The Development and Validation of Expert Systems for Predicting Toxicity

The Report and Recommendations of an ECVAM/ECB Workshop (ECVAM Workshop 24)\textsuperscript{1,2}

John C. Dearden,\textsuperscript{3} Martin D. Barratt,\textsuperscript{4} Romualdo Benigni,\textsuperscript{5} Douglas W. Bristol,\textsuperscript{6} Robert D. Combes,\textsuperscript{7} Mark T.D. Cronin,\textsuperscript{8} Philip N. Judson,\textsuperscript{8} Martin P. Payne,\textsuperscript{9} Ann M. Richard,\textsuperscript{10} Milon Tichy,\textsuperscript{11} Andrew P. Worth\textsuperscript{12} and Jeffrey J. Yourick\textsuperscript{13}

\textsuperscript{3}School of Pharmacy and Chemistry, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, UK; \textsuperscript{4}Environmental Safety Laboratory, Unilever Research, Colworth House, Sharnbrook, Bedford MK44 1LQ, UK; \textsuperscript{5}Istituto Superiore di Sanità, Viale Regina Elena 299, 00161 Rome, Italy; \textsuperscript{6}NIH/NIEHS, Research Triangle Park, NC 27709, USA; \textsuperscript{7}FRAME, Russell & Burch House, 96–98 North Sherwood Street, Nottingham NG1 4EE, UK; \textsuperscript{8}Heather Lea, Blund Hill, Norwood, Harrogate HG3 1TE, UK; \textsuperscript{9}Health & Safety Laboratory, Broad Lane, Sheffield S3 7HQ, UK; \textsuperscript{10}NHEERL, Environmental Protection Agency, Research Triangle Park, NC 27711, USA; \textsuperscript{11}Predictive Toxicology Laboratory, National Institute of Public Health, Srobarova 48, 100 42 Prague 10, Czech Republic; \textsuperscript{12}ECVAM, JRC Environment Institute, 21020 Ispra (VA), Italy; \textsuperscript{13}Cosmetics Toxicology Branch, Food & Drug Administration, 8301 Muirkirk Road, Laurel, MD 20708, USA

Preface

This is the report of the twenty-fourth of a series of workshops organised by the European Centre for the Validation of Alternative Methods (ECVAM). ECVAM’s main goal, as defined in 1993 by its Scientific Advisory Committee, is to promote the scientific and regulatory acceptance of alternative methods which are of importance to the biosciences and which reduce, refine or replace the use of laboratory animals. One of the first priorities set by ECVAM was the implementation of procedures which would enable it to become well-informed about the state-of-the-art of non-animal test development and validation, and the potential for the possible incorporation of alternative tests into regulatory procedures. It was decided that this would be best achieved by the organisation of ECVAM workshops on specific topics, at which small groups of invited experts would review the current status of various types of tests and their potential uses, and make recommendations about the best ways forward (1).

The joint ECVAM/ECB (European Chemicals Bureau) workshop on The Development and Validation of Expert Systems for Predicting Toxicity was held in Angera, Italy, on 1–4 October 1996, under the chairmanship of John Dearden (Liverpool John Moores University, Liverpool, UK). The workshop participants reviewed the current status of a variety of systems for predicting toxicity, focusing primarily on the prediction of endpoints relevant to mammalian (including...
human) toxicity. As a result of the discussions which took place, recommendations were made for the further development of expert systems. In addition, general principles were sought for the validation of expert systems, as were criteria for the acceptability of data to be used in their development and validation. In this report, the expert systems discussed at the workshop are critically reviewed; specific conclusions and recommendations, referring to the individual systems, are presented in the relevant sections of the report, whereas general conclusions and recommendations are summarised in the final section.

Introduction

The phrase “expert system for predicting toxicity” is used in various ways in the literature, so it is defined, for the purposes of this report, as follows:

“An expert system for predicting toxicity is considered to be any formalised system, not necessarily computer-based, which enables a user to obtain rational predictions about the toxicity of chemicals. All expert systems for the prediction of chemical toxicity are built upon experimental data representing one or more toxic manifestations of chemicals in biological systems (the database), and/or rules derived from such data (the rulebase).”

The database and rulebase associated with an expert system enable the user to rationalise individual model predictions, and to define the limitations associated with them; however, they are not supplied with all expert systems.

Individual rules within the rulebase are generally of two main types. Some rules are based on mathematical induction, that is, by the extraction of correlations from a particular data set, whereas other rules are based on existing knowledge and expert judgement. Rules of the former type, “induced rules”, offer the advantage of extending existing knowledge without being biased toward particular mechanisms of toxic action. Their disadvantage, however, is that they may be nothing more than empirical relationships, devoid of biological meaning. In contrast, rules of the latter type, “expert rules” or “knowledge-based rules”, are likely to have a strong mechanistic basis, but they are expressions of existing knowledge rather than of new knowledge. Typically, induced rules are (quantitative) structure-activity relationships ([Q]SARs), whereas expert rules are often based on reactive chemistry. Rules can also be formulated on the basis of biokinetic models and molecular models of ligand-receptor interactions, although such rules do not fit neatly into the above categories, and should perhaps be classified separately.

Expert systems are sometimes characterised according to the nature of the rules in their rulebase. An expert system based on induced rules is called an “automated rule-induction system” or a “correlative system”, whereas a system based on expert rules is referred to as a “knowledge-based system”. The differences between the two types of expert system are discussed in more detail elsewhere (2, 3). In practice, distinguishing between the two types of expert systems may be difficult, since both types of rules can be present in their rulebases.

One of the aims of this workshop was to assess the strengths and weaknesses of various expert systems which are either commercially available or are in a stage of pre-market development. It is hoped that the critiques, which are summarised in different sections of this report, will feed back into the ongoing development of these systems. Three presentations were given on DEREK, focusing on different endpoints: Martin Barratt (Unilever Environmental Safety Laboratory, UK) discussed its performance in the prediction of skin sensitisation potential; Martin Payne (Health and Safety Executive, UK) considered irritancy and corrosivity; and Bob Combes (FRAME, UK) concentrated on mutagenicity and carcinogenicity. The CASE technology, with particular reference to ToxAlert, was discussed by Mark Cronin (Liverpool John Moores University, UK), who also presented information about COMPACT. Jeffrey Yourick (Food and Drug Administration [FDA], MD, USA) assessed the merits of TOPKAT, while Milon Tichy (National Institute of Public Health, Czech Republic) assessed those of Hazardexpert. Ann Richard (Environmental Protection Agency [EPA], NC, USA) discussed Oncologic and the Purdy Method. An introduction to the StAR system was given by Philip Judson (Judson Consulting, UK), who also
described recent research on toxicophore discovery systems (REX, DTOX and PRO-GOL). Moving away from individual systems, the relative performances of various approaches, as they have emerged from the evaluation undertaken under the auspices of the US National Toxicology Program (NTP), were described by Romualdo Benigni (ISS, Italy). Finally, the integrated use of predictive techniques in chemical risk assessment was addressed both by Douglas Bristol (National Institute of Environmental Health Sciences, NC, USA), who focused on different kinds of models, and by Ann Richard, with emphasis on expert systems.

DEREK

Deductive Estimation of Risk from Existing Knowledge (DEREK) is a knowledge-based system for the qualitative prediction of a range of toxic endpoints (4, 5). It runs on a VAX or UNIX system. It was originally devised at Schering, but is now being developed and marketed by LHASA UK (School of Chemistry, University of Leeds, UK) and Harvard University (Boston, MA, USA). The following endpoints are covered: skin sensitisation, respiratory sensitisation, irritancy, corrosivity, mutagenicity, carcinogenicity, teratogenicity, neurotoxicity, lachrymation, methaemoglobinemia, and anticholinesterase activity.

DEREK has several rulebases, consisting of descriptions of molecular substructures (structural alerts), which have been associated with toxic endpoints on the basis of existing knowledge. The rules are generic in nature; that is, they are based on sets of related chemicals rather than on specific chemicals, and most of them are derived from mechanistic organic chemistry. The ongoing development of the rules is monitored by the DEREK Users Group.

The user communicates with DEREK by drawing the two-dimensional topographical structure of the query molecule on the screen. Then, the rulebases are searched, any structural alert located within the query structure is highlighted, and a message indicating the nature of the toxicological hazard is provided. For many of the structural alerts, the hazard evaluation is justified with relevant literature references.

Performance of the DEREK skin sensitisation rulebase

If a chemical is capable of reacting with a nucleophilic moiety on a protein, either directly or after appropriate (bio)chemical transformation, it has the potential to be a skin sensitiser (contact allergen), providing that it is able to locate in the appropriate epidermal compartment (6). The potential of a chemical to act as a contact allergen is modulated by its ability to penetrate the skin, as supported by QSAR studies (for example, 7) which show that skin sensitisation potential depends on physicochemical parameters which are equally important determinants of percutaneous absorption (8). One such parameter is log P (where P is the octanol/water partition coefficient). In general, the higher the log P value (that is, the greater the lipophilicity), the greater the skin permeability, although at very high log P values the permeability starts to decrease again.

Originally, the DEREK skin sensitisation rulebase contained about 40 structure-activity rules (9, 10); it now contains 50 rules, as a result of subsequent additions and refinements. The performance of the rulebase has been tested on various data sets, including the list of substances proposed by the European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC) for evaluating novel tests (11), as described previously in the report of the ECVAM workshop on skin sensitisation testing (12).

Another data set which has been processed through the DEREK skin sensitisation rulebase (M.D. Barratt & J.J. Langowski, poster presentation at the 36th Annual Meeting of the Society of Toxicology, Cincinnati, OH, USA; March 1997) is the list of chemicals published by the Bundesinstitut für gesundheitlichen Verbraucherschutz und Veterinärmedizin (BgVV; 13). This list contains 84 contact allergens (65 strong sensitisers, 13 moderate sensitisers and six weak sensitisers). When the structures were first processed through DEREK, structural alerts were found in 71 of them, representing a “hit rate” of 85%. Of the remaining 13 structures, 12 had no structural alert for skin sensitisation, and one, an organomercury compound, could not be processed. A mechanistic rationale can be found for ten of the 12 contact allergens which did not contain structural alerts; these ten chemicals comprised three photoallergens, four organic hydroperoxides,
one phenolic derivative, one thiol-exchange agent and one hydroxylamine precursor. The reactivity of the other two contact allergens, diphenylcyclopropenone and N-vinylcarbazole, has yet to be explained. Minor modifications and additions to the rulebase have been proposed (M.D. Barratt & J.J. Langowski, poster presentation at the 36th Annual Meeting of the Society of Toxicology, Cincinnati, OH, USA; March 1997), which would increase the coverage of the rulebase to 81 out of the 84 contact allergens, giving a hit rate of 96%.

Performance of the DEREK rulebase for irritancy and corrosivity

The DEREK rulebase has nine rules for the prediction of irritancy (4); two rules are specific to eye irritancy, but none is specific to skin irritancy or corrosivity. To assess the performance of this rulebase, a combined data set of about 300 chemical structures has been employed (J. Morley, MSc research project, Health and Safety Laboratory, Sheffield, UK). The irritancy information was taken from the following sources: a) published EC classifications of new or existing chemicals; b) the ECE-TOC reference data banks for eye and skin irritation (14, 15), which have been recommended for use in the validation of alternative methods; and c) a large set of eye irritancy data obtained in a single laboratory (16). Most of the chemicals (75%) were labelled/identified either as significant skin or eye irritants, or as corrosives, on the basis of irritation test scores. The remaining chemicals were either not classified, or were identified as weak irritants, on the basis of irritation scores.

Of the chemicals labelled as skin and eye irritants, which constituted distinct, but overlapping, subsets, only 28% and 30%, respectively, were predicted by DEREK to be irritants. Of the new chemicals not labelled as skin or eye irritants, 19 out of 45 were predicted to be irritants. For many of the structurally simple chemicals in the combined data set, irritancy could be predicted or explained on the basis of chemical reactivity, although, in certain cases, other mechanisms had to be invoked, such as lipid extraction. For the structurally complex chemicals which DEREK failed to predict as irritants, putative toxicophores were recognisable in less than half of them.

DEREK predicted only 20–40% of the corrosive chemicals to be irritants. Chemicals which were not predicted included: a) chemicals known to be highly reactive with water or biological molecules, such as anhydrides, aluminium alkyls and peroxy acids; and b) chemicals for which the corrosivity is attributable to surfactant properties. Rules for the first group, but not the second, could easily be written into DEREK, thereby improving its performance.

In summary, the predictive capability of DEREK for irritancy and corrosivity is poor compared to its predictive capacity for skin sensitisation. Although the mechanisms of irritancy and corrosivity are only partially understood, the critical (rate-limiting) step for these endpoints (that is, the step which needs to be modelled) may be an initial physical effect on cell membranes (such as an alteration of their electrical properties), rather than a receptor-mediated interaction. Therefore, given that DEREK cannot take physicochemical properties into account, it is questionable whether this system is an appropriate platform for predictions of irritancy and corrosivity.

Performance of DEREK in predicting mutagenicity and carcinogenicity

Many structural factors affect the mutagenicity and carcinogenicity of chemicals including: a) intrinsic reactivity (electrophilicity); b) the electron density in, and near, reactive centres; c) substituent effects (for example, steric hindrance); d) susceptibility to metabolic activation and detoxification (for example, the balance between C-hydroxylation and N-hydroxylation); e) the stability of reactive forms of chemicals; f) the ability of chemicals and their metabolites to traverse biological membranes; g) the size and shape of molecules which control access to target sites on DNA; h) the type and conformation of adducts formed between the chemical and DNA; i) the susceptibility of the adduct to DNA repair (error-prone or error-free); and j) the ability to reach and react with other cellular targets (in the case of non-genotoxic carcinogens). Ideally, expert systems should take all of these factors into account when assessing the activities of mutagens and carcinogens. For example, systems which only incorporate rules for the identification of electrophilicity have a limited ability to predict the carcinogenicity of non-genotoxic chemicals.
The ability of DEREK to predict mutagenicity has been evaluated by using the NTP database (17). The performance has increased with successive versions of the system, as a result of improvements made to the rulebase by the DEREK Users Group (5). This is illustrated by the fact that the latest version of DEREK predicted 98% of the mutagens and 70% of the non-mutagens, whereas the original version identified only 41% of the mutagens and 76% of the non-mutagens.

The performance of DEREK in predicting mutagenicity has also been assessed by GlaxoWellcome, using 311 chemical intermediates for which Salmonella data were available (91 mutagens; 220 non-mutagens) and all of the relevant rules in the DEREK rulebase (D.J. Wedd, personal communication). For many of the false positive assessments made by DEREK, the identified toxicophore was located in a ring structure, and not in a side chain. For some of the false negatives, the Ames data were either weakly positive or equivocal, so the bacterial mutagenicity in these cases may have been due to sample contaminants, which were not considered by DEREK. The overall accuracy (sensitivity and specificity) of the DEREK predictions was 61%.

DEREK has been assessed for its potential as a screen for predicting the mutagenicity and carcinogenicity of certain chemicals found in foods (18). Examples of toxicophores identified included: a) for aflatoxins, a bisfuranoid substructure centred around the 2,3-vinyl ether bond; and b) for polyaromatic hydrocarbons, a six-bond substructure, based on the “bay” region of phenanthrene. The effects on mutagenicity of substituents and extra rings were also considered.

The ability of DEREK to predict carcinogenicity has been assessed by Achampong et al. (unpublished information), using 498 of the 522 rodent carcinogens in the carcinogenic potency database of Gold and colleagues (for example, 19, 20). Two different rulebases were used for the predictions: a) the in-house carcinogenicity rulebase, in which the rules are based on the structural alerts of Ashby & Paton (21); and b) the US FDA rulebase. The overall sensitivities were 40% for the in-house rulebase and 75% for the FDA rulebase. A combination of the two rulebases resulted in a sensitivity of 80%. It was concluded that new DEREK rules are required to predict the carcinogenicity of the following chemical groups: a) safrole and its derivatives; b) halogenated alkanes and halogenated alkenes; and c) non-genotoxic carcinogens in general. In addition, modifications to existing rules are needed for the following chemical groups: a) hydrazine and its derivatives; b) thioamides and thiouracils; and c) \( \alpha,\beta \)-unsaturated carbonyl compounds.

**Strengths of DEREK**

1. The users of DEREK are also its developers.
2. It is readily customised (for example, with respect to the alteration of on-screen colours, highlighting of toxicophores, and representation of molecular structures).
3. The system is “transparent”: on-screen messages provide information as to why a rule is fired, and there are plans to incorporate explanations when rules are not fired.
4. There is considerable flexibility for adding data to the database.
5. The user can access the information used to formulate the rulebase, that is, the database, references, and other supporting information (including relevant statements).
6. Rule modification can be undertaken by someone with the necessary programming skills, and there are safeguards against the illicit alteration of information.
7. It can be used in batch mode.

**Limitations of DEREK**

1. The lack of default display of C and H atoms can be confusing to the non-chemist.
2. The calculation of descriptors for overall size, connectivity and steric effects is not automatic. Although chemicals are not represented in three dimensions, their processing is topologically based, and particularly important structural features can be represented, for example, in the form of wedge bonds.
3. Physicochemical properties are not automatically calculated. However, it is possible for these to be determined by other systems (see recommendation 1, below).

4. It does not account for the activating and detoxifying effects of metabolism.

5. The application of three-dimensional QSARs, reflecting receptor-based mechanisms, is currently outside the scope of DEREK. Nevertheless, the use of an extended knowledge-base could enable the identification of close analogues to receptor-based toxicants, thereby alerting to the possibility of toxicity in such instances. In addition, the StAR system could provide a means of using conformation-based SARs, should these become available (see later).

Recommendations

1. The usefulness of DEREK rules for identifying the toxicity of novel, untested structures could be enhanced by supplementing them with physicochemical data and other information, such as that provided by computational chemistry.

2. New rules are required to predict the irritancy of: a) highly reactive chemicals, such as anhydrides, aluminium alkyls and peroxy acids; and b) corrosive surfactants.

3. For certain classes of reactive chemicals, irritancy rules could be formulated on the basis of reported effects in humans, but the applicability of such rules should be restricted to the range of known active chemicals.

4. The current irritancy rule for anilines needs to be reformulated by relating the irritancy of these aromatic amines to their polarity, rather than to their basicity.

5. For more-accurate predictions of irritancy and corrosivity, new rules are needed to express variations in potency due to the effects of solubility and molecular size on skin and eye penetration, particularly for less reactive chemicals which exert their effects on non-superficial cells. The formulation of such rules could be guided by structure-permeability relationships derived for skin absorption (for example, 8, 22).

The CASE Technology

The Computer Automated Structure Evaluation (CASE) technology refers to a range of different programs supplied by MULTICASE Inc. (Cleveland, OH, USA), known as ToxAlert, CASE, Multi-CASE, and CASE-TOX. ToxAlert is a relatively simple, PC-based program, running under Microsoft Windows, whereas CASE and its derivatives are more complex VAX-based programs, having access to quantum chemical and graph indices, and geometry-based variables which ToxAlert cannot access.

The CASE methodology is described in detail elsewhere (23–25), and examples of its applicability have been published (26, 27). Briefly, it requires a large number of high quality, heterogeneous biological data for the endpoint in question, covering a wide range of activity. Typically, the biological data are obtained from published databases, such as the NTP database. The data are then converted by the CASE system into scalar units, in the range of 10–99; values 10–19 indicate inactivity, 20–29 indicate marginal activity, and 30–99 indicate increasing activity. This scaling is used even for qualitative endpoints (that is, endpoints reported as positive or negative).

The unique aspect of CASE is the creation and detection of structural alerts. To achieve this, the structure of each molecule is divided up into all possible fragments, from two heavy (non-hydrogen) atoms in length (for example, CH₂-OH for an aliphatic hydroxy group) to potentially any number of atoms (although, in practice, the use of fragments greater than six atoms is unwieldy). Statistical methods are then used to classify the fragments as biophores or biophobes, according to whether they are associated with the biological activity of interest, or no activity, respectively. The fragments are then combined to give an equation of the following form (25):

\[
\text{CASE units} = \text{constant} + a \text{[Fragment 1]} + b \text{[Fragment 2]} + \ldots
\]

(Equation 1)

In original versions of the CASE programs, the fragments were restricted to linear chains of atoms and bonds. In more recent versions, it has been possible to include branched fragments in the analysis. Indirectly, the CASE programs take some
account of overlap (and hence interaction) between fragments, because larger fragments automatically encompass smaller ones (for example, N-C-C-O-C-C-S contains N-C-C-O and O-C-C-S). Nevertheless, the CASE programs often fail to distinguish, for example, molecules containing several small chains within one complex fragment from other molecules containing the same fragments distributed separately.

**Strengths of the CASE approach**

1. CASE identifies structural fragments which alert for toxicity without reference to mechanisms of action. This may be useful for: a) the analysis of data sets with diverse and unknown mechanisms of action; and b) the automated analysis of large data sets, for example, during high-throughput screening.

2. It can be used to detect biophores for non-genotoxic carcinogens.

3. It can predict metabolic pathways by identifying structural features predisposing molecules to particular reactions.

**Limitations of the CASE approach**

1. CASE attaches no mechanistic significance to the fragments, although it is possible for the user to do so retrospectively. Some of these fragments may be chance correlations and thus irrelevant to the toxicity.

2. The biological data are converted into unitless values (CASE units), which are difficult to interpret in a truly quantitative manner.

**ToxAlert**

ToxAlert provides predictions of a number of toxicity endpoints, including chromosomal aberration, micronucleus induction, *Salmonella* mutagenicity, rodent carcinogenicity and rat teratogenicity. It also predicts biodegradability and calculates physicochemical parameters (log P and aqueous solubility). The ToxAlert system uses models developed according to the CASE methodology and can therefore be considered to be a front end for Equation 1. Once a molecular structure has been entered in the form of a simplified molecular input line system entry (SMILES) code, ToxAlert identifies the appropriate structural fragments and calculates a toxicity value from the relevant equation. The conclusions which follow are based on results obtained with a Beta version of the program, kindly provided by Professor Gilles Klopman (Case Western Reserve University, Cleveland, OH, USA).

**Strengths of ToxAlert**

1. ToxAlert is easy to use compared to the other CASE programs.

2. It predicts endpoints which can be used for the assessment of genotoxicity, such as *Salmonella* mutagenicity, micronucleus induction and chromosomal aberration.

3. It can be operated in batch mode.

**Limitations of ToxAlert**

1. ToxAlert does not indicate the types of molecules from which the structural fragments may have originated. This would be useful in assigning some level of confidence to the predictions, that is, in checking whether the molecular environments of the structural fragments are the same. These features can be accessed in the full CASE and Multi-CASE programs.

2. Although the fragments identify structural features associated with particular toxicity endpoints, they are not sufficiently refined to take account of the molecular environment. Thus, no account is taken of the modulation of activity resulting from steric hindrance, or from the presence of electron-withdrawing or electron-donating groups. In contrast, Multi-CASE distinguishes between fragments which are responsible for the basic activity and those which merely modulate this activity.

3. The database and rulebase cannot be searched.

4. It does not enable the user to modify its models, or to perform a CASE analysis on new data sets.

**Recommendations**

1. The artificial intelligence approach of the CASE methodology should form part of a battery of expert systems for predicting toxicity.
2. Greater attention should be paid to the meaning of the structural fragments generated, especially for endpoints which are reasonably well understood, such as mutagenicity and skin sensitisation. Structural fragments could be cross-referenced with those from other systems (for example, mechanistic approaches such as DEREK, and automated procedures such as TOPKAT).

3. The ToxAlert system could be extended in the following ways: a) by providing information on given structural fragments (such as the identities of the molecules which contain them); and b) by incorporating QSARs relating to specific fragments, thereby including a means to allow for any modulation of the toxicity.

COMPACT

Computer-Optimised Molecular Parametric Analysis of Chemical Toxicity (COMPACT) is a methodology developed by Lewis et al. (School of Biological Sciences, University of Surrey, UK). It can be used on a variety of platforms (including PC, VAX and UNIX systems) to predict the potential of a chemical to act as a substrate for one or more of the cytochromes P450 (P4501, P4502B, P4502E, and P4504 [28, 29]). COMPACT can also be used to identify chemicals which have the potential to bind to receptors involved in: a) the induction of cytochromes P450 (for example, binding to the Ah receptor induces P4501); and b) peroxisome proliferation (for example, the ppar receptor).

In general, oxidative metabolism by the cytochromes P450 results in detoxification, since the insertion of oxygen into a molecule makes it more polar and more susceptible to conjugation, and therefore it is more likely to be eliminated from the body by virtue of its increased water solubility. However, in certain cases, metabolic activation occurs, giving rise to a carcinogen, pro-carcinogen, or some other kind of reactive intermediate. For example, P4501 converts polycyclic hydrocarbons to epoxides, which act as pro-carcinogens.

Basis of the toxicity prediction

The COMPACT methodology is based on the hypothesis that the structural characteristics of a molecule determine its ability to fit into the appropriate binding site on an enzyme (cytochrome P450), whereas its electronic characteristics determine whether the resulting enzyme-substrate complex can be activated for oxidative metabolism. Thus, a combination of structural and electronic parameters is used to predict whether a particular molecule is likely to act as a P450 substrate. The toxicological consequences of the P450-mediated oxidation will depend on the toxic properties of the oxidation product, and on those of any subsequent metabolites.

The basis of the COMPACT methodology is supported by molecular modelling studies which display graphically the interaction between a given cytochrome P450 and one of its substrates. To do this, it is first necessary to model the three-dimensional structure of the cytochrome P450 by amino acid sequence homology with a bacterial cytochrome (for example, 30) whose crystal structure is known. In a similar way, it has been possible to model interactions with receptors involved in the induction of peroxisome proliferation (31).

The structural parameter “molecular planarity” is defined as:

\[
molecular\ planarity = \frac{a}{d^2}\]

(Equation 2)

where \(a\) is the molecular cross-sectional area and \(d\) is the molecular depth. High values for molecular planarity are characteristic of planar molecules. The parameter is considered to be an important determinant of binding to the Ah receptor and to the active sites of P450 isozymes; it has been used to explain the differences in activity exhibited by isomeric pairs, such as benzo[a]pyrene and benzo[e]pyrene, or 2-acetylaminofluorene and 4-acetylaminofluorene (32).

The electronic parameter used in COMPACT studies is the “electronic activation energy” \((\Delta E)\). This is defined as the difference in energy between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO), that is, \(E(LUMO) - E(HOMO)\). The \(E(LUMO)\) of a molecule represents its ability to accept electrons, whereas the \(E(HOMO)\) value represents its ability to donate electrons. The \(\Delta E\) value of a molecule, expressed in electron volts (eV), is a measure of its propensity for metabolic activation: the smaller the (positive) value, the more susceptible the mole-
molecule is to activation. Metabolic activation by P4502E, for example, requires $\Delta E \leq 15.5$. The $\Delta E$ value may also indicate the ability of a molecule to be activated by oxygen (29), and has been suggested to be a gross measurement of molecular stability (33).

The “COMPACT radius”, derived from Equation 2 and $\Delta E$, provides a means of discriminating P4501 substrates from substrates of other P450 cytochromes, and thus carcinogens from non-carcinogens (assuming that most carcinogenic chemicals are likely to be activated by P4501). The COMPACT radius is defined by Equation 3:

\[
\text{COMPACT radius} = \sqrt{(\Delta E - 9.5)^2 + \left(\frac{a}{d^2} - 7.8\right)^2}
\]  
(Equation 3)

P4501 specificity is predicted for COMPACT radii $\leq 15.5$. P4501 specificity can also be determined with the “COMPACT ratio”, defined by Equation 4:

\[
\text{COMPACT ratio} = \frac{a}{d^2} \times \Delta E
\]  
(Equation 4)

COMPACT ratios above 0.25 are indicative of P4501 specificity, whereas values below 0.15 are indicative of other P450 specificities. All of these equations and cut-off values are dependent on the theoretical method used.

The combined use of the structural and electronic parameters is illustrated in Figure 1, in which the molecular planarity is plotted against $\Delta E$ (28). The substances above the arc are substrates for P4501, and are therefore considered to be carcinogens, whereas the substances below the arc are substrates for the other cytochrome P450 isozymes.

Several successful correlations have been obtained by using the above mentioned para-

\[TCDD = \text{tetrachlorodibenzo-p-dioxin}\]

Adapted from Parke et al. (28).
meters as QSAR descriptors; for example, the ∆E value is proportional to the bacterial mutagenicity of amines found in cooked foods (34), and the COMPACT radius and log P correlate with the induction of known P4501 inducers, several of which are carcinogens (35).

**Performance of COMPACT**

The ability of COMPACT to predict carcinogenicity has been assessed in several studies by using data obtained from rodent carcinogenicity bioassays. For example, in a prospective evaluation of 40 chemicals tested by the NTP, COMPACT gave a concordance of 72% (36), and in a study of 100 chemicals from the NTP database the degree of correlation was 92% (37). In a later study, using 80 NTP chemicals (56 carcinogens; 24 non-carcinogens), COMPACT predicted both carcinogens (71%) and non-carcinogens (67%) with a similar level of accuracy, leading to an overall concordance of 70% (38). Finally, a recent validation study based on more than 200 NTP chemicals also found a concordance of 70% (D.F.V. Lewis, personal communication). The correct prediction of the carcinogenicities of some of these chemicals indicates that the COMPACT approach can be used to identify direct-acting carcinogens as well as chemicals requiring metabolic activation by the P450 system.

**Strengths of COMPACT**

1. The COMPACT methodology is based on a sound mechanistic background and is supported by molecular modelling studies.
2. It is based on easily and quickly calculated physicochemical properties.
3. It models the whole molecule, rather than fragments.
4. COMPACT is applicable to all classes of organic chemicals, up to and including those with 150 atoms per molecule.
5. It is a potentially transferable technology; for example, it is currently being imported into the SYBYL molecular modelling software.

**Limitations of COMPACT**

1. COMPACT is not a stand-alone toxicity prediction system. It can detect chemicals which are metabolised by cytochrome P450 isozymes, but these predictions need to be supplemented with additional information to assess the toxicological significance of the oxidative metabolism. This requires a degree of expert knowledge.
2. It can identify some direct-acting carcinogens, but not all of them.

**Recommendations**

1. COMPACT is best used within a battery of expert systems for predicting toxicity.
2. It should be further investigated for use as a prescreen to identify those chemicals which may be activated to carcinogens and other types of toxicants.

**TOPKAT**

Toxicity Prediction by Computer Assisted Technology (TOPKAT), developed by Health Designs Inc. (Rochester, NY, USA), is a PC-based system for the prediction of a range of acute and chronic toxicity endpoints. Each TOPKAT module consists of a specific database and several validated chemical subclass QSAR regression models for predicting a specific toxicity endpoint. Currently available modules comprise: rodent carcinogenicity, Ames mutagenicity, developmental toxicity potential, skin and eye irritation, acute oral toxicity LD50, acute inhalation toxicity LC50, acute toxicity LC50, acute toxicity EC50, maximum tolerated dose (MTD), chronic lowest-observable-adverse effect level (LOAEL), skin sensitisation, and log P.

A TOPKAT prediction is generated through several sequential steps. The user enters the chemical structure as a SMILES code and selects the relevant prediction module. The program screens the test structure against the substructural library for that model, determines whether the test molecu-
lar structure is “covered” by the model, and then formulates the toxicity prediction.

**Basis of the toxicity prediction**

TOPKAT predictions are derived by using the concept of linear free energy relationships in a statistical regression analysis structure (39). The TOPKAT models were developed by using topology-based structural descriptors which were shown by TOPKAT’s developers to be statistically comparable to the molecular orbital methods used in the more traditional QSAR approaches (39). Specific descriptors include: a) electronic properties (charge, electron density, residual electronegativity, effective polarisability); b) connectivity descriptors quantifying topological features; c) shape descriptors (kappa shape indices); and d) substructure descriptors selected from a library of about 3000 molecular fragments. TOPKAT uses these descriptors to define a toxic measure associated with a specific chemical substructure (one and two atom fragments; 40).

The TOPKAT program uses either continuous or dichotomous measures to predict toxicity endpoints. For continuous measures, such as LD50, LC50, EC50, MTD and LOAEL, TOPKAT uses linear multiple regression equations to generate predictions. For dichotomous measures, such as carcinogenicity, mutagenicity, developmental toxicity potential, skin irritation, eye irritation, and oral LD50, TOPKAT uses two-group linear discriminant regression functions to generate endpoint predictions. Toxicity predictions are reported in weight/weight or weight/volume units for continuous measures, and as a probability of a positive test result of 0–1 for dichotomous measures (40).

A judgement on the confidence of individual TOPKAT predictions can be made by performing a similarity search in combination with a check of whether the prediction is within the optimum prediction space (OPS). TOPKAT uses the degree of structural coverage as a parameter in determining the confidence of the prediction. In TOPKAT 3.0, a parameter defined as the OPS enables the user to ascertain whether the test structure is contained in the model descriptor space (Computational Toxicology NEWS 18, Health Designs Inc.). TOPKAT assigns little confidence to a toxicity prediction which lies outside the OPS. To evaluate further the confidence which can be placed on a TOPKAT prediction, the user can search the database. A “similarity search” enables the user to check the performance of TOPKAT in predicting the effects of a chemical which is structurally similar to the test structure. The user is also given literature references to the original sources of information. This is particularly useful if a discrepancy exists between the TOPKAT prediction and the experimental findings. In TOPKAT 1.5, the degree of confidence which can be placed on a particular prediction is determined by a more manual and subjective process; the user determines confidence from the degree of chemical structure coverage (41).

**Performance of TOPKAT**

The quality of the data (literature-derived) which are available from toxicity studies undertaken in the past is an important consideration in the determination of TOPKAT’s reliability. The FDA’s Office of Cosmetics and Colors has initiated a pilot project to evaluate six of the TOPKAT toxicity modules: carcinogenicity, mutagenicity, developmental toxicity potential, skin irritation, eye irritation, and oral LD50. Prior to the inclusion of study data in the evaluation test sets for this project, the “quality of data” was classified according to pre-established criteria. These criteria were developed from the proposed guidelines for toxicity testing described in the FDA publication Toxicological Principles for the Safety Assessment of Direct Food Additives and Color Additives used in Food, commonly referred to as Red Book II (42). No Red Book II guidelines exist for skin or eye irritation. Each study was scored (A, B or C) according to how well it conformed to the Red Book II guidelines: a grade A study met at least 80% of the proposed guidelines for a particular toxicity test; a grade B study met at least 50% of the proposed guidelines; and a grade C study satisfied fewer than 50% of the guidelines. Only grade A studies were selected from the literature for inclusion in the evaluation test sets. Once the prediction was made by TOPKAT, the user had to determine the confidence (similarity search and OPS) in the prediction. The results for TOPKAT’s performance were compiled as module summaries, by noting whether or not the experimental results were consistent with the TOPKAT predictions.

Good progress has been made on the evaluation of several TOPKAT toxicity
modules. To date, the majority of work has been with cosmetic ingredients and colour additives, although non-cosmetic chemicals are now being added to the test sets. Only a few of the initial chemicals selected from the literature for inclusion in the evaluation test sets have actually been retained in the final test sets, because: a) determination of TOPKAT performance (that is, specificity and sensitivity percentage calculations) has been limited to grade A studies; b) many of the chemicals selected from literature sources were found to be already present in the TOPKAT model databases; and c) many of the TOPKAT predictions were found to be outside the OPS. As a result of these factors, relatively few positive results from toxicity studies could be included in the evaluation test sets, so it was relatively easy to determine specificity percentages for the modules, but difficult to determine sensitivity percentages.

**Strengths of TOPKAT**

1. TOPKAT offers several different toxicity prediction modules.
2. It requires little time to run several different endpoint predictions, once the chemical structure has been entered in the form of a SMILES code.
3. The TOPKAT modules contain several different chemical structure sub-models. The specific chemical class sub-model is automatically chosen by the program, as part of the initial prediction run.
4. After running a prediction, TOPKAT informs the user whether the prediction is within the OPS or not, reflecting the degree of confidence to assign to that prediction.
5. TOPKAT enables the user to access the databases used in generating the models. Once a prediction has been made for a query chemical, the user can do a similarity search for chemical structures which were used to define the specific model employed. The program ranks the TOPKAT database chemicals according to their degree of similarity with the query structure, enabling the user to see how TOPKAT performs in predicting the effects of chemicals which are similar to the query structure.

**Limitations of TOPKAT**

1. Some chemical structure assessments (long-chain aliphatics, polymers and complex ring structures) are not well covered by certain TOPKAT modules. As the coverage is completely model-dependent, a structure (and its associated prediction) may be outside the OPS for one endpoint, but within the OPS for another. It is not possible to generate TOPKAT predictions for chemical structures comprising salts, inorganics or enol-keto forms, or for iodinated chemicals.
2. The applicability of TOPKAT predictions is limited to the criteria originally selected by TOPKAT's developers for model development, that is, their criteria for inclusion of toxicity studies (for example, species, sex, dose, route, and endpoints measured). For the purpose of validation, it is difficult to find carcinogenicity and developmental toxicity studies on chemicals not already contained in the TOPKAT databases (especially those with positive outcomes) of sufficiently high quality to meet the original model development criteria.
3. TOPKAT assumes that substructural features contribute independently to biological activity, which is not always the case.

**Hazardexpert**

Hazardexpert, and its sister program Metabolexpert, are PC-based systems developed by CompuDrug Chemistry Ltd (Budapest, Hungary). Hazardexpert predicts a range of health hazards relating to organic chemicals (43), whereas Metabolexpert predicts likely metabolites. Although they evolved as separate packages, a shortened version of Metabolexpert has been incorporated into Hazardexpert.

Predictions can be obtained for the following endpoints: mutagenicity, carcinogenicity, teratogenicity, irritation, sensitisation, immunotoxicity, and neurotoxicity. The user then selects the chemical of interest from the database; if the chemical is missing, the user must enter it into the database before retrieving it. Then, the user defines the species, route of administration, dose level, and duration of exposure. Each endpoint is
predicted as one of four levels of toxicity, taking into account the effects of bioavailability and bioaccumulation. Substructures which may exert a positive or negative modulatory effect are identified. In addition, several physicochemical parameters are calculated, with reasonable accuracy: molecular weight, pKa (the negative log of the acid dissociation constant), log P, and log D (where D is the distribution coefficient at a specified pH). Finally, the user can search abstracts published in the journal *Quantitative Structure-Activity Relationships* by entering keywords.

Hazardexpert works by searching the query structure for known toxicophores; these are held in the “Toxic Fragments Knowledge Base”, which is based on literature in the QSAR field and on reports by the US EPA. The identification of a toxicophore leads to estimates of the toxicity endpoints by triggering rules in the knowledge-bases. The rules describe toxic segments and their effects on various biological systems, and are based on the combined use of toxicological knowledge, expert judgement, QSAR models, and fuzzy logic (which simulates the effects of different exposure conditions).

The chemicals database is accessible to the user. Substances already in the database can be modified, renamed or deleted; new substances can be added by using a graphical interface or by incorporating the metabolites predicted by Metabolexpert. Similarly, the knowledge-bases on metabolic transformation, toxic fragments and references can be modified, by using the “Knowledge Maintenance Module”. In contrast, the log P and pKa databases cannot be altered.

**Performance of Hazardexpert**

As a test of its ability to predict human and animal carcinogenicity, 192 agents evaluated in the *IARC Monographs* (Volumes 1–42) were processed through Hazardexpert. The difference between the classification in the IARC list and that assessed by Hazardexpert was calculated and used for the analysis. About 55% of the compounds were predicted within ± 1 of the IARC classification when the “high” exposure condition was chosen, compared to about 75% when the “low” exposure condition was chosen.

Important fragments were found to be missing from the toxic fragments database. For example, benzene and various polyaromatic hydrocarbons were not found to be carcinogenic by Hazardexpert, but some of their metabolites, such as the epoxide derivative of benzene, were correctly predicted to be carcinogenic. In addition, Hazardexpert was unable to make predictions about: a) vinyl chloride; b) organophosphates; c) organometallic compounds; d) isocyanates; and e) conjugated (Phase II) metabolites.

In a separate evaluation study based on 80 NTP chemicals (56 rodent carcinogens; 24 non-carcinogens), Hazardexpert was found to be good at identifying non-carcinogens (the specificity was 81%), but poor at identifying carcinogens (the sensitivity was 36%); the overall concordance was 51% (38).

**Strengths of Hazardexpert**

2. It provides estimates of bioavailability and bioaccumulation.
3. It provides semi-quantitative predictions of a range of toxicity endpoints. As always, it is important to ensure that the predicted values are meaningful.
4. Hazardexpert can predict the likely metabolites of a compound, and their toxicities.
5. It provides an estimation of IARC human carcinogenic risk categories.
6. The database can be inspected and modified.
7. It is possible to examine the rules underlying the predictions.

**Limitations of Hazardexpert**

1. Hazardexpert gives no indication of the relative probabilities with which different metabolites are formed.
2. It does not provide estimates of acute toxicity.

**Recommendations**

1. The knowledge-base in Hazardexpert should be expanded, especially with respect to benzene, vinyl chloride and other vinyl derivatives, organophosphorus compounds, and organometallics. In addition, the knowledge-base for certain polyaromatic hydrocarbons should be re-evaluated.
2. The bioaccumulation and bioavailability modules should be further developed.

OncoLogic

OncoLogic is a knowledge-based expert system developed by LogiChem Inc. (Boyertown, PA, USA) for the prediction of chemical carcinogenicity (44). Its purpose is to make available the human expertise which is used routinely in a regulatory setting when screening chemicals for potential carcinogenicity. This expertise is represented in the monograph series Chemical Induction of Cancer (45–49), which spans more than 20 years and deals with the structural bases and mechanisms of chemical-induced carcinogenesis. The authors are members of the Structure-Activity Team of the EPA’s Office of Pollution Prevention and Toxics. This expertise in carcinogenicity has been applied in a predictive mode during the premanufacture-notification screening process mandated by the Toxic Substances Control Act (50), and has been used to estimate the chemical carcinogenicities of, or sufficiency of data for, more than 20,000 chemicals (51). The experts have had centralised access to large amounts of supporting data, both public and confidential, for the development and refinement of carcinogenicity evaluation rules.

In its current form, OncoLogic is a PC-based program consisting of four independent subsystems for estimating the carcinogenicity of: a) fibres; b) metals or metal-containing compounds; c) polymers; and d) organics. Each subsystem has a fixed, hierarchical, decision-tree structure, consisting of rules of the “if-then-else” type, and proceeds to a carcinogenicity evaluation based on the structural information entered, and on answers to program queries provided by the user. The subsystems vary considerably in user interface, function, and information content. The fibres subsystem consists primarily of a database of known fibres, as well as estimation rules based on simple physical characteristics (average length and diameter of fibres) and on observations (for example, swelling in water). The metals subsystem is capable of taking into account factors such as exposure, usage, oxidation state, physical state, solubility, and dissociation products. The polymers subsystem performs an evaluation based on parameters such as exposure route, segmental distance, the characteristics of subunits, and the presence of reactive functional groups. In the case of residual monomers, the user is referred to the organics subsystem for further evaluation.

The organics subsystem is the largest and best developed of the four subsystems. It consists of over 40,000 discrete program rules representing expert knowledge and generalisations derived from the examination of more than 10,000 organic chemicals. It also includes updated tables of known carcinogenicity results, and/or expert evaluations, for approximately 1100 frequently encountered compounds. In principle, this subsystem is applicable to a virtually unlimited number of chemicals, and covers a wide range of non-congeneric chemicals. However, it relies strictly on a chemical class, mechanism-based, approach to prediction. It contains separate and distinct modules for nearly 50 chemical classes, including, for example, acrylates, aldehydes, aromatic amines, carbamyl halides, epoxides, ethyleneimines, halogenated nitroaromatics, organophosphates, polyaromatic hydrocarbons, sulphur mustards, and triazenes. These chemical class modules vary considerably in coverage, information content, and the reliability of the resulting predictions. For example, the heterocyclic polyaromatic hydrocarbon class contains a relatively large reference database of 77 known carcinogens, but has no expert system capabilities. In contrast, the aromatic amines subsystem contains the most extensive and well-validated set of rules for prediction, but has no compiled database. As an approximate indicator of the extent of the knowledge-base, discrete program rules range from more than 8000 for the aromatic amines, to a relatively small number (fewer than 500) for more narrowly defined classes, such as the carbamates, siloxanes, and sulphones.

The carcinogenicity evaluation of an organic chemical begins when the user assigns the query structure to one of the predefined chemical classes, either by selecting structural templates or by drawing in structures within the constraints of the chosen class. The user is guided through program queries by documentation and help features. Finally, the program produces a detailed justification report which conveys the mecha-
nism-based expert reasoning underlying the evaluation. In particular, this dialogue communicates: a) general considerations applicable to the class; b) rules specific to the query structure within the class; and c) various metabolic and reactivity considerations pertinent to the chemical evaluation.

**Strengths of OncoLogic**

1. OncoLogic rules are underpinned by a wealth of expertise in the evaluation of carcinogenicity.
2. The structure entry process is faithful to the rationalisation process used by the experts, in terms of progressively narrowing the structural requirements for chemical classification and prediction.
3. It takes a relatively conservative approach to carcinogenicity evaluation, by delimiting strict chemical class domains for which adequate knowledge is available to make a prediction, and by confining predictions to within these domains.
4. It provides a mechanism-based justification for each carcinogenicity evaluation, which accurately reflects the rationale of the EPA’s Structure-Activity Team in the premanufacture-notification review process.
5. It is the only expert system for carcinogenicity prediction which attempts to evaluate both chemical structure and non-structural factors, such as the route of exposure and physical properties.

**Limitations of OncoLogic**

1. The burden of structural classification is placed entirely on the user, owing to a lack of structural recognition and processing capability. Prior to an evaluation, the user must classify the query chemical into one of the predefined chemical classes. Even if the user has the necessary expertise in organic chemistry, classification can be arbitrary if the structure does not fit neatly into one of the predefined classes.
2. If a chemical cannot be evaluated by OncoLogic, it is not straightforward to determine why. The omission of a common class may be deliberate, for example, due to known inactivity or lack of concern. Lack of chemical coverage could be due to inadequate knowledge on the part of the experts, or to the incomplete status of program development. This is a problem shared by other expert systems for toxicity prediction.
3. Although OncoLogic makes use of some physical property data, wherever available, it makes no use of QSAR models in its evaluations, and it is unable to calculate physicochemical properties to support such models. For example, it cannot estimate electronic or physical properties, such as atomic charges or solubilities.
4. The user cannot modify the rulebase.

Future enhancements of OncoLogic are planned, which should improve its chemical carcinogenicity evaluation capabilities by considering other relevant toxicity data, such as genotoxicity, oncogene activation, and P450 induction, in the overall decision-tree evaluation (52). Finally, while there have been considerable internal validation and EPA-sponsored external review of the scientific approach and rationale used by OncoLogic, there has been limited external evaluation of the system’s prospective predictive capabilities. Clearly, there are variable levels of knowledge and expertise being applied in the various subsystems and modules, resulting in a wide disparity in the levels of confidence associated with the predictions. Hence, the process of overall validation will be a slow, iterative process, depending on: a) the availability of new data and increased knowledge applicable to each subsystem or class; b) the general availability of the program; and c) accessibility to the database used.

**Purdy’s Method for Carcinogenicity Prediction**

Another approach to structure-based carcinogenicity prediction, which is at an early stage of development, is the method of Purdy (53). This decision-tree approach was developed by heuristic induction rather than by formal objective means, and relies entirely on structural features and physicochemical properties (that is, no bioassay data). It is based on known or hypothetical chemical reactivity considerations, and on the mecha-
nisms of chemical interactions in biological systems.

In its current form, which is not automated, the approach relies on a set of 11, sequentially applied, major classification rules, which are based primarily on feature identification. Examples of major classifiers include: a) the presence of an ester or azo bond; b) polyaromatic hydrocarbon; c) primary aromatic amine; and d) the presence of a chlorine atom covalently bonded to an sp³-hybridised carbon atom (sp³-[C]-Cl). Each rule feeds either directly into an activity assignment (positive or negative), another feature identification query, or a QSAR based on calculated properties. The QSARs are relatively simple in form, ranging from an allowed range or cut-off of a computed property value (for example, log P, molecular volume, ELUMO, partial atomic charges, and superdelocalisabilities), to a three-dimensional constraint specifying a fixed distance between lone pairs of electrons.

**Performance of Purdy’s method**

The training and test sets used in the development of this method comprised 306 and 301 chemicals, respectively. The data for these chemicals were taken from several sources, including the FDA carcinogenicity database and NTP technical reports. The chemicals were classified in an unambiguous manner, either as “confirmed carcinogens” or as “clear negatives”. The method performed well when applied to the test set of chemicals (correctly classifying over 90%), and has been used for prospective predictions, although the results for these are not yet known (53).

**Strengths of Purdy’s method**

1. Purdy’s method has the potential for generating useful insights into the structural basis of chemical carcinogenicity.
2. It supplements feature identification with QSAR considerations.
3. It relies entirely on computed structural properties.

**Limitations of Purdy’s method**

1. It is not automated.
2. It relies on external software packages for the computation of molecular properties.

3. Many of the component QSARs do not have clear mechanistic interpretations, and have not been adequately validated for general use.

**StAR**

Standardised Argument Report (StAR) is a PC/Windows-based system for the assessment of toxicological hazard and risk (54). It is the main product of a three-year project which ended in December 1996, and which involved the following collaborators (all UK-based): Imperial Cancer Research Fund, LHASA UK, Logic Programming Associates, and City University. The project was supported by the UK Department of Trade and Industry, and by the Engineering and Physical Sciences Research Council. The first fully commercial version of StAR is expected to be released during the second half of 1997.

Initially, the knowledge-base was limited to carcinogenicity (55), although other endpoints are now being added, such as skin sensitisation (using equations for the assessment of skin penetration developed by the UK Health and Safety Laboratory, Sheffield, UK and by Unilever ESL, Sharnbrook, UK). Generalised rules are being developed by LHASA UK, both alone and in collaboration with others. Databases are being developed which contain the toxicological properties of specific chemicals, and which link with the knowledge-bases to support and illustrate the generalised rules.

**Basis of the toxicity prediction**

The recognition of potential toxicological hazards is based primarily on structural alerts, but StAR is not restricted to this. For example, rules could be written in StAR to recognise hazards associated with fibres having particular dimensions and surface properties. Having recognised a hazard, StAR assesses the likely significance of the hazard by applying other rules, and/or by calling on other programs.

StAR has a “reasoning engine” which uses quantitative, semi-quantitative or qualitative information to derive predictions of corresponding precision: if the information is adequate for a quantitative estimation of risk, StAR will give one, but if some or all of the information is inadequate, StAR will restrict its output to a qualitative prediction.
To discover which uncertainty words are most consistently used and understood by different people, psychological research is being carried out by City University, so that the most appropriate words can be used in the StAR system to express qualitative probability (for example, “possible”, “probable”, “certain”, “plausible”, “doubtful”).

A key feature of StAR is that its reasoning is based on the Logic of Argumentation, which enables arguments to be made for and against a proposition (56). Furthermore, StAR can trace back through its reasoning to give detailed explanations of its decisions. Since arguments can be of different strengths, some arguments may overrule others. For example, if one piece of evidence were merely consistent with a proposition being false, and a different piece of evidence proved conclusively that the proposition was true, StAR would conclude that the proposition was true, overruling the weak evidence to the contrary. Where there is ambiguity, StAR will not make arbitrary decisions, but will report the evidence to be equivocal or contradictory. In all cases, the user can examine the arguments for and against a conclusion, in order to judge its significance.

The system automatically recognises interactions between different rules. For example, it can follow through the chain of reasoning which leads from the observation of a positive Ames test result to a prediction of mutagenicity, and from that to a prediction of potential carcinogenicity. It can also recognise when a prediction of carcinogenicity in humans should be weakened, for example, when an observed activity in rats is associated with peroxisome proliferation.

A simple text editor enables rule writers to enter rules such as “substances which are mutagenic are likely to be carcinogenic”, or “substances which are carcinogenic in several mammalian species are likely to be carcinogenic in others”. Information about chemical structural alerts is entered graphically in the form of standard chemical representations, the so-called “Markush diagrams”.

StAR can make predictions which take account of aromaticity and stereochemistry (for example, chirality), as well as steric and electronic effects. Input is required in the form of a conventional chemical diagram or a standard chemical connection table, such as a molfile.

**Links to other systems**

In addition to using data provided by the user, StAR can collect data from other databases, or can call on other applications to do calculations. Links to other databases serve two purposes. Firstly, if the user enquires about a substance for which toxicity data are already contained in a database, StAR will draw the user’s attention to those data. Secondly, the user can assess the validity and importance of StAR rules by accessing a database containing the information upon which the rules are based. For example, a rule may state that substances containing unsaturated aldehyde fragments and having certain physical properties are likely to be skin sensitisers, and the supporting database may contain specific examples of both sensitising and non-sensitising aldehydes accompanied by relevant data. Links to other applications enable StAR to make judgements about rules which refer to physical properties, such as log P, and to make use of QSAR equations which are valid for a substance in a user query.

**Toxicophore Discovery Systems: REX, DTOX and PROGOL**

The identification of toxicophores is an important step in the processing of chemical structures by many expert systems. Therefore, some recent research on the recognition of toxicophores may lead to new products which could be incorporated into expert systems. REX was designed to generate rules suitable for DEREK, by producing output directly in PATRAN, the program language for the DEREK rulebase. The DTOX project for the Ministry of Agriculture, Fisheries and Food (MAFF) assessed the potential of different data-mining techniques for finding toxicophores. PROGOL is a toxicophore recognition system based on inductive logic programming.

**REX**

REX has two unique features (57–59). Firstly, the descriptors it uses are so-called “atom pairs” rather than specific fragments or atom and bond chains. An atom pair is made up of identifiers for two atom types and the distance between them (expressed as the number of bonds). For example, the
sequence N-C-C-O-C-C-S in a molecule would generate atom pairs such as N..3..O (a nitrogen atom three bonds away from an oxygen atom), N..4..C, N..6..S, O..3..S, etc. Various generic groups are included as atom types; for example, the halogens are recognised as being related as well as being treated individually, and special treatment is given to hydrogen atoms and lone pairs on heteroatoms.

The reason for choosing atom pairs, rather than chains, is that typical biological interactions are with binding centres at certain distances apart in space; what is between the centres does not matter. Although the term “atom pair” is used, multiple bonds are included in the analyses; for example, S-C-C(=O)-O-C would generate descriptors such as “S..2..=” (where “=” signifies a double bond). Arguably, topological distances (number of bonds) are as effective as, and possibly better than, three-dimensional through-space distances (58). Nevertheless, the REX technology can use three-dimensional through-space distances, if necessary.

The second unique feature of REX is that once atom pairs which appear to be associated with activity have been identified, they are mapped back onto the structures of active molecules. Where they overlap consistently, they are fused together to build up complete, complex toxicophores. For example, atom pairs like N..2..C, N..3..O, N..4..P, N..4..=, N..5..O, N..5..S, N..6..C, P..3..C, O..2..S, O..2..O can be mapped together to create the fragment N-C-C-O-P(=S)-O-C, which is part of an insect toxicophore requiring metabolic activation (only a subset of the relevant atom pairs is listed).

DTOX

DTOX refers to a research project carried out for MAFF by Integral Solutions Ltd (Basingstoke, UK), the suppliers of the neural network analysis package “Clementine”, in collaboration with LHASA UK, to see whether the data-mining techniques included in Clementine could be used to discover toxicophores from data about chemical structures (60).

Several thousand descriptors of various types were selected so that their behaviour could be compared: atom and bond chains, augmented atoms, ring indices, topological atom pairs (including pairs of augmented atoms), pairs of binding centres with three-dimensional through-space distances, and three-dimensional three-centre pharmacophores. The three-dimensional descriptors were generated by using Chem-X, in collaboration with Chemical Design Ltd (Chipping Norton, Oxon, UK). The data used were taken from the Genetox database and from NTP publications.

Correlations between sets of descriptors and activity were found, but it was not clear how reliable these were. There was no clear evidence that one kind of descriptor was better than another. More encouragingly, it was found that Clementine could process sets containing thousands of descriptors at an acceptable speed (data-mining tasks more usually involve sets of objects having only 10–20 attributes).

PROGOL

PROGOL is intended to be a general learning engine using inductive logic, but in a collaboration between Ross King (University of Oxford, UK) and Mike Sternberg (Imperial Cancer Research Fund, UK) its use for discovering toxicophores or pharmacophores in sets of active chemicals was explored. Unlike other systems for achieving this task, it uses inductive logic and therefore it does not depend on chemical structure descriptors of the conventional kind — it requires only connection tables. Thus, the need to generate descriptors is avoided, and with it the need to make arbitrary decisions about descriptor types. The collaborative study showed that PROGOL is able to discover toxicophores (61, 62), but a clear assessment of the limitations of the system is still needed.

The US NTP Exercise on the Prediction of Rodent Carcinogenicity

As a result of the public and political pressure to have tools capable of predicting the widest possible range of chemical carcinogens, most efforts have focused on systems trained on non-congeneric chemicals (63). A unique opportunity for evaluating the merits of different approaches was presented by a
recent comparative exercise which challenged systems with a common set of chemicals. Crucial to interpreting the outcome of this exercise is that the carcinogenic potentials of the chemicals were unknown at the time the predictions were made. The exercise considered 44 chemicals which were in the process of being tested for rodent carcinogenicity by the NTP. At the end of the study, only 40 chemicals had actually been tested experimentally; the actual (experimental) and estimated carcinogenicities for these were compared.

The predictive approaches

Some of the predictive approaches included in the comparative exercise were (Q)SAR methods, whereas others were so-called “activity-activity relationships” (AARs), which seek to establish relationships between carcinogenesis and shorter-term biological events. The Tennant et al. (64) AAR method used different types of information: general and target organ toxicity in rodents, Salmonella mutagenicity, and the presence of structural alerts according to the list compiled by Ashby et al. (65). This information was combined to yield a final prediction of carcinogenicity by means of expert judgement. The approach also used previous carcinogenicity data, wherever available in the literature, to modulate the predictions. A second AAR approach, called Rapid Screening of Hazard (RASH; 66), consisted of a modification of the Tennant et al. approach, in that it adopted an alternative means for ranking toxic potencies. Among the (Q)SAR approaches there were CASE, which was used in conjunction with Multi-CASE (67), TOPKAT (68), and COMPACT (36). Predictions were also made by using DEREK.

Other predictions were based on the experimental measurement of electrophilicity (Ke): chemicals below a certain Ke threshold were predicted to be non-carcinogenic, and vice versa (69). Benigni (70) made predictions based on a combination of two types of information: a) the theoretically estimated electrophilicity of a chemical (Ke), according to Benigni et al. (71); and b) the presence of potentially alerting substructures in a chemical, according to Ashby et al. (65). Weisburger and Lijinski submitted two sets of predictions based on expert intuition (72), which used the structures of the NTP chemicals as their starting information (International Workshop, 1993; NIEHS, NC, USA). The results of the NTP exercise have been discussed in several papers (for example, 72, 73).

The various approaches used were very different in nature. The Tennant et al. and RASH methods were AAR approaches; Weisburger also incorporated some biological information. Some of the (Q)SAR approaches were more quantitative in character (Benigni et al., CASE, COMPACT, TOPKAT), while others were more qualitative (DEREK, Lijinski, Weisburger). Three approaches (RASH, Tennant et al. and Weisburger) were not “pure” prediction systems, but were mixed prediction/evaluation systems because they incorporated previous experimental carcinogenicity results wherever these were available. The Ke approach was based on the experimental measurement of a physicochemical property of the test chemicals.

Performance of the predictive approaches

A comparison of the different predictions has been undertaken by Benigni (74), by using multivariate data analysis methods. The AAR approaches gave similar prediction profiles for the individual chemicals, as did most of the (Q)SAR systems (including Ke). Thus, the main difference in the prediction profiles was between those systems which relied on chemical structure as input information and those which were based on biological information. This was particularly surprising in view of the very different rationales and technologies employed by the various (Q)SAR approaches.

The percentage of agreement between each set of predictions and the actual rodent carcinogenicity results was calculated once the rodent data (obtained in two-sex, two-species experiments) had been summarised as an overall dichotomous carcinogenicity classification (Table I). When considering the figures in Table I, it should be noted that these percentages derive from only 40 chemicals (and even fewer for some systems). Thus, a difference in accuracy of 5%, for example, amounts to only two chemicals; even though this may appear large, it could be entirely due to chance. In addition, the figures in Table I refer specifically to the 40 NTP chemicals, and may have been different if another database had been used. In a more recent analysis (75), the test set was
increased to 44 chemicals, which represents an increase of 10% in sample size and a reduction of 20% (from 5% to 4%) in relative probabilistic error.

The accuracy of the (Q)SAR approaches was in the range 50–65%, which is rather limited; the biologically based approaches attained 75% accuracy. When judging this result, it should be borne in mind that a difference of 10% is equivalent to only four chemicals, and also that the Tennant et al. approach used previous carcinogenicity data if these were available. This approach cannot be applied when neither in vitro nor in vivo biological data are available. For these reasons, no real comparison is possible between the (Q)SAR and AAR methods. Furthermore, when a carcinogenicity classification maintaining some of the original gradation was considered, the (Q)SAR methods showed an improved performance of about 70%, whereas the performance of the Tennant et al. approach declined. The lack of stability in these results stresses the importance of expressing carcinogenicity results in the most appropriate manner.

The 40 NTP chemicals were broken down into the following classes: A) the most powerful carcinogens (positive in all four rodent systems); B) chemicals with mixed carcinogenicity profiles; and C) non-carcinogens. Most of the prediction systems were concordant in the identification of class A chemicals. The chemicals with mixed carcinogenicity profiles were often predicted to be positive, although the number of positive predictions was lower than that found with the most powerful carcinogens. Serious problems were encountered with the non-carcinogens, since many of them were predicted to be positive by the various approaches. Thus, the clear limitation of almost all of the prediction systems, irrespective of whether they represented AAR or (Q)SAR approaches, was their excessive sensitivity (63).

An inspection of the results for individual chemicals showed that the various (Q)SAR approaches acted as gross "class-identifiers"; they pointed to the presence or absence of alerting chemical functionalities, but were not able to make gradations within each potentially harmful class (63). Therefore, the large-scale source of variation, represented by the alerting substructures, obscured the small-scale sources of variation (that is, differences among chemicals belonging to the same potentially alerting class). Consequently, all of the systems tended to overestimate toxicity and the potential differences were not highlighted, their prediction profiles being similar to a large extent.

The limited performance of the (Q)SAR approaches in predicting rodent carcinogenicity is due to several factors. Firstly, the exercise was conducted with non-congeneric data sets, which included different chemical classes possibly acting by different mechanisms. The goal appeared to be the construction of a super-model, capable of incorporating various chemical class-specific models; however, the latter can be very different from each other, making it difficult, if not impossible, to derive a good general model.

A second, more subtle, reason, which is often overlooked, is that the relationships which hold on the large scale may differ from those apparent on the small scale (63). This can happen when a set of non-congeneric chemicals spans a large range of descriptor values, while individual, congeneric, chemical classes each span a portion of this range. It is also conceivable that no relationships exist within the individual classes, but when the entire range of descriptor values is con-
sidered, an overall relationship emerges. In such a case, the (Q)SAR may be purely empirical, devoid of any relationship with an underlying mechanism of action. Such a (Q)SAR would not enable accurate predictions to be made for individual chemicals, but only averaged estimates for each sub-class of chemicals.

A third reason, which accounts for the limited performance of the (Q)SAR approaches to non-congeneric chemicals, is the way in which biological activity is classified. The outcome of the rodent carcinogenicity experiments was complicated. There were four different experimental groups (rat and mouse, male and female), various types of tumours, and various rates of induction. The outcome for each of the four groups is usually summarised as a classification into one of four categories: no evidence (NE), equivocal evidence (EE), some evidence (SE), and clear evidence (CE). In a subsequent step, a chemical classified as SE or CE in any of the four rodent systems is considered to be a potential carcinogen in humans. While this dichotomous (positive/negative) carcinogenicity classification may be useful for the purpose of a conservative risk assessment, the intrinsic qualitative differences in the rodent results are lost when they are summarised in such a way. The inappropriate classification of biological activity can prevent the derivation of reliable (Q)SAR models or can bias the results obtained with such models. For example, if the analyses undertaken by Benigni (74) had been performed by omitting the bioassay studies classified as EE, rather than by equating them to a negative classification, it is possible that this would have reduced the uncertainty associated with the results.

The adequacy of the existing carcinogenicity database should be taken into account, since reliable (Q)SARs can only be formulated when the initial information is of high quality. The database established by Gold and colleagues (for example, 19, 20) includes about 1000 chemicals tested in carcinogenicity studies. This number may seem large, but the individual mechanisms for each chemical class split the overall data set into many small subsets. In addition, the carcinogenicity studies are not usually planned to meet the requirements of SAR studies. For example, even the systematic investigation being conducted by the NTP (76) follows criteria which are not particularly suitable in this respect; the test chemicals were chosen to represent a sample of those on the market. Since commercial chemicals often have complicated structures, characterised by the presence of several different functional groups, it is difficult to study the effects of individual functionalities.

**Conclusions and recommendations**

1. By highlighting the strengths and limitations of individual predictive approaches, the NTP comparative exercise has illustrated where improvements should be sought and how these should be achieved.

2. Overall, the evidence has reaffirmed the notion that (Q)SAR models are context-specific. In this respect: a) further development of the predictive systems would be strongly aided by undertaking toxicity studies on individual chemical classes; and b) the (Q)SAR models incorporated into expert systems should, ideally, refer to classes of compounds acting through the same mechanism (and possibly with the same rate-limiting step).

3. The predictive systems would benefit from careful checking of the data included in existing databases on toxic chemicals, and from the availability of biological activity classifications appropriate for (Q)SAR modelling.

4. Progress in chemical theory (for example, improved ways for representing chemicals and for defining chemical similarity) will enable more reliable predictive systems to be developed.

**The Integrated Use of Computational Prediction Techniques**

The overall process of chemical risk assessment is usually performed as a series of steps involving hazard identification, hazard evaluation, exposure assessment and, finally, a full evaluation of risk. In a previous ECVAM workshop report on the integrated use of alternative approaches (77), the importance of combining the use of *in vitro* data with other predictive methods was emphasised. It is unlikely that in the future risk assessment will rely on predictions made by a single computational prediction technique (CPT), that
is, a single model or expert system for the prediction of toxicity, since different CPTs have been conceived for different purposes, and therefore exhibit their own particular strengths and weaknesses.

Since different kinds of information are required for each step of the risk assessment process, a variety of models, each of which has been tailored to satisfy the special requirements of a particular step, can be used in combination to meet the overall needs of risk assessment. An ordered continuum of models which manages information within the limits of each type of modelling approach can eliminate uncertainties, thereby making the information generated more reliable. Such a continuum of models was presented at the workshop by Douglas Bristol, as illustrated in Figure 2.

The advantage of such an approach for integrating the use of models is that aspects specifically associated with each step can be identified, and can therefore be modelled appropriately. Ideally, a sequence of models can be developed, so that the output information from one model serves as the input to the next. The logical starting point is the development of reliable hazard identification models because, if there is no activity, the risk assessment need not proceed further. In addition to providing a reliable classification of activity, hazard identification models can

Figure 2: Continuum of models needed for risk assessment

PB-PK/PD = physiologically based-pharmacokinetic/pharmacodynamic.
also associate activity predictions with relevant mechanistic pathways or modes of action. The output from a good classification model is suitable as the input for QSAR models, which provide mechanistic insight into important biological processes, such as uptake, transport, metabolism, target-site binding, reactivity and elimination. The output from QSAR models can serve as the input to pharmacokinetic and pharmacodynamic models, which provide a rationale for dose-response effects and enable interspecies extrapolations. Conceivably, this approach could provide all of the information needed to perform safety evaluations or risk assessments for chemicals, with a minimum of animal testing being required (78).

A general scheme for integrating the use of expert systems (Figure 3) was presented at the workshop by Ann Richard. On the left of the scheme are statistical or multivariate approaches, such as CASE and TOPKAT, which can be employed for the non-biased discovery of correlations between chemical structure and biological activity. Such methods, when applied to non-congeneric chemicals, generally act as feature identifiers and classifiers, detecting large-scale variations in biological activities between chemical classes. Other QSAR approaches, such as Hansch-type analysis and molecular modelling approaches, are usually applied in a more focused way to determine the conditions for activity, or potency variation, within narrowly defined chemical classes. Crucial to the success of these approaches is the appropriate classification of chemicals according to assumed mechanism of action. Such a classification can be based on: a) biological induction (based on the analysis of effects found in biological systems); b) the results of statistical SAR models; or c) chemical reactivity considerations. In each of these

Figure 3: The integrated use of expert systems in toxicity prediction

![Diagram of expert system integration](#)
approaches, a final model is derived by using a combination of mathematical and heuristic techniques for processing various types of chemical, structural and biological information. After having been validated to an acceptable degree, each model is static, and can be represented subsequently in the form of one or more rules.

Current knowledge-based expert systems, such as DEREK, Hazardexpert and Onco-Logic, rely primarily on chemical feature identification supplemented by expert knowledge and judgement. However, in principle, such approaches are capable of incorporating rules of any sort, derived from any type of SAR method, providing that there is a means for calculating the necessary parameters. Hence, knowledge-based expert systems such as these represent the final stage of a natural progression toward hybrid approaches which are capable of considering and incorporating all types of relevant information into a particular prediction of toxicity.

General Comments on the Development and Validation of Expert Systems

The reliability of any expert system for predicting toxicity is crucially dependent on the quality of the database and the rulebase. Thus, minimal criteria should be established for the acceptability of data to be used in the development and evaluation of expert systems. For example, information based on qualitative assessments should be deemed unacceptable unless backed up by quantitative data, and published data should be obtained by using test compounds of known purity. It is strongly recommended that the compilation of databases is undertaken in close collaboration with a suitably experienced toxicologist. Indeed, the database and the rulebase should be subject to continual review by such toxicologists. Rules should not be based on limited experimental data, although their usefulness in predicting the toxic effects of novel chemicals could be enhanced if other kinds of information (for example, from computational chemistry and model building) were also used in their formulation.

Access to large amounts of high quality toxicological data is not always possible because of the confidentiality restrictions placed on proprietary data. As a result of its existing reputation and status as a non-profit organisation, LHASA UK receives confidential data from companies, but these can only be used for internal research purposes at LHASA UK. The data supplied under secrecy arrangements can only be made public if a time comes when the company supplying the data no longer regards them as confidential. In the meantime, new rules about toxicity can be developed and made public, but the supporting data on which they are based cannot be divulged.

The validation of expert systems could be approached by employing a two-stage process. In the first stage, the relative performances of entire systems could be assessed by using a common set of reference chemicals. In the second stage, the performances of individual systems could be evaluated in a more comprehensive manner, by assessing the validity of the various models on which the systems are based. This stage would require adherence to criteria developed for the validation of CPTs. In the case of (Q)SARs, such criteria could be formulated by analogy with those adopted by the EU for validating (Q)SARs for risk assessment purposes (79). In the case of biokinetic models, several issues relevant to their validation were addressed in a previous ECVAM workshop report (80).

Conclusions and Recommendations

Development and validation of expert systems

1. Ideally, an expert system should incorporate all of the available mechanistic information pertaining to the type of toxicity being modelled, and should take into consideration not only the route of exposure and bioavailability, but also kinetic processes, particularly metabolism.

2. An expert system should state, for a given endpoint, the species and strain of animal for which a prediction is being made, as well as other relevant toxicological parameters, such as the route of administration.

3. In the case of systems employing SARs based on two-dimensional chemical fragment representations, it must be
remembered that structural alerts are merely indicator variables which often have to be progressively simplified to make the systems more generally applicable, but this is done at the expense of a more accurate mechanistic basis.

4. The evolution of expert systems toward quantitative modelling should be encouraged; for example, by incorporating QSAR models derived from individual groups of congeneric chemicals.

5. Knowledge-based rules in most expert systems are usually based on reactive chemistry. To predict certain endpoints, such as non-genotoxic carcinogenesis, there is an urgent need for additional rules reflecting, for example, receptor-based mechanisms.

6. In the case of knowledge-based systems, problems with impurities are best left to the discretion of rule writers, who should judge the likelihood of contaminants affecting the endpoint in question. For example, positive test results may occur with the non-carcinogen 1-naphthylamine due to traces of the carcinogen 2-naphthylamine. In such cases, it is important that the effect of the impurity on the outcome of the prediction is reported to the user.

7. The rule writer has to decide between a large number of very specific rules, with few or no exceptions, and a smaller number of general rules, coupled with a consideration of exceptional cases. The latter approach is preferable in that there is more likely to be a default for handling an unknown structure, even though this may prove to be anomalous.

8. An expert system should enable the user to access the information upon which a prediction is based, from its database or rulebase, so that the limitations associated with the prediction can be defined.

9. An expert system should enable the user to modify its rules and model algorithms, as well as the contents of its database. Nevertheless, a standard version should also be made available for the purposes of reproducibility and regulatory acceptance. Changes to the standard version should be peer-reviewed by an independent set of experts, perhaps including representatives from regulatory agencies.

10. A process should be set up for the formal and on-going, independent, validation of expert systems. The principles underlying such a process could be formulated by analogy with the ones adopted by the EU for validating (Q)SARs for risk assessment purposes.

11. There is a critical need for reliable data to be made available for use in the development and validation of expert systems. Efforts should be made to obtain these data, either by searching the literature or by gaining access to data which are currently considered to be confidential. In this respect: a) ECVAM could act as a forum for representatives from industry and regulatory bodies to discuss practical ways of making confidential data available. If, for commercial reasons, such data cannot be made widely available, it may still be possible to allow access in a restricted manner; for example, by allowing developers to use the data in training sets, provided that they respect the confidentiality and search non-confidential sources for confirmation of the rules/models derived; b) criteria are needed for determining the acceptability of data. The recommendations contained in previous ECVAM workshop reports are endorsed, particularly those referring to the chemical purities of test substances and knowledge concerning likely contaminants; c) consensus is needed on how to deal with conflicting data from various tests and with inconsistent results from the same tests; and d) a database containing reliable data should be established and maintained by a neutral organisation, such as ECVAM, and a procedure should be set up whereby missing data can be generated.

12. The successful development of any expert system requires close collaboration between experts in relevant disciplines (toxicology, chemistry, statistics, and computer science).

13. Developers should be encouraged to work with each other, especially if plans for an integrated scheme of expert systems are ever to be realised (see point 22).
14. Before any validation exercises take place, there should be clarification as to how the systems are to be applied, to minimise the chances of bias or misinterpretation of the results.

15. When assessing the performance of an expert system, it must be borne in mind that the complete verification of numerical models of biological systems is inherently impossible; at best, confirmations of predictivity can only be partial.

16. Given the inherent variability of biological systems, the predictions made by expert systems should not be expected to be more accurate than successive measurements of the same endpoint would be.

17. When evaluating an expert system, it is important to distinguish between a predictive approach and one which classifies substances.

18. A validation exercise involving one or more expert systems is more likely to be convincing to the scientific and regulatory communities if the predictions are made prospectively, rather than retrospectively. In the case of purely correlative approaches, such as the CASE technologies, the scientific validity of the predictions is actually the same, irrespective of whether they are made prospectively or retrospectively. However, in cases where expert judgement is used for the construction of rules, it is impossible to eliminate prior knowledge, so the predictions will be more convincing if they are made prospectively.

19. An interesting question for future research is whether knowledge-based expert systems can progress beyond human expert judgement, for example, by employing artificial intelligence techniques to recognise hidden patterns.

20. The predictions which can be obtained with knowledge-based systems are restricted by the particular knowledge-base employed. The predictions which can be obtained with automated rule-induction systems are restricted by the quality of the data in the training set.

21. The user of an expert system should be able to judge the value, reliability and context of the toxicity predictions produced by the system, in order to use the prediction in a considered manner.

22. Since each system has its own particular strengths and weaknesses, it seems unlikely that any one system will come to be regarded as the best and only choice for chemical risk assessment purposes; different systems have been conceived for different purposes and, therefore, are applicable in different areas of predictive toxicology. This raises the practical difficulty as to which system(s) should be invested in by organisations and regulatory authorities. The ideal situation would be an integrated approach which exploits the strengths of all of the systems available. Bearing costs in mind, realistic attempts should be made to devise a hierarchical or battery approach to the use of appropriate expert systems, in combination with other sources of relevant information, such as experimental and epidemiological data.

23. In view of the possible presence of unknown contaminants with potent toxicities in a sample, expert systems can be used for screening purposes, but not for providing a complete assurance of safety.

24. For quality control purposes, users should check their operation of an expert system by using a reference set of chemical structures; these should be supplied by the system developer.

25. When using fragment-based SAR systems, it must be borne in mind that particular fragments may occur in active (or inactive) molecules by chance. Therefore, the user needs to consider all of the evidence available, to avoid being misled by substructures presenting as toxicophores (or toxicothorobes).

26. The judicious use of expert systems should reduce the need for animal experimentation. For example, during the development of pharmaceuticals, expert systems provide a means of high-throughput screening which is potentially useful for early compound prioritisation. Similarly, expert systems provide a means of prioritising chemicals during toxicological risk assessment, and enable missing data to be generated.
27. Expert systems are useful for educational purposes, for example, by illustrating the biological effect of changing a particular substructure within a molecule, and by conveying the mechanistic rationale used by experts when evaluating toxicity.

Acknowledgements

The authors would like to thank Julia Fenem (ECVAM) and Walter Karcher (ECB, Joint Research Centre, Ispra, Italy) for organising this workshop, and would like to acknowledge the participation of Bjørn Hansen (ECB) and Sharon Munn (ECB) in the final session of the workshop.

References


35. Lewis, D.F.V. (1997). Quantitative structure-activity relationships in substrates, inducers and inhibitors of cytochrome P4501 (CYP1). Drug Metabolism Reviews, in press.


## Appendix 1

### Information Sources for Expert System Software

<table>
<thead>
<tr>
<th>System</th>
<th>Supplier/Developer</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEREK, StAR</td>
<td>LHASA UK Ltd</td>
</tr>
<tr>
<td></td>
<td>School of Chemistry</td>
</tr>
<tr>
<td></td>
<td>University of Leeds</td>
</tr>
<tr>
<td></td>
<td>Leeds LS2 9JT</td>
</tr>
<tr>
<td></td>
<td>UK</td>
</tr>
<tr>
<td>ToxAlert, CASE, Multi-CASE, CASETOX</td>
<td>MULTICASE Inc.</td>
</tr>
<tr>
<td></td>
<td>P.O. Box 22517</td>
</tr>
<tr>
<td></td>
<td>Cleveland</td>
</tr>
<tr>
<td></td>
<td>OH 44122</td>
</tr>
<tr>
<td></td>
<td>USA</td>
</tr>
<tr>
<td>COMPACT</td>
<td>Dr David Lewis</td>
</tr>
<tr>
<td></td>
<td>School of Biological Science</td>
</tr>
<tr>
<td></td>
<td>University of Surrey</td>
</tr>
<tr>
<td></td>
<td>Surrey GU2 5XH</td>
</tr>
<tr>
<td></td>
<td>UK</td>
</tr>
<tr>
<td>TOPKAT</td>
<td>Health Designs Inc.</td>
</tr>
<tr>
<td></td>
<td>183 East Main Street</td>
</tr>
<tr>
<td></td>
<td>Rochester</td>
</tr>
<tr>
<td></td>
<td>NY 14604</td>
</tr>
<tr>
<td></td>
<td>USA</td>
</tr>
<tr>
<td>Hazardexpert, Metabolexpert</td>
<td>CompuDrug Chemistry Ltd</td>
</tr>
<tr>
<td></td>
<td>Hollán Ernö utca 5</td>
</tr>
<tr>
<td></td>
<td>P.O. Box 405</td>
</tr>
<tr>
<td></td>
<td>1395 Budapest</td>
</tr>
<tr>
<td></td>
<td>Hungary</td>
</tr>
<tr>
<td>OncoLogic</td>
<td>LogiChem Inc.</td>
</tr>
<tr>
<td></td>
<td>Boyertown</td>
</tr>
<tr>
<td></td>
<td>PA 19512</td>
</tr>
<tr>
<td></td>
<td>USA</td>
</tr>
<tr>
<td>Clementine</td>
<td>Integral Solutions Ltd</td>
</tr>
<tr>
<td></td>
<td>Berk House</td>
</tr>
<tr>
<td></td>
<td>Basing View</td>
</tr>
<tr>
<td></td>
<td>Basingstoke</td>
</tr>
<tr>
<td></td>
<td>Hampshire RG21 4RG</td>
</tr>
<tr>
<td></td>
<td>UK</td>
</tr>
</tbody>
</table>