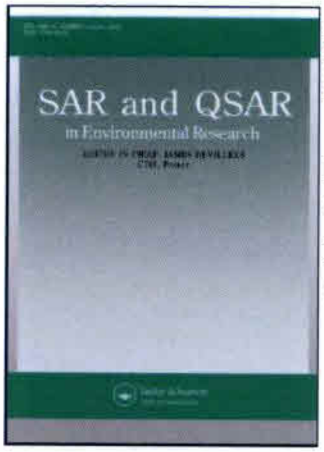

SAR and QSAR in Environmental Research
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t716100694

Development of an information-intensive structure-activity relationship model and its application to human respiratory chemical sensitizers

A. R. Cunningham[a]; S. L. Cunningham[a]; D. M. Consoer[a]; S. T. Moss[a]; M. H. Karol[b]
[a] Department of Environmental Studies, Louisiana State University, Baton Rouge, LA, USA [b] Department of Environmental and Occupational Health, University of Pittsburgh, Pittsburgh, PA, USA

To cite this Article Cunningham, A. R. , Cunningham, S. L. , Consoer, D. M. , Moss, S. T. and Karol, M. H.(2005) 'Development of an information-intensive structure-activity relationship model and its application to human respiratory chemical sensitizers', SAR and QSAR in Environmental Research, 16: 3, 273 — 285

To link to this Article: DOI: 10.1080/10659360500036976

URL: http://dx.doi.org/10.1080/10659360500036976

SAR and QSAR in Environmental Research, Vol. 16, No. 3, June 2005, 273–285

Taylor & Francis
Taylor & Francis Group

Development of an information-intensive structure–activity relationship model and its application to human respiratory chemical sensitizers

A. R. CUNNINGHAM†*, S. L. CUNNINGHAM†, D. M. CONSOER†,
S. T. MOSS† and M. H. KAROL‡

†Department of Environmental Studies, Louisiana State University, Baton Rouge, LA 70803, USA
‡Department of Environmental and Occupational Health, University of Pittsburgh,
Pittsburgh, PA 15261, USA

(Received 20 May 2004; in final form 9 October 2004)

Structure–activity relationship (SAR) models are recognized as powerful tools to predict the toxicologic potential of new or untested chemicals and also provide insight into possible mechanisms of toxicity. Models have been based on physicochemical attributes and structural features of chemicals. We describe herein the development of a new SAR modeling algorithm called cat-SAR that is capable of analyzing and predicting chemical activity from divergent biological response data. The cat-SAR program develops chemical fragment-based SAR models from categorical biological response data (e.g. toxicologically active and inactive compounds). The database selected for model development was a published set of chemicals documented to cause respiratory hypersensitivity in humans. Two models were generated that differed only in that one model included explicate hydrogen containing fragments. The predictive abilities of the models were tested using leave-one-out cross-validation tests. One model had a sensitivity of 0.94 and specificity of 0.87 yielding an overall correct prediction of 91%. The second model had a sensitivity of 0.89, specificity of 0.95 and overall correct prediction of 92%. The demonstrated predictive capabilities of the cat-SAR approach, together with its modeling flexibility and design transparency, suggest the potential for its widespread applicability to toxicity prediction and for deriving mechanistic insight into toxicologic effects.

Keywords: Structure–activity relationship (SAR); *In silico* modeling; Respiratory sensitizer; Predictive toxicology; Chemical fragments; Categorical SAR (cat-SAR) program

1. Introduction

The task of identifying toxic agents is not a small or trivial challenge. One approach has been to use mathematical models that relate biological activity to chemical structure. Benfenati and Gini [1] describe modern structure–activity relationship (SAR) and quantitative SAR (QSAR) methods as typically involving three parts: (1) the chemical part, (2) the biological part (i.e. activity) and (3) the methodology for relating parts 1 and 2. The main premise for these methods is that recurring and identifiable attributes of chemicals are associated with, or responsible for, particular biological effects. The attributes can take many forms including

*Corresponding author. Email: arc@lsu.edu

SAR and QSAR in Environmental Research
ISSN 1062-936X print/ISSN 1029-046X online © 2005 Taylor & Francis Group Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/10659360500036976

chemical structures, chemicophysical or quantum mechanical properties and graph indices, to name a few. There are numerous methods that relate chemical structure with activity such as those based on human expertise like Ashby's "structural alerts" for potential carcinogenicity [2–4] to statistical QSAR methods like Hansch analysis (see e.g. [5]), comparative molecular field analyses (CoMFA) [6] and MCASE [7–9].

Advances in computing and chemoinformatics, standardized biological or toxicological testing, and the subsequent development of large libraries of test results have ushered in the era of computational or *in silico* SAR. Computational SAR models have gained recent acceptance in the regulatory community for both human health [10] and ecological endpoints [11]. Dearden succinctly summarized the field of computational SAR or *in silico* toxicity prediction to include QSAR models of congeneric and noncongeneric datasets and "expert systems" [12]. The utility and application of some important expert system toxicology prediction methods have been reviewed by Richard [13,14]. Through the use of various techniques, the overall goal is to identify meaningful associations between activity and chemical structure. These associations can then be used to investigate the underlying mechanisms of toxicity, or be extended to estimate or predict the toxicity of untested compounds.

With today's fast CPUs, abundant amounts of computer memory, and the availability of chemical informatics and graphics software we have aimed to readdress the challenge of computer-based SAR expert systems for modeling large and chemically diverse datasets. We describe herein the first generation of a new data and information-intensive approach to toxicological SAR modeling. The program is based on the well-established premise in SAR modeling that like structure begets like activity and employs chemical substructures to differentiate between categories of biologically active and inactive compounds for toxicological endpoints. We have named the new program cat-SAR for categorical SAR.

The cat-SAR program uses 2-dimensional chemical fragments generated by the Sybyl HQSAR module. We chose early in the development process of cat-SAR to use the Sybyl platform which already possessed the needed utilities of *in silico* chemical fragmenting, molecular graphics, and chemical informatics and database requirements associated with our modeling goals. Of importance, the HQSAR module is used solely to generate molecular fragments and is not used for further model development or statistical analysis.

Briefly, the HQSAR module is used to generate a list of chemical fragments associated with compounds in a learning set and produce a data matrix of compounds and fragments. In the data matrix, the rows are the chemicals and the columns are the molecular fragments. Thus for each chemical, a tabulation of all its fragments are recorded across the table rows and for each fragment all chemicals that contain in it are tabulated down the table columns. The compound-fragment matrix is then analyzed, in conjunction with the known biological activity category of each compound, by the cat-SAR program. The cat-SAR program identifies structural features associated with the biologically active and inactive categories. The cat-SAR program, the respiratory sensitizer learning set (described below), and the compound-fragment matrix are available through the corresponding author.

Since cat-SAR modeling is independent of the biological data used in the process we anticipate that it can be generally applied from the study of drugs to environmental toxicants. Moreover, the models can be used for either mechanistic studies of biological phenomena or for the prediction of biological activity for untested compounds.

The cat-SAR program stands alone from other computerized SAR expert systems in its openness, flexibility, routine for identifying important attributes of biological activity or inactivity, and its method for predicting the activity of untested compounds. Several commercially available computational SAR expert systems including MultiCASE, TOPKAT, and Oncologic are relatively closed systems where proprietory (and unknown) routines are used to generate the final model. On the other hand, cat-SAR is completely open with every detail of modeling transparent to the user. As for inflexibility, many of the commercially available expert systems maximally only allow the user to alter the makeup of the learning sets (users cannot alter the parameters for model development). The cat-SAR approach allows the user to select and/or adjust many parameters during the model process from learning set makeup, to selection of types of fragment attributes to consider, to ultimately what numerical or statistical considerations are employed in developing the final model. These are described in detail below.

The cat-SAR approach is also a very data- and information-intensive SAR expert system. During model development and the creation of the final model, all fragments associated with the categories are presented. This leaves the user with an unbiased view of all important features associated with the biological endpoint. Consider the fact that the published MCASE model of the same respiratory sensitizer learning set used herein produced a model based on eight biophores and no biophobes [15]. One of the models developed with the cat-SAR program produced 1213 fragments associated with activity and 92 associated with inactivity. Similarly, the prediction of the activity of compounds outside the model's learning set presents the user with a *complete* correspondence between all the fragments in the model (e.g. 1213 active and 92 inactive) and those in the compound being predicted. Again considering the published MultiCASE report for this dataset, MultiCASE predicted the activity of methyldopa and presented the user with two reasons (i.e. biophores) for why the compound was predicted active. The cat-SAR program provided 22 reasons.

The approach we have taken in developing cat-SAR clearly diverges from existing SAR expert systems and is more in tune with modern QSAR techniques. For instance, the user is presented with a number of selectable and adjustable modeling parameters. The notion of having selectable and adjustable modeling parameters facilitates that ability to rigorously explore the relationships between chemical structure and biological activity.

We chose to test the method on a previously published respiratory sensitization model due to its small size (i.e. 80 compounds) and good modeling potential that was previously demonstrated using CASE-MultiCASE [15]. This model has recently been reviewed by Rodford *et al.* [16].

2. Materials and methods

2.1 Description of the cat-SAR SAR program

The cat-SAR models are built through a comparison of structural features found amongst the active and inactive compounds in the model's learning set. A categorical approach is used with, in this instance, compounds designated as active or inactive. For this exercise, active compounds were chemical respiratory sensitizers and inactive compounds were nonsensitizers. The modeling process began with the compilation of a set of chemicals and their biological activity (described below). Using the Tripos Sybyl HQSAR module, each chemical was fragmented into all possible fragments. HQSAR allows the user to select attributes for fragment determination including atom size, bond types, atomic connections, inclusion of hydrogen atoms, chirality and hydrogen bond donor and acceptor atoms. Moreover, fragments can be linear, branched or cyclic moieties.

We developed two sets of fragments from the model's learning set. The first (fragment set ABC) contained fragments between three and seven atoms in size and considered Atoms, Bond types, and atomic Connections (i.e. the arrangement of atoms in the fragment). The second (fragment set ABCH) included the same descriptors as the previous set plus associated Hydrogen atoms. A compound-fragment matrix was produced for both sets of fragments.

A measure of each fragment's association with biological activity was next determined. This step is controlled by the user. To ascertain an association between each fragment and activity (or lack of activity) a set of rules is established to choose "important" active and inactive fragments. It should be noted that in this generation of the program we are using a common-sense approach, rather than statistical analysis, to select "significant" fragments.

The first selection rule is the number of times a fragment is identified in the learning set. For this exercise, it was arbitrarily set at three compounds (or 3.75% of the compounds in the learning set). This was a reasonable decision considering that if a fragment is found in only one or two compounds in the learning set it may be a chance occurrence. We do, however, note that fragments found in only one or two compounds may not be outliers but rather underrepresented descriptors of activity. On the other hand, since the learning set is composed of only 40 active and 40 inactive compounds (see next section), if we required fragments to be found in more than three compounds, we would expect to miss important features.

The second rule relates to the proportion of active or inactive compounds that contain each fragment. For both the ABC and ABCH fragment sets, we set the proportion at 0.90. We reasoned that even if a particular fragment is associated with activity, there may yet be other reasons (i.e. fragments) for its being inactive, thus it would not be expected to be found in 100% of the active compounds. Likewise is true for inactive fragments. Thus, if we considered only those fragments found exclusively in active or inactive compounds we would rarify the fragments pool to an unreasonable level and risk losing valuable information. On the other hand, we expected that fragments found to be present approximately equally in the active and inactive fragment sets would not be associated with biological activity. Such fragments may serve as chemical scaffolds holding the biologically active features and are not directly related to activity or inactivity.

In summary, fragments were considered "significant" if they were found in at least three compounds in the learning set and also found in at least 90% of the active or inactive compounds that derived them.

The resulting list of fragments can then be used for mechanistic analysis, or to predict the activity of an unknown compound. In the latter circumstance, the model determines which, if any, fragments from the model's learning set the compound contains. If none are present, no prediction of activity is made for the compound. If one or more fragments are present, the number of active and inactive compounds containing each fragment is determined. The probability of activity or inactivity is then calculated based on the total number of active and inactive compounds containing the fragments.

The probability of activity of a test chemical is calculated from the average probability of active and inactive fragments. For example, if a test compound contains two fragments, one present in 9/10 active compound (i.e. 90% active) and one in 3/3 inactive one (i.e. 100% inactive), the unknown compound will be predicted to be *active* based on the higher probability of activity derived from chemicals containing these fragments.

In this manner, the probability of activity or inactivity is determined by comparison of the structure of the unknown compound with the entire structural information present in the model.

It requires noting that cat-SAR predictions are based on what can be conceived as two separable models: The inactive fragment model and the active fragment model. By so doing, cat-SAR predictions are based on information that is associated with biological activity and inactivity. The cat-SAR program does not employ the use of default predictions wherein, as in the case of MultiCASE, if no biophores are present in an unknown chemical it is predicted by default to be inactive. This, of course, presents the situation wherein the cat-SAR program will not make predictions on some chemicals. Although this may seem like a drawback to the program by appearing less universal, the user of the program always has the option to simply define chemicals that are not predictable by cat-SAR with a default value.

2.2 Respiratory sensitization databases

The dataset of respiratory sensitizers has been reported by Graham *et al.* [15]. Briefly, chemical sensitizers were identified through a search of the medical literature. Selection criteria were in accordance with the US Department of Health and Human Services "Guidelines for Diagnosis and Treatment of Asthma" [17]. The search criteria included chemicals with inhalation challenge followed by a drop of $>20\%$ in forced expiration volume at 1 s within 24 h of challenge. Forty compounds were identified. No reports were identified of chemicals tested as described and found to be nonsensitizers in humans except for the often-used control substance, lactose. Since, as discussed, the cat-SAR method requires a comparison of biologically active with inactive compounds, we designated as "negative" a set of 40 chemicals previously selected as respiratory nonsensitizers by Graham *et al.* [15]. These 40 compounds were randomly selected from a dataset of chemicals tested for human allergic contact sensitizing ability via patch testing and were found to be nonsensitizers [18]. The assumption was made that dermal nonsensitizers would also be respiratory nonsensitizers. In general, chemicals were relatively small organic compounds that did not include salts, metals, mixtures, or polymers.

3. Results and discussion

3.1 Predictive performance of the cat-SAR respiratory sensitization models

To evaluate the predictive ability of the models, a leave-one-out cross-validation test was conducted. For each chemical in the learning set, one at a time, its chemical fragments were removed from the total fragment set, and the probability of activity or inactivity associated with each fragment was recalculated. Using the criteria described above to estimate activity of unknown compounds, the activity of the removed chemical was predicted.

Overall, the ABC and ABCH models correctly classified 91 and 92% of the chemicals they were capable of predicting (table 1). The predicted activity for each chemical is listed in table 2. The cat-SAR program, using the n-1 cross-validation learning sets (i.e. models built on 79 compounds), was unable to make predictions for five chemicals in the ABC model and

Table 1. Predictive performance of ABC and ABCH respiratory sensitization models. The ABC model was based on fragments of size between three and seven heavy atoms and considered atoms, bonds, and atom connection. The ABCH model also included consideration of hydrogen atoms.

Model	*Total. Fragments**	*Model Fragments†*	*Active Fragments‡*	*Inactive Fragments¶*	*Sensitivity§*	*Specificity‖*	*OCP#*
ABC	5737	1305	1213	92	0.94	0.87	0.91
ABCH	14424	3356	2926	430	0.89	0.95	0.92

*number of fragments derived from learning set.
†number of fragments meeting specified rules of the model.
‡number of fragments meeting specified rules to be considered as active.
¶number of fragments meeting specified rules to be considered as inactive.
§number of correct positive predictions / total number of positives.
‖number of correct negative predictions / total number of negatives.
#Observed Correct Predictions: Number of correct predictions / total number of predictions.

three in the ABCH (table 2). The reason for this is that each of these compounds did not possess any structural features that the n-1 models could base a prediction upon. A previous CASE/MultiCASE model of the same data reported an overall correct classification of 95%. This was based on the Bayesian combination of four CASE/MultiCASE submodels that individually had sensitivities ranging from 72–80% and specificities ranging from 95–98% [15]. In a separate published model based on chemicophysical parameters, a sensitivity of 85% and a specificity of 74% was achieved [19]. Interestingly, the individual ABC and ABCH cat-SAR models are quite balanced with respect to sensitivity and specificity (table 1). This is not the case with the previous CASE/MultiCASE and chemicophysical models. The individual CASE/MultiCASE models tended to have a better ability to predict the inactive chemicals and the chemicophysical model was better able to predict the active ones.

The question arises as to why the program produced wrong predictions. In the case of any of the previously mentioned respiratory sensitizing models, the simplest explanation lies in the possibility that some of the information on which the models were built is not correct. Consider the National Toxicology Program's *Salmonella* mutagenicity database. The *Salmonella* database is derived from a standardized protocol and, more importantly, has been analyzed for reproducibility and accuracy by replicate analyses of chemicals [20]. The interlaboratory reproducibility of the *Salmonella* mutagenicity assay is only 85% [20]. Therefore, the databases may contain some incorrect information.

However, other explanations should be considered. The incorrect ABC model prediction for hexamethylene diisocyanate and the incorrect ABC and ABCH model predictions for isophorone diisocyanate are of interest. They both contain the isocyanate moiety which is clearly associated with biological activity. The cat-SAR program also identifies this moiety in these two compounds. However, the compounds contain a number of inactivating fragments that counterbalance the isocyanate-related ones. At this time, a complete understanding of the inaccurate predictions is not possible, but further development of both the models and the databases should lead to a more comprehensive analysis.

3.2 Respiratory sensitization model analysis

As described above, two models were developed using the same set of 80 compounds. These models can be considered as independent since they are built upon different fragment bases. The ABC model started with a total fragment set of 5737 and the ABCH model with a set of

Table 2. Model validation for respiratory sensitizers. Compounds with values above 50% were predicted to be active compounds and those below 50% were predicted to be inactive.

		Model 3-7/3/0.90	
Chemical	*Experimental Activity*	*ABC % Active*	*ABCH % Active*
1,5-Napthalene diisocyanate	+	1.00	1.00
2-(*N*-Benzyl-*N*-*tert*-butylamino)-4′-hydroxy-3′-hydroxymethyl acetophenone diacetate	+	0.63	0.59
2,4-Toluene diisocyanate	+	1.00	1.00
2,6-Toluene diisocyanate	+	1.00	1.00
6-Amino penicillanic acid	+	1.00	1.00
7-Amino cephalosporanic acid	+	0.99	0.99
Ampicillin	+	1.00	1.00
Azocarbonamide	+	1.00	0.98
Benzylpenicillin	+	1.00	1.00
Brilliant orange GR	+	1.00	1.00
Carminic acid	+	0.57	0.54
Cephalexin	+	1.00	1.00
Chlorhexidine	+	1.00	0.96
Dichlorvos	+	*	*
Dimethyl ethanolamine	+	1.00	1.00
Diphenyl methane-4,4′-diisocyanate	+	1.00	1.00
Epigallocatechin gallate	+	0.57	0.60
Ethanolamine	+	1.00	1.00
Ethyl cyanoacrylate	+	*	$0.03^{\dagger}$
Ethylenediamine	+	1.00	1.00
Fenthion	+	0.91	0.96
Hexamethylene diisocyanate	+	1.00	$0.38^{\dagger}$
Isononanoyl oxybenzene sulfonate	+	0.98	0.82
Isophorone diisocyanate	+	$0.22^{\dagger}$	$0.17^{\dagger}$
Maleic anhydride	+	1.00	1.00
Methyl-2-cyanoacrylate	+	*	*
Methyldopa	+	0.99	0.95
Phenylglycine acid chloride	+	1.00	1.00
Phthalic anhydride	+	1.00	1.00
Piperacillin	+	1.00	1.00
Piperazine	+	1.00	1.00
Plicatic acid	+	0.53	0.74
Reactive orange 3R	+	1.00	1.00
Rifafix red BBN	+	1.00	1.00
Rifazol black GR	+	1.00	1.00
Tetrachloroisophthalonitrile	+	*	*
Tetrachlorophthalic anhydride	+	1.00	1.00
Triethylenetetramine	+	1.00	1.00
Trimellitic anhydride	+	1.00	1.00
Tylosin	+	$0.14^{\dagger}$	$0.14^{\dagger}$
1,1,3,3,5-Pentamethyl-4,6-Dinitroindane	−	0.00	0.00
1,4-Cineole	−	0.00	0.04
1-Hexanol	−	*	0.07
2,4-Dimethylbenzyl acetate	−	0.00	0.02
2-Butyl-4,4,6-trimethyl-1,3-dioxane	−	$1.00^{\dagger}$	0.50
2-*tert*-Amylcyclohexyl acetate	−	0.03	0.06
3,6-Dimethyloctan-3-yl acetate	−	0.05	0.06
3-Butyl phthalide	−	0.03	0.06
4-Acetyl-6-*tert*-butyl-1,1-dimethylindane	−	0.00	0.06
5-Methyl α-ionone	−	0.12	0.09
9-Decenyl acetate	−	0.05	0.05
Acetyl ethyltetramethyltetralin	−	0.00	0.00
Allyl heptylate	−	0.10	0.05
Benzyl butyrate	−	0.10	0.06
Butyl isobutyrate	−	0.06	0.07
Camphene	−	0.00	0.04
cis-3-Hexenyl anthranilate	−	$0.65^{\dagger}$	0.35

Table 2 – *continued*

		Model 3-7/3/0.90	
Chemical	*Experimental Activity*	*ABC % Active*	*ABCH % Active*
cis-4-Decen-1-al	–	0.03	0.04
Citronellyl nitrile	–	0.03	0.05
Cyclohexylethyl alcohol	–	0.00	0.06
Dibutyl sulphide	–	1.00[†]	0.93
Dihydro-isojasmone	–	0.03	0.04
Dimethylheptenol	–	0.03	0.05
Ethyl acetoacetate ethylene glycol ketal	–	0.27	0.19
Ethyl lactate	–	0.09	0.07
Eugenyl phenylacetate	–	1.00[†]	0.81[†]
γ-Dodecalactone	–	0.05	0.07
Geranyl benzoate	–	0.03	0.06
Heptyl butyrate	–	0.06	0.06
Hexane	–	0.00	0.09
Hexyl tiglate	–	0.04	0.06
Isoamyl butyrate	–	0.06	0.06
Lactoscatone	–	0.04	0.05
l-Carvyl propionate	–	0.04	0.04
Methyl tiglate	–	0.09	0.07
Musk xylol	–	0.00	0.00
Phenylethyl acetate	–	0.77[†]	0.32
p-Isopropylcyclohexanol	–	0.00	0.04
Rhodinyl formate	–	0.03	0.05
Undecenyl acetate	–	0.05	0.05

* no prediction was made for the compound.

[†] wrong prediction was made for the compound.

14424 fragments (table 1). In both models, approximately 23% of the total number of fragments met the criteria to be considered "significant" (i.e. 1307 significant /5753 total = 22.7% for ABC and 3356 significant /144424 total = 23.2%) (table 1). The remaining fragments were either not present in a sufficient number of compounds (i.e. found in <3 or 3.75% of compounds in the learning set), or the fragments did not come from compounds that were predominately (i.e. >90%) active or inactive.

Overall, both models performed similarly. However, when considering the sensitivity and specificity of the models, the distinction was not clear-cut. The ABC model was better able to correctly predict the active chemicals while the ABCH model was better able to predict the inactive ones. At this point, we chose to focus on the ABC model. This decision was based on several criteria: (1) Both models have nearly equivalent correct prediction rates (table 1) and make similar predictions on the majority of compounds in the validation set (table 2), (2) Considering the law of parsimony, the ABC model is based on fewer fragments and (3) The models are constructed from a set of 40 chemicals *tested* and found to be respiratory sensitizers, whereas the set of 40 chemicals designated as "inactive" are *presumed* to lack activity. Therefore, based on the quality of information of these active and inactive sets, we favored a model with better ability to predict activity as compared with inactivity.

Although beyond the scope of this report, we bring attention to the finding that the cat-SAR method derives multiple independent models for the same endpoint. The observation that the ABC and ABCH models do not predict the same activity for each chemical suggests that the models may be capable of describing different attributes of the activity. This suggests the possibility of development of a consensus model using a Bayesian technique similar to those previously reported using CASE/MultiCASE [15].

3.3 Examples of the cat-SAR model predictions

Methyldopa and 2,4-dimethylbenzyl acetate were selected to demonstrate the predictive ability of the cat-SAR modeling method for an active and inactive chemical, respectively. For this demonstration, we used the ABC model for reasons just described. Tables 3 and 4 list the significant fragments derived from the two compounds. Figures 1 and 2 illustrate the intact compounds and their associated fragments. The predictions presented for the two compounds are based on results obtained from the leave-one-out validation exercise. Therefore, the compounds themselves are not contributing to the fragment set of the model and are thus not influencing their own prediction of activity or inactivity.

Table 3 lists and figure 1 shows all the significant fragments used in the leave-one-out validation exercise to predict the activity of methyldopa. Methyldopa was predicted to have a probability of activity of 0.988. This represents the average probability of activity of the 22 fragments used in the prediction (table 2). No fragments associated with methyldopa were considered inactive.

Likewise, table 4 and figure 2 shows all the significant fragments used in the validation exercise to predict the activity of 2,4-dimethylbenzyl acetate. 2,4-Dimethylbenzyl acetate was predicted to have a probability of inactivity of 1.0.

As indicated, the prediction for the respiratory sensitizing ability of methyldopa and 2,4-diemthylbenzyl acetate were based on the complete correspondence of significant fragments

Table 3. Fragments from the ABC model leave-one-out validation analysis used to predict the activity of the respiratory sensitizer methyldopa

Fragment	*No. Active**	*No. Inactive*†	*Total*‡	*% Active*	*% Inactive*
frag258	10	1	11	0.909	0.091
frag283	10	1	11	0.909	0.091
frag308	10	1	11	0.909	0.091
frag348	8	0	8	1.000	0.000
frag357	8	0	8	1.000	0.000
frag400	14	0	14	1.000	0.000
frag471	6	0	6	1.000	0.000
frag522	6	0	6	1.000	0.000
frag914	4	0	4	1.000	0.000
frag915	4	0	4	1.000	0.000
frag920	4	0	4	1.000	0.000
frag921	4	0	4	1.000	0.000
frag2378	3	0	3	1.000	0.000
frag2401	3	0	3	1.000	0.000
frag2415	3	0	3	1.000	0.000
frag2416	3	0	3	1.000	0.000
frag2463	3	0	3	1.000	0.000
frag2471	3	0	3	1.000	0.000
frag2472	3	0	3	1.000	0.000
frag2507	3	0	3	1.000	0.000
frag2509	3	0	3	1.000	0.000
frag2706	3	0	3	1.000	0.000
	Probability of activity			0.988	0.012

* number of active compounds that contain the fragment.
† number if inactive compounds that contain the fragment.
‡ number of compounds in the dataset that contain the fragment.

Table 4. Fragments from the ABC model leave-one-out validation analysis used to predict the activity of the respiratory nonsensitizers 2,4-Dimethylbenzyl acetate.

Fragment	*No. Active**	*No. Inactive†*	*Total‡*	*% Active*	*% Inactive*
frag4970	0	3	3	0.000	1.000
frag4979	0	3	3	0.000	1.000
frag4982	0	3	3	0.000	1.000
frag5003	0	4	4	0.000	1.000
frag5011	0	4	4	0.000	1.000
frag5032	0	4	4	0.000	1.000
frag5033	0	4	4	0.000	1.000
frag5073	0	4	4	0.000	1.000
	Probability of activity			0.000	1.000

See table 3 footnotes for reference.

from the model's validation set to all the fragments identified in the compound. Methyldopa was predicted to be active based on 22 fragments from its validation set of fragments. Inspection of these fragments revealed several major themes. Fragment 348 leads to a series of complimentary moieties covering the amine to carboxylic acid portion of the molecule. Fragment 283 covers the *para* unsubstituted phenol and accounts for four other validation fragments. Fragment 2706

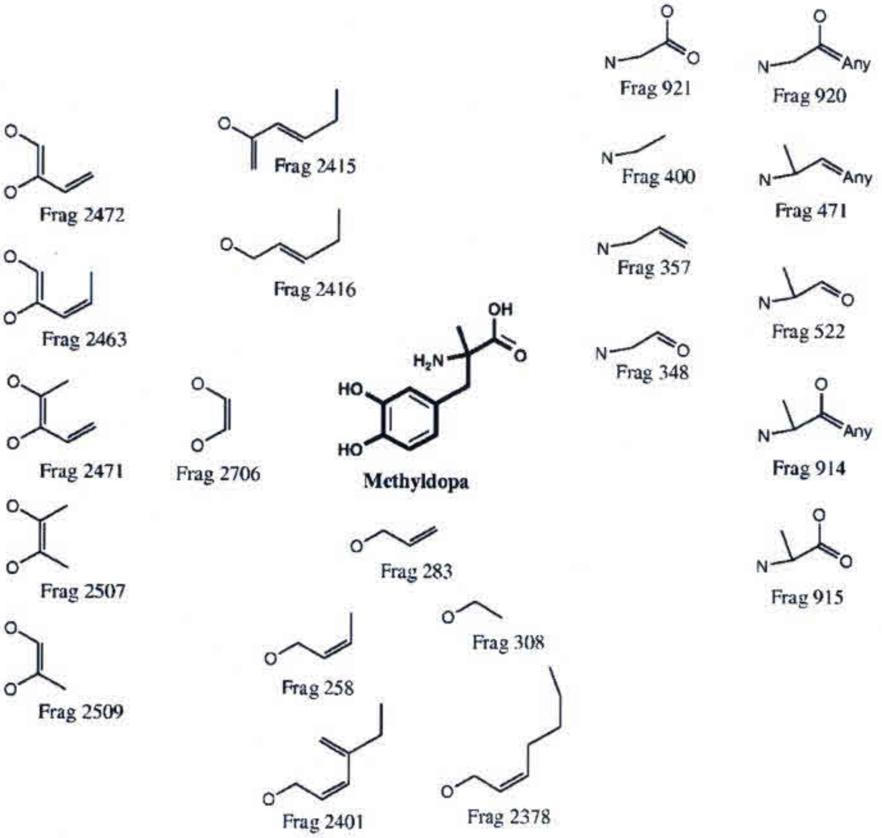

Figure 1. Illustration of the 22 significant fragments contributing to the active validation prediction of methyldopa.

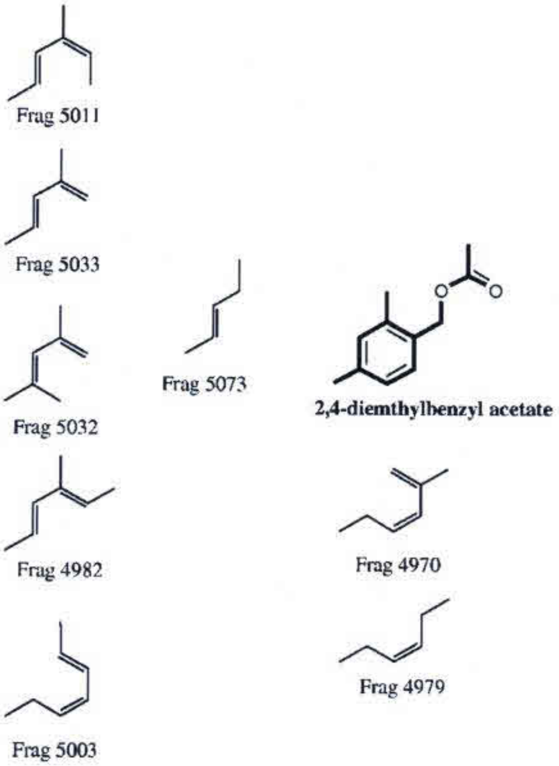

Figure 2. Illustration of the eight significant fragments contributing to the inactive validation prediction of 2,4-dimethylbenzyl acetate.

covers the 3,4-diol and accounts for five other validation fragments. Fragments 2415 and 2416 are closely related to Fragment 2706 but cover just the 3-hydroxyl.

For 2,4-dimethylbenzyl acetate, Fragments 4970 and 4979 cover the *para* substituted methyl section of the molecule. Moreover, Fragment 5073 covers the 2,4-methyl substitution and can account for four similar fragments.

From a prediction point-of-view, any one fragment would have been sufficient for the accurate prediction in these examples. From a mechanism point-of-view, for methyldopa, just the four major fragment families (i.e. from fragments 348, 283, 2706, and 2416) would have covered the major identified structural themes relating to activity. The same is true for 2,4-dimethylbenzyl acetate where two sets of similar fragments (i.e. from fragments 5073–4970) described the compound. In this model, the fragment redundancy is obvious. However, we speculate that this may not be the case with other toxicological endpoints. In models for other endpoints, where fragments are similar but not exact, each fragment may contribute novel mechanistic and predictive information to the model.

Clearly, from the results of the validation exercises, the cat-SAR program in not performing at 100% accuracy. To judge the predictive performance of our models, we compared them to two previously developed MCASE models. One model is based on the National Toxicology Program's *Salmonella* mutagenicity database. The *Salmonella* database is derived from a standardized protocol and, more importantly, has been analyzed for reproducibility and accuracy by replicate analyses of chemicals [20]. As previously indicated, the interlaboratory reproducibility of the *Salmonella* mutagenicity assay is only 85% [20].

4. Conclusions

The new cat-SAR modeling approach described herein has a predictive ability in line with other respiratory sensitization models developed by us [15,19]. This clearly suggests its utility and warrants further development. It is applicable to toxicological or pharmacological SAR modeling. The cat-SAR program uses a binary approach to identify structural features associated with biological activity or inactivity. This is straightforward when the toxicologic endpoint is categorical (e.g. sensitizers vs. nonsensitizers, carcinogens vs. noncarcinogens or mutagens vs. nonmutagens). However, for other endpoints, where a continuous scale of activity is measured, the dichotomy can be imposed between highly active and less active compounds (e.g. extremely toxic vs. nontoxic as in the case of LD_{50} values or high or low receptor affinity as in the case of estrogen receptor ligands).

The cat-SAR method has two main areas of strength when compared with other 2-dimensional modeling systems. The first is the transparency of the method. The derivation of model fragments and decision rules are open for inspection. The entire compound-fragment matrix and the identified model fragments are all easily inspected. The second strength is the amount of user-selectable parameters available for adjustment. For the fragment development part of the program, the user can select fragments of different size and choose other fragment attributes including the consideration of atoms, bond, and hydrogen atoms. Moreover, when identifying important or significant fragments the user can manipulate the selection process by altering the requirements for how many compounds in the learning set contain each fragment and also what proportion of active or inactive compounds in the learning set contain the fragment.

Thus, the cat-SAR method is transparent with regard to the overall modeling process. Users of the program have the opportunity to optimize the process for their own needs. Considering the fact that toxicologic endpoints differ in their mechanisms, it makes sense that the modeling algorithm should be transparent to meet the requirements of the endpoint being modeled.

Overall, in prediction mode, this method presents the user with a *complete* correspondence of fragments in the model and the unknown chemical. In model analysis mode, the method provides the user with a complete listing of all interesting fragments. It should be noted that there is no hierarchy of fragments or filtering of "significant" fragments other than what the user chooses. There are no hidden or proprietary rules in the process. All fragments that meet the user-specified structural requirements and the rules of association with activity or inactivity are included in the model. This leads to the identification of many (e.g. 1000 s) fragments, some with great structural similarity. This clearly presents difficulty in being able to succinctly describe the model. However, important information is retained and accessible to the user.

The cat-SAR program of course has some drawbacks and limitations. Like so many other expert systems in toxicology, it is applicable only to organic chemicals. Metals, mixtures, and polymeric compounds are not suitable for analysis. Moreover, as mentioned, the cat-SAR program presents the final SAR model, in terms of all relevant fragments. This lead to a model that may contain 1000 s of fragments which may lead to difficulty in model interpretation.

Acknowledgements

We gratefully acknowledge support for the development of the cat-SAR program from the Department of Defense Congressionally Directed Medical Research Program for Breast Cancer Idea Award DAMD17-01-0376.

References

[1] E. Benfenati, G. Gini. *Toxicology*, **119**, 213 (1997).
[2] J. Ashby, D. Paton. *Mutat. Res.*, **286**, 3 (1993).
[3] J. Ashby. *Environ. Mutagen.*, **7**, 919 (1985).
[4] J. Ashby, R.W. Tennant. *Mutat. Res.*, **257**, 229 (1991).
[5] C. Hansch, A. Leo. *Exploring QSAR Fundamentals and Applications in Chemistry and Biology*, American Chemical Society, Washington, D.C. (1995).
[6] R.D. Cramer, D.E. Patterson, J.D. Bunce. *J. Am. Chem. Soc.*, **110**, 5959 (1988).
[7] G. Klopman. *J. Am. Chem. Soc.*, **106**, 7315 (1984).
[8] G. Klopman. *Quant. Struct. Act. Relat.*, **11**, 176 (1992).
[9] G. Klopman, H.S. Rosenkranz. *Mutat. Res.*, **305**, 33 (1994).
[10] M.T.D. Cronin, J.S. Jaworska, J.D. Walker, M.H.I. Comber, C.D. Watts, A.P. Worth. *Environ. Health Perspect.*, **111**, 1376 (2003).
[11] M.T.D. Cronin, J.D. Walker, J.S. Jaworska, M.H.I. Comber, C.D. Watts, A.P. Worth. *Environ. Health Perspect.*, **111**, 1376 (2003).
[12] J.C. Dearden. *J. Comput. Aided Mol. Des.*, **17**, 119 (2003).
[13] A.M. Richard. *Toxicol. Lett.*, **102–103**, 611 (1998).
[14] A.M. Richard. *Knowl. Eng. Rev.*, **14**, 307 (1999).
[15] C. Graham, H.S. Rosenkranz, M.H. Karol. *Regul. Toxicol. Pharmacol.*, **26**, 296 (1997).
[16] R. Rodford, G. Patlewicz, J.D. Walker, M.P. Payne. *Environ. Toxicol. Chem.*, **22**, 1855 (2003).
[17] USDHHS. *U.S. Department of Health and Human Services, National Institutes of Health, Publication No. 90–3042* (1991).
[18] C. Graham, R. Gealy, O.T. Macina, M.H. Karol, H.S. Rosenkranz. *Quant. Struct. Act. Relat.*, **15**, 224 (1996).
[19] M.H. Karol, O.T. Macina, A.R. Cunningham. *Ann. Allergy. Asthma. Immunol.*, **87**, 28 (2001).
[20] W.W. Piegorsrch, E. Zeiger. In *Statistical Methods in Toxicology*, L. Hotham (Ed.), pp. 35, Springer-Verlag, Heidlberg (1991).