organic chemical class and functional group hierarchies, chemical properties, and biological activities, including carcinogenicity and mutagenicity as contained within RTECS, NCI/NTP and CPDB (150,000+ chemicals).

[a] Website urls were active and current at the time of submission of this review; if a url becomes inactive, we suggest referring to the top-level url of the company or organization to relocate specific information.
[b] See Table 5.1 for definitions of abbreviations and description of data sources.
[c] Contains structure-searchable relational content, allowing a user to independently explore the toxicity databases contained therein; databases within MultiCASE products are not accessible by relational searching independent of the prediction algorithm functions.

database report all experiments yielding either positive or negative responses in a given assay system (e.g., NCI/NTP)? Are the results reported as quantitative experimental measures of activity (e.g., slope of the dose response curve of revertants/nmol in *Salmonella typhimurium* [SAL] TA100 strain, standard Ames reversion assay), or as a categorical assignment of summary activity, either positive or negative (e.g., clearly above or below a chosen threshold of activity)? To what degree does the final reported activity represent the results of a clearly defined experimental system (e.g., with respect to species, strain, target organ, assay)? Alternatively, to what degree has the reported activity been averaged or combined with other activities to produce a summary result, or considered with other information to produce a weight-of-evidence conclusion? Since each of these data considerations has the potential to significantly influence SAR modeling outcomes, they must be acknowledged and openly confronted in any analysis of SAR model significance and predictive applicability.

5.3.2 Summary Toxicity Results

A number of summary toxicity measures are commonly employed in SAR modeling studies (see, for example, currently available TOPKAT and CASE/M-CASE SAR models in Table 5.2). An example of a summary toxicity result is a "positive SAL mutagenicity" result for a chemical listed in EPA Gene-Tox if a positive result was reported in any of the five standard SAL strains: TA98, TA100, TA1535, TA1537, TA1538[20]. A second example is a "positive carcinogenicity" result in the NTP rodent bioassay if a significant tumor outcome is observed at a single tissue or organ site in any one of the 4 tested rodent species- and gender-specific models[21,22]. Several motivations to focus on summary toxicity results as opposed to individual bioassay results transcend the particulars of the SAR method or model approach. The first is practical: to create a training set spanning the largest diversity of chemicals and descriptor space as is possible for the purpose of adding statistical weight to putative SAR associations. In general, the more targeted the bioassay (e.g., Strain A, male mouse, liver tumors), the smaller the database that is available. The second motivation pertains to the ultimate use of the bioassay results and associated SAR model, such as in hazard identification for assessing potential effects in humans. In the latter case, one is less interested in the particular strain or species- or gender-specific effect of a chemical, and more interested in encompassing general and varied mechanisms of mutagenicity or carcinogenicity that are confirmed in multiple assays, and that could have potential relevance to humans. In contrast, a weight-of-evidence call generally involves consensus of an expert committee that has taken into account other information besides the explicit bioassay results (e.g., knowledge of species-specific mechanisms of bioactivation, experience with analogs, epidemiological evidence), an example being an IARC classification of a NTP rodent carcinogen as a probable or possible human carcinogen.

Why are the above distinctions important? The further an SAR model is removed from a biologically relevant experimental test outcome, and presumed common mechanisms of action within activity classes, the less theoretical underpinning is provided and the more heuristic the model becomes[2,23]. If the goal of an SAR study is to provide mechanistic insight into the activity under consideration, then it is paramount that the experimental data under consideration provide a clear

and objective measure of a chemical-induced biological activity of interest[23,24]. On the other hand, if the goal is to create an SAR model for use in hazard assessment or screening, then an ability to reproduce less objective historical "activity calls" or hazard assessments is of greater interest. The evolution of carcinogenicity SAR prediction models created with the commercial MultiCASE (CASE/M-CASE) and TOPKAT systems over the past several years exhibits a trend towards increased reliance on more biologically refined models. Current TOPKAT and MultiCASE commercial offerings (see Table 5.2) include several species- and gender-specific rodent carcinogenicity submodels, as well as models at the species (rat or mouse) and rodent level (rat and mouse combined) and, in the case of TOPKAT, multi-site vs. single site tumorgenesis within each species and gender model. Not surprisingly, the more focused submodels (e.g., male rat) are more uniquely characteristic and predictive of the submodel bioassay results, and are potentially more informative of species- and gender-specific mechanisms[25-27]. A corollary, however, is that these "less averaged" models are more tied to the peculiarities of the species- and gender-specific data and are more influenced by singular and spurious results in that data[25-28], in that they are attempting to faithfully replicate the actual bioassay results.

Building species- and gender-specific SAR submodels for rodent carcinogenicity allows for potentially greater flexibility and transparency in prediction strategies. One can attempt to either mirror the process of expert heuristic evaluation of rodent carcinogenicity (e.g., by combining rodent submodel results in various ways to yield a summary result) or one can model the heuristics directly (e.g., by modeling the summary rodent carcinogenicity calls directly), with different possible outcomes. Rosenkranz and coworkers[25,29] have reported strategies for combining CASE/M-CASE rodent species-specific carcinogenicity submodels and summed models using Bayesian statistics to optimize overall prediction performance measures (sensitivity, specificity, concordance). Because each of these SAR models is derived from a different set of data, each model contains a different profile of biophores (i.e. structural fragments significantly associated with active chemicals) that presumably captures different information relative to the SAR prediction problem. For example, Cunningham et al.[27] have reported only 36% overlap in CASE/M-CASE biophores derived from the CPDB rat and mouse summary tables, implying significantly different structural drivers for carcinogenicity in the two species.

Rodent bioassay data resolved to the species or species- and gender-specific level, in principle, can be further resolved to tumor site (e.g., liver, kidney, etc.)[30,31]. This focus can be more problematic from an SAR modeling standpoint due to limited numbers of chemicals for which data are available relative to any particular tumor site. Hence, virtually no reported SAR models of rodent carcinogenicity data resolved to tumor site have been reported. The most prevalent tumor site observed in the NCI/NTP rodent bioassay experiments is the liver, and yet this tumor site was observed in only 15% of experiments [30]. In addition, the biological significance of tumor site-specific information is an issue of some controversy. In one of the few quantitative analyses of tumor site-specific rodent bioassay information, Benigni and Pino[32] reported that species specificity generally overcame organ specificity in the majority of tumor site categories (e.g., liver tumors are nearly exclusive to mice and rarely occur in rats). Moreover, the species specificity was remarkably stronger than the trans-species sex specificity (e.g., females of both rats and mice shared few

common tumor sites). Finally, within a few chemical classes believed to be most clearly associated with common carcinogenic activation mechanisms (e.g., aromatic amines), no obvious association of chemical structure with tumor profile was discerned; that is tumors were produced at a wide range of sites for chemicals within each class. These results suggest that stochastic elements in the carcinogenic process are likely to play a role in the intervening steps to tumor formation, subsequent to the initial chemical bioactivation step (e.g., nitrenium ion formation in aromatic amines). The implication for future SAR study is that tumor site-specific information may not prove useful for improving mechanism-based categorizations of rodent carcinogenicity data, and by inference, tumor site is unlikely to be a viable target for SAR prediction.

5.3.3 NCI/NTP and CPDB Rodent Carcinogenicity Summary Results

The vast majority of SAR models developed to date for carcinogenicity prediction have been built upon one of two main public sources of rodent bioassay data, i.e. from summary tables of the NCI/NTP and the CPDB rodent bioassay databases (see Table 5.1). These include the varied SAR models that participated in the NTP Predictive Toxicity Evaluation (PTE-1 and PTE-2) exercises, discussed by Benigni in Chapter 9 of this volume and in published studies[9,11]. To understand the distinctions among published SAR models derived from these two data sources requires understanding of the major differences in the summary rodent carcinogenicity tables from the two databases. The NCI/NTP rodent bioassay database provides data on over 400 chemicals, generated in a number of laboratories, but using a standard experimental protocol with respect to numbers of animals, strains, dosing regimens, pathology, and statistical analysis of results[21,22]. Although there have been some changes in these protocols over time, this database is considered to be relatively consistent in terms of experimental design. The CPDB contains a larger diversity of chemical structures (over 1300), and includes tumor data reproduced from all of the NCI/NTP rodent bioassay *Technical Reports* as well as additional data extracted from over 1200 literature sources subjected to extensive review[33-36]. In addition, the CPDB includes bioassay results from species other than rat and mouse and incorporates a wider variety of experimental protocols from the general literature that meet well-defined, but generally less stringent inclusion criteria when compared to the NCI/NTP protocols (http://potency.berkeley.edu/text/methods.html).

A further distinction between the summary rodent carcinogenicity tables derived from the NCI/NTP and CPDB databases, that is sometimes overlooked by SAR modelers, is that different summary calls may be listed for the same chemical[34]. When literature experiments are also factored in the assignment of species/gender positivity, the CPDB summary table occasionally lists a positive species/gender call for a chemical listed as negative in the NCI/NTP. In addition, a quantitative measure of carcinogenic potency is included in the CPDB, but not the NCI/NTP summary table. This potency measure, termed a TD_{50}, is defined as: "that dose-rate in mg/kg body wt/day which, if administered chronically for the standard lifespan of the species, will halve the probability of remaining tumorless throughout that period"[37]. The TD_{50} takes into account a number of experimental details (such as length of experiment, conversion factors, and estimate of dose) and is computed for

species/gender/tissue/tumor site in each experiment. The CPDB carcinogenicity summary table reports the harmonic mean TD_{50} value from positive experiments for each species[34,37]. Significant documentation and details pertaining to the inclusion criteria used for incorporating a study result into the CPDB and the computation of a TD_{50} are available at the CPDB website (see Table 5.1).

Specifically because it provides a quantitative and comparable measure of relative carcinogenic potency among CPDB chemicals, the TD_{50} poses an alluring modeling challenge for traditional quantitative structure-activity relationship (QSAR) study of carcinogenicity. Benigni and Passerini[38] have reported successful development of predictive QSAR equations for rat and mice, based on the species-level rat and mouse TD_{50} values, for a well-defined chemical class, i.e., aromatic amines. Similarities between the forms of these QSAR equations and those derived earlier for *Salmonella* mutagenic potency of aromatic amines, as well as the mechanistic relevance of individual QSAR parameters, increase confidence in the validity of these equations[38]. The success of these QSAR modeling efforts, further demonstrated in objective statistical terms, lends independent support to the contention that the species-averaged TD_{50} potency measure has some biological relevance in the context of a mechanistically well-defined chemical class.

5.3.4 Data Quality and Reproducibility of Rodent Bioassay Results

An interesting corollary to the above discussion concerns issues of data quality and reproducibility associated with the rodent carcinogenicity bioassays, and the potential impact on SAR model success. The rodent carcinogenicity bioassay, as performed by the standard protocols of the NCI/NTP, is very costly and time-consuming. As a result, full replicate experiments are not performed by the NCI/NTP and are seldom performed by others. Although reproducibility is assumed under the strict guidelines of the NCI/NTP protocol, the true reproducibility of these experiments, as well as other experiments operated under less strict protocols, is largely unknown. And because error associated with experimental reproducibility places an upper limit on the absolute predictivity achievable by any SAR model, this limit of predictivity is also unknown.

Based on analyses of a relatively small set of 38 replicate experiments from the literature (testing the same route, sex and strain of rodent), Gold et al.[33] have estimated overall reproducibility of the rat bioassay at 85% and the mouse bioassay slightly less, at 80%. A more recent analysis by Gottmann et al.[39] makes the provocative assertion that "rodent carcinogenicity assays are much less reproducible than previously expected" and because of this "rodent carcinogenicity assays should be treated as unreliable, which has consequences for SAR modelers and the risk assessment process". These conclusions were derived from analysis of a larger set of 121 chemicals for which replicate rodent bioassay results for the same chemicals, but tested under different protocols, were available from both the NCI/NTP rodent bioassay database and additional rodent bioassay experiments contained in the CPDB. These authors estimated concordance of only 57% in overall rodent carcinogenicity classifications (i.e. positive or negative) from both sources, with comparably poor concordances found with respect to species-, gender-, strain-, and target organ-specific test results across laboratories. Interestingly, however, the

results of Gottmann et al.[39] agreed with those of Gold and coworkers[33] in the finding that rat bioassay results were considerably more reproducible than mouse bioassay results (62% vs. 46%) and that rats were significantly more sensitive to carcinogens than mice (i.e. a larger percentage of chemicals are found to cause tumors in rats than in mice).

Gottmann et al.[39] note that a large proportion of the replicate experiments (34/47) examined in the earlier Gold et al.[33] analysis were published by the same authors. Among the number of other significant differences in these two replicate studies, is the larger and more varied set of chemicals considered in the Gottmann et al.[39] study; however, differences in experimental protocol in what are considered replicate experiments cannot be ruled out as the main reason for observed lack of concordance. Given that the NCI/NTP experimental protocols are generally stricter and more uniformly applied than in the majority of literature rodent bioassay studies, in our view the more variable literature studies cannot be used as a reliable judge of the reproducibility of the NCI/NTP experiments. The most that can be concluded from the Gottmann et al.[39] analyses is that estimating the reproducibility of rodent bioassay results is indeed problematic given current data constraints and that adherence to strict experimental protocols (such as the NCI/NTP) may be essential for achieving reproducibility in results, but that this assertion remains unconfirmed.

The above analysis of "replicate" bioassay results highlights experimental protocol as one of the most important confounding factors. It is reasonable to expect that the lack of concordance observed for chemicals tested by both the NCI/NTP and additional studies included in the CPDB would be representative of more chemicals in the CPDB if more replicate data from the NCI/NTP were available. Hence, beyond differences in chemical coverage due to the larger number of chemicals represented in the CPDB, it is anticipated that the different information content in rodent carcinogenicity summary tables derived from these two databases will yield significant differences in SAR models. Indeed, this has been reported in various CASE/M-CASE published analyses[25-27] and is manifested in the commercial availability of separate NCI/NTP and CPDB SAR models (see Table 5.2).

5.4 DATA DEPENDENCE OF SAR MODELS: CASE/M-CASE EXAMPLES

5.4.1 Database Informatics Analyses

The CASE/M-CASE approaches consist of computer-based algorithms for automated SAR analysis and prediction that can, in principle, be applied to any sort of data in which organic chemicals with known structures are linked with corresponding activities in biological systems. Details of the CASE/M-CASE approaches are provided in Chapter 6 of this volume and in published studies[6-8]. In brief, the methodology is primarily based upon the deconstruction of chemical structures into all possible composite structural fragments of length 2-10 heavy atoms. Each of these fragments is assigned a CASE activity unit (based on categorical or potency assignments) reflecting the activity of the corresponding parent structure, and fragments from the entire database are then pooled into gross activity categories, i.e. positive, marginal or negative. A structural fragment is labeled as a *biophore*, in CASE parlance, only if it has significantly skewed statistical representation in the

active category (i.e., is represented in many more active than inactive parent compounds). The older CASE technology operates in this fashion on the entire database of chemicals, without prior or subsequent classification. The newer M-CASE technology adopts a hierarchical classification process in which biophores of greatest statistical significance are extracted from an initial CASE analysis and used to define major biophore-containing classes. These classes are separately analyzed by a subsequent CASE analysis to discern substructural modifiers to the activity of the major biophore class (an example would be different patterns of methyl substitution modifying activity within the class of aromatic amines, each member of which contains the aromatic amine functionality). The CASE and M-CASE approaches operate on the same dataset in different ways and, hence, will often yield a somewhat different set of biophores and related, but distinct prediction models.

The CASE/M-CASE approaches represent unbiased, *de novo* SAR analyses in the sense that, once CASE activity units are assigned to each molecule in the database (a point for human intervention and some subjective judgment), the derived prediction model is fully determined by automated and objective analysis of the data. A corollary is that the CASE/M-CASE model outcomes will be determined solely by the nature of the data, and will be intimately tied as well to the quality, extent (i.e. numbers and types of chemicals included), and biological representation of the data[40,41]. It is acknowledged that any number of alternative SAR approaches could be taken to analyzing the same set of biological data, using different chemical descriptors, types of information, and functional algorithms, thus producing different model outcomes and predictive capabilities. It is also recognized that the CASE/M-CASE approaches have inherent limitations tied to the nature of the chemical representations and algorithms employed (for comparisons of different SAR approaches applied to predictive toxicology, the reader is referred to a number of reviews on the topic[2,9-14]). For purposes of this discussion, we are primarily interested in the ability of the automated CASE/M-CASE technology to shed new light back onto the toxicology databases used in model development. In large part, this is due to the transparency and interpretability of the formulation of CASE/M-CASE results (i.e., consisting of discreet substructural fragments).

A number of CASE/M-CASE publications have demonstrated this general informatics capability, effectively highlighting the intimate relationship between modeled data and model outcome. A novel method has been described for assessing the informational content of toxicity databases used to train CASE models by applying these models to predicting on a large external dataset of 5000 compounds, designed to approximate the "universe" of chemicals from a structural standpoint[41,42]. The proportion of CASE model predictions that are accompanied by a warning of the presence of an unknown structural feature (i.e., a fragment not previously seen in the model training data set), provides an objective measure of the informational content of the training data set relative to the external dataset. The informational content is quantitatively estimated as (100 - % predictions accompanied by warning). This approach has been applied to evaluating and proposing strategies for increasing the informational content of existing databases for *Salmonella* mutagenicity and clastogenicity[41]. Increasing informational content of a toxicity database involves targeting molecules containing unknown functionalities for testing and subsequent incorporation into an expanded training data set. It follows that the optimal size of a toxicity database, from the CASE modeling perspective, is the stage at which the

informational content of the database no longer increases significantly with increasing size[43]. For a *Salmonella* mutagenicity database, Liu et al.[43] found this plateau to occur at a training database size of approximately 400 chemicals. Prior to this number, the indices of CASE model predictivity (i.e., sensitivity, specificity, and concordance between experimental and predicted results) increased with increasing size of the database. Note that, because the CASE informational content measure does not depend on fragment activity assignments but only on single fragment incidences in the database, it can only serve as an approximate measure of informational content relative to the biological activity under study. For example, if the same approach applied to *Salmonella* mutagenicity were applied to evaluating databases of rodent carcinogenicity, an endpoint of greater biological complexity, it is likely that a larger optimal database size (i.e. beyond 400 chemicals), having approximately the same CASE measure of informational content, would be necessary to achieve comparable measures of CASE model predictivity. This conclusion has been borne out in subsequent studies[26-28].

In other studies, Rosenkranz and coworkers have used CASE model biophores that reflect both fragment representation and biological activity considerations within the database as an objective means for assessing mechanistic similarity (or dissimilarity) between two or more toxicological endpoints[26,27,44]. Here, the assumption is that CASE biophores represent a distillation of the mechanistic informational content of the toxicological database, capturing the main drivers for predicting the structural basis of the particular toxicological activity under study. Two databases for different toxicological endpoints might contain entirely different chemical structures that have undergone testing, yet some proportion of the CASE biophores associated with activity could be the same, indicating common drivers for the two toxicities. Equally informative could be CASE biophores that differ between two models, indicating possible mechanistic divergences between the two test systems. Analyses have been reported indicating significant commonalities, for example between mutagenicity in *Salmonella* and carcinogenicity in mice (approx. 40% overlap in identical or embedded biophores)[26]. In addition, these types of analyses have proven useful for assessing mechanistic informational content and overlap between cytotoxicity endpoints and rodent carcinogenicity, and endpoints reflecting genotoxic vs. non-genotoxic modes of carcinogenic activation[26,27].

In addition to the above informatics applications, the CASE technology has been used to examine the effect on model performance of varying the ratios of actives and inactives within the database[43,44], and to suggest procedures for objective validation of models[29] and for assessing model predictivity[45]. These varied applications demonstrate utility of an SAR approach that goes beyond toxicity prediction for individual chemicals, illustrating the application of objective data analysis methods to illuminating characteristics of toxicity databases that impact on the larger toxicity prediction problem.

5.4.2 Rodent Carcinogenicity Prediction Models

We have devoted significant discussion elsewhere in this chapter to highlighting differences in content and activity representations within the NCI/NTP and CPDB rodent carcinogenicity summary tables. These differences are clearly manifested in reports of CASE/M-CASE models for carcinogenicity in mice and rats derived from these two databases[25-27]. Although the first published CASE/M-CASE species- and gender-specific rodent carcinogenicity models were based exclusively on the NCI/NTP summary calls, motivations for deriving models based on the CPDB summary calls included the larger numbers and diversity of chemical structures and the species-averaged TD_{50} as a measure of relative potency. The TD_{50} was used to calibrate more finely CASE significance of structural fragments in association with activity[26,27]. Hence, the CASE/M-CASE models derived for the four rodent experiments (male and female rat and mouse) represented in the two datasets differed not only in terms of the chemicals included, but also in terms of the means used for categorizing carcinogenic activity. A quantitative indication of the profound differences between these model training sets is reflected in the mere 28% overlap in biophores reported for the CPDB and NCI/NTP rat models[27]. This slight overlap is even more remarkable considering that the structural information pertaining to the NCI/NTP chemicals is completely contained within the CPDB; it is only the activity assignments that potentially differ. This significant lack of concordance between models for the two rat carcinogenicity datasets shed some doubt on the significance of either model result. As a result, neither CPDB rat model was incorporated into the CASE/M-CASE rodent carcinogenicity prediction models in two reported studies[25,29].

Overall concordances of rat and mouse (species level) CASE/M-CASE models for the CPDB were reported as 64% and 70%, respectively[27]. Interestingly, a number of other performance indicators by which the rat models were judged less significant than the mouse models, included a similar lower concordance of rat compared to mouse for the NCI/NTP models. Given the evidence, independently corroborated in two reproducibility studies[33,39], that rat carcinogenicity data are significantly more experimentally reproducible than the mouse carcinogenicity data, the lower performance indicators for the rat models are somewhat surprising. Cunningham et al.[27] point out that the rat data are significantly more robust than the mouse data in terms of having 92 more carcinogenic chemicals in the CPDB for the rat than for the mouse, and in terms of the significantly smaller number of different tested strains (74 for rats vs. 101 for mice). However, they also point to the distinction that reproducibility represents repeated challenges of the same chemical, whereas the more varied response in the rat is with respect to different chemicals that can act by possibly more varied mechanisms. Hence, the authors speculate, "the lesser predictivity of the rat CPDB SAR model may be indicative of a more variable response to chemical carcinogens for rats than for mice". [27]

5.4.3 Influence of Toxicity Protocol on SAR Models

A final point is made concerning the ability of SAR models to comparatively assess databases meeting different quality criteria or employing different protocols for formulating and classifying experimental data, again referencing relevant examples employing the CASE/M-CASE technologies. In the first example, CASE analyses were applied to modeling cytogenetic endpoint data extracted from both the EPA/Gene-Tox database (see Table 5.1) and the NTP database[46]. Models that allowed thorough analyses of the structural features of the cytogenetic endpoints were successfully derived for the NTP dataset, whereas the CASE technology was applied without success to the EPA/Gene-Tox dataset[45]. It was concluded that the standard protocol and quality control ensured in the NTP dataset could not be assessed for the literature-abstracted data collated within the EPA/Gene-Tox dataset, and that greater experimental variability and poorer data quality within the EPA/Gene-Tox dataset likely accounted for the failure to derive CASE models[45]. Although this conclusion might have been anticipated based solely on data quality control considerations, the explicit failure to derive CASE models gives independent and objective credence to this assessment.

In a second example, CASE/M-CASE analyses effectively contrasted two distinct protocols for activity classification of mouse lymphoma forward mutational assay (MLA) results[47]. The first database consisted of MLA results generated and evaluated under a defined protocol of the NTP (MLA/NTP). The second dataset, consisting largely of different chemicals, resulted from an in-depth reevaluation of literature studies that were judged according to a significantly different protocol for activity assignment than used by the NTP; this analysis was carried out by an EPA Gene-Tox working group (MLA/GT)[47]. It was reported that CASE/M-CASE models for the MLA/GT dataset were significantly more predictive than for the MLA/NTP dataset[47]. Additionally, the MLA/GT models were reportedly far simpler than the MLA/NTP models, containing fewer, more statistically significant biophores. In this example, it appears that the effect of significantly different protocols for activity assignments outweighed possible quality control issues in determining SAR modeling success. These SAR model results also independently suggest that the MLA/GT protocol for activity assignment possibly provides a more biologically coherent and meaningful measure of activity than the MLA/NTP protocol.

In the third example, Matthews and Contrera[28] report different calibration and application of rodent carcinogenicity models in development of optimized M-CASE modules, with the objective of better replicating the heuristics of the carcinogenicity review process for pharmaceuticals of the U.S. Food and Drug Administration (FDA). One of the most important changes in this M-CASE implementation was the assignment of a potency weight factor that ranks carcinogens and biophores (i.e., active fragments) according to FDA regulatory importance: trans species>trans-site/single species>single-site/species[28]. This is a more specific designation of carcinogens than the activity designation that was used in deriving the NTP/M-CASE model; that is the latter assigning equal weight to trans or single site carcinogens, labeling both as positive. The second major modification was that the FDA/M-CASE system was trained on a larger data set (n=934) that included a significantly larger percentage of pharmaceuticals extracted from the CPDB and FDA files. The FDA/M-CASE optimized model identified over twice as many

biophores as the default M-CASE model that was trained on a smaller NCI/NTP data set (n=316)[28]. In addition, in application to a beta test set containing a significant percentage of pharmaceutical-type chemicals, this model performed significantly better than prior M-CASE models that had been trained on the NCI/NTP dataset, the latter containing mostly industrial and environmental chemicals and few pharmaceuticals. The optimized FDA/M-CASE model was exceedingly accurate in predicting carcinogens correctly in the beta test set, achieving 98% specificity, whereas a relatively large percentage of carcinogens were also falsely predicted to be negative (over 40%)[28]. The latter performance indicator is most likely a reflection of generally greater ignorance (i.e., fewer examples in the training dataset) pertaining to the more varied activity-conferring structural moieties in larger pharmaceutical-type chemicals. This example illustrates, once again, the strong reliance of the M-CASE prediction model and performance statistics on the training dataset and the activity designations used in model derivation.

5.5 TOXICITY DATABASE TOOLS TO AID SAR MODEL DEVELOPMENT

5.5.1 Commercial Relational and Data-mining Applications

The ability to relationally search across public toxicity databases using both biological and chemical criteria represents a potentially powerful approach for SAR hypothesis generation, model development, and model validation. This paradigm offers maximum flexibility to an informed user and empowers the concept of analog searching, in both chemical and biological domains. Large pharmaceutical and chemical companies, in particular, have invested heavily in relational database platforms and data-mining tools for managing, exploring, and providing widespread corporate access to large internal libraries of chemical and biological test information. In government, the FDA's Center for Drug Evaluation (FDA-CDER) is emulating this corporate approach by creating a relational database, searchable by chemical structure, for pharmaceuticals submitted for registration and approval[48]. In addition, they are coupling this technology to the M-CASE SAR predictive software to add *in-silico* toxicity prediction capabilities across a variety of endpoints of concern, including mutagenicity and carcinogenicity[28]. These two technologies -- relational searching and automated toxicity prediction -- are being used hand-in-hand within the FDA-CDER program to facilitate and improve initial hazard assessments of reviewed chemicals[48].

Examples of commercial relational database applications containing extensive compilations of field-delimited mutagenicity and carcinogenicity data linked with chemical structure information include the MDL, Inc. Toxicity database and SciVision's ToxSys software (see Table 5.2). The version of TOPKAT currently marketed under Accelrys also allows, as a complement to its SAR prediction modules, relational structure-based searching across TOPKAT mutagenicity and carcinogenicity databases used in model development.

Examples of commercial data-mining applications that have been applied to analysis of mutagenicity and carcinogenicity data, primarily for pharmaceutical drug development, are offered by LeadScope, Inc. (see Table 5.2) and Bioreason, Inc (see Table 5.3). Data-mining applications differ from commercial toxicity prediction

programs, such as TOPKAT and CASE/M-CASE, in that they provide a user with automated tools for interactive data exploration, rule-extraction, and *de novo* SAR hypothesis generation pertaining to mutagenicity and carcinogenicity endpoints. LeadScope's ToxScope product includes large stores of public mutagenicity and carcinogenicity data primarily abstracted from RTECS (see Tables 5.1 and 5.2)[49]. The unique feature of this application is the ability to interactively visualize activity patterns across hierarchically displayed organic substructural classes, coupled with the ability to filter activities according to multiple structure-based criteria. It is envisioned that a corporate user of this product would merge public toxicity data stores with proprietary toxicity databases, if available, to customize and enhance data-mining capabilities. Bioreason's ClassPharmer suite of programs similarly provides users with interactive computational algorithms for organizing, classifying, and generating SAR hypotheses from structure-linked toxicity databases, although in this case, databases must be provided by the user. Bacha et al.[50] have demonstrated use of this technology for analyzing *Salmonella* mutagenicity data, illustrating the ability to simultaneously explore classifications of chemicals that incorporate features potentially relevant to both the desired pharmacological activity as well as the undesired toxicity.

Both relational database applications and data-mining applications add valuable functionality to existing, historical toxicity data records, to enable more sophisticated use and exploration of these data. However, since they rely primarily on the same publicly available stores of toxicity data, it follows that these applications will be bound by the same data availability, representation and quality constraints that strongly influence other types of SAR modeling endeavors.

5.5.2 Public Toxicity Database Initiatives

Two new public database initiatives, in early stages of development, will be briefly described. Both are aimed at improving public accessibility to structure-linked toxicity data across a variety of endpoints, test systems, and data sources. In addition, shared objectives of both efforts are (1) to add chemical structures to existing public toxicity data to aid SAR model development, (2) to standardize the format of chemical and toxicological information to facilitate relational searching across diverse chemical and biological information fields, and (3) to enter into partnerships with persons and entities that use and maintain these public toxicity data stores to expand these efforts.

A consortium of industry and government sponsors has charged the International Life Sciences Institute (ILSI) with development of an SAR toxicity database (see Table 5.3). The stated mission of the ILSI Structure Activity Relationship (SAR) Database Subcommittee is to "utilize the vast collection of toxicology that has been developed by the international government, industry, and academic community to establish a centralized database of toxicity testing results, including structure-activity relationships, which will be useful for predictive toxicology" (www.ilsi.org, Table 5.3). The relational database application chosen for this effort is a modified version of IUCLID (see Table 5.3), an application currently endorsed as the primary toxicity data exchange tool for the European Union Risk Assessment Program and the Organization for Economic Cooperation and Development (OECD) Existing Chemicals Program. LHASA Limited, working in

TABLE 5.3
Miscellaneous Websites for Commercial Data-Mining and Relational Database Applications Requiring User-Supplied Data, and Websites for Public Toxicity Database Development Efforts

Website URL [a]	Company/ Application	Type	Description
http://www.bioreason.com	Bioreason, Inc./ClassPharmer	Data-mining application software	Provides application tools for data management, and structure-driven knowledge discovery based on algorithms for organizing, classifying, and generating SAR hypotheses.
http://www.ilsi.org	International Life Sciences Institute (ILSI) SAR Database Subcommittee	SAR toxicity database	Nonprofit organization collaborating with LHASA, Ltd. and consortium of industry and government groups to develop an SAR database of toxicity information for use in predictive toxicology.
http://ecb.ei.jrc.it/Iuclid/	European Chemicals Bureau/ IUCLID database system	Relational database	Database application used for data collection and evaluation within the European Union Risk Assessment Program; does not contain chemical structure field in current form.
http://www.chem.leeds.ac.uk/luk/	LHASA, Ltd.	SAR expert knowledge -based technologies	Markets DEREK and METEOR expert systems for toxicity and metabolism prediction, but with no databases included. Added structure field and structure-searching capabilities to IUCLID for building ILSI SAR toxicology database prototype.
http://www.dsstox.net or http://www.epa.gov/nheerl/ dsstox/[b]	EPA/Distributed Structure-Searchable Toxicity (DSSTox) database network	Standard format files of chemical structures and toxicity data	Central website will contain general information, tools and guidance for Sources in constructing new DSSTox files, central field definitions file, and links to DSSTox source websites containing DSSTox standardized toxicity data files available for free download by the public.

TABLE 5.3 (*Continued*)
Miscellaneous Websites for Commercial Data-Mining and Relational Database Applications Requiring User-Supplied Data, and Websites for Public Toxicity Database Development Efforts

Website URL [a]	Company/ Application	Type	Description
http://www.acdlabs.com	Advanced Chemistry Development/ ChemFolder	Chemical relational database application	Low-cost, PC-based chemical relational database application, allows structure, substructure, property, text searching of data, linked to chemical drawing program, ChemSketch; allows for searching across multiple separate databases.
http://chemfinder.cambridgesoft.com	CambridgeSoft/ ChemFinder	Chemical relational database application	Low-cost, PC-based chemical relational database application, allows structure, substructure, property, text searching of data, linked to chemical drawing program, ChemDraw; databases must be imported and merged for into single database for searching.
http://www.mdli.com/products/framework. html	MDL, Inc./ Integrated Scientific Information System	Chemical relational database application	Provides information on the SDF standard import/export format; also provides a variety of integrated information management products using the ISIS base, ISIS draw, and ISIS host applications
http://www.oracle.com/ and http://www.accelrys.com/accord/	Oracle and Accelrys/Accord	Chemical relational database application	Accord application runs on top of Oracle system to provide chemical structure fields and structure-searchability functions.; typical of larger corporate centralized databases managed by a central server and administrator.

[a] Website urls were active and current at the time of submission of this review; if a url becomes inactive, we suggest referring to the top-level url of the company or organization to relocate specific information.
[b] Public launch of this website (reached from either url) is anticipated for early 2003.

collaboration with ILSI, has incorporated structure fields and structure-searchability into the IUCLID application to extend its capabilities for use in development of a centralized SAR toxicity database. LHASA has also been primarily charged with coordinating efforts to obtain data from public sources for populating the database. The initial pilot project has completed the incorporation of databases for *Salmonella* mutagenicity and carcinogenicity from public sources (e.g., NTP, CPDB), and is planning to expand efforts to collect public toxicity data from other sources[16]. In addition, a more ambitious and longer term goal is to move toxicity data that no longer must be confidential from the private records of government regulatory agencies and industry into the public domain. The affiliation of government and industry members in this data collection effort represents a major distinction of this toxicity database project over other commercial and non-commercial efforts.

A second public toxicity database effort, also in development, is the EPA-sponsored Distributed Structure-Searchable Toxicity (DSSTox) database network. Details of this proposal have been published[51], and the launch date of the public website is planned for early 2003 (see Table 5.3). The proposal is distilled into the following three major elements: (1) an application-independent, standard SDF file format adopted for public toxicity databases that supports inclusion of chemical structures; (2) a distributed source approach to enable decentralized, free public access to DSSTox SDF data files; and (3) community involvement in contributing to and expanding the DSSTox public database network. Structure Data File (SDF) format, developed by MDL, Inc. (Table 5.3), is a public, ASCII flat file format that stores field-delimited structure, text and property information for any number of molecules. SDF was chosen for the DSSTox effort because it is a *de facto* standard data import/export feature of virtually all commercially available chemical relational database applications[51]. The latter include low-cost PC-based applications, such as ChemOffice's ChemFinder, ACD's ChemFolder, and Accelrys' ACCORD (see Table 5.3), in addition to applications with higher-end functionality (e.g., nested fields, reaction fields), such as MDL's ISIS, and Oracle-backed systems (See Table 5.3) that are typically employed in corporate situations. Each DSSTox SDF file will contain a set of standard chemical identifier fields that includes the 2D structure, followed by toxicity information fields. DSSTox SDF files are being created for a wide variety of public toxicological databases, including a number of the main public sources of carcinogenicity and mutagenicity data listed in Table 5.1. These files will be offered for free public download from either the DSSTox Central website or DSSTox Source websites, and will be easily convertible to data tables or importable into any commercial or private chemical relational database application. A DSSTox Central Website (see Table 5.3) will serve as the hub of the DSSTox project, providing general information, a central index of field names, links to DSSTox Source websites containing DSSTox SDF files, and public tools and resources of general interest to the DSSTox community[51]. Another crucial role of this website will be to connect the DSSTox user community members and to enlist their help in propagating the DSSTox recommended standards, reporting DSSTox SDF file errors to the Sources, offering enhancements to existing DSSTox SDF files, and aiding in the construction of new DSSTox SDF files.

The DSSTox proposal is distinguished in two important respects from those capabilities and initiatives previously discussed: (1) the complete DSSTox SDF files,

including chemical structures, will exist entirely in the public domain and be freely available for download, allowing for completely customized use in database development; and (2) the distributed network of DSSTox data files will be a community-supported, application-independent effort, as opposed to a centralized effort creating a large application-specific database. Complementarities exist, however, in that DSSTox SDF files will be directly importable into the central ILSI SAR toxicity database effort to expand data contained within the latter. Another clear advantage of the DSSTox approach is that SDF files will be faithful representations of existing databases, circumventing difficult value judgments on data quality or superiority of one data measurement over another, and deferring these judgments to the toxicological domain experts[16]. The ultimate success of the DSSTox project will depend on the active cooperation and involvement of both the toxicity database Sources and the larger DSSTox user community. The DSSTox database network will allow a much larger community of academics, government researchers and regulators, and small to medium-sized industries access to powerful chemical relational database structure-searching capabilities, and open and complete access to public toxicity databases. This, in turn, will serve to enhance communication and collaboration between toxicologists and the SAR modeling community, and will facilitate SAR modeling efforts across a wide range of public databases and toxicity endpoints.

5.6 CONCLUSIONS

Issues pertaining to the experimental reproducibility, and hence quality of rodent carcinogenicity data are currently unanswerable in the most direct sense, and are likely to remain so for the foreseeable future. However, it is important to realize that the upper limit of predictivity of an SAR model (but not the lower limit) is bound by the same data quality constraints as are assessed directly by experiment. Hence, the most stringent assessments of SAR model predictivity, such as provided by the NTP prospective prediction exercises for rodent carcinogenicity (44 and 30 chemicals in the PTE-1 and PTE-2, respectively), can, in turn provide some independent and objective assessment of data quality and reproducibility. Benigni, in a summary analysis of the results of the PTE-2, concludes that the upper limit of 67% predictivity of rodent carcinogenicity is achieved only when SAR considerations were combined with expert judgment[2]. In this exercise, the pure SAR methods that relied solely on chemical structure, such as CASE/M-CASE, performed poorly, although many reasons, such as NTP bias towards more “difficult” chemicals already suspected of carcinogenicity and small test set, can be enumerated for this result[2]. However, when SAR modeling is confined to the structurally homogeneous set of aromatic amines, improved activity discrimination accuracies are reported in the range of 80-90% (see Chapter 4 in this volume and Benigni and Passerini[38]). It can be argued that this result places limits on possible experimental variability and error of the rodent carcinogenicity results to within a manageable range of 10-20%, at least for this chemical class and species- and gender-specific data. The FDA/M-CASE results of Matthews and Contrera[28] also indicate that improvements in model performance can be achieved with enriched training sets and refinements in weighting and categorization of rodent carcinogenicity information. In the case of

the *Salmonella* mutagenicity assay, inter-laboratory reproducibility has been estimated at 82%, with CASE/M-CASE models reportedly achieving respectable predictive concordances of 77%[25].

Data availability, quality and representation issues pertaining to mutagenicity and carcinogenicity endpoints clearly have a profound influence on SAR model development and predictive capabilities. With increasing interest in predictive toxicology technologies and new initiatives to enhance public data availability linked with chemical structure, an appreciation of the fundamental limitations and potential capabilities of SAR models in this area of toxicological study is all the more pressing. This requires some understanding of the nature of the biological data under study and the myriad ways in which these data can be pooled, categorized, and interpreted. A number of examples relative to rodent carcinogenicity data for use in SAR models, and application of the CASE/M-CASE technology, have been presented in this review to illustrate some important concepts that transcend the particulars of the toxicity endpoint or SAR technology being applied. For the SAR model developer and user alike, it is hoped that this discussion has provided some cautionary guidance in the application of SAR technologies, as well as presented an expanded view of the informatics capabilities of SAR technologies.

ACKNOWLEDGEMENTS

The authors thank Carl Blackman, Patricia Schmieder, Russell Owen, and Julian Preston for helpful comments in review of this manuscript. The editor of this volume is also thanked for his exceptional patience and faith in the ultimate completion of this chapter. Finally, we are very grateful to Lois Swirsky Gold for alerting us to some inaccurate and misleading statements made in reference to the CPDB that were addressed in an errata statement in the first printing and corrected in this and subsequent printings. This manuscript has been reviewed by the US Environmental Protection Agency and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

REFERENCES

1. Helma C., Gottmann E., and Kramer S., Knowledge discovery and data mining in toxicology, *Statis. Meth. Medical Res.*, 9, 1, 2000.
2. Richard, A.M. and Benigni, R., AI and SAR approaches for predicting chemical carcinogenicity: survey and status report, *SAR QSAR Environ. Toxicol.*, 13, 1, 2002.
3. Moudgal, C.J., Lipscomb, J.C., and Bruce, R.M., Potential health effects of drinking water disinfection by-products using quantitative structure toxicity relationship, *Toxicology*, 147, 109, 2000.
4. Enslein, K. et al., A structure-activity prediction model of carcinogenicity based on NCI/NTP assays and food additives, *Toxicol. Ind. Health*, 3, 267, 1987.
5. Woo, Y.T. et al., Use of mechanism-based structure-activity relationships analysis in carcinogenic potential ranking for drinking water disinfection by-products, *Environ. Health Perspect.,*110(suppl. 1), 75, 2002.

6. Klopman, G., Artificial intelligence approach to structure-activity studies. Computer automated structure evaluation of biological activity of organic molecules, *J. Am. Chem. Soc.*, 106, 7315, 1984.
7. Klopman, G., MULTICASE 1. A hierarchical computer automated structure evaluation program, *Quant. Struct.-Act. Relat.*, 11, 176, 1992.
8. Klopman, G. and Rosenkranz, H.S., Approaches to SAR in carcinogenesis and mutagenesis. Prediction of carcinogenicity/mutagenicity using MULTI-CASE, *Mutat. Res.*, 305, 33, 1994.
9. Benigni, R., Predicting chemical carcinogenesis in rodents: the state of the art in light of a comparative exercise, *Mutat. Res.* 334, 103, 1995.
10. Benfenati, E.and Gini G., Computational predictive programs (expert systems) in toxicology, *Toxicology*, 119, 213, 1997.
11. Benigni, R., The first US National Toxicology Program exercise on the prediction of rodent carcinogenicity: definitive results, *Mutat. Res.*, 387, 35, 1997.
12. Benigni, R., Richard, A.M., Quantitative structure-based modeling applied to characterization and prediction of chemical toxicity, *Methods* 14, 264, 1998.
13. Richard, A.M., Structure-based methods for predicting mutagenicity and carcinogenicity: are we there yet?, *Mutat. Res.*, 400, 493, 1998.
14. Greene, N., Computer systems for the prediction of toxicity: an update, *Adv. Drug Deliv. Rev.*, 54, 417, 2002.
15. Brinkhuis, R.P., Toxicology information from US government agencies, *Toxicology*, 157, 25, 2001.
16. Richard, A.M., Williams, C.R., and Cariello, N.F., Improving structure-linked access to publicly available chemical toxicity information, *Curr. Opin. Drug Discov. Devel.*, 5, 136, 2002.
17. Wexler, P., Introduction to special issue (part II) on digital information and tools, *Toxicology*, 173, 1, 2002.
18. Young, R.R., Genetic toxicology: web resources, *Toxicology*, 173, 103, 2002.
19. Junghans, T.B., Cancer information resources: digital and online sources, *Toxicology*, 173, 13, 2002.
20. Auletta, A.E. et al., Current status of the Gene-Tox Program. *Environ. Health Perspect.*, 96, 33, 1991.
21. Ashby, J. and Tennant, R.W., Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the U.S. NTP, *Mutat. Res.*, 257, 229, 1991.
22. Selkirk, J.K. and Soward, S.M., Compendium of abstracts from long-term cancer studies reported by the National Toxicology Program from 1976 to 1992, *Environ. Health Perspect.*, 101, 1, 1993.
23. Richard, A.M., Application of artificial intelligence and computational methods to predicting toxicity, *Knowledge Eng. Rev.*, 14, 307, 1999.
24. Dearden, J.C. et al., The development and validation of expert systems for predicting toxicity: the report and recommendations of ECVAM/ECB workshop 24, *Alternatives to Laboratory Animals (ATLA)*, 25, 223, 1997.
25. Macina, O.T., Zhang, Y.P., and Rosenkranz, H.S., Improved predictivity of chemical carcinogens: the use of a battery of SAR models, in *Carcinogenicity: Testing, Predicting, and Interpreting Chemical Effects*, Kitchin, K., Ed., Marcel Dekker Inc., New York, 1999, chap. 7.
26. Cunningham, A.R. et al., Identification of 'genotoxic' and 'non-genotoxic' alerts for cancer in mice: the carcinogenic potency database, *Mutat. Res.* 398, 1, 1998.

27. Cunningham, A.R. et al., Identification of structural features and associated mechanisms of action for carcinogens in rats, *Mutat. Res.* 405, 9, 1998.
28. Matthews, E.J. and Contrera, J.F., A new highly specific method for predicting the carcinogenic potential of pharmaceuticals in rodents using enhanced MultiCASE QSAR-ES software, *Regulat. Pharmacol. Toxicol.*, 28, 242, 1998.
29. Zhang, Y.P. et al., Prediction of the carcinogenicity of a second group of chemicals undergoing carcinogenic testing, *Environ. Health Perspect.*, 104, 1045, 1996.
30. Huff, J. et al., Chemicals associated with site-specific neoplasia in 1394 long-term carcinogenesis experiments in laboratory rodents, *Environ. Health Perspect.,* 93, 247, 1991.
31. Ashby, J. and Paton, D., The influence of chemical structure on the extent and sites of carcinogenesis for 522 rodent carcinogens and 55 different human carcinogen exposures, *Mutat. Res.*, 286, 3, 1993.
32. Benigni, R. and Pino, A., Profiles of chemically-induced tumors in rodents: quantitative relationships, *Mutat. Res.* 421, 93, 1998.
33. Gold, L.S. et al., Reproducibility of results in 'near-replicate' carcinogenesis bioassays, *J. Natl. Cancer Inst.,* 78, 1149, 1987.
34. Gold, L.S., Sloan, T.H., and Bernstein, L., Summary of carcinogenic potency and positivity for 492 rodent carcinogens in the Carcinogenic Potency Database. *Environ. Health Perspect.*, 79, 259, 1989.
35. Gold L.S. et al., Supplement to the Carcinogenic Potency Database (CPDB): results of animal bioassays published in the general literature in 1993 to 1994 and by the National Toxicology Program in 1995 to 1996, *Environ. Health Perspect.,* 107(suppl. 4), 527, 1999.
36. Gold, L.S. et al., Compendium of chemical carcinogens by target organ: results of chronic bioassays in rats, mice, hamsters, dogs, and monkeys, *Toxicol. Pathol.*, 29, 639, 2001.
37. Peto, R., The TD_{50}: a proposed general convention for the numerical description of the carcinogenic potency of chemicals in chronic-exposure animal experiments, *Environ. Health Perspect.*, 58, 1, 1984.
38. Benigni, R. and Passerini, L., Carcinogenicity of the aromatic amines: from structure-activity relationships to mechanisms of action and risk assessment, *Mutat. Res.*, 511, 191, 2002.
39. Gottmann, E., et al., Data quality in predictive toxicology: reproducibility of rodent carcinogenicity experiments, *Environ. Health Perspect.*, 109, 509, 2001.
40. Klopman, G. and Rosenkranz, H.S., Structure-activity relations: maximizing the usefulness of mutagenicity and carcinogenicity databases, *Environ. Health Perspect.*, 96, 67, 1991.
41. Rosenkranz, H.S. et al., Development, characterization, and application of predictive-toxicology models, *SAR QSAR Environ. Res.*, 10, 277, 1999.
42. Takihi, N. et al., An approach for evaluating and increasing the informational content of mutagenicity and clastogenicity data bases, *Mutagenesis*, 8, 257, 1993.
43. Liu, M. et al., Estimation of the optimal database size for structure-activity analyses: the *Salmonella* mutagenicity data base, *Mutat. Res.*, 358, 63, 1996.
44. Liu, M. et al., Structure-activity and mechanistic relationships: the effects of chemical overlap on the structural overlap in databases of varying size and composition, *Mutat. Res.*, 372, 79, 1996.
45. Klopman, G. and Rosenkranz, H.S., Quantification of the predictivity of some short-term assays for carcinogenicity in rodents, *Mutat. Res.*, 253, 237, 1991.

46. Rosenkranz, H.S. et al., Significant differences in the structural basis of the induction of sister chromatid exchanges and chromosomal aberrations in Chinese hamster ovary cells, *Environ. Mol. Mutagen.*, 16, 149, 1990.
47. Grant, S.G. et al., Modeling the mouse lymphoma forward mutational assay: the Gene-Tox program database, *Mutat. Res.,* 465, 201, 2000.
48. Matthews, E.J., Benz, R.D., and Contrera, J.F., Use of toxicological information in drug design, *J. Mol. Graph. Model.,* 18, 605, 2000.
49. Roberts, G. et al., Leadscope: Software for exploring large sets of screening data, *J. Chem. Inf. Comput. Sci.*, 40, 1302, 2000.
50. Bacha, P.A. et al., Rule extraction from a mutagenicity data set using adaptively grown phylogenetic-like trees, *J. Chem. Inf. Comput. Sci.*, 42, 1104, 2002.
51. Richard, A.M. and Williams, C.R., Distributed structure-searchable toxicity (DSSTox) Public Database Network: A Proposal, *Mut. Res.*, 499, 27, 2002.