Structural Alerts of Mutagens and Carcinogens

Romualdo Benigni* and Cecilia Bossa

Experimental and Computational Carcinogenesis, Environment and Health Department, Istituto Superiore di Sanita', Viale Regina Elena 299 00161, Rome, Italy

Abstract: This paper summarizes the evidence on the Structural Alerts of mutagenicity and carcinogenicity. The Structural Alerts are molecular substructures or reactive groups that are related to the carcinogenic and mutagenic properties of the chemicals, and represent a sort of "codification" of a long series of studies aimed at highlighting the mechanisms of action of the mutagenic and carcinogenic chemicals. The identification of the Structural Alerts has had a great value both in terms of understanding mechanisms, and of assessing the risk posed by chemicals. This mini-review illustrates a number of case studies where the Structural Alerts have played a fundamental role in risk assessment, and describes recent work aimed at expanding or refining the knowledge on the Structural Alerts through the use of Artificial Intelligence and Data Mining approaches.

Keywords: Risk assessment, artificial intelligence, chemical description, data mining, software, action mechanism.

INTRODUCTION

In a brilliant review paper entitled "Domestication of chemistry by design of safer chemicals: Structure-Activity Relationships", E.J. Ariens described how the concept and practice of Structure-Activity Relationships (SAR) was intervening in the field of toxicology in the mid 1980's [1]. He wrote: "...Biological –including toxic- effects are the result of an interaction of the xenobiotic molecules with particular molecules, usually biopolymers in the biological objects. The chemical properties of the xenobiotics therefore are a determining factor. A relationship between chemical structure or chemical properties and biological action, SAR, therefore is in the nature of things and undeniable, notwithstanding the fact that it is not always easily recognized...". Then Ariens showed that there were two approaches to SAR: "..1. The functional group approach: This takes into account the significance of particular groups in the molecule for particular aspects, part processes, in the biological action. Examples are groups described as pharmacophores or toxicophores; …. 2. The integral approach: In this case the overall properties of the molecules count. The various correlative methods are examples, such as the Hansch regression analysis...".

As a matter of fact, the 1980's were the years during which the Quantitative Structure-Activity Relationships (QSAR) approach was having a dramatic development, with an exponential increase in methods and computerized technologies proposed, and the more qualitative approach based on the simple recognition of toxicophores was entering into shade. This is even more so today. However, the knowledge of the toxicophores, as recognition and classification of the molecular substructures and reactive groups responsible for the toxic effects, is still at the basis of the mechanistic science of toxicology and provides powerful means of intervention to "domesticate" the chemicals. This mini-review is a survey of the field of toxicophores, or Structural Alerts (SA) for mutagenicity and carcinogenicity. It shows the evolution of the field, and presents the practical implementation of such knowledge in today toxicology.

*Address correspondence to this author at the Experimental and Computational Carcinogenesis, Environment and Health Department, Istituto Superiore di Sanita', Viale Regina Elena 299 00161, Rome, Italy; E-mail: rbenigni@iss.it

THE IDENTIFICATION OF THE STRUCTURAL ALERTS

The electrophilic theory of chemical carcinogenesis developed by James and Elizabeth Miller [2, 3] enabled the activity of the large majority of animal carcinogens known by the 1970's to be tentatively rationalized. Equally, the activity of chemicals as mutagens to Salmonella almost always seems plausible within the context of the Miller's hypothesis [4]. Historically, in the 1960's the Miller's noted the electrophilicity of the carcinogenic alkylating agents. Since then, a number of acylating agents were found to be carcinogenic, and these chemicals were also electrophilic as administered. The Miller's were also much impressed by the variety of chemical carcinogens of rather different structures for which metabolism to electrophilic reactants had been demonstrated. Overall, this evidence led them to suggest "that most, if not all, chemical carcinogens either are, or are converted *in vivo* to, reactive electrophilic derivatives which combine with nucleophilic groups in crucial tissue components, such as nucleic acids and proteins" [3].

After a number of decades, the hypothesis of the electrophilic reactivity of (many) chemical carcinogens maintains its validity, and has been incorporated into a more general theory on the chemical carcinogens. From the point of view of the mechanism of action, the carcinogens are classified into: a) genotoxic carcinogens, which cause damage directly to DNA. Many known mutagens are in this category, and often mutation is one of the first steps in the development of cancer [5]; and b) epigenetic carcinogens, that do not bind covalently to DNA, do not directly cause DNA damage, and are usually negative in the standard mutagenicity assays [6]. Whereas the epigenetic carcinogens act through a large variety of different and specific mechanisms, the genotoxic carcinogens have the unifying

1573-4099/06 $50.00+.00

feature that they are either electrophiles per se or can be activated to electrophilic reactive intermediates, as originally postulated by the Miller's. Following this hypothesis, several investigators studied the mechanisms of action and the metabolic fate of a large number of carcinogens; this led to the identification of several chemical functional groups and substructures (Structural Alerts, SA) for genotoxic carcinogens. On the contrary, the recognition of SAs for the nongenotoxic carcinogens is far behind, also because no unifying theory provides scientific support [6].

ASHBY'S COMPILATION OF STRUCTURAL ALERTS

Following the seminal work of the Miller's, a distinguished contribution to the advancement and dissemination of the knowledge on the SAs for carcinogenicity has been provided by John Ashby. He has also shown that the activity of chemicals as mutagens to Salmonella can explained by the electrophilicity theory of the Miller's. As discusses above, this applies to the so-called genotoxic carcinogens. A very effective popularization of the SAs has been provide by Ashby in the form of a graphical display (poly-carcinogen), which represents a hypothetical chemical made of most of the known SAs. Fig. (**1**) is an adaptation of the well-known Ashby's poly-carcinogen, presented originally in [7], and in revised form in [4].

The SAs represented in Fig. (**1**) are the following:

a) alkyl esters of either phosphonic or sulphonic acids;

b) aromatic nitro groups;

c) aromatic azo groups (because of the possible reduction to aromatic amines);

d) aromatic rings *N*-oxides;

e) aromatic mono- and dialkylamino groups;

f) alkyl hydrazines;

g) alkyl aldehydes;

h) *N*-methylol derivatives;

i) monolakenes;

j) β-haloethyl mustards;

k) *N*-chloroamines;

l) propiolactones and propiosultones;

m) aromatic and aliphatic aziridinyl derivatives;

n) aromatic and aliphatic substituted primary alkyl halides;

o) derivatives of urethane (carbamates);

p) alkyl *N*-nitrosoamines;

q) aromatic amines (including their *N*-hydroxy derivatives and the derived esters);

r) aliphatic and aromatic epoxides.

Each of the SAs is a "code" for a well-characterized chemical class, with its own specific mechanism of action. However, there are also general factors that may influence the potential reactivity of a chemical, i.e., one could expect to observe compounds with structurally alerting features but which are biologically inactive because of a number of reasons. Among the physicochemical factors that modulate and may hinder the potential biological activity of the chemicals with SAs are: 1) Molecular Weight (MW): chemicals with very high MW and size have little chance of being absorbed in significant amounts; 2) physical state, which influences the capability of the compounds to reach critical targets; 3) solubility: in general highly hydrophilic compounds are poorly absorbed and, if absorbed, are readily excreted; 4) chemical reactivity: compounds which are "too reactive" may not be carcinogenic because they hydrolize or polymerize spontaneously, or react with noncritical cellular constituents before they can reach critical targets in cells. Another critical factor is the geometry of the chemical compounds: many potent carcinogens and mutagens (e.g. polycyclic aromatic hydrocarbons, aflatoxin B1, etc…) are planar molecules, with an electrophilic functional group and favorable size, so that they can intercalate properly into DNA [8].

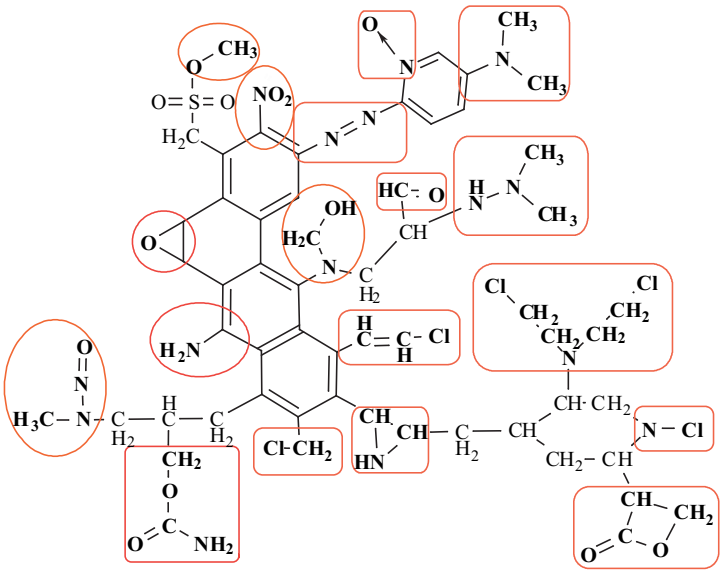

Fig. (1). The figure displays the Ashby's Structural Alerts (adaptation form the original drawing in [4]) (see details in the text).

To take into account the above modulating factors and the empirical evidence provided by the experimental results, Ashby further elaborated on the description of SAs [4]. Some examples of his considerations are the following:

1) All aromatic amino, substituted amino, and nitro compounds have been considered as positive, except in cases of *ortho*-di-substitution or where a carboxylic acid is present *ortho* to the nitrogen substituent. These factors are expected to hinder metabolic activation of the adjacent nitrogen substituent;

2) Aromatic –NR2 groups have been scored as positive unless R = C3 or greater, or extensive steric crowding of the substiteunts exists;

3) Chlorinated olefins have been scored as positive only where a sterically accessible epoxide derivative could be formed, i.e., where at least one hydrogen or alkyl group is attached to each carbon atom.

Large-scale applications of the SAs to classify chemicals in terms of their toxic propensity have been provided by Ashby and co-authors in a number of papers, e.g., [4, 9-13]. Another interesting contribution was provided in a paper [14] where Ashby participated to a challenge on the prospective prediction of the Salmonella mutagenicity of 100 chemicals which were on the way of being tested by the US National Toxicology Program, but their results had not yet been published at the time of the predictions. For the results of this prediction exercise, see also [15]. In his contribution, Ashby first assessed the chemical structures for actual or potential electrophilic centers according to the mega-structure described in [4]. As a secondary exercise, these classifications were re-assessed for the likelihood that the Salmonella assay would detect SA-containing agents as mutagens. This secondary expert judgment was based on such considerations as prior experience with a wide range of structurally diverse chemicals and mechanistic inferences [14].

APPLICATION OF SAS IN TOXICOLOGY: CASE STUDIES

The recognition of SAs and of the critical structural factors has been a very important scientific advancement, since it has contributed to the design of safer chemicals [1], and to the assessment of the toxic potential of chemicals devoid of appropriate toxicological data [8]. The use of the SAs has been extensive in the assessment of the risk posed by environmental chemicals (see, e.g., [16]). The use of the knowledge on the SAs is increasing in other application fields as well. The following is a very recent description of the use of SAs by the US Food and Drug Administration (FDA) [17].

The FDA food contact notification (FCN) process is the primary method of authorizing new uses of food additives that are food contact substances (FCS) in the U.S. The organism responsible for ensuring the safe use of U.S. food ingredients and food packaging, is the FDA Office of Food Additive Safety (OFAS), that administers the program that evaluates safety information in industry submissions for the use of various categories of food substances. In this program a SAR analysis is required for the chemicals, even if the use of the substance would result in a very low dietary concentration. Among the various classification schemes developed, most of which are based on classification of

Table I. Structural Alerts for Carcinogenicity According to Bailey *et al.* [17]

Aryl and heterocyclic ring substituted amino- and nitro-derivatives:
Primary and secondary aromatic amines (with methyl or ethyl, or activated methyl or ethyl, substituents)
Tertiary aromatic amines (with methyl or ethyl substituents)
Secondary aromatic acetamides and formamides
Nitroarenes
Nitrosoarenes
Arylhydroxylamines
Nitroso compounds:
N-nitroso-*N*-dialkylamines
N-nitroso-*N*-alkylamides
N-nitroso-*N*-alkylureas
N-nitroso-*N*-alkylcarbamates (aka urethanes)
N-nitrosos-*N*-alkylnitriles
N-nitroso-*N*-hydroxylamines
Hydrazo derivatives:
Hydrazines
Azoxy alkane
Natural electrophiles:
Aliphatic halides
Benzylic halides
Oxiranes and aziridines
Propiolactones
Alkyl esters of sulfonic and sulphuric acids (with methyl or ethyl substituents)
Alkyl esters of phosphonic and phosphoric acids (with methyl or ethyl substituents)
Mixed alkyl esters of phosphoric with methyl or ethyl substituents)
Haloethylamines
Haloalkylethers (ethyl and methyl)
α-Halocarbonyl or a-halohydroxy
Haloamines
α,β-unsaturated carbonyls (aldehyde, ketone, ester, or amide group)
Allylic halides and alkoxides (Cl, Br or I)
Other alerting groups:
Halogenated methanes
Vinyl halides (Cl, Br or I)
Polycyclic aromatic hydrocarbons
Isocyanate
Isothiocyanate
Azoarenes (sulfonic group on both rings non-alerting)

either mutagens or carcinogens into broad categories, the Ashby and Tennant classification scheme for SAs was found one of the most useful schemes to assess carcinogenic potential of an untested substance. Their list (Table **I**) of functional groups associated with DNA reactivity (genotoxicity) is based on Ashby's composite ''model structure'' [4] and a related functional groups list compiled by Munro *et al.* [18].

A second case study refers to the application of the SAs to the definition of the so-called Threshold of Toxicological Concern (TTC). This is an approach aimed at reducing extensive toxicity evaluations. This approach refers to the establishment of a generic human exposure threshold value for (groups of) chemicals below which there would be no appreciable risk to human health. The underlying principle is that such a value can be identified for many chemicals, including those of unknown toxicity, when considering their chemical structures and the known toxicity of chemicals which share similar structural characteristics. In the meantime, the concept that there are levels of exposure that do not cause adverse effects is strictly related to the possibility of setting acceptable daily intakes for chemicals with known toxicological profiles. A general TTC approach, mainly based on carcinogenicity data, was the scientific basis of the U.S. Food and Drug Administration Threshold of Regulation for indirect food additives (Threshold of Regulation for Food contact materials). Further developments [19, 20] were based on an extensive analysis of available chronic toxicity data of substances, which were divided into three chemical classes on the basis of their structure using the Cramer decision tree [21]. The cumulative distributions of NOELs (no observed effect levels) for the compounds in each Cramer structural class were plotted, and a log-normal distribution was fitted. The fifth percentile NOEL values were calculated and converted to corresponding human intakes by dividing by the usual 100-fold uncertainty factor and then multiplied by 60 to scale to the adult human body weight. These analyses gave thresholds of toxicological concern of 1800, 540, and 90 μg per person per day for Cramer structural classes I, II, and III.

Cheeseman *et al.* [22] presented an approach for extending the principle of a single threshold of regulation, applied by FDA to components of food-contact articles, to a range of dietary concentrations, by using structure-activity relationships, genotoxicity, and short-term toxicity data. In particular, in order to identify structural alerts useful to support higher threshold levels, the authors examined the most potent substances in a set of 709 carcinogens, extracted from the Gold carcinogenic potency database (http://potency.berkeley.edu/). Structural alerts similar to those utilized by Ashby and Tennant [4], were identified and correlated with the TD50s. This resulted in the identification of eight more complex, less generalized structural alerts, that include a majority of the more potent of the 709 carcinogens (Table **II**). This study shows that the inclusion of structural alerts as criteria for substances proposed for approval under a threshold of regulation process, can significantly increase the safety assurance margin. Substances that do not belong to any of the structural alert classes are likely to have much lower carcinogenic potencies, and therefore may qualify for a higher threshold level.

The scheme of structural alerts proposed by Ashby and Tennant [4] and by Cheeseman *et al.* [22] was re-examined by Kroes *et al.* [23] in order to identify the structural groups of most concern at the lowest dietary concentrations. They analyzed a database of 730 compounds (including the 709 set by Cheeseman *et al.* [22]), focused on identifying the structural alerts that would give the highest calculated risks if present at very low concentrations in the diet. In this study, five structural groups were identified to be of such high potency that if a TTC were to be established it would need to be set at a much lower dietary concentration than a TTC for other structural groups. These are three high potency genotoxic carcinogens (aflatoxin-like compounds, N-nitroso-compounds, azoxy-compounds), and two non-genotoxic carcinogens (steroids, and polyhalogenated dibenzo-p-dioxins and-dibenzofurans). In conclusion, it is suggested that a TTC would not be appropriate for chemicals with the structural alerts for high potency carcinogenicity.

IMPLEMENTATION OF THE SAS INTO SOFTWARE PROGRAMS

The need to speed up, and make as automatic as possible, the use of the knowledge on SAs in the assessment of the risk posed by the chemicals has stimulated the implementation of SAs into software programs. Several of these programs are expert system approaches that attempt to

Table II. Structural Alerts for the TTC Approach According to Cheeseman *et al.* [22]

SAs	Underlying action mechanism
N-Nitroso compounds	Bioactivated to produce highly reactive electrophile
Endocrine disruptors	Potential hormonal mechanism
Strained heteronuclear rings	Activation-independent and bioactivated to produce highly reactive electrophile
Heavy metal compounds	Neurotoxins
Alpha-nitro furyl compounds	Bioactivated to produce highly reactive electrophile
Hydrazines/triazenes/azides/azoxy	Bioactivated to produce highly reactive electrophile compounds
Polycyclic amines	Bioactivated to produce highly reactive electrophile
Organophosphorous compounds	Neurotoxins

codify existing knowledge, derived from human expert judgement, bioassay data, or any other modeling approach, into generalized rules for use in prediction. Notable examples include the DEREK system, applicable to the prediction of multiple toxicity endpoints, and the OncoLogic system, restricted to chemical carcinogenicity. Each of these expert systems serves as a repository for existing knowledge, and each rule conveys an explicit SAR hypothesis that can be refined and modified as further information becomes available. Hence, these expert systems do not discover new associations, but rather can be considered the end-stage of the model development process.

OncoLogic [24;25] was developed for the express purpose of capturing, and making available for outside use, expertise in the field of structure-based mechanisms of chemical carcinogenesis routinely being applied in regulatory setting by the US Environmental Protection Agency (EPA). Mechanism-based SAR analysis has been effectively used by EPA for many years to assess the potential carcinogenic hazard of new chemicals, for which there are no or scanty data, under the Premarketing / Premanufacturing Notification program of the Toxic Substance Control Act. Essentially, mechanism-based SAR analysis involves comparison of an untested chemical with structurally related compounds for which carcinogenic activity is known [16]. OncoLogic is a computer program consisting of four independent subsystems for estimating the carcinogenicity of fibers, metals or metal-containing compounds, polymers, and organics. Each subsystem has a hierarchical, decision-tree construction, consisting of "IF-THEN-ELSE" rules that attempt to mimic the reasoning of the human experts. This reasoning goes beyond the recognition of specific structure alerts, to consider general reactivity properties of the chemical class, structural modulators to activity, metabolic activation, and mechanisms of chemical carcinogenesis. An enhancement allows for consideration of functional, non-cancer toxicity data (e.g. genotoxicity, oncogene activation, P450 induction, etc.) in the overall decision tree to improve the chemical carcinogenicity evaluation capabilities. The organics subsystem in OncoLogic is by far the largest and most well developed of the four subsystems, with separate and distinct modules for nearly 50 chemical classes (examples include acrylates, aldehydes, and aromatic amines), although these modules vary considerably in coverage and information content. OncoLogic differs from other prediction systems in that a query molecule is not entered at the start of the analysis. Rather, a carcinogenicity evaluation of an organic chemical begins with user assignment of the chemical to one of the predefined chemical classes, and proceeds through selection of structural templates, or user-drawn entry of structures within the constraints of the chosen class. Finally, the program produces, as its primary output, a detailed justification report in which the discreet program rules are converted into a dialogue that intelligibly conveys the mechanism-based expert reasoning underlying the semi-quantitative evaluation. OncoLogic rules are all based on qualitative associations with chemical structure, i.e. the program has no capabilities for computing physical chemical properties.

Another rule-based expert system exploiting the knowledge on SAs is DEREK ("Deductive Estimation of Risk from Existing Knowledge") [26, 27], which is the result of a non-profit collaboration among the University of Leeds, and various other educational and commercial institutions, who contribute to the review and evolution of the toxicity rule bases. Also confidential in-house information from industries is used. In DEREK, rules (of the type "IF-THEN-ELSE") associate particular chemical functional groups, or SAs, with various forms of toxicity. The rules are not chemical-specific; rather they are generalizations with respect to chemical structure (e.g. halogen-containing, acid, or alkylating agent). The resulting generalized structural features used in prediction are termed toxophores. The toxicological end points currently covered by the DEREK system include carcinogenicity, mutagenicity, skin sensitization, irritancy, teratogenicity, and neurotoxicity. Each toxicity endpoint consults a different rule base, and a set of toxophores. To interrogate the system, the query structure is entered, and the rulebase is searched by comparing structural features in the target compound with the toxophores described in its knowledge base. Any Structural Alert located within the query structure is highlighted, and a message indicating the nature of the toxicological hazard is provided. In the most recent versions, the use of physical chemical properties (loP and logKp) to support the predictions about likely harmful and harmless effects of the chemicals has been also included, and the extent and quality of the supporting information has been improved. This includes declaring the mechanistic rules used by DEREK to generate its assessments, and the matches with structurally similar chemicals from a large database of chemicals with known toxicological data [28].

DEVELOPMENTS OF SAS KNOWLEDGE THROUGH ARTIFICIAL INTELLIGENCE AND DATA MINING APPROACHES

In the case studies shown above, the mechanistic knowledge on the SAs developed by human experts was implemented into computer programs e.g., OncoLogic, DEREK, in order to make it available to a wider circle of experts and practitioners of chemical risk assessment. This section of the paper regards a number of attempts aimed at expanding the knowledge on SAs through the use of a range of Artificial Intelligence / Data Mining methods applied to large databases of chemicals tested for toxicity.

A first example is provided by CASE / MULTICASE. At odds with DEREK and OncoLogic, this approach does not use a priori knowledge on the mechanisms of action or on SAs, but re-analyzes the database of chemicals in order to develop its own rules (i.e., SAs) liking the toxicity of chemicals to their structures. This approach has been implemented into a commercial computer program [29-37].

CASE and MULTICASE are a very characterized SAR approach, which is distinguished from other approaches by its central reliance on computer-generated substructural fragments, which are its major type of molecular descriptors, and the completely automated and unbiased manner in which these descriptors are generated and chosen for inclusion in SAR model. In this the CASE technology relies on previous research, whose first examples can be find in [38, 39]. In CASE (Computer Automated Structure Evaluation), each of the molecular structures belonging to the training database is

decomposed by the program into all possible constituent fragments of length 2-10 contiguous heavy atoms, with attached hydrogens and one possible side chain. The statistical analysis of the set of fragments generated by the decomposition of all molecules in the training set involves examination of the distribution of each unique fragment among active and inactive molecules, and identification of fragments whose distribution deviates from an ideal symmetrical binomial distribution: each of the fragments significantly deviating from the reference distribution is labeled either a biophore (activating fragment) or a biophobe (inactivating fragment). Biophores and biophobes are the primary molecular descriptors of the CASE (Q)SAR model: this may be expressed either as a (Bayesian) activity / inactivity probability, or as a linear regression relating a potency to the substructural descriptors. MULTICASE (MULTIple Computer Automated Structure Evaluation) is a development of the CASE program, which evolved from the recognition of problems found in the course of CASE analyses. In particular, MULTICASE responds to the problem of distinguishing between SAs that provoke the activity, and other fragments or molecular determinants that modulate the activity. In more general terms, it attempts to face the presence of hierarchy and nonlinearity within SAR models as applied to noncongeneric sets of chemicals. As CASE, MULTICASE starts by creating its own dictionary of descriptors directly from the database. At this point, and in contrast to CASE, MULTICASE selects the statistically most important of these fragments as a biophore, believed to be responsible for the observed activity of those molecules that contain it, and separates out all the molecules containing this biophore from the remaining database. This process is repeated on the remaining database with the next most significant biophore, and so on, until the database is segmented into major biophore-containing chemical classes. CASE analysis is then applied to each biophore class separately to determine substructural modifiers to the biophore activity. As final result of the analysis, a model based on SAs identified in an unbiased way by CASE / MULTICASE is provided. It should be remarked that the final set of SAs used for the model are context-dependent, and for each set of chemicals analyzed a different set of SAs is generated. In addition, if a data set with multiple recorded biological activities is studied, a specific set of SAs is obtained for each biological activity considered.

The next case study regards an approach to expand and refine the Ashby's SAs through the application of modern data mining techniques to the historical database of chemicals tested for their mutagenicity in Salmonella [40]. The procedure adopted was a combination of mechanistic knowledge, statistical tests and data mining. A dataset of compounds with available Ames data was assembled from the Carcinogenic Potency Database (http://potency.berkeley.edu/) and from other public toxicity databases (i.e., EPA/IARC Genetic Activity Profile database at http://www.epa.gov; and Developmental Therapeutics Program at dtp.nci.nih.gov/webdata.html). The data were filtered applying a series of quality criteria. A dataset of 4337 compounds with corresponding molecular structures and toxicity categorizations (2401 mutagens and 1936 nonmutagens) was extracted. In a first analysis eight historical SAs from the Ashby's compilation, able to identify correctly 75% of all mutagens in the dataset, were selected. Each of these substructures, called "general toxicophores", detects 70 or more mutagens with at least 70% of accuracy. These general toxicophores are listed in Table **III**. Among them, the aromatic nitro and amine, the azo-type groups and the three-member heterocycles are moieties well recognized toxicophores for mutagenicity. Two other simple substructure representations that perform a satisfactory detection of mutagens are the aliphatic halide group (excluding the fluorine atom) and the unsubstituted heteroatom-bonded heteroatom group (a substructure that contains an unsubstituted heteroatom that is attached with a single bond to another heteroatom). Finally, another general toxicophore was represented by large polycyclic aromatic systems, i.e., systems of three or more fused aromatic rings, whose corresponding substructure representation consists of one aromatic atom that is connected to at least two atoms belonging to multiple aromatic rings.

Table III. Basic Structural Alerts According to Kazius *et al.* [40]

1.	aromatic nitro
2.	aromatic amine
3.	three-membered heterocycle
4.	nitroso
5.	unsubstituted heteroatom bonded heteroatom
6.	azo-type
7.	aliphatic halide
8.	polycyclic aromatic system

The second step of the study consisted of improving the specificity of these simple, general toxicophores by increasing their structural complexity. General toxicophores were used to organize the data set into different subsets. Each of these subsets was then separately analyzed to derive more specific toxicophores. A final set of 29 toxicophores (Table **IV**) containing new substructures was assembled that classified the mutagenicity of the investigated dataset with 18% classification error. In this set of toxicophores, the accuracy of the "general" toxicophore (i.e. aromatic nitro, aromatic amine) was increased by the identification and the incorporation of detoxifying substructures (such as the trifluoromethyl, the sulfonamide, the sulfonic acid, and the arylsulfonyl derivatives) that were present in *ortho*, *meta*, and/or *para* position(s) with respect to the toxicophore. In addition, some general toxicophores were split into different specific toxicophores: for example the nitroso group was replaced by an aromatic nitroso, an alkyl nitrite, and a nitrosamine substructure; the azo-type group, by an azide, a diazo, a triazene, and an aromatic azo (with the incorporation of sulfonic acid group as detoxifying substructure) toxicophores. In some cases, general toxicophores were flanked by specific ones: aliphatic halides by carboxylic acid halide group and nitrogen and sulfur mustard groups, and polycyclic aromatic system by polycyclic aromatic hydrocarbons with and without bay- or K-regions. Finally, some additional toxicophores were finally identified for those mutagenic compounds (~600 compounds) that did not

contain any general toxicophore. These new groups were inserted as specific toxicophores even though the shortage of available data prohibited the *p*-value criterion.

Table IV. Extended Structural Alerts list according to Kazius *et al.* [40]

1.	specific aromatic nitro
2.	specific aromatic amine
3.	aromatic nitroso
4.	alkyl nitrite
5.	nitrosamine
6.	epoxide
7.	aziridine
8.	azide
9.	diazo
10.	triazene
11.	aromatic azo
12.	unsubstituted heteroatom-bonded heteroatom
13.	aromatic hydroxylamine
14.	aliphatic halide
15.	carboxylic acid halide
16.	nitrogen or sulfur mustard
17.	bay-region in polycyclic aromatic hydrocarbons
18.	K-region in polycyclic aromatic hydrocarbons
19.	polycyclic aromatic system
20.	sulfonate-bonded carbon (alkyl alkane sulfonate or dialkyl sulfate)
21.	aliphatic *N*-nitro
22.	α,β-unsaturated aldehyde (including R-carbonyl aldehyde)
23.	diazonium
24.	*β*-propiolactone
25.	α,β-unsaturated alkoxy group
26.	1-aryl-2-monoalkyl hydrazine
27.	aromatic methylamine
28.	ester derivative of aromatic hydroxylamine
29.	polycyclic planar system

To test the relevance of the derived list of toxicophores, the authors performed an external validation exercise by collecting a second, independent dataset of Ames test data generated with standardized protocols by either the National Toxicology Program (NTP) or the Environmental Protection Agency (EPA). On this validation set of 535 compounds (including 342 mutagens (64%) and 193 nonmutagens (36%)), the set of 29 specific toxicophores was able to identify correctly the mutagens / nonmutagens with 15% error percentage.

Finally, we will present a work (including one of the authors of the above work, J. Kazius) that tests the ability of a new approach –based on recent developments of Chemical Data Mining and Chemical Representation- to automatically identify SAs from the same database of Salmonella mutagens previously used. To exploit the chemical information that exists in a set of molecule, additional features were considered, with respect to substructure shape and atom type, to represent compounds. A graph based chemical representation was developed that allows better detection of any shape substructure, and increase the level of chemical detail considered. In addition, the authors exploited recent developments in graph mining algorithms in order to efficiently detect substructures in a database of thousands of compounds [41].

The method proposed, uses an elaborated graph-based representation of compounds. Commonly, atoms are labelled with atom type, or with a wildcard label used to indicate the presence of any atom, irrespective of its atom type. In this study, atoms were also represented with atomic hierarchies, consisting in small tree-shaped structures with one central atom label, the root, to which further atom labels are attached. The root of an atomic hierarchy is labelled with a general label, that describes a property shared by multiple atom types (such as aromatic atoms, halogen, acidic groups, or hydrogen donors or acceptors). Additional labels, called specifiers, describe more atom-specific chemical information (i.e. atom type, its formal charge and number of connected hydrogens). The advantage of using atomic hierarchies instead of standard atom types is that the resulting substructures are able to describe different degrees of chemical detail including both general and specific features in the substructure mining process. To efficiently search for two dimensional fragments (substructures of any size, shape, and level of chemical detail) among the chemicals of the data set, a novel graph-based substructure mining algorithm, named Gaston (http://www.liacs.nl/~snijssen/gaston/), was used. This algorithm iteratively performs a step that consists of both substructure generation and the corresponding substructure search, thus determining which molecules this substructure detects. The substructures of potential interest may be filtered by applying some constraints: a minimum number of chemicals that have to be detected; a size constraint on atoms and bonds; a limit on the maximum structural complexity level.

After using different scenarios of different complexity, a final list of six substructures able to satisfactorily discriminate between mutagens and nonmutagens was identified. The first substructure extracted, that is the is most discriminative for mutagenicity, is a highly branched substructure that contains 11 planar atoms connected with planar bonds. In practice, it describes a polycyclic planar system consisting of at least three rings. The second substructure contains a nitrogen atom that is connected through a double bond to a nitrogen or oxygen atom. It comprises three previously described structural alerts (the aromatic nitro, the nitroso, and the azo-type group), but also detects different chemical compounds previously not classified or predicted as mutagens.

The subsequent three substructures, select aliphatic epoxides and aziridines, aliphatic halogens (Cl, Br, I) and aromatic primary amine respectively, already known as toxicophores. The last substructure describes an heteroatom-

bonded heteroatom that detects heteroatom bound secondary amines, hydroxylamines, primary peroxides and primary amines that are connected to secondary amines. The authors emphasize that this final decision list of only six substructures -generated in a completely automatic way with no a priori knowledge about mechanisms of action- had an error of 20% in classifying the mutagenicity of the data set, which is comparable to that obtained in the previous work [40] based on the use of a priori knowledge on the SAs.

CONCLUSIONS

The main lesson of this mini-review is that the research in the field of modeling the structural properties of mutagens and carcinogens is a highly interdisciplinary work. In particular, mechanistic research based on experimental systems together with human ingenuity in the interpretation of the results has provided the essential basis for the identification of the Structural Alerts that characterize the mutagens and carcinogens. More recently, *in silico* methods have provided the means to verify, refine and implement into computer programs the knowledge on the Structural Alerts. At the same time, *in silico* methods permit the dissemination of this knowledge to a wider circle of experts and practitioners of the risk assessment procedures.

REFERENCES

[1] Ariens, E.J. *Drug Metab. Rev.,* **1984,** *15,* 425.
[2] Miller, J.A.; Miller, E.C. *Origins of human cancer*; Cold Spring Harbor Laboratory: Cold Spring Harbor, **1977,** pp. 605.
[3] Miller, E.C.; Miller, J.A. *Cancer,* **1981,** *47,* 2327.
[4] Ashby, J.; Tennant, R.W. *Mutat. Res.,* **1988,** *204,* 17.
[5] Arcos, J.C.; Argus, M.F. Chemical induction of cancer. Modulation and combination effects; Birkhauser: Boston, **1995,** pp. 1.
[6] Woo, Y.T. Quantitative Structure-Activity Relationship (QSAR) models of mutagens and carcinogens.; CRC Press: Boca Raton, **2003;** Chapter 2, pp. 41.
[7] Ashby, J. *Environ. Mutagen.,* **1985,** *7,* 919.
[8] Woo, Y.T.; Arcos, J.C. *Carcinogenicity and pesticides: Principles, issues, and relationships*; American Chemical Society: **1989;** Chapter 11, pp. 175.
[9] Ashby, J.; Paton, D. *Mutat. Res.,* **1993,** *286,* 3.
[10] Ashby, J.; Tennant, R.W.; Zeiger, E.; Stasiewicz, S. *Mutat. Res.,* **1989**, *223*, 73.
[11] Ashby, J.; Tennant, R.W. *Mutat. Res.,* **1991,** *257,* 229.
[12] Brown, L.P.; Ashby, J. *Mutat. Res.,* **1990,** *244,* 67.
[13] Tennant, R.W.; Ashby, J. *Mutat. Res.,* **1991,** *257,* 209.
[14] Zeiger, E.; Ashby, J.; Bakale, G.; Enslein, K.; Klopman, G.; Rosenkranz, H.S. *Mutag.,* **1996,** *11,* 474.
[15] Benigni, R. *Chem. Revs.,* **2005,** *105,* 1767.
[16] Woo, Y.T.; Lai, D.Y.; McLain, J.L.; Ko Manibusan, M.; Dellarco, V. *Environ. Health Perspect.,* **2002,** *110,* 75.
[17] Bailey, A.B.; Chanderbhan, N.; Collazo-Braier, N.; Cheeseman, M.A.; Twaroski, M.L. *Regulat. Pharmacol. Toxicol.,* **2005,** *42,* 225.
[18] Munro, I.C.; Ford, R.A.; Kennepohl, E.; Sprenger, J.G. *Drug Metab. Rev.,* **1996,** *28,* 209.
[19] Munro, I.C.; Ford, R.A.; Kennepohl, E.; Sprenger, J.G. *Food Chem. Toxicol.,* **1996,** *34*, 829.
[20] Munro, I.C.; Kennepohl, E.; Kroes, R. *Food Chem. Toxicol.,* **1999,** *37,* 207.
[21] Cramer, G.M.; Ford, R.A.; Hall, R.L. *Food Cosmet. Toxicol.,* **1978,** *16,* 255.
[22] Cheeseman, M.A.; Machuga, E.J.; Bailey, A.B. *Food Chem. Toxicol.,* **1999,** *37,* 387.
[23] Kroes, R.; Renwick, A.G.; Cheeseman, M.A.; Kleiner, J.; Mangelsdorf, I.; Piersma, A.; Schilter, B.; Schlatter, J.; van Schothorst, F.; Vos, J.G.; Wurtzen, G. *Food Chem. Toxicol.,* **2004,** *42,* 65.
[24] Woo, Y.T.; Lai, D.Y.; Argus, M.F.; Arcos, J.C. *Toxicol. Lett.,* **1995,** *79,* 219.
[25] Woo, Y.T.; Lai, D.Y.; Argus, M.F.; Arcos, J.C. *J. Environ. Sci. Health. C. Environ. Carcinog. Ecotoxicol. Revs.*, **1998,** *C16*, 101.
[26] Sanderson, D.M.; Earnshaw, C.G. *Hum. Exp. Toxicol.,* **1991,** *10,* 261.
[27] Ridings, J.E.; Barratt, M.D.; Cary, R.; Earnshaw, G.G.; Eggington, C.E.; Ellis, M.K.; Judson, P.N.; Langowski, J.J.; Marchant, C.A.; Payne, M.P.; Watson, W.P.; Yih, T.D. *Toxicol.,* **1996,** *106,* 267.
[28] Greene, N.; Judson, P.N.; Langowski, J.J.; Marchant, C.A. *SAR QSAR Environ. Res.,* **1999,** *10,* 299.
[29] Klopman, G. *J. Amer. Chem. Soc.,* **1984,** *106,* 7315.
[30] Rosenkranz, H.S.; Klopman, G. *Mutag.,* **1990,** *5,* 333.
[31] Klopman, G.; Rosenkranz, H.S. *Mutat. Res.,* **1992,** *272,* 59.
[32] Klopman, G. *Quant. Struct.-Act. Relat.,* **1992,** *11,* 176.
[33] Klopman, G.; Rosenkranz, H.S. *Mutat. Res.,* **1994,** *305,* 33.
[34] Cunnigham, A.R.; Rosenkranz, H.S.; Zhang, Y.P.; Klopman, G. *Mutat. Res.,* **1998,** *398,* 1.
[35] Cunnigham, A.R.; Klopman, G.; Rosenkranz, H.S. *Mutat. Res.,* **1998,** *405,* 9.
[36] Cunningham, A.R.; Rosenkranz, H.S. *Environ. Health Perspect.,* **2001,** *109,* 953.
[37] Rosenkranz, H.S. Quantitative Structure-Activity Relationship (QSAR) models of chemical mutagens and carcinogens.; CRC Press: Boca Raton, **2003;** Chapter 6, pp. 175.
[38] Hodes, L.; Hazard, G.F.; Geran, R.I.; Richman, S. *J. Med. Chem.,* **1977,** *20,* 469.
[39] Tinker, J. *J. Chem. Inf. Comput. Sci.,* **1981,** *21*, 3.
[40] Kazius, J.; McGuire, R.; Bursi, R. *J. Med. Chem.,* **2005,** *48,* 312.
[41] Kazius, J.; Nijssen, S.; Kok, J.; Back, T.; Ijzerman, A.P. *J. Chem. Inf. Model.,* **2006,** *in press*.

Received: March 24, 2006 **Revised: March 27, 2006** **Accepted: March 27, 2006**