AD_____

Award Number: DAMD17-01-1-0376

TITLE: Investigating the Mechanism of Action and the Identification of Breast Carcinogens by Computational Analysis of Female Rodent Carcinogens

PRINCIPAL INVESTIGATOR: Albert R. Cunningham, Ph.D.

CONTRACTING ORGANIZATION: Louisiana State University
Baton Rouge LA 70803-3701

REPORT DATE: August 2006

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01-08-2006	Final	15 Jul 01 - 14 Jul 06

4. TITLE AND SUBTITLE
Investigating the Mechanism of Action and the Identification of Breast Carcinogens by Computational Analysis of Female Rodent Carcinogens

5a. CONTRACT NUMBER

5b. GRANT NUMBER
DAMD17-01-1-0376

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Albert R. Cunningham, Ph.D.

E-Mail: arc@lsu.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Louisiana State University
Baton Rouge LA 70803-3701

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This project investigated the potential that environmental chemicals may be involved in the etiology of breast cancer. We hypothesized that specific features of chemicals can be identified that are significantly associated with female and breast carcinogens and that these features are related to mechanisms of chemical carcinogenesis. Our overall scientific objective was to investigate the hypothesized relationship between environmental chemicals, xenoestrogens, and the development of breast cancer. With the success of the project, we published several papers, have one accepted pending revision but temporarily withdrawn specific to our rat mammary carcinogen model, and we will be preparing others for later publication. We also developed a novel SAR approach that allowed us to address the question of "why do some carcinogens cause cancer in the breast?" which is a very different question than that posed in older SAR studies of "why do some chemicals cause cancer?" Two graduate students have been awarded MS degrees based (supported) on this project. We have also used this project to obtain an appointment at the University of Louisville's Brown Cancer Center, significant funding for an associated project in conjunction with its NIH-funded Molecular Targets Program, and another BCRP IDEA award studying a novel approach to discover highly specific breast cancer drugs.

15. SUBJECT TERMS
structure-activity relationship (SAR), computer modeling, mechanisms and etiology of breast cancer, environmental carcinogens

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	120	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

Investigating the Mechanism of Action and the Identification of Breast Carcinogens by Computational Analysis of Female Rodent Carcinogens

DAMD17-01-1-0376

Albert R. Cunningham, Ph.D.

Introduction

The well-established breast cancer risk factors may account for only 47% of the breast cancer incidence in the United States. This leaves a considerable portion of breast cancer from undetermined origin. This project is investigating the potential that environmental chemicals and particularly those with estrogenic activity may be involved in the etiology of breast cancer. We hypothesize that specific features of chemicals can be identified that are significantly

associated with female and breast carcinogens and that these features are related to mechanisms of chemical carcinogenesis. Our overall scientific objective is to investigate the hypothesized relationship between environmental chemicals, xenoestrogens, and the development of breast cancer. The successful completion of this project will provide mechanistic information related to chemical-induced breast cancer as well as structure-activity relationship (SAR) models capable of estimating the likelihood that chemicals with unknown carcinogenic activity may be breast carcinogens.

Body

Notes on Final Report

As mentioned in previous reports, we anticipated during our move from the University of Pittsburgh to Louisiana State University that we would require a one year no cost extension and that the final report would be completed in 2006. July of 2005 was the start of the final year of the project. At about that time, I started looking for employment/research opportunities in order to continue and extend the work of this project. I discussed this with my Grants Manager Dr. Moore and she kindly allowed me to extend the due date for this final report.

I was able to successfully use the results described herein to obtain an appointment offer as an Associate Professor of Medicine from the University of Louisville and I will be a member of its James Graham Brown Cancer Center. Regarding the Brown Cancer Center, in September 2003 it was awarded a five-year, $11 million Center of Biomedical Research Excellence (COBRE) grant from the National Center for Research Resources at the National Institutes of Health under the directorship of Dr. Donald Miller. The grant established the Molecular Targets Program, which I will be part of by providing for the recruitment of researchers from a variety of disciplines to identify and develop new molecular targets for anti-cancer drugs and therapies using the techniques of modern structural biology.

My appointment at the University of Louisville is tentatively set to start April 1, 2007 and will include a three year COBRE-funded project in Molecular Targets. My recruitment to the University of Louisville's Brown Cancer Center and my proposed COBRE project in Molecular Targets will allow me to continue and extend the line of study initially supported through the Congressionally Directed Medical Research Program's Breast Cancer Research Program.

Since our focus was on establishing a suitable research environment elsewhere, we did not finalize three manuscripts we recently submitted for publication. They are:

1. Cunningham AR, Moss ST, and Cunningham SL. A predictive and mechanistically insightful structure-activity relationship analysis of rat mammary carcinogens. Chemical Research in Toxicology (accepted pending revision—temporarily withdrawn).
2. Rosenkranz HS, Cunningham SL, Cunningham AR. Practical aspects of SAR modeling of rodent carcinogenicity: Congeneric *vs.* non-congeneric learning sets. Mutation Research (accepted pending revisions).
3. Rosenkranz HR, Cunningham SL, Cunningham AR. Practical aspects of SAR modeling of rodent carcinogenicity: Congeneric vs non-congeneric datasets. Mutation Research (accepted pending revisions—revisions returned).

We anticipate the publication of these three papers and several others after my appointment at the University of Louisville begins this coming April.

Specific Aim Accomplishments

Specific aim 1: Development and validation of SAR models for female breast carcinogens (months 1-12).

a. Identify chemicals tested in female rodents from the Carcinogenic Potency Database and the National Toxicology Program (month 1).
b. Enter chemical structures and potency values into MCASE program (months 2-8).
c. Validate models using 10-fold cross validation (months 9-12)
d. Summarize and interpret models and prepare publication.

As previously discussed, these models have been developed and validated (i.e., a-c) as planned in MCASE and then with the cat-SAR program (described below). We have also updated rodent carcinogenicity models so that all models (mouse and rat, as well and female specific version) have been built on the same datasets and analyzed with the cat-SAR program. We will revise and resubmit for publication the manuscript for rat mammary carcinogens and finish preparing and submit another, specifically with mouse mammary carcinogens shortly.

Female and Mammary Carcinogen Models
As mentioned, we have had success in developing the female and mammary carcinogen models and have devoted a significant effort in "assuring" their appropriateness. Basically, we have now developed several different female and mammary carcinogen models.

1. Rat Mammary Carcinogen Models: In previous reports, we speculated that our original model protocol might not have been optimal. The common approach of most SAR studies entails a comparison of structural features between biologically "active" and "inactive" compounds. Thus, when considering carcinogenesis, the categories are clearly carcinogens and noncarcinogens. However, when considering organ-specific carcinogenesis as in the case of breast cancer, we asked the question "what are the appropriate inactive compounds?" Should they be noncarcinogens or carcinogens that are just not carcinogenic to the organ under study? As we proceeded we considered both options. We note again this important aspect of the project was not considered in the original proposal. Moreover, as we investigated this approach, it became evident that this modeling paradigm is to the best of our knowledge, novel.

We developed two separate models for rat mammary carcinogens: The mammary carcinogen - noncarcinogen model and the mammary carcinogen - non-mammary carcinogen model. The details were described in the previous Annual Report. We are happy to report that this new approach to modeling organ-specific carcinogenicity was successful, although we had to withdraw the manuscript after it had been accepted in *Chemical Research in Toxicology* pending revisions. The manuscript entitled "A predictive and mechanistically insightful structure-activity relationship analysis of rat mammary carcinogens" is included in the appendices.

One set of models for this manuscript was developed based on a comparison of rat mammary carcinogens to noncarcinogens (MC-NC) and the second and novel method compared mammary

carcinogens to non-mammary carcinogens (MC-NMC). The best rat MC-NC model achieved an 82% observed correct prediction (OCP) rate with a sensitivity of 77% and a specificity of 88%. The best rat MC-NMC model achieved a 79% OCP rate with a sensitivity of 83% and a specificity of 74%. As mentioned, the MC-NMC model was based on a learning set that contained carcinogens in both the active (i.e., mammary carcinogens) and inactive (i.e., carcinogens to sites other than the mammary gland) categories and was able to distinguish between different types of carcinogens (i.e., tissue specific), not simply between carcinogens and noncarcinogens. Based on a structural comparison between this model and one for *Salmonella* mutagens, there was no observed relationship between the two phenomena since both the active and inactive categories contained a high proportion of mutagens. Overall, these findings suggest that the MC-NMC model is identifying structural attributes to address the specific question of "why do some carcinogens cause cancer in the breast?" which is a significantly different question than "why do some chemicals cause cancer?'.

2. Mouse Mammary Carcinogen Models:
A similar group of analyses as listed above for rat carcinogens is being completed for mouse carcinogens. The validation results are promising. These models are based on 24 mouse mammary carcinogens from the published CPDB target site summary.

Specific aim 2: Identify chemical and biological attributes of female and/or breast carcinogens to provide evidence to test the hypothesis that xenoestrogens are involved in breast cancer (months 13-36).

a. Compare and identify Structural Feature Overlap Method of female and breast carcinogens to those of other available toxicological SAR models (see Facilities and Equipment for a complete list of available models) (months 13-16).
b. As above using Joint Prevalence Method (months 16-24).
c. Identify the exact features of female and breast carcinogen models that are responsible for predicted similar activities identified above (months 25-26).
d. Conduct QSAR and CoMFA analyses with chemicals containing these structures using biological data from appropriate assays (months 28-36).
e. Conduct metabolism experiments on identified outliers to see whether metabolic activation is required for activity and update models if required (months 28-36).
f. Summarize and interpret data and prepare publications (months 28-36).

We have concluded the migration of a set of approximately 20 MCASE toxicological SAR models to cat-SAR and the models have been validated. This was required for Specific Aims 2a and 2b. We are currently using these models to prepare the Joint Prevalence Method datasets for the 20 endpoints.

As mentioned in previous reports, we began to develop three estrogen cat-SAR models that will be directly applicable to testing the relationship between estrogenicity and mammary carcinogenicity. The manuscript describing the cat-SAR analysis of 122 chemicals tested for estrogenicity in the E-Screen assay is nearly complete.

We also noted previously that in particular SA 3 requires the MCASE module META. Moreover, some of the other sub-SA could be most easily accomplished with MCASE.

However, as noted in the previous Annual Report, we did not have current access to a working copy of MCASE, though Professor Rosenkranz was working on the issue. Moreover, Professor Rosenkranz passed away in November of 2004. Due to significant legal issues between himself and Professor Klopman (his co-developer of MCASE), I did not request to have access to MCASE or META.

Specific Aim 3: As previously reported, in retrospect, SA3, while generally important, is not a practical aim for this project. The reason is that since metabolism is not considered during model development, metabolic products cannot be used to identify outliers. In other words, since the models are built on the parent molecules (i.e., what the animals are dosed with) they do not explicitly consider metabolic products. Thus even if we were to analyze known metabolites of a carcinogen, the models would likely identify the parent structures as relating to carcinogenesis, not the additional metabolism-related moieties.

Looking Forward

With our appointment to the Brown Cancer Center we are in the initial stages of preparing a COBRE Molecular Targets Project with a goal to extend the computational work of Specific Aims of this project to the discovery of novel *in vivo/in vitro* molecular targets associated with breast cancer.

The project will be based on our finding that carcinogenic chemicals possess identifiable structural attributes associated with breast carcinogenicity and not carcinogenicity to other tissues. In other words, since the results of this project supported or original hypothesis that a) specific features of chemicals can be identified that are significantly associated with female and breast carcinogens and b) these features are related to mechanisms of chemical carcinogenesis, we are now formulating a new hypothesis (i.e., and project) to study chemical-induced breast carcinogenesis. Tentatively, the new hypothesis for our COBRE Molecular Targets Project will be that this specific and structurally-defined carcinogenicity (i.e., SAR results from our IDEA award) is due to an interaction of the carcinogen (i.e., specific structural features of the carcinogen) with a specific molecular target found only in the sensitive tissue (i.e., breast cells).

We propose to employ a unique combination of further SAR modeling to focus mass spectrometric-based proteomics and/or genomics analyses where we foresee the identification of these molecular targets that are specifically and exquisitely associated with chemical-induced carcinogenicity in breast cells. The potential outcome of this work will be 1) a better insight about the interactions of carcinogens and molecular targets and why certain chemicals induce breast cancer (i.e., molecular targets associated with toxicity/carcinogenicity) and 2) new drugs based on this understanding may therefore only target, and be effective against, breast cancer cells possessing the particular molecular target. Moreover, these drugs will be anticipated to be innocuous (i.e., or have minimal side effects) to other cells since they do not have the proper target of toxicity. The potential for success of this project lies in a synergism obtained through the use of biological and computational techniques that refines information on chemical carcinogenicity to discover information about cell type-specific molecular targets of activity.

Key Research Accomplishments

1. Developed structurally-based computational models that identify structural features of carcinogens associated with their potential to specifically induce breast cancer in rodents.

These accomplishments, their significance, and the forthcoming manuscripts have been described above.

2. Developed new SAR modeling algorithm called cat-SAR.

Originally, the structure-activity relationship (SAR) modeling was proposed to be conducted with the MCASE program. However, for multiple reasons we decided to switch platforms to Tripos Sybyl. This change did not alter the project and the SOW was updated.

During the early part of the project it was becoming evident that MCASE was not developing models for this project (details discussed below) that were of stellar predictivity. On account of successful SIMCA modeling of aromatic amine Salmonella mutagens and skin sensitizing agents for a project supported by Proctor & Gamble, we spent some time investigating whether Sybyl could be employed to produce adequate models relating to this project.

Briefly, although Sybyl and MultiCASE are different modeling packages, the Sybyl family of SAR modules allows for a similar type of analysis of toxicant. As described in the proposal, MCASE takes a binary approach to analyzing toxicants by comparing structural features (2-dimensional biophores) found in active and inactive compounds. Similarly, for the Sybyl analyses the HQSAR (hologram QSAR) routine calculates 2-dimensional holograms(i.e., linear fragments comparable to biophores) and the Advanced QSAR module uses the soft independent modeling of class analogy (SIMCA) algorithm to perform statistical analysis of the holograms. SIMCA is a regression-type analysis that develops predictive models based on categorical data (i.e., carcinogens and noncarcinogens). Overall, the HQSAR-SIMCA models appear to be superior to MCASE models.

Our initial HQSAR-SIMCA models included an analysis of potential carcinogens identified using of ~1600 diverse Salmonella mutagens and 122 compounds tested for estrogenicity using the E-SCREEN assay for which MultiCASE models existed. The mutagenicity MultiCASE model and its analysis have been published (1, 2) and the E-SCREEN environmental estrogen model has been accepted for publication with minor revisions (3). In both instances, Sybyl has been able to develop models comparable to the predictivity of MultiCASE. We anticipate a manuscript from this work describing how HQSAR-SIMCA can be successfully employed for computational analysis and prediction of environmental mutagens and potential carcinogens.

At this juncture, all projects in my laboratory are being transferred to Sybyl. Needless to say, Tripos Sybyl software is in my estimation superior to MCASE. Notably, the Tripos family of software is growing and is on the cutting edge of technology. The literature is replete with Sybyl-based investigations. On the other hand, MCASE is owned by an individual and has the real potential to become a legacy or obsolete system in the coming years.

We originally thought that SIMCA combined with HQSAR (hologram quantitative SAR) models appeared to be superior to MCASE models. The HQSAR-SIMCA approach utilized categorical biological data (i.e., carcinogen vs. non-carcinogen) and molecular fragments as SAR descriptors. Therefore, this seemed a reasonable substitute SAR approach for MCASE.

However, upon consultation with the developers of Sybyl HQSAR-SIMCA, we learned that there was a large degree of random assignment of SAR descriptors. Basically, as it turned out, although the modeling software was able to produce models that could predict the activity of unknown chemicals, they were not very mechanistically insightful. In other words, we would not be able to interpret these models in order to understand the structural attributes of breast carcinogens. Moreover, without an SAR model having a solid and understandable mechanistic foundation, we were troubled that even the "predictive" models may have more to do with chance occurrences than true accurate predictions.

Concerned about the completion of the project, we developed a new SAR system that we named cat-SAR (categorical-SAR). I discussed the development of this new program with Dr. Moore and we were in agreement that this would be an appropriate path to take in order to achieve the overall goals of the project. The program has been developed with guidance from Prof. Herb Rosenkranz. Dr. Rosenkranz was co-PI of this project and a co-developer of MCASE.

A description of the program was published in *SAR and QSAR in Environmental Research*. This manuscript described the program in detail. We note the publication was on respiratory sensitizers, not breast carcinogens. The reason for this was 1) it was a small and manageable dataset and 2) a previous MCASE analysis of this data yielded a very good model. As such, this was a suitable dataset on which to develop and test the cat-SAR program. A copy of the manuscript detailing the cat-SAR program is included in the appendices.

3. Developed predictive/mechanistically insightful SAR models for rat and mouse carcinogens and mammary carcinogens

We developed shareable databases/learning sets of carcinogens, their molecular structure and associated activity values. We discussed the exact method of "sharing" this information with Dr. Ann Richard of the Environmental Carcinogenesis Division of the U.S. EPA. Dr. Richard has developed and maintains the EPA's DSSTox Public Toxicity Database Network. As soon as my lab is operational at the University of Louisville, I will contribute these databases/learning sets to Dr. Richard's program.

Reportable Outcomes

Graduate Students

Shanna Moss and Daniel Consoer both graduated with M.S. degrees from the Department of Environmental Studies at LSU with thesis research based on this project. Both Ms. Moss and Mr. Consoer were supported for their studies through this project.

1. Ms. Moss's thesis is titled *Identification of 'structural alerts' and associated mechanisms of action of mammary gland carcinogens in female rodents.*

2. Mr. Consoer's thesis is titled *Evaluation of a Novel Method of Predicting Estrogen Activity of a Group of Structurally Diverse Compounds*.

Manuscripts Published

1. Cunningham AR, Cunningham SL, Rosenkranz HR. Structure activity approach to the identification of environmental estrogens: The MCASE approach. SAR and QSAR in Environmental Research 15:55-67(2004).
2. Cunningham AR, Cunningham SL, Consoer DM, Moss ST, Karol MH. Development of an information-intensive structure- activity relationship model and its application to human respiratory chemical sensitizers. SAR and QSAR in Environmental Research 16:273-285(2005).

Manuscripts to be published

1. Rosenkranz HS, Cunningham SL, Mermelstein R, Cunningham AR. The challenge of testing chemicals for potential carcinogenicity using multiple short term assays. An analysis of a proposed test battery for hair dyes. Mutation Research (accepted pending revisions).
2. Rosenkranz HR, Cunningham SL, Cunningham AR. Practical aspects of SAR modeling of rodent carcinogenicity: Congeneric vs non-congeneric datasets. Mutation Research (accepted pending revisions).
3. Cunningham AR, Moss ST, and Cunningham SL. A predictive and mechanistically insightful structure-activity relationship analysis of rat mammary carcinogens. Chemical Research in Toxicology (accepted pending revisions but withdrawn).

Funded Proposal Applied for Based on Work Supported by this Award

We note that the below listed proposals all relate to the discovery of novel antibreast cancer therapeutics. Given that the estrogen receptor is involved in the etiology, cure, and prevention of breast cancer, this IDEA Award has allowed us to pursue new avenues of research into drug discovery.

1. We are happy to report that we received a BCRP IDEA award titled *A novel approach for the identification of pharmacophores through differential toxicity analysis of estrogen receptor positive and negative cell lines* (PI, $372,542). Our success of achieving this award lies clearly in the work of the project we are now completing. Specifically, the current award has provided us with a good set of models on which to understand breast carcinogens and has allowed us to develop the basis of a new SAR modeling approach wherein it is possible to differentiate "different" actions of toxicant or carcinogens. As the current IDEA

award is now focusing on why particular *carcinogens* are carcinogenic to breast tissue, this new award will focus on understanding and identifying why toxic agents are only specific to certain types of breast tumor cells. The goal is to identify novel and highly selective new chemotherapeutic agents and pharmacophores for their development.

2. Also, as mentioned, I have been offered a position as an Associate Professor of Medicine at the University of Louisville's James Graham Brown Cancer Center. This appointment includes the extension of this BCRP project into the area of molecular targets and my proposed COBRE Molecular Targets project will be funded with $ for equipment, $75,000 for yearly equipment and supplies, along with three years of funding for a post-doctoral fellow.

3. Carcinogenic assessment of antimicrobial isothiazolines, The Arch Chemical Company (ARC PI, ongoing).

Submitted but not Funded Proposals

1. Louisiana Environmental Hazard Survey and Breast Cancer Analysis, The Coypu Foundation (ARC PI, $).
2. Estimating exposure to environmental carcinogens and breast cancer, The Susan G. Komen Breast Cancer Foundation (ARC PI, $.

   These first two proposals are based directly on the work of this project wherein we are proposing to apply the predictive capabilities of the models to Geographic Information System (GIS) overlays of breast cancer mortality and morbidity and environmental chemical transport data from Toxic Release Inventory release sites. Only the Komen proposal was returned with review comments and we are pleased with the fact that on this first attempt for funding of this new initiative, the comments were positive and useful. We are hopeful that with further refinement of the proposal, it will become fundable in the near future.

3. Pharmacophore discovery by differential toxicity studies, The Susan G. Komen Breast Cancer Foundation, (Billy Day PI, $).

4. Pharmacophore discovery by differential toxicity studies, National Institutes of Health (PI, $).

   These proposals were versions of the just discussed IDEA award.

Patent/copyright

We have submitted a patent and copyright application to LSU's Office of Intellectual Property for the cat-SAR computational toxicology expert system. As noted, this system was developed to replace the MCASE system described in the original proposal and SOW.

Presentations

1. Louisiana State University Environmental Lecture Series, 2003, Structure-Activity Relationships: Estrogenic Mimics and Endocrine Disruptors.
2. Tulane University, 2003, Structure-Activity Relationships: Estrogenic Mimics and Endocrine Disruptors.
3. Advisory Council to the Louisiana State University School of the Coast and Environment, 2004, Predictive Toxicology: Estimating Human and Environmental Health Consequences of (Hazardous) Pollutants.
4. The University of Arkansas, 2006, Predictive Toxicology: Estimating the Biological Activity of Chemicals.
5. The University of Texas at El Paso, 2006, Predictive Toxicology: Estimating the Biological Activity of Chemicals.
6. University of South Carolina, 2006, Predictive Toxicology: Estimating the Biological Activity of Chemicals.
7. Tulane University, 2006 Predictive Toxicology: Estimating the Biological Activity of Chemicals.
8. University of Louisville, 2006, Structure-Activity Relationship Modeling: Carcinogens, Chemotherapeutics, and Molecular Targets.
9. East Carolina University, 2006 Structure-Activity Relationship Modeling: Carcinogens, Chemotherapeutics, and Molecular Targets.

Conclusions

With the success of the rat mammary carcinogen models we are preparing a similar manuscript describing mouse mammary carcinogens. We are also completing work on a general chemical carcinogen manuscript and another manuscript describing female-specific carcinogens. Also of importance, we are working on several xenoestrogen and toxicological models that, although not detailed in the project proposal, will be of great importance for studying the receptor- and mutagenesis-based mechanisms of breast carcinogenesis.

To date, we have developed most of the proposed models set forth in the proposal and anticipate developing several more as mentioned in specific aim 2. Due to the time required to develop the cat-SAR program and employment issues related to Louisiana State University we have not yet been able finish some of the work nor publish several papers. However, with my recent appointment offer at the University of Louisville and participation in its COBRE Molecular Targets program, and other projects in my laboratory, we look forward to continuing and extending this project.

Finally, in our discussion with friends knowledgeable in SAR and carcinogenesis and many including my mentor, the late Professor Rosenkranz, we are very excited about the continuation of this project and the applicability of its results in the search for new molecular targets associated with the etiology of breast cancer and which also might serve as new chemotherapeutic targets for the treatment of the disease. Specifically by being able to address the issue of “why certain chemical carcinogens target the breast?” rather than previous SAR-based questions of “why do chemicals cause cancer?” we anticipate taking SAR-based analyses of breast carcinogens to a new level of detail and understanding as well as publish several other manuscripts from the work of this IDEA award.

Appendix

Manuscripts and publications

1. Cunningham AR, Moss ST, and Cunningham SL. A predictive and mechanistically insightful structure-activity relationship analysis of rat mammary carcinogens. Chemical Research in Toxicology (accepted pending revision—temporarily withdrawn).
    a. Specific Aim 1
2. Rosenkranz HS, Cunningham SL, Cunningham AR. Practical aspects of SAR modeling of rodent carcinogenicity: Congeneric *vs.* non-congeneric learning sets. Mutation Research (accepted pending revisions).
    a. Specific Aim 2
3. Cunningham AR, Cunningham SL, Consoer DM, Moss ST, Karol MH. Development of an information-intensive structure- activity relationship model and its application to human respiratory chemical sensitizers. SAR and QSAR in Environmental Research 16:273-285(2005).
    a. Specific Aim 1
4. Cunningham AR, Cunningham SL, Rosenkranz HR. Structure activity approach to the identification of environmental estrogens: The MCASE approach. SAR and QSAR in Environmental Research 15:55-67(2004).
    a. Specific Aim 2

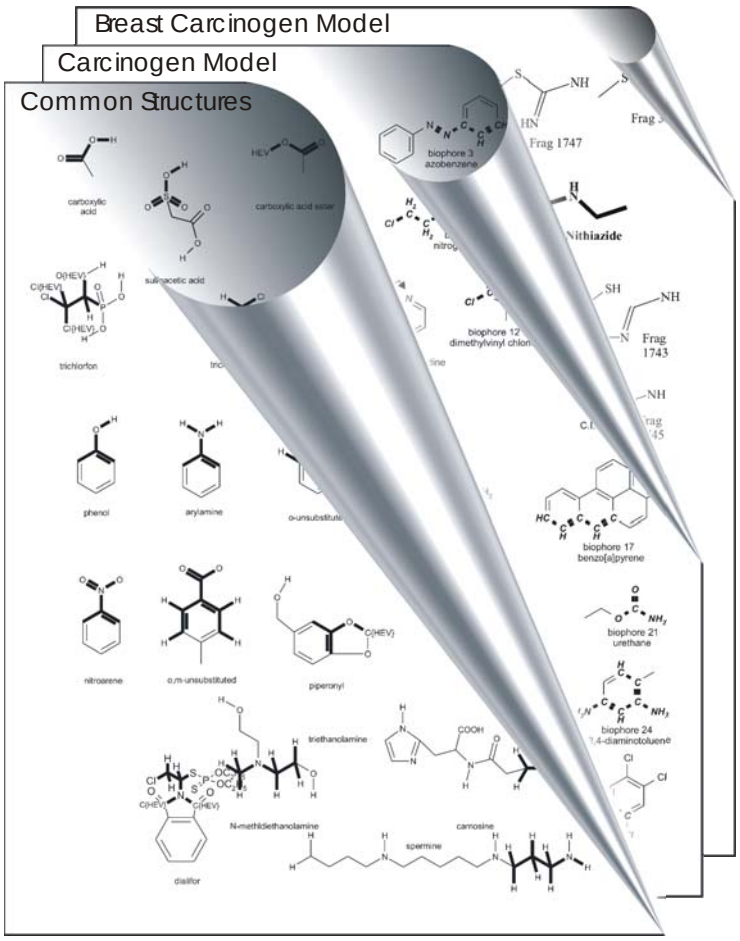
A predictive and mechanistically insightful structure-activity relationship analysis of rat mammary carcinogens

Albert R. Cunningham*, Shanna T. Moss, and Suzanne L. Cunningham

Department of Environmental Studies
Louisiana State University
Baton Rouge, LA 70803

*Corresponding author
1285 Energy, Coast & Environment Building
Baton Rouge, LA 70803
225-578-9422
arc@lsu.edu

Running title: **structure-activity analysis of rat mammary carcinogens**



TOC Graphic Illustration of how SAR models can remove layers of chemical information in order to study specific aspects of the process. Consider chemical carcinogenesis: A SAR model developed from many chemicals, that have been categorized as carcinogens and noncarcinogens removes common chemical structures (top layer) to reveal features associated with carcinogens (middle layer). The SAR model described herein was subsequently developed from carcinogens that have been categorized as breast carcinogens and non-breast carcinogens. This later model removes carcinogen-related structures (middle layer) to reveal a set of features associated with breast-specific carcinogens (bottom layer).

Abstract

Structure-activity relationship (SAR) models are powerful tools to investigate the mechanisms of action of chemical carcinogens and to predict the potential carcinogenicity of untested compounds. We describe herein the application of the recently developed cat-SAR algorithm to two learning sets of rat mammary carcinogens. One set of models developed was based on a comparison of rat mammary carcinogens to noncarcinogens (MC-NC) and the second compared mammary carcinogens to non-mammary carcinogens (MC-NMC). The best rat MC-NC model achieved an 82% observed correct prediction (OCP) rate with a sensitivity of 77% and a specificity of 88%. The best rat MC-NMC model achieved a 79% OCP rate with a sensitivity of 83% and a specificity of 74%. The MC-NMC model was based on a learning set that contained carcinogens in both the active (i.e., mammary carcinogens) and inactive (i.e., carcinogens to sites other than the mammary gland) categories and was able to distinguish between different types of carcinogens (i.e., tissue specific), not simply between carcinogens and noncarcinogens. Based on a structural comparison between this model and one for *Salmonella* mutagens, there was no observed relationship between the two phenomena since both the active and inactive categories contained a high proportion of *Salmonella* mutagens. Overall, these findings suggest that the MC-NMC model is identifying structural attributes to address the specific question of "why do some carcinogens cause cancer in the breast?" which is a significantly different question than "why do some chemicals cause cancer?'.

Introduction

The identification of human carcinogens is a difficult and complex task. Only a limited number of high-quality epidemiological studies have been conducted that identify particular agents that induce cancer in humans. In lieu of such data, rodent cancer bioassays or short-term tests for genotoxicity have been used to estimate the likelihood that particular chemicals will be human carcinogens.

However, it is evident that not all chemicals in use today will be tested for carcinogenesis. There are approximately 75,000 industrial chemicals on the Toxic Substance Control Act's Chemical Substance Inventory *(1)* and the National Institute of Environmental Health Sciences estimates that there are over 80,000 chemicals registered for use in the United States *(2)*. A complete two-year bioassay as conducted by the National Toxicology Program (NTP) including planning, evaluation, and review takes about five years to complete, costs between $2-4 million, and uses 400 animals *(3)*. To test all chemicals in this manner is thus prohibitive.

In fact, the NTP has only tested over 500 chemicals for rodent carcinogenicity in standardized 2-year rodent bioassays. Furthermore, the Carcinogenic Potency Database (CPDB) analyzes and consolidates into a single resource the world's diverse literature and NTP Technical Reports of chronic long-term animal cancer bioassays *(4)*. To date, analyses of 6073 experiments on only 1458 chemicals are available on the CPDB's web site *(5)*. Fortunately, the consolidation, standardization, and analyses of cancer bioassay data by the CPDB provides a comprehensive resource for investigating chemical carcinogenesis including analyses by structure-activity relationship (SAR) modeling and predictive toxicological methods.

SAR modeling and other predictive toxicological methods provide a means to estimate toxicological properties of chemicals based on information from previously tested compounds. We have reported predictive and mechanistically insightful SAR models for mice *(6)* and rats *(7)* using the CASE/MULTICASE SAR expert system and chemical carcinogenicity data from the first five plots of the CPDB *(8-12)*. Depending upon validation methods, these models had an observed correct prediction (OCP) rate for chemicals removed from the model's learning set of between 64% and 78% *(6,7)*. Many others have also demonstrated varying degrees of success modeling chemical carcinogens. The utility and application of some important toxicologically-focused predictive methods have been reviewed in-depth by Richard *(13,14)*.

The SAR models of rat and mouse carcinogens developed by us, while being predictive, also provided insight into the structural underpinnings for species-specific carcinogenesis. Many, though not all, of the readily explainable attributes of these models corresponded with the genotoxic or electrophilic paradigm of carcinogenesis *(15)*. In retrospect, this is not surprising given the large numbers of electrophilic or proelectrophilic carcinogens used to build the models and the *a priori* acceptance of the electrophilic theory. These findings did however provide solid evidence that many of the features developed for the models were justifiable and mechanistically sound.

Of note, even in light of this bias toward electrophilicity, we were able to glean an interesting relationship between estrogenicity and carcinogenicity. We identified a 2-dimensional feature of rodent carcinogens that dichotomizes the so-called "beneficial" (e.g., phytoestrogens) from

"harmful" (e.g., pesticides and industrial chemicals) xenoestrogens *(16)*. Further investigation of this feature showed that differences in regional lipophilicity were evident between phytoestrogens and other man-made xenoestrogens. We speculated at the time that these differences in chemical features of estrogen could induce different biological responses *(16-18)*. During this same time, the estrogen receptor alpha (ERα) ligand binding domain was crystallized and its atomic coordinates resolved with those of bound estradiol and raloxifene *(19)*, genistein *(20)*, and 4-hydroxytamoxifen and diethylstilbestrol *(21)*. It was noted that the lipophilic cavity is nearly twice the size of estradiol, which may explain in part the ER's promiscuity *(22)*. Most importantly, it was observed by these authors that estrogen antagonists induce a different conformational change in the AF-2 region compared to that for the natural ligand. These analyses demonstrated the utility of SAR analysis to not only generate predictive models that are explainable by current knowledge but also their ability to provide hypothetical (and testable) information regarding the mechanistic action of toxicants.

The study described herein uses a new SAR algorithm to analyze chemicals that specifically induce mammary cancer in rats. Environmental risk factors, including chemical exposure, may play a role in the development of breast cancer. Unfortunately, many of these factors remain largely unknown. We note that one group of chemicals that has received considerable attention with regards to breast cancer is the environmental endocrine disruptors, with specific attention paid to the xenoestrogens *(23,24)*. For instance, many industrial chemicals (e.g., PCBs and pesticides), consumer products (e.g., plasticizers and phenols), and plant products (e.g., phytoestrogens such as genistein and coumestrol) have been shown to possess estrogenic activity in a number of *in vitro* and *in vivo* assays. As such, xenoestrogens warrant vigorous attention,

especially in light of conflicting epidemiological data and expert opinions regarding their role in the development of this disease *(25-28)*. Overall, in a thorough review of the literature, it was concluded that the available data do not support *or* reject the relationship between exposure to organochlorine compounds and breast cancer *(25)*. As for the role of xenoestrogens in general, the National Research Council states that in fact most studies have been limited primarily to DDT, DDE, TCDD, and PCBs with other compounds receiving little or no attention *(29)*.

Although xenoestrogens have received much public attention, there is a sizable majority of rodent mammary carcinogens that are not estrogenic. Thus, although environmental estrogens may play an important role in the development of breast cancer, other nonestrogenic chemicals may also contribute to the disease. As such, any screening approach designed to identify environmental estrogens, although useful, will not allow for the identification of all potential mammary carcinogens.

The Food and Drug Administration's (FDA) National Center for Toxicological Research recently noted that FDA reviewers are interested in organ-specific carcinogenicity to aid in evaluating new chemicals *(30)*. As such, they have undertaken the task of building an organ-specific database of chemical carcinogens from CPDB data. In their preliminary SAR analyses of liver carcinogens, they obtained a correct prediction rate of 63%, with a sensitivity of 30% and a specificity of 77% *(30)*. Their efforts in developing this comprehensive dataset of organ-specific toxicological data will provide a needed resource for the development and validation of organ-specific SAR models.

Aside from the practical needs of entities like the FDA for estimating organ-specific toxicity, the development of organ-specific carcinogenicity models is also technically appealing. SAR models developed from whole-animal carcinogenicity data attempt to deal with many underlying and often competing mechanisms. As such, it is quite possible that information is lost in the modeling process. Therefore by focusing on specific organs, we hypothesize the development of a clearer picture of the chemical requirements for carcinogenesis in that organ.

For the analyses described herein, we used a newly developed SAR expert system to analyze the set of rat mammary carcinogens reported in the CPDB *(31)*. The system is called cat-SAR for categorical-SAR. Basically, the cat-SAR approach is a computational SAR or *in silico* toxicity prediction "expert system" as classified by Dearden *(32)*. In a previous analysis of human respiratory sensitizers, the cat-SAR program was able to achieve an overall correct prediction rate of 92% with sensitivities between 89 and 94% and specificities between 87 and 95% *(33)*.

The approach we have taken in developing the cat-SAR program clearly diverges from existing commercial 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 control and selection of modeling parameters facilitates the ability to rigorously explore the relationships between chemical structure and biological activity. Ultimately, this rationale negates any *a priori* requirements that a given set of data must fit the attributes of a predefined and often proprietary modeling process.

The cat-SAR models are built through a comparison of structural features found amongst categorized compounds in the model's learning set. Generically, these categories are biologically active and inactive compounds. When just considering whole animal carcinogenesis, the categories are simply carcinogens and noncarcinogens. However, when considering organ-specific carcinogenesis, the question arises as to the selection of the inactive or noncarcinogenic compounds. Should they be whole animal noncarcinogens or carcinogens that are just not carcinogenic to the organ under consideration? For this exercise, we considered both options and developed predictive SAR models comparing rat mammary carcinogens to noncarcinogens (MC-NC model) and rat mammary carcinogens to non-mammary carcinogens (MC-NMC model).

Materials and Methods

Mammary Gland Carcinogen Learning Sets

The CPDB standardizes the experimental results (whether positive or negative for carcinogenicity), including qualitative data on strain, sex, route of compound administration, target organ, histopathology, and the author's opinion and reference to the published paper, as well as quantitative data on carcinogenic potency, statistical significance, tumor incidence, dose-response curve shape, length of experiment, duration of dosing, and dose rate *(8)*. Moreover, a potency value for carcinogens, the TD_{50} is available. The TD_{50} is "that dose rate (in mg/kg body weight/day) which, if administered chronically for the standard lifespan of the species, will halve the probability of remaining tumorless throughout that period" *(8)*.

The rat mammary carcinogen learning sets were developed from the published CPDB carcinogen target site summary *(31)*. This reference listed 102 rat mammary carcinogens. We excluded norlestrin and dimethylaminoethylnitrosoethyl urea nitrite salt. Norlestrin is a mixture, while the second compound is an organic complex. Therefore a total of 100 rat mammary carcinogens were included in the learning sets.

As discussed below, the cat-SAR program derives SAR models through the comparison of structural features associated with categorical responses (e.g., active *vs.* inactive compounds or carcinogens *vs.* noncarcinogens). When just considering whole animal carcinogenesis, the categories are simply carcinogens and noncarcinogens. However, when considering organ-specific carcinogenesis, selection of the inactive or noncarcinogenic compounds could be whole animal noncarcinogens or carcinogens that are not carcinogenic to the organ under consideration. We considered both options for this analysis. Hence, we developed two separate sets of models for rat mammary carcinogens: The mammary carcinogen - noncarcinogen (MC-NC) model and the mammary carcinogen - non-mammary carcinogen model (MC-NMC).

Since we had sufficient numbers of noncarcinogens and non-mammary carcinogens to include as "inactives", we made triplicate inactive datasets (designated Sets 1, 2 and 3 in Table 1) of 100 chemicals each. By so doing we were able to assess the stability of the derived models. Statistical comparison of the each of the models' fragment sets and predictivity was conducted to determine whether the three sets were statistically different. Moreover, this approach prevented the chance of selecting 100 inactive compounds that produced a "good" model. For the MC-NC model three random sets of 100 noncarcinogens were randomly selected from the 449 rat

noncarcinogens listed in the CPDB. Likewise, for the MC-NMC model three random sets of 100 carcinogens were selected from the 395 rat carcinogens in the CPDB that did not induce mammary cancer.

The Categorical-SAR or cat-SAR Expert System Methodology

The Learning Set

The cat-SAR models are built through a comparison of structural features found amongst two designated categories of compounds in the model's learning set. As mentioned, for these analyses the categories for the first model were mammary carcinogens *vs.* noncarcinogens (MC-NC) and mammary carcinogens *vs.* non-mammary carcinogens (MC-NMC) for the second model. The cat-SAR learning set consists of the chemical name, its structure as a .MOL file, and its categorical designation (e.g., one or zero). Organic salts are included as the freebase. Simple mixtures and technical grade preparations are included as the major or active component. Metals, metaloorganic compounds, polymers, and mixtures of unknown composition are not included.

Since the cat-SAR program requires a number of user-specified options, there is not an *a priori* determination of the final model. In other words, the user is allowed to explore and optimize the modeling program. As such, we have developed and reported herein several different cat-SAR MC-NC and MC-NMC models.

In Silico Chemical Fragmentation and the Compound-Fragment Data Matrix

Using the Tripos Sybyl HQSAR module, each chemical was fragmented *in silico* into all possible fragments meeting user-specified criteria. HQSAR allows the user to select attributes for fragment determination including atom count (i.e., size of the fragment), bond types, atomic connections (i.e., the arrangement of atoms in the fragment), inclusion of hydrogen atoms, chirality and hydrogen bond donor and acceptor groups. Fragments can be linear, branched or cyclic moieties.

The first sets of models developed contained fragments between three and seven atoms in size and considered Atoms, Bond types, and atomic Connections. These are referred to as ABC fragment set models. The second set included the same descriptors as the previous set plus associated Hydrogen atoms. These are referred to as ABCH fragment set models.

Upon completion of the fragmentation routine, a Sybyl HQSAR add-on procedure produces the compound-fragment data matrix as a text file. In the matrix, the rows are intact chemicals and the columns are 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 it are tabulated down each column.

The HQSAR module is not used for statistical analysis or model development. The compound-fragment matrix is then analyzed, using the cat-SAR programs we have developed in order to identify structural features associated with active and inactive compounds. The programs, mammary carcinogen models, and the compound-fragments matrix are available through the corresponding author.

Identifying "Important" Fragments of Activity and Inactivity

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 inactivity) 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 in the learning set (i.e., 1.5%). 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 sets are composed of 100 active and 100 inactive compounds, 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. We derived models for two proportions, 0.75 and 0.90, for both the ABC and ABCH fragment sets. In general, we reasoned that even if a particular fragment is associated with activity, there may 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. A similar argument can be made 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 depending on model, also found in at least 75% or 90% of the active or inactive compounds that derived them. The models developed are listed in Table 1.

Predicting Activity

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 predicted chemical is calculated from the average probability of the active and inactive fragments contained in it. For example, if a compound contains two fragments, one being found in 9/10 active compounds in the learning set (i.e., 90% active) and the other being found in 3/3 inactive compounds (i.e., 100% inactive), the unknown compound

will be predicted to be inactive based on the higher probability of inactivity 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.

"Validating" the Model

The cat-SAR program contains a leave-one-out cross-validation routine. For each chemical in the model's 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 remaining fragment was recalculated. Using the same criteria described above, the activity of the removed chemical was then predicted using the reduced fragment set.

The cat-SAR predictions are based on two separable fragment sets: The inactive fragments and the active fragments. As mentioned, the predicted activity of a chemical is based on the average probability of all the active and inactive compounds contributing to its fragments. As such, the user can "decide" at what predicted probability of activity or inactivity to categorize the test compound as a carcinogen or noncarcinogen.

To address this, we have adapted a routine from our previous MultiCASE work in which we identify a cut-off point that optimally separates the prediction of active and inactive compounds. This is based on the results of the validation exercise. In other words, since the prediction of activity or inactivity is a probability, we allow the validation exercise to guide us in determining, based on a probability of activity, what is ultimately classified as an active or inactive prediction.

Results and Discussion

Overview of Predictive Performance of the Cat-SAR Mammary Carcinogen Models

The best rat MC-NC model achieved an 82% observed correct prediction (OCP) rate with a sensitivity of 77% and a specificity of 88% (ABC 3/90 Model 1, Table 1). This model made predictions on 145 of the 200 chemicals in the learning set. The best rat MC-NMC model achieved a 79% OCP rate with a sensitivity of 83% and a specificity of 74% (ABC 3/90 Model 2, Table 2). This model made predictions on 124 of the 200 chemicals in the learning set.

In order to better judge how well these two models performed, we can consider the "accuracy" or reproducibility of *in vivo* or *in vitro* toxicological tests themselves. In general, surrogate tests and carcinogen bioassays are not reproducible with 100% concordance. For instance, the NTP's *Salmonella* mutagenicity database, which is derived from a standardized protocol, has been estimated to be 85% reproducible *in vitro (34)*. Moreover, it was found that based on "near-replicate" experiments in the CPDB, there was also a degree of non-reproducibility *(4,35)*. For example, 11 out of 54 chemicals tested in similar experiments for their ability to induce cancer in mice were discordant (i.e., 80% reproducible) and 16 out of 104 chemicals tested for cancer in rats were discordant (i.e., 85% reproducible) *(4)*. This does not imply that the CPDB data is flawed, only that there is variability in results. However, based on these findings, and since the majority of results in the CPDB have not been subjected to replication, it does suggest that SAR models built on this data will not, nor should be expected to achieve 100% accuracy. The cat-SAR mammary carcinogen models appear to be in the same neighborhood of predictivity as the bioassays themselves.

Analysis of Random Subsets of Inactives Models

Statistical analysis of each set of three models derived from the random selection of non carcinogenic or nonmammary carcinogenic compounds indicated the models had approximately the same OCP rate. The most variable set of OCP values was for the rat MC-NMC ABCH 3/90 model where it varied from 72-79% (Table 2). All other models showed a closer spread of OCP values. This provides a degree of confidence that the accurate predictions made by the models were not spurious events based on the fortuitous selection of "good" compounds for the learning set. In other words, this provides assurance that the models are based on a sound foundation and are not providing arbitrary predictions or mechanistic assertions.

Comparison of 75% vs. 90% Models

In order to best compare the 75% and 90% models as well as comparisons described below between the ABC and ABCH models, we chose to consider the average values for fragment counts, OCP, sensitivity, and specificity. These were calculated for each set of three models derived from different sets of randomly selected inactive compounds. This makes discussion of the data more straightforward and also provides for a more robust analysis. However, similar observations can be found when comparing individual models as well.

As the criteria for selecting important fragments increased from those found in 75% of active or inactive compounds to those found in 90%, the model's OCP, sensitivity, and specificity increased. For example, when looking at the ABC MC-NC models, as the fragment selection criteria increased from 75 to 90% the average OCP rose from 74 to 81% (Table 1). The

sensitivity and specificity also rose from 71 to 81% and from 78 to 80%, respectively (Table 1). Similar increases can be seen for all other rat MC-NC and MC-NMC models (Tables 1 and 2).

The trend in improved OCP, sensitivity, and specificity indicates that as the requirements for selecting important fragments are tightened (e.g., increasing proportion from 75% to 90%) the accuracy of predictions made from the resultant more stringent model increased. This is not unexpected. There is a cost associated with this increased accuracy, however. The more stringent models do not contain as many important fragments as the less stringent ones. The rat MC-NC ABC 3/75 model was based on an average of 1484 fragments while the 3/90 model contained 1152 (Table 1). Similar trends can be found for the other models.

Comparison of ABC vs. ABCH Models

Upon comparison of the ABC to ABCH models, the ABC models typically performed better. In general, the ABCH models contained about twice as many fragments. For example, the MC-NC ABC models on average were based on 16979 fragments and the set of ABCH models were based on 36947 fragments (Table 1). Similar increases were also observed for the total number of important fragments as well as the pool of active and inactive fragments. The reason for this is that, as expected, the fragments in the ABCH models are more specific by explicitly considering hydrogen atoms.

Interestingly, these more specific ABCH fragments did not enhance the overall predictivity of the models, but lowered it. For example, the rat MC-NC ABC 3/90 had an average OCP of 81% while the ABCH 3/90 had an OCP of 76% (Table 1). This trend is evident in other comparisons

between the ABC and ABCH models. However, in general the ABCH models were able to make predictions on a greater number of compounds. Again, for example, the rat MC-NC ABC 3/90 model made predictions on 139 chemicals while the ABCH 3/90 made predictions on 164 compounds (Table 1). Similar results were seen for other MC-NC as well as MC-NMC models (Tables 1 and 2).

Examples of cat-SAR Predictions

Atrazine and fenaminosulf were selected to illustrate cat-SAR predictions of mammary carcinogens based on the MC-NC model. Nithiazide and 1-phenyl-3,3-dimethyltriazene were selected to illustrate the MC-NMC models. These "prediction" examples 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. Additionally, the following discussion of the compounds with consideration of their tumor sites and *Salmonella* mutagenicity are based on their classification in the CPDB.

Atrazine is a male mammary carcinogen in rats and does not induce tumors in any other sites in the rat male. Atrazine also induces cancer of the hematopoietic system and uterus in female rats. Atrazine has been tested in male and female mice and has been determined not to be a mouse carcinogen. It is also not a *Salmonella* mutagen. Atrazine was predicted to be a rat mammary carcinogen based on the possession of 10 fragments associated with mammary cancer in other compounds. (Table 3 and Figure 1). Each of these fragments was found in three other

compounds in the learning set, all of which were mammary carcinogens (Table 3). Atrazine is therefore predicted to have a 100% chance of being a mammary carcinogen.

Fenaminosulf has been tested in male and female mice and rats and has not been observed to induce cancer in any of the four groups. Fenaminosulf however, is classified as a *Salmonella* mutagen. Fenaminosulf was predicted to be a rat mammary noncarcinogen based on seven fragments (Table 4 and Figure 2). Each of the seven fragments was found in four other chemicals in the learning set, none of which were mammary carcinogens. Fenaminosulf is therefore predicted to have a 100% chance of not being a rat mammary carcinogen.

Nithiazide has been tested in male and female mice and rats. In rats, it only induces mammary gland cancer in females. It is a noncarcinogen in male rats. Alternately, it only produces liver tumors in male mice and is a noncarcinogen in female mice. It is also a *Salmonella* mutagen. Nithiazide was predicted to be a rat mammary carcinogen based on the possession of 29 fragments. All 29 fragments were predominately found in other mammary carcinogens. However, six fragments (i.e., frags 329 to 756) also included some inactive members. As such, Nithiazide was predicted to have a 98% probability of being a rat mammary carcinogen (Table 5 and Figure 3).

Lastly, 1-phenyl-3,3-dimethyltriazene has been tested in male and female rats and was found to induce cancer of the nervous system. It has not been tested in mice. The compound is a *Salmonella* mutagen. 1-Phenyl-3,3-dimethyltriazene was predicted to be a rat nonmammary carcinogen based on the possession of eight fragments. All of the fragments except fragment

1552 were derived from other rat nonmammary carcinogens. This compound was predicted to have a 98% probability of being a nonmammary carcinogen (Table 6 and Figure 4).

Since cat-SAR predictions are based on the complete correspondence of important fragments in the model to fragments contained in the test compound, redundant reasons for activity or inactivity are observed. For example, the 10 fragments used to predict the activity of atrazine can be grouped into those just representing triazine moieties and those including part of the ethylamino and isopropylamino groups. This presents the program user with a challenge in understanding the predictions, given the redundant fragments. However, we feel that this is a strength of the cat-SAR program given the fact that it is capable of simultaneously handling thousands of fragments in the prediction and analysis processes. As such, no important features are lost due to filter mechanisms which may or may not be understood and accessible to the user. Consider that a previous SAR analysis of rat carcinogens from the CPDB based on 745 chemicals using the MultiCASE program yielded only 26 major biophores *(7)*.

Preliminary Mechanistic Analysis

Comparisons between the rat mammary carcinogen models and three other set(s) of cat-SAR models were conducted to assess the likelihood that these models were related and thus have a common underlying mechanism(s) of action. For these analyses we considered models based on rat carcinogens, female rat carcinogens, and *Salmonella* mutagens (unpublished models).

The extent of relating features between two SAR models can be taken to be indicative of the extent of mechanistic overlap between models and the underlying biological phenomena they

describe. We used the Chemical Diversity Approach (CDA) previously described by us to investigate the possible interrelationships between these models. Briefly, the CDA consists of using a random sample of 10,000 chemicals representing the "universe of chemicals". Then, using validated SAR models we predicted the activity of these chemicals. The prevalence of chemicals predicted to possess simultaneously greater than chance the ability to induce two or more toxicological effects should then provide a measure of the mechanistic relatedness of these phenomena.

The first set of CDA analyses considers the relationship between the two rat mammary carcinogen models (i.e., MC-NC and MC-NMC) and two models recently built from CPDB data for all rat carcinogens and female rat carcinogens (unpublished data). For these analyses, the level of significance was set at $p<0.001$). The rat carcinogen and female rat carcinogen models showed 43.6% greater than expected overlap (Analysis 1, Table 7). This significant overlap is expected since the learning set for the female rat carcinogen model is a subset of rat carcinogens. There was also a 72.4 % significant overlap between the rat carcinogen and rat MC-NC models (Analysis 2, Table 7). Comparison of the female rat carcinogen model to the rat MC-NC model shows a 138.0% overlap (Analysis 3, Table 7). We note that the rat MC-NC model is not a perfect subset of the female rat model since several male mammary carcinogens are included in it. However, the majority of mammary carcinogens in the MC-NC model are female mammary carcinogens. Overall, this very high degree of overlap underlines the close relationship between female carcinogens and mammary carcinogens. Taken together, these analyses suggest that the rat carcinogen, female rat carcinogen, and rat MC-NC models are all closely related.

On the other hand, the degree of overlap between the rat MC-NMC model with the rat carcinogen, female rat carcinogen, and rat MC-NC models is not high. The rat MC-NMC and rat carcinogen model had a significant but modest 19.3% overlap (Analysis 4, Table 7). The rat MC-NMC model and the female rat carcinogens model had a nonsignificant (p=0.032) overlap of only 11.4% (Analysis 4, Table 7) and the rat MC-NC and MC-NMC models also had a nonsignificant (p= 0.010) overlap of 13.6% (Analysis 5. Table 7). These last two analyses suggest that the rat MC-NMC model is significantly different than the rat MC-NC and female rat carcinogen models.

The last set of analyses considered the relationships between the carcinogen models and *Salmonella* mutagenicity. Analyses 7, 8, and 9 (Table 7) indicate a strong and expected overlap between mutagenicity and carcinogenicity. Interestingly however, there was no overlap observed between the rat MC-NMC and *Salmonella* models (Analysis 10, Table 7) (p=0.961). Since the rat MC-NMC model contained carcinogens in both its "active" (i.e., mammary carcinogens) and "inactive" (i.e., carcinogens at other sites than the mammary gland) categories, the model was standardized for carcinogens. In other words, the MC-NMC is based on mechanistic attributes that describe how carcinogens may act as breast carcinogens—not how chemicals are carcinogens.

Although the MC-NMC model was standardized for carcinogens, it also took into account a basic mechanism of carcinogenicity, that being mutagenicity. Analysis of the *Salmonella* mutagenicity of compounds in the rat MC-NC and MC-NMC models showed of the 73 mammary carcinogens with accompanying mutagenicity data from the CPDB, used for both the

MC-NC- and MC-NMC models, 61 (83.6%) were mutagens. This is consistent with findings by Gold and colleagues who reported for chemicals tested in both mice and rats, that 79% of mutagens were carcinogens and only 49% of nonmutagens were carcinogens *(4)*. Considering the MC-NC model, of the 66 noncarcinogens included in the "inactive" category that had mutagenicity data, only 14 (21%) are mutagens. This again is consistent with Gold *et al.* where they report 25% of noncarcinogens are mutagens *(4)*. On the other hand, when considering the nonmammary carcinogens used to make the "inactive" category of the MC-NMC model, of the 75 compounds with mutagenicity data, 43 (57.3%) were mutagenic. In other words, the MC-NMC model, while having carcinogens in both its "active" (i.e., mammary carcinogens) and "inactive" (i.e., nonmammary carcinogens) categories, both categories also had a high prevalence of mutagens. Thus mutagenic features were represented in both categories and therefore were not identified as being associated with activity of either category. As above, since the phenomena of mutagenesis was not considered in the modeling process, the MC-NMC describes how carcinogens may act as breast carcinogens—not how mutagens induce cancer.

Conclusions

Currently, the NTP designates 228 substances "known" or "reasonably anticipated" to pose a cancer risk *(36)*. Unfortunately, this number is based on the analysis of only a few of upwards of 70,000 chemicals manufactured and used in this country. Computational structure-activity relationships (SAR) have gained recent acceptance in the regulatory community for both human health *(37)* and ecological endpoints *(38)*. The present investigation consisted of a SAR analysis of a subset of the CPDB that included mammary carcinogens, noncarcinogens, and carcinogens at sites other than the mammary gland. Cat-SAR analysis of the MC-NC- and MC-NMC

datasets produced two sets of models with balanced sensitivity and specificity and OCP values between 70 and 82%.

Interestingly, the MC-NMC model was based on a learning set that contained carcinogens in both the active (i.e., mammary carcinogens) and inactive (i.e., carcinogens to sites other than the mammary gland) categories. The best of these models was able to achieve an OCP of 82%, indicating the ability to distinguish between different types of carcinogens (i.e., tissue specific), not simply between carcinogens and noncarcinogens. Moreover, based on a structural comparison between this model and a model for *Salmonella* mutagens, there was no observed relationship between the two phenomena. Likewise, in an analysis of the proportion of *Salmonella* mutagens contained in the models learning set, both the active (i.e., mammary carcinogens) and inactive (i.e., carcinogens to sites other than the mammary gland) categories had a high prevalence of mutagens. These findings suggest that the MC-NMC model is identifying structural attributes of chemicals that impart on them the ability to induce breast cancer which are separable from those generally associated with carcinogenic potential (e.g., DNA-reactivity).

By including carcinogens in the active (i.e., breast carcinogens) and inactive (i.e., non-breast carcinogens) categories of the MC-NMC, we hypothesize that we have removed a "layer" of explainable mechanisms associated with chemical carcinogenesis and have revealed another layer for studying why carcinogens target the breast (Figure 5). In other words, most SAR models, by analyzing chemicals with a known and specific biological activity, remove the majority of chemical structures and identify a subset of features associated with the given biological activity. In fact, we have demonstrated that even traditional organic chemical

categories are often removed since both the active or inactive groups being modeled contain many of these traditional features *(39)*. Likewise, since we have populated both the active and inactive categories with carcinogens, features of chemical carcinogens have been removed from the MC-NMC model allowing for features associated with how carcinogens target the breast to be identified.

Finally, the cat-SAR expert system used herein is a knowledge based one (i.e., knowledge contained in the learning set) and is not hypothesis driven. Thus, the toxicophores identified are not dependent upon previous knowledge or assumptions regarding a mechanism of action. As such, the identified attributes of breast carcinogens can be used to explore previously established or hypothesized mechanisms or more importantly, in the case of the MC-NMC model, to develop new testable hypotheses relating to the chemical induction of breast cancer.

Acknowledgments
This research was supported by the Department of Defense Breast Cancer Research Program under award number DAMD17-01-0376. Views and opinions of, and endorsements by the author(s) do not reflect those of the US Army or the Department of Defense.

Table 1. Predictive performance summary for rat mammary carcinogen – noncarcinogen (MC-NC) SAR model. 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*
ABC3/0.75							
Model 1	18021	1336	758	578	0.73(66/90)	0.78(69/88)	0.76(135/178)
Model 2	17369	1486	786	700	0.71(67/95)	0.80(72/90)	0.75(139/185)
Model 3	15547	1629	737	892	0.69(62/91)	0.76(67/88)	0.72(129/179)
Average	16979	1484	760	723	0.71	0.78	0.74(134/181
ABC3/0.90							
Model 1	18021	1016	642	374	0.82(62/76)	0.78(47/60)	0.80(109/136)
Model 2	17369	1129	617	512	0.77(56/73)	0.88(63/72)	0.82(119/145)
Model 3	15547	1311	624	687	0.83(63/76)	0.73(44/60)	0.79(107/136)
Average	16979	1152	628	524	0.81	0.80	0.81(112/139)
ABCH3/0.75							
Model 1	38797	3859	1790	2069	0.72(68/94)	0.76(68/90)	0.74(136/184)
Model 2	37636	4293	2007	2286	0.71(70/98)	0.77(75/97)	0.74(145/195)
Model 3	34407	4093	1785	2308	0.73(71/97)	0.65(62/95)	0.69(133/192)
Average	36947	4082	1861	2221	0.72	0.73	0.72
ABCH3/0.90							
Model 1	38797	2746	1434	1312	0.76(63/83)	0.78(61/78)	0.77(124/161)
Model 2	37636	2923	1392	1531	0.75(63/84)	0.78(67/86)	0.77(130/170)
Model 3	34407	2949	1372	1577	0.74(66/89)	0.71(52/73)	0.73(118/162)
Average	36947	2873	1399	2210	0.75	0.76	0.76(124/164

Footnotes:
Total Fragments: number of fragments derived from learning set.
Model Fragments: number of fragments meeting specified rules of the model.
Active Fragments: number of fragments meeting specified rules to be considered as active.
Inactive Fragments: number of fragments meeting specified rules to be considered as inactive.
Sensitivity: number of correct positive predictions / total number of positive predictions.
Specificity: number of correct negative predictions / total number of negative predictions
OCP: Observed Correct Predictions: number of correct predictions / total number of predictions.

Table 2. Predictive performance summary for rat mammary carcinogen–nonmammary carcinogen (MC-NMC) SAR model. 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*
ABC3/0.75							
Model 1	13868	1349	849	500	0.80(70/88)	0.66(53/80)	0.73(123/168)
Model 2	14461	1330	861	469	0.72(63/87)	0.72(59/82)	0.72(122/169)
Model 3	14427	1245	767	478	0.68(59/87)	0.74(64/86)	0.71(123/173)
Average	14252	1308	826	482	0.73	0.71	0.72(123/170)
ABC3/0.90							
Model 1	13868	1102	731	371	**0.83(58/70)**	**0.74(40/54)**	**0.79(98/124)**
Model 2	14461	1086	723	363	0.82(54/66)	0.72(44/64)	0.75(98/130)
Model 3	14427	847	520	327	0.82(51/62)	0.72(41/57)	0.77(92/119)
Average	14252	1308	826	482	0.82	0.73	0.77(96/124)
ABCH3/0.75							
Model 1	32235	3679	2081	1598	0.81(78/96)	0.62(55/89)	0.72(133/185)
Model 2	32374	3921	2088	1833	0.70(66/94)	0.64(59/92)	0.67(125/186)
Model 3	32627	3497	1928	1569	0.75(70/93)	0.69(65/94)	0.72(135/187)
Average	32412	3699	2032	1667	0.75	0.65	0.70
ABCH3/0.90							
Model 1	32235	2750	1642	1108	0.81(65/80)	0.76(50/66)	0.79(115/146)
Model 2	32374	2947	1637	1310	0.75(55/73)	0.69(53/77)	0.72(108/150)
Model 3	32627	2241	1170	1071	0.81(63/78)	0.70(52/74)	0.76(115/152)
Average	32412	3699	2032	1667	0.79	0.72	0.76

Footnotes: see table 1

Table 3. Fragments from the ABC 3/0.90 mammary carcinogen noncarcinogen (MC-NC) model leave-one-out validation analysis used to predict the rat mammary carcinogen atrazine.

Fragment	*No. Active**	*No. Inactive†*	*Total‡*	*% Active*	*% Inactive*
Frag 3662	3	0	3	1.000	0.000
Frag 3663	3	0	3	1.000	0.000
Frag 3664	3	0	3	1.000	0.000
Frag 3665	3	0	3	1.000	0.000
Frag 3666	3	0	3	1.000	0.000
Frag 3667	3	0	3	1.000	0.000
Frag 3668	3	0	3	1.000	0.000
Frag 3670	3	0	3	1.000	0.000
Frag 3671	3	0	3	1.000	0.000
Frag 3677	3	0	3	1.000	0.000
	Probability of activity			1.00	0.00

Table 4. Fragments from the ABC 3/0.90 mammary carcinogen noncarcinogen (MC-NC) model leave-one-out validation analysis used to predict rat noncarcinogen fenaminosulf.

Fragment	*No. Active**	*No. Inactive†*	*Total‡*	*% Active*	*% Inactive*
Frag 6443	0	4	4	0.000	1.000
Frag 6446	0	4	4	0.000	1.000
Frag 6447	0	4	4	0.000	1.000
Frag 6451	0	4	4	0.000	1.000
Frag 6452	0	4	4	0.000	1.000
Frag 6455	0	4	4	0.000	1.000
Frag 6461	0	4	4	0.000	1.000
	Probability of activity			0.00	1.00

Table 5. 29 Fragments from the ABC 3/90 rat nonmammary (MC-NMC) model leave-one-out validation analysis used to predict nithiazide as being a mammary carcinogen.

Fragment	*No. Active**	*No. Inactive†*	*Total‡*	*% Active*	*% Inactive*
Frag328	11	1	12	0.917	0.083
Frag352	11	1	12	0.917	0.083
Frag361	12	1	13	0.923	0.077
Frag508	21	2	23	0.913	0.087
Frag746	12	1	13	0.923	0.077
Frag756	12	1	13	0.923	0.077
Frag1739	10	0	10	1.000	0.000
Frag1740	10	0	10	1.000	0.000
Frag1741	10	0	10	1.000	0.000
Frag1742	9	0	9	1.000	0.000
Frag1743	9	0	9	1.000	0.000
Frag1744	9	0	9	1.000	0.000
Frag1745	9	0	9	1.000	0.000
Frag1746	9	0	9	1.000	0.000
Frag1747	9	0	9	1.000	0.000
Frag1754	9	0	9	1.000	0.000
Frag1758	10	0	10	1.000	0.000
Frag1759	9	0	9	1.000	0.000
Frag1765	10	0	10	1.000	0.000
Frag1775	11	0	11	1.000	0.000
Frag1776	11	0	11	1.000	0.000
Frag1777	10	0	10	1.000	0.000
Frag1778	10	0	10	1.000	0.000
Frag1779	10	0	10	1.000	0.000
Frag1780	10	0	10	1.000	0.000
Frag1781	10	0	10	1.000	0.000
Frag1782	10	0	10	1.000	0.000
Frag1801	10	0	10	1.000	0.000
Frag1831	10	0	10	1.000	0.000
	Probability of activity			0.977	0.023

Table 6. 8 Fragments from the ABC 3/90 rat nonmammary (MC-NMC) model leave-one-out validation analysis used to predict 1-phenyl-3,3-dimethyltriazene as being a nonmammary carcinogen.

Fragment	*No. Active**	*No. Inactive†*	*Total‡*	*% Active*	*% Inactive*
Frag1552	1	11	12	0.083	0.917
Frag1557	0	5	5	0.000	1.000
Frag1558	0	5	5	0.000	1.000
Frag1559	0	5	5	0.000	1.000
Frag1561	0	5	5	0.000	1.000
Frag1572	0	6	6	0.000	1.000
Frag1576	0	6	6	0.000	1.000
Frag1577	0	5	5	0.000	1.000
	Probability of activity			0.02	0.98

Table 7. Mechanistic relationships between the cat-SAR rat mammary carcinogen, other carcinogen, and *Salmonella* mutagen models.

Analysis	*Observed*	*Expected*	*p-value*	*Δ*	*100Δ/Expected*
1. Rat + F-Rat	1166	812	<0.0001	354	43.6
2. Rat + Rat MC-NC	1410	818	<0.0001	592	72.4
3. F-Rat + Rat MC-NC	1202	505	<0.0001	697	138.0
4. Rat + Rat MC-NMC	1335	1119	<0.0001	216	19.3
5. F-Rat + Rat MC-NMC	769	690	0.032	79	11.4
6. Rat MC-NC + Rat MC-NMC	791	696	0.010	95	13.6
7. Rat + Salm	1537	1021	<0.0001	516	50.5
8. F-Rat + Salm	1648	692	<0.0001	956	138.2
9. Rat MC-NC + Salm	1595	697	<0.0001	898	128.8
10. Rat MC-NMC + Salm	933	935	0.961	-2	-0.2

Notes:
Δ: Observed Prevalence – Expected Prevalence
100Δ/Expected: 100*Δ/Expected Prevalence

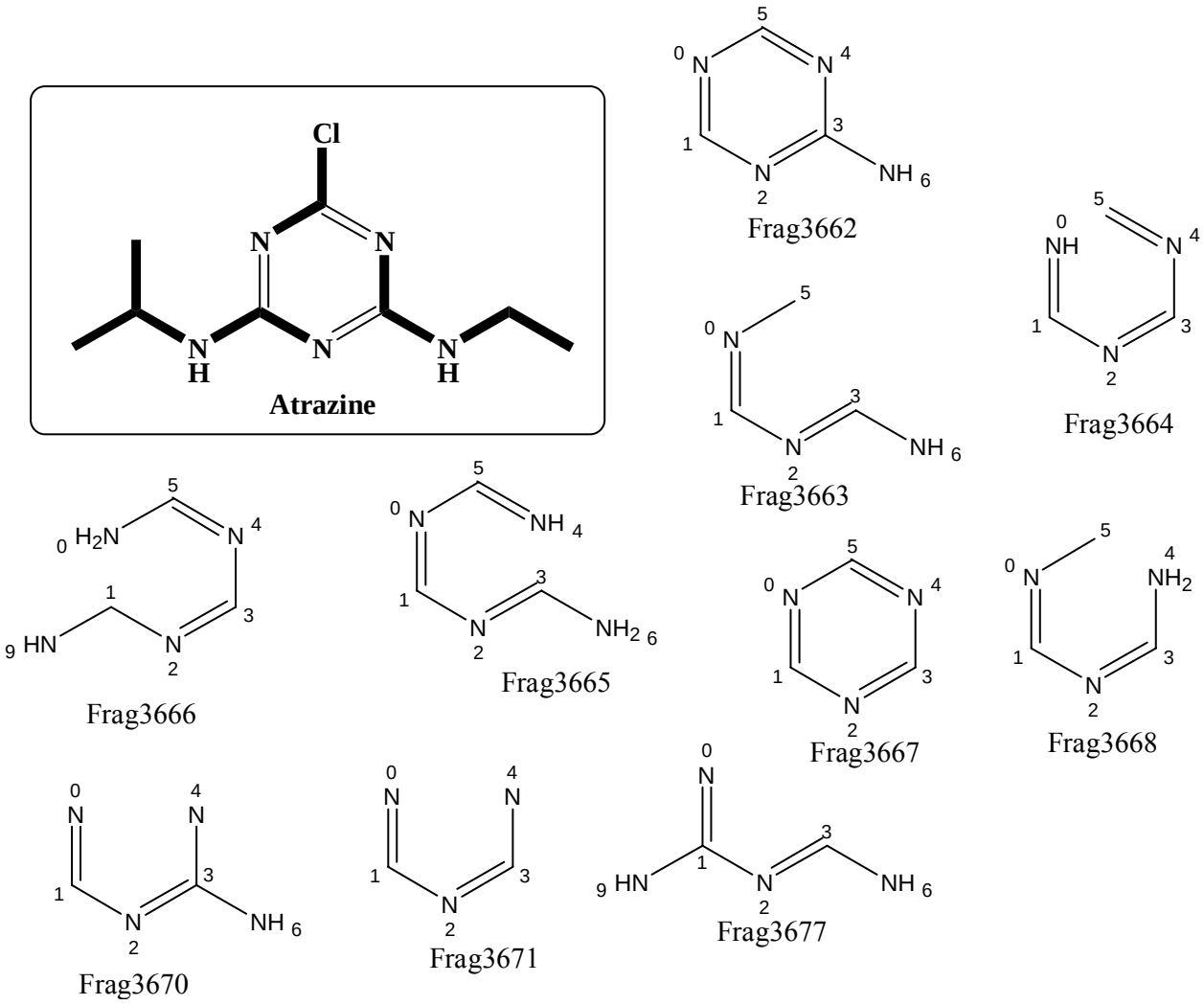


Figure 1. Illustration of the 10 significant fragments contributing to the active validation prediction of atrazine.

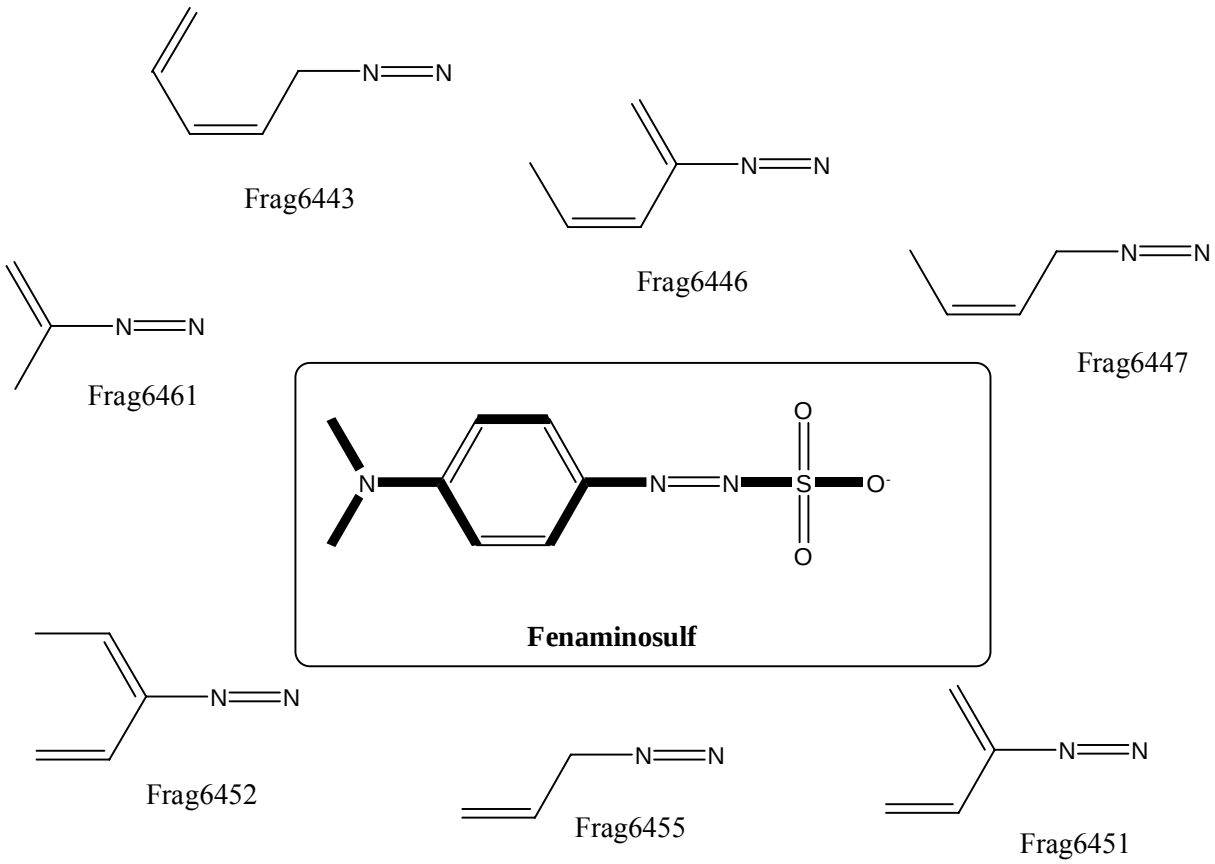


Figure 2. Illustration of the 7 significant fragments contributing to the inactive validation prediction of the fungicide fenaminosulf.

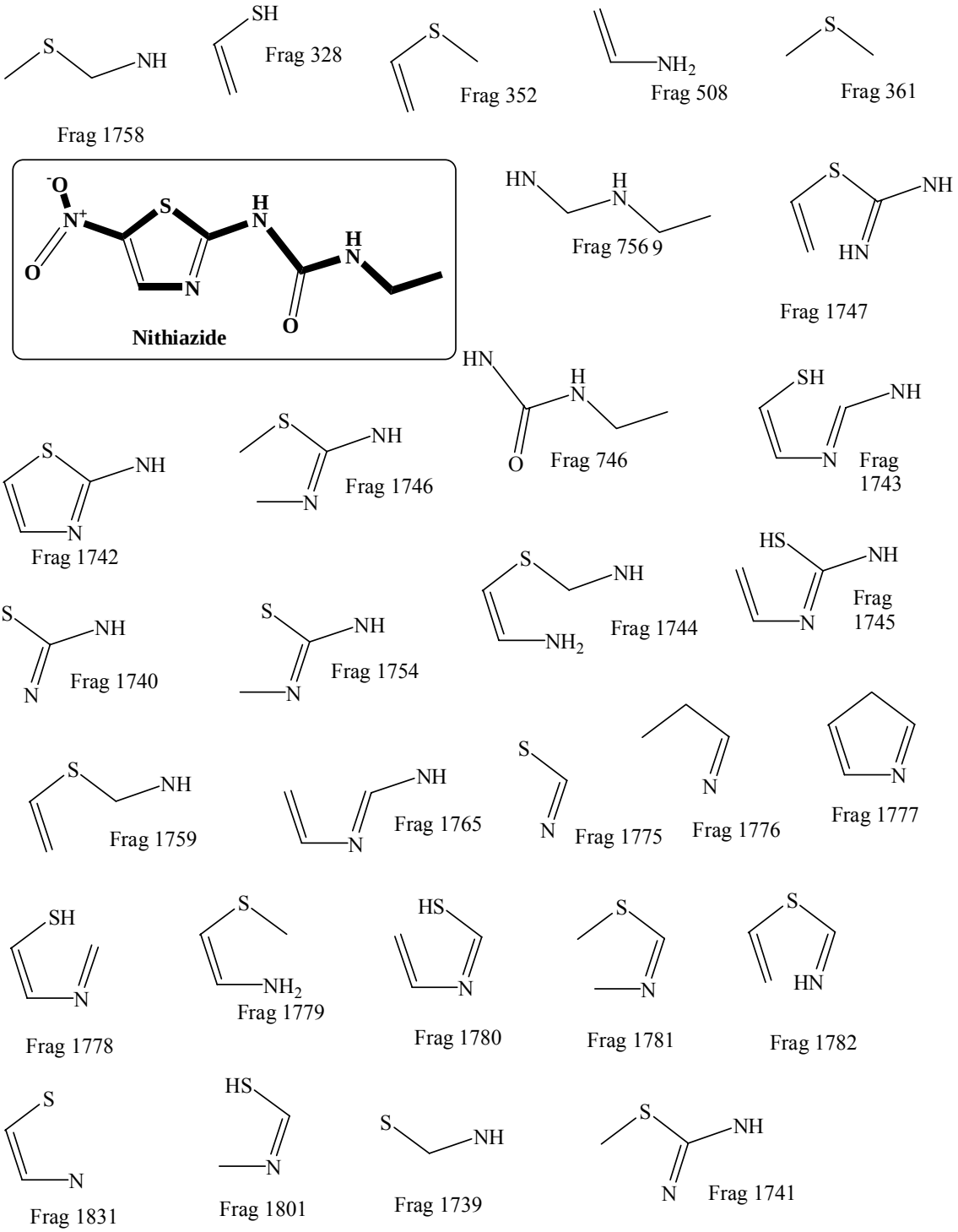


Figure 3. Illustration of the 29 significant fragments contributing to the active validation prediction of nithiazide.

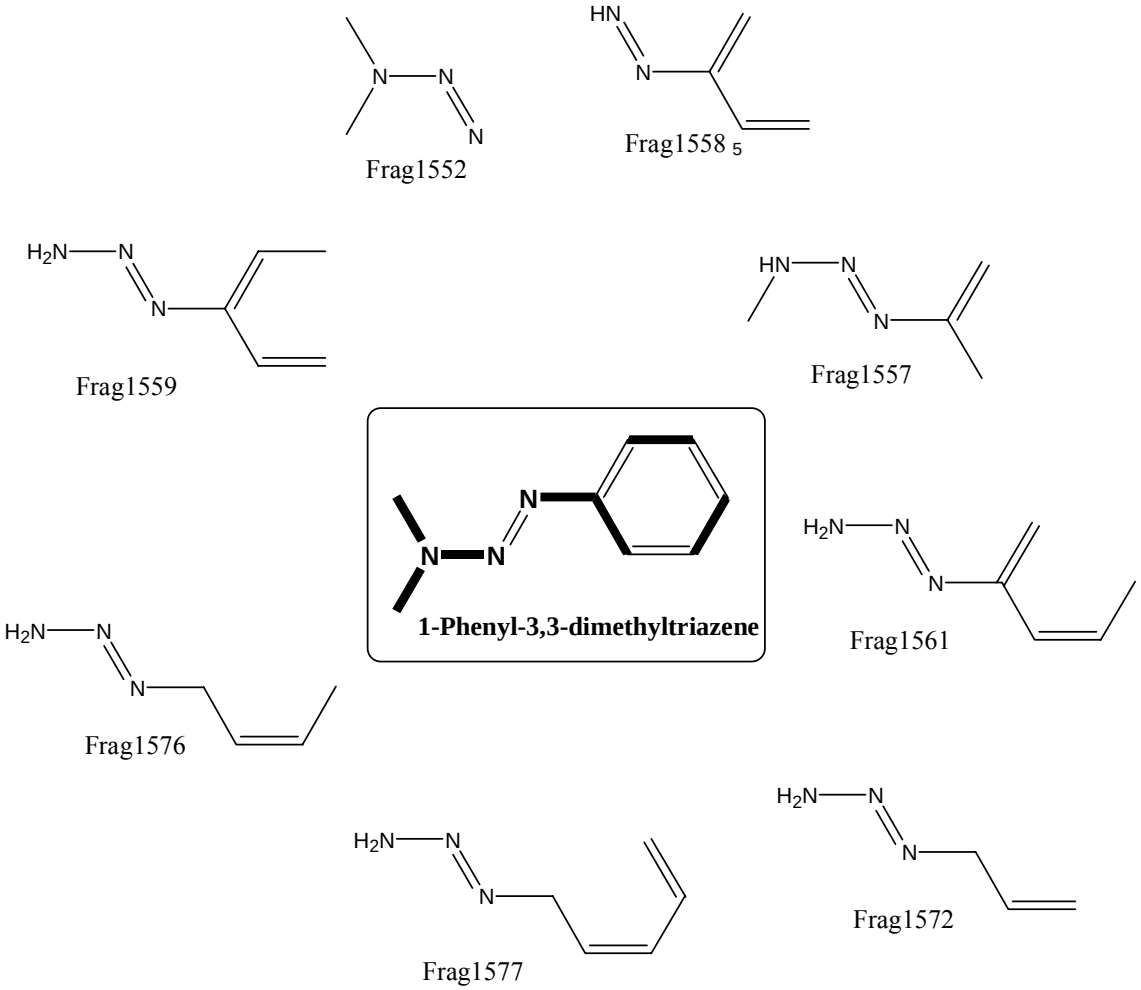


Figure 4. Illustration of the eight fragments contributing to the inactive validation prediction of 1-phenyl-3,3-dimethyltriazene.

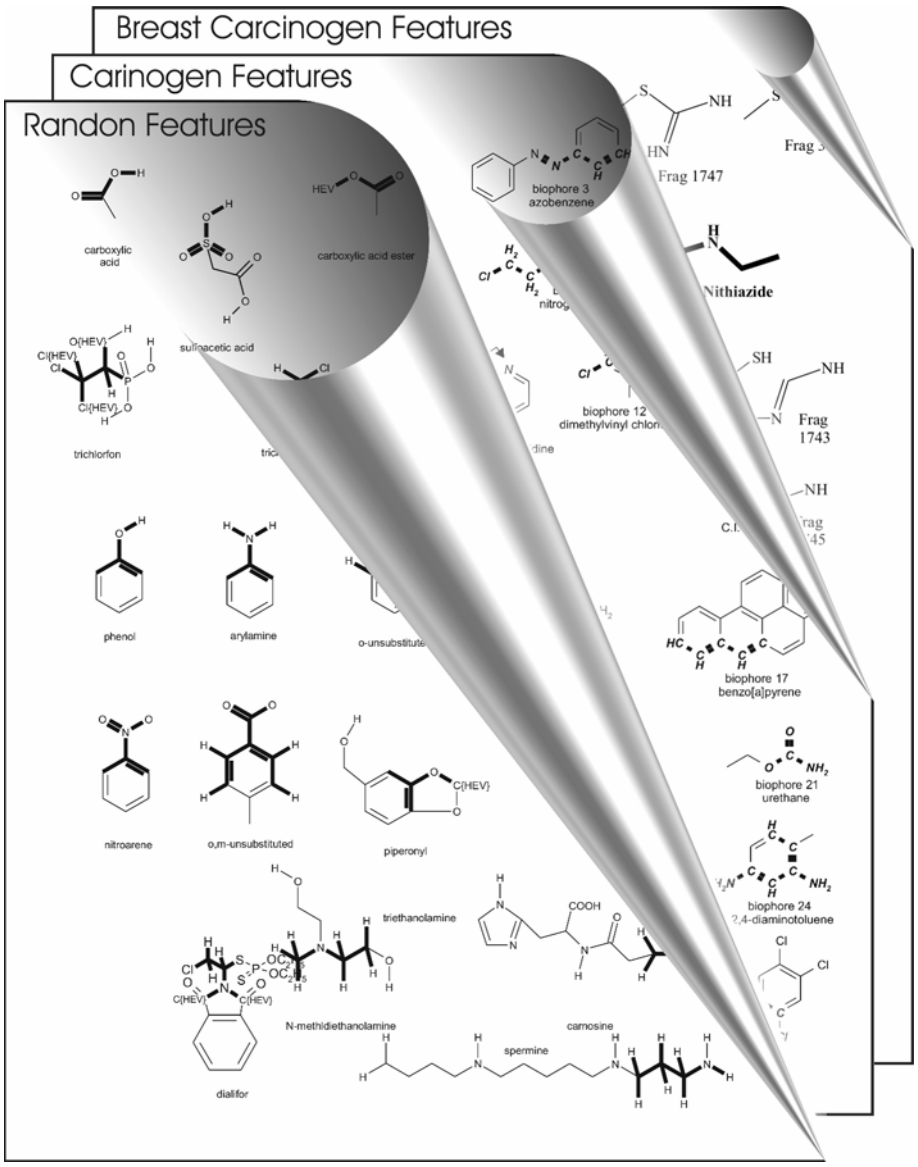


Figure 5. Illustrative nature of how SAR models can remove mechanistic layers of chemical carcinogenesis in order to study specific aspects of the process. A typical SAR model developed from categories of carcinogens and noncarcinogens removes many common chemical structures (top layer) reveals a set of features associated with carcinogenesis (middle layer). The SAR model developed from categories that both contained carcinogens (middle layer) reveals a set of features associated with breast-specificity (bottom layer).

References

(1) EPA (2004) What is the TSCA Chemical Substance Inventory? *http://www.epa.gov/oppt/newchems/invntory.htm*, *last accessed 10/21/04*.

(2) NIEHS (2004) About the NTP. *http://ntp-server.niehs.nih.gov/index.cfm?objectid=7201637B-BDB7-CEBA-F57E39896A08F1BB*, *last accessed 10/21/04*.

(3) NTP (1996) NIEHS Fact Sheet #3 The National Toxicology Program. *http://www.niehs.nih.gov/oc/factsheets/fsntp.htm*, *last accessed 10/21/04*.

(4) Gold, L.S., Sloan, T.H. and Ames, B.N. (1997) Overview and update of analyses of the carcinogenic potency database. In *Handbook of Carcinogenic Potency and Genotoxicity Databases* (Gold, L.S. and Zeiger, E., Eds.) pp 661-693, CRC Press, New York.

(5) CPDB (2004) Carcinogenic Potency Database. *http://potency.berkeley.edu*, *last accessed 10/21/04*.

(6) Cunningham, A.R., Rosenkranz, H.S., Zhang, Y.P. and Klopman, G. (1998) Identification of "genotoxic" and "non-genotoxic" alerts for cancer in mice: The carcinogenic potency database. *Mutat. Res.*, *398*, 1-17.

(7) Cunningham, A.R., Rosenkranz, H.S. and Klopman, G. (1998) Identification of structural features and associated mechanisms of action for carcinogens in rats. *Mutat. Res.*, *405*, 9-28.

(8) Gold, L.S., Sawyer, C.B., Magaw, R., Backman, G.M., deVeciana, M., Levinson, R., Hooper, N.K., Havender, W.R., Bernstein, L., Peto, R., Pike, M.C. and Ames, B.N. (1984) A carcinogenic potency database of the standardized results of animal bioassays. *Environ. Health Perspect.*, *58*, 9-319.

(9) Gold, L.S., deVeciana, M., Backman, G.M., Lopipero, M., Smith, M., Blumenthal, R., Levinson, R., Bernstein, L. and Ames, B.N. (1986) Chronological supplement to the carcinogenic potency database: Standardized results of animal bioassays published through December 1982. *Environ. Health Perspect.*, *67*, 161-200.

(10) Gold, L.S., Slone, T.H., Backman, G.M., Magaw, R., DaCosta, M., Lopipero, P., Blumenthaul, M. and Ames, B.N. (1987) Second Chronological supplement to the Carcinogenic Potency Database: Standardized results of animal bioassays published through December 1984 and by the National Toxicology Program through May 1986. *Environ. Health Perspect.*, *74*, 237-329.

(11) Gold, L.S., Slone, T.H., Backman, G.M., Eisenberg, S., DeCosta, M., Wong, M., Manley, N.B., Rohrbach, L. and Ames, B.N. (1990) Third chronological supplement to the Carcinogenic Potency Database: Standardized results of animal bioassays published through December 1986 and by the National Toxicology Program through June 1987. *Environ. Health Perspect.*, *84*, 215-286.

(12) Gold, L.S., Manley, N.B., Slone, T.H., Garfinkel, G.B., Rohrbach, L. and Ames, B.N. (1993) Fifth plot of the Carcinogenic Potency Database: Results of animal bioassays Published in the general literature through 1988 and by the National Toxicology Program through 1989. *Environ. Health Perspect.*, *100*, 65-168.

(13) Richard, A.M. (1998) Commercial toxicology prediction systems: a regulatory perspective. *Toxicol. Lett.*, *102-103*, 611-616.

(14) Richard, A.M. (1999) Application of artificial intelligence and computer-based methods to predicting chemical toxicity. *Knowl. Eng. Rev.*, *14*, 307-317.

(15) Miller, J.A. and Miller, E.C. (1977) Ultimate chemical carcinogens as reactive mutagenic electrophiles. In *Origins of Human Cancer* (Hiatt, H.H., Watson, J.D. and Winsten, J.A., Eds.) pp 605-627, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
(16) Cunningham, A.R., Klopman, G. and Rosenkranz, H.S. (1997) A dichotomy in the lipophilicity of natural estrogens/xenoestrogens and phytoestrogens. *Environ. Health Perspect. Suppl.*, *105(Suppl3)*, 665-668.
(17) Rosenkranz, H.S., Cunningham, A. and Klopman, G. (1996) Identification of a 2-D geometric descriptor associated with non-genotoxic carcinogens and some estrogens and antiestrogens. *Mutagenesis*, *11*, 95-100.
(18) Cunningham, A.R., Rosenkranz, H.S. and Klopman, G. (1998) Structural analysis of a group of phytoestrogens for the presence of a 2-D geometric descriptor associated with non-genotoxic carcinogens and some estrogens. *Proc. Soc. Exp. Biol. Med.*, *217*, 288-292.
(19) Brzozowski, A.M., Pike, A.C., Dauter, Z., Hubbard, R.E., Bonn, T., Engstrom, O., Ohman, L., Greene, G.L., Gustafsson, J.A. and Carlquist, M. (1997) Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature*, *389*, 753-758.
(20) Pike, A.C., Brzozowski, A.M., Hubbard, R.E., Bonn, T., Thorsell, A.-G., Engstrom, O., Ljunggren, J., Gustafsson, J.-A. and Carlquist, M. (1999) Structure of the ligand-binding domain of oestrogen receptor beta in the presense of a partial agonist and a full antagonist. *The EMBO Journal*, *18*, 4608-4618.
(21) Shiau, A.K., Barsted, D., Loria, P.M., Cheng, L., Kushner, P.J., Agard, D.A. and G.L., G. (1998) The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell*, 297-237.
(22) Hihi, A.K. and Wahli, W. (1999) Structure and function of the estrogen receptor. In *Estrogens and Antiestrogens I* (Oettel, M. and Schillinger, E., Eds.) pp 111-126, Springer, Berlin.
(23) Davis, D.L., Bradlow, H.L., Wolff, M., Woodruff, T., Hoel, D.G. and Anton-Culver, H. (1993) Medical hypothesis: Xenoestrogens as preventable causes of breast cancer. *Environ. Health Perspect.*, *101*, 372-377.
(24) Davis, D.L., Axelrod, D., Bailey, L., Gaynor, M. and Sasco, A.J. (1998) Rethinking Breast Cancer Risk and the Environment: The Case for the Precutionary Principle. *Environ. Health Perspect.*, *106*, 523-529.
(25) Ahlborg, U.G., Lipworth, L., Titus-Ernstoff, L., Hsieh, C.C., Hanberg, A., Baron, J., Trichopoulos, D. and Adami, H.O. (1995) Organochlorine compounds in relation to breast cancer, endometrial cancer, and endometriosis: An assessment of the biological and epidemiological evidence. *Crit. Rev. Toxicol.*, *25*, 463-531.
(26) Ashby, J., Houthoff, E., Kennedy, S.J., Stevens, J., Bars, R., Jekat, F.W., Campbell, P., Van Miller, J., Carpanini, F.M. and Randall, G.L.P. (1997) The challenge posed by endocrine-disrupting chemicals. *Environ. Health Perspect.*, *105*, 164-169.
(27) Safe, S.H. (1995) Environmental and dietary estrogens and human health: Is there a problem? *Environ. Health Perspect.*, *103*, 346-351.
(28) Falck, F.J., Ricci, A.J. and Wolfe, M.S. (1992) Pesticides and polychlorinated biphenyl residues in human breast lipids and their relation to breast cancer. *Arch. Environ. Health*, *47*, 143-146.
(29) NRC (1999) *Hormonally Active Agents in the Environment.* National Academy Press Washington, D.C.

(30) Young, J.F., Tong, W., Fang, H., Xie, Q., Pearce, B., Hashemi, R., Beger, R.D., Cheeseman, M.A., Chen, J.J., Chang, Y.-c.I. and Kodel, R.L. (2004) Building an organ-specific carcinogenic database for SAR analyses. *J. Toxicol. Environ. Health. A*, *67*, 1363-1389.

(31) Gold, L.S., Manley, N.B., Slone, T.H. and Ward, J.M. (2001) Compendium of chemcal carcinogens by target organ: Results of chronic biassays in rats, mice, hamsters, dogs, and monkeys. *Toxicol. Pathol.*, *29*, 639-652.

(32) Dearden, J.C. (2003) *In silico* prediction of drug toxicity. *J. Comput. Aided Mol. Des.*, *17*, 119-127.

(33) 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 QSAR Environ. Res.*, *16*, 273-285.

(34) Piegorsrch, W.W. and Zeiger, E. (1991) Measuring Intra-assay Agreement for the Ames *Salmonella* Assay. In *Statistical Methods in Toxicology* Statistical Methods in Toxicology ed. (Hotham, L., Ed. pp 35-41, Springer-Verlag, Heidlberg.

(35) Gold, L.S., Wright, C., Bernstein, L. and deVeciana, M. (1987) Reproducibility of results in near-replicate carcinogenesis bioassay. *J. Natl. Cancer Inst.*, *78*, 1149-1158.

(36) NTP "Report on Carcinogens, Tenth Edition," U.S. Department of Health and Human Services, Public Health Service, National Toxicology Program, 2002.

(37) Cronin, M.T.D., Jaworska, J.S., Walker, J.D., Comber, M.H.I., Watts, C.D. and Worth, A.P. (2003) Use of quantitative structure-activity relationships in international decision-making frameworks to predict health effects of chemical substances. *Environ. Health Perspect.*, *111*, 1376-1390.

(38) Cronin, M.T.D., Walker, J.D., Jaworska, J.S., Comber, M.H.I., Watts, C.D. and Worth, A.P. (2003) Use of quantitative structure-activity relationships in international decision-making frameworks to predict ecological effects and environmental fate of chemical substances. *Environ. Health Perspect.*, *111*, 1376-1390.

(39) Rosenkranz, H.S. and Cunningham, A.R. (2001) Chemical categories for health hazard identification: A feasibility study. *Regul. Toxicol. Pharmacol.*, *33*, 313-318.

PRACTICAL ASPECTS OF SAR MODELING OF RODENT CARCINOGENICITY: CONGENERIC *VS.* NON-CONGENERIC LEARNING SETS.

Herbert S. Rosenkranz[1],
Suzanne L. Cunningham[2], and Albert R. Cunningham[2]

Running Title: Congeneric *vs*. Non-Congeneric Carcinogenicity Learning Sets

* The authors declare that they have no conflict of interest

ABSTRACT

SAR approaches to predict the potential health effects associated with chemicals are achieving regulatory acceptance. In order to be effective and to continue receiving recognition, it is important to gain an understanding of the role of the nature of the database and its interaction with specific SAR paradigms.

One of the most controversial toxicities to predict is the potential of chemicals to induce cancers in rodents and the extrapolation of such an effect to human risk. In the present study, we report on the interaction between MULTICASE SAR and datasets derived from rodent carcinogenicity bioassays. Specifically, we compare the predictive performance of models derived from the total (non-congeneric) database with class-specific subsets that share significant toxicophores. Unexpectedly, models derived from such subsets did not show greatly improved predictive performances when compared to those derived from non-congeneric sets. Part of the reason for this derives from the fact that in non-congeneric data sets, moieties derived from toxicants containing toxicophores other than those derived from the specific subsets also contribute to overall carcinogenic potential. It is demonstrated that when the predictions derived from the SAR models of the total non-congeneric database as well as from toxicophore-specific subsets are combined using a Bayesian weight-of-evidence approach, the predictivity is improved.

Key words: *Congeneric, non-congeneric, databases, chemical classes, SAR, predictivity, carcinogenicity.*

Address for correspondence: H. S. Rosenkranz, Department of Biomedical Sciences, Florida Atlantic University, 777 Glades Rd., P. O. Box 3091, Boca Raton, FL 33431-0991. USA. Email: *rosenkra@fau.edu*

Introduction

The increased acceptance of SAR (structure-activity relationship) approaches to predict, regulate as well as understand toxicological phenomena [1-8] has resulted in increased emphasis on the actual performance of the various available SAR programs as well as on the role of the database (*i.e.* learning sets) for optimal performance [9-11].

One of the characteristics that has made recently developed SAR paradigms acceptable for the study of toxicological phenomena is their ability to process and analyze non-congeneric data sets, *i.e*. learning sets that contain different classes of chemicals [12,13]. The ability to work with non-congeneric data is one of the requirements for the application of SAR to the study of toxicological phenomena. This is in contrast to the use of SAR in medicinal chemistry [12]. In the latter application, most frequently one deals with a specific receptor or the active site of an enzyme and consequently the chemicals being modeled are congeneric in nature (*i.e.* they belong to a single chemical class or subclass). In toxicology, on the other hand, an untoward effect (*e.g.* carcinogenicity, lethality, allergic manifestation, hepatotoxicity) can be caused by a multitude of mechanisms and involve multiple targets. This results in a situation wherein the population of causative agents is non-congeneric and yet must be handled by a single SAR model specific for a particular toxicity.

The determination of the carcinogenic potential of a chemical has been controversial [14], especially the reliability of the rodent carcinogenicity bioassay as a predictor of human risk and the role of mechanistic information in human carcinogenic hazard identification. However, the demonstration that highly predictive substructure-based SAR models of rodent carcinogenicity could be developed [15] highlights the potential usefulness of SAR models in estimating carcinogenic risk. Moreover, the concurrent analyses of these models with SAR models of ancillary phenomena (*e.g.* genotoxicity, cell and systemic toxicity, inhibition of gap

junctional intercellular communication) could be used to ascertain human risk [16] further emphasize the potential usefulness of SAR models in understanding the mechanisms involved in chemical-induced carcinogenesis. Because the carcinogenic process is so complex, it would seem that the choice of the dataset used to derive the SAR model may be an important determinant of the nature and predictivity of the model.

In the present study we analyze the role of congeneric subsets of chemicals within a non-congeneric data set on the predictivity of SAR models of rodent carcinogenicity. For this investigation we used the MULTICASE SAR system [11,17,18]. This is one of the substructure-based SAR methods that is being used in predictive toxicology [3]. Specifically, we compare SAR models derived from non-congeneric, congeneric and "hybrid" data sets. A parallel study [19] extends the analysis to a consideration of SAR models of human contact allergenicity. The latter presents different sets of problems and applications.

Materials And Methods

SAR methodology

The MULTICASE SAR methodology has been described previously [11,17,18]. An SAR model's performance was ascertained by its ability to accurately predict active compounds (*i.e.*, sensitivity), inactive one (*i.e.,* specificity) or both (*i.e.*, overall correct predictions) was determined by its ability to predict chemicals external to the datasets used to generate the initial model [4,9,10,20,21].

Databases

For the purpose of SAR modeling of carcinogenicity in rodents, we selected the Carcinogenic Potency Data Base (CPDB) [22]. In that dataset the results of assays for chemicals tested in both mice and rats were included. A chemical was considered a carcinogen if it induced

tumors in either rats or mice. A non-carcinogen is defined as a chemical that did not cause cancers in either rats or mice.

In CPDB, for chemicals judged to be carcinogenic, a potency value (*i.e*. TD_{50}) was estimated. The TD_{50} is the dose in the two-year cancer bioassay that is estimated to result in 50% of the animals remaining tumor-free at the end of the standard lifespan. The TD_{50} value accounts for the occurrence of spontaneous cancers [23,24].

For SAR modeling we chose cut-off values of 8 and 28 mmol/kg/day between carcinogens and marginal carcinogens and between marginal carcinogens and non-carcinogens, respectively. This resulted in the relationship

$$\text{SAR Activity} = (18.328 \log 1/TD_{50}) + 46.55 \qquad \text{Eq. 1}$$

between TD_{50} values and arbitrary SAR units. SAR models associated with this database have been described previously [25-28]. The composition of the various datasets used herein are described elsewhere [29].

Selection of SAR concensus models

The joint predictivity of rodent carcinogenicity of combinations of SAR models was determined by application of Bayes' theorem [30]. Briefly, it relates the sensitivity (α^+) and specificity (α^-) of an SAR model to the probability (θ^+_1) that the test chemical is a carcinogen based upon the SAR predictions.

For a single SAR prediction P_1, θ^+_1 is defined as:

$$\theta^+_1 = \frac{\theta^+_0 \alpha^+_1}{\theta^+_0 \alpha^+_1 + (1 - \theta^+_0)(1 - \alpha^-_1)} \qquad \text{Eq. 2}$$

if the prediction is positive, and

$$\theta^+_1 = \frac{\theta^+_0 (1 - \alpha^+_1)}{\theta^+_0 (1 - \alpha^+_1) + (1 - \theta^+_0) \alpha^-_1} \qquad \text{Eq. 3}$$

if the prediction is negative.

Iteration of these equations can be used to calculate the results of batteries of SAR models. θ^{+}_{0} is the prior probability and in the present situation it was assigned a value of 0.50 because that is the prevalence of carcinogens in the data base (Table 1). It has been shown previously that the actual assignment of the value of θ^{+}_{0} does not affect qualitatively the θ^{+}_{1} value [30] as essentially the Bayesian approach calculates the increase in knowledge gained by the testing result over that which was available initially [31,32]. Thus, assuming that θ^{+}_{0} is 0.50, finding that θ^{+}_{1} is 0.80 after testing represents a significant gain in knowledge, while going from 0.50 to 0.52 would not represent an acceptable increase in knowledge.

The values for α^{+} and α^{-} of the various SAR models are listed in Table 1. It must be stressed that these values were obtained with chemicals external to the SAR model.

Results And Discussion

As mentioned earlier, one of the most difficult SAR models to develop is the one describing carcinogenicity in rodents [14,33]**.** This most probably reflects the complexity of the phenomenon and the multiplicity of mechanisms that can lead to it (*e.g.* genotoxic *vs*. non-genotoxic; initiation *vs*. promotion, promotion *vs*. progression). Indeed it has been shown previously that the predictive performance of the MULTICASE SAR program is a function of the complexity of the biological phenomenon being modeled [11]. Still, when attention is focused on the nature of the data and the mechanistic significance of the experimental results, a highly predictive MULTICASE SAR model of rodent carcinogenicity could be developed [15]. Given that initial success, we investigated herein additional aspects of the SAR model building process for rodent carcinogenicity and the possibility that intermediary "human" input might improve the final model as well as provide a deeper understanding of the model building process itself.

The "parental" non-congeneric carcinogenicity data base derived from Gold *et al*. [22] consisted of 437 chemicals approximately equally divided between carcinogens and non-carcinogens [29] (Table 1, Model 1). Previous studies had shown that a ratio of unity was optimal for predictive performance [11,34-36]. This observation was consistent with the assumption of a binary distribution implicit in the SAR paradigm [37]. Because, in most instances SAR models are based upon published experimental results that were obtained for reasons other than SAR studies, the available data rarely have a ratio of unity. Accordingly we have studied the consequences of deviation from ratio of unity. In fact, we have shown that the resulting models, within limits, tolerated such deviations [35,36,38].

One of the major toxicophores identified by MULTICASE was the aromatic amine moiety NH_2-c=cH (Table 2, Toxicophore 1), *i.e.* an aromatic amine that contains one unsubstituted *ortho*-position (Figure 1A). There are 65 chemicals in the "learning set" that contain that toxicophore. Of these, 48 are rodent carcinogens, 2 are marginal carcinogens and 15 are non-carcinogens. The fact that 23% of the aromatic amines are non-carcinogens reflects the situation that not all aminoarenes are carcinogens [39]. Still, the presence of that toxicophore is associated with a 75% probability of carcinogenicity and a basal carcinogenic potency of 50.3 SAR units which corresponds to a TD_{50} value of 0.62 mmol/kg/day, (see Equation 1) (Figure 1). The modulators associated with specific toxicophores determine whether the probability of carcinogenicity will be expressed and, if so, to what extent, *i.e*. the potency (Table 3).

It is to be noted that while the overall learning set has a ratio of actives/inactives near unity (*i.e.*1.1) (Table 1, Model 1), that ratio is not maintained among the chemicals that contribute to the NH2-c=cH toxicophore, *i.e*. the ratio for this subgroup is 3.3 (*i.e*. Table 2, Toxicophore 1).

Moreover, with respect to this rodent carcinogenicity SAR model, it is to be noted that some of the toxicophores identified in an aromatic amine may be derived from non-arylamine-

containing carcinogens in the non-congeneric learning set. Thus 2-toluidine, in addition to containing toxicophore NH_2-c=cH- (Figure 2, Toxicophore A), also contains another one (CH_3-c=cH-cH=cH) (Table 2, Toxicophore 18; Figure 2, Toxicophore C) derived from a different set of carcinogens. The question now is whether this second toxicophore is a confounder unrelated to the carcinogenicity of 2-toluidine or whether it is a contributor and thereby increases the probability of carcinogenicity. *A priori*, it is of interest to note that this second toxicophore (Fig. 2C) is in fact closely related to the NH_2-c=cH- associated modulator No. 10 (Table 3 and Figure 2B) that augments the carcinogenic potency by 12.1 SAR units. This is close to the contribution of this additional toxicophore 18 (*i.e*. 14 SAR units (Figure 2C)).

This finding would argue for the fact that the ability of MULTICASE to develop SAR models using non-congeneric data bases expands the information pool. In the case of 2-toluidine that might actually provide a clue related to the broader carcinogenic spectrum of the former compared to the more restricted one of 2, 4, 5-trimethylaniline [40]. Similarly, in predicting the lack of carcinogenicity of methyl anthranilate (Figure 3), using the "parental" non-congeneric data base (Database 1, Table 1), in addition to recognizing the NH_2-c=cH-toxicophore (Figure 3A), MULTICASE identifies a deactivating moiety (Figure 3B) derived from non-aminoarenes non-carcinogens that negates the potential for carcinogenicity.

On the other hand, unexpectedly, the SAR model based upon the non-congeneric data base (Database 1, Table 1) predicts *p*-aminobenzoic acid (pABA)(Figure 4), a widely-used sunburn preventative and, in fact, a physiological chemical expected to be non-carcinogenic, to be a carcinogen. Based upon the chemicals in the learning set, that model does not identify within pABA an inactivating modulator which would abolish the potential associated with the NH_2-c=cH-toxicophore (Table 3). Although when wishing to be risk averse, a false positive prediction (as for pABA) is more acceptable than a false negative one [41], still such inaccurate predictions are disconcerting and cast doubt on the validity of the derived SAR model.

The overall predictive performance of this model for all carcinogens and non-carcinogens external to the model, albeit unremarkable, is, however, statistically highly significant, (*i.e.*, sensitivity: 0.67; specificity: 0.65; concordance between predicted and experimental results: 0.66; $\chi^2 = 44.8$). On the other hand, the predictivity (0.85) of the model for arylamines external to the model is considered very good (Table 1, Models 1 and 1A).

It is to be noted that it has been previously demonstrated that SAR models of this database can be greatly improved by specific calibration of the input activity [15] or by combination [28] of rodent carcinogenicity models derived from different datasets, *e.g.* CPDB [22] and the U.S. National Toxicology Program [40,42].

These findings regarding differing predictivities derived from subsets present in a non-congeneric database used to derive an SAR model of a complex biological phenomenon (*e.g.* carcinogenesis) indicates the need to use the resulting models cautiously.

In view of the unexpected prediction of the carcinogenicity of pABA and the dichotomy in the overall predictivity of the model towards non-arylamines and arylamines, we investigated whether the predictivity of the SAR model could be enhanced by isolating the chemicals sharing previously identified toxicophores and performing an SAR analysis on this subset. It is realized, of course, that when this is done, this will segregate the SAR model from the predictive contribution of toxicophores and their modulators that are derived from non-aminoarene-containing chemicals in the data set (*e.g.* see Figure 2, Toxicophore C, or Figure 3 deactivating moiety B). Moreover, in deriving these new models, it must be borne in mind that the toxicophores selected by MULTICASE are not necessarily the simplest ones (see Table 2). Thus, the aminoarene toxicophore selected from the non-congeneric database (Database 1, Table 1) was NH_2-c-cH=, *i.e.* an *ortho* unsubstituted moiety rather than the simpler and more common NH_2-c= moiety (Table 2). The choice of the *ortho*-substituted toxicophore resulted in an enrichment with respect to arylamine carcinogens [39].

In order to examine the effect of sequestering the subset of chemicals sharing a common toxicophore, we adopted two strategies: (a) we performed a MULTICASE SAR analysis on the 65 chemicals containing the NH_2-c=cH- toxicophore (Table 1, Model 2) and (b) we also supplemented these 65 chemicals with 33 randomly selected non-carcinogens such as to achieve the preferred ratio of carcinogens/non-carcinogens of unity (Database 3, Table 1). The chemicals constituting databases Nos. 2 and 3 are listed elsewhere [29]. In following the first strategy, *i.e.* Database 2, it has to be realized that since all of the chemicals in the learning set, inactives as well as actives, contain the NH_2-c=cH arylamine toxicophore, that it, or its parent NH_2-c=, will not necessarily be either identified or be the major toxicophore.

As a matter of fact, MULTICASE SAR analysis of Database No. 2 identified a number of new toxicophores (Table 4). While intuitively we would expect that the toxicophore of that SAR model would resemble the modulators associated (Table 3) with the parental toxicophore (NH_2-c=cH-) of the non-congeneric data base (Database 1, Table 1), in fact this was not uniformly so. Only one of the major toxicophores (Table 4, No. 1) of this subset resembled previously identified modulators (Table 3, Nos. 10-14, 16-19). Only one of the major toxicophores for this enriched subset included the aromatic amino moiety (Table 4, No. 3) but in contrast to the toxicophore of the model derived from the non-congeneric database (*i.e*. NH_2-c=cH, an *ortho*-unsubstituted compound [see Figures 1-2]), it identified an *ortho*-substituted aromatic amine (See Figure 5A); obviously this toxicophore is derived from the portion of the molecule containing the other *ortho*-position. The predictivity (0.91) of this model for arylamines external to the model is very good (Table 1).

In order to evaluate the performance of this SAR model based upon the NH_2-c=cH-containing subset of the non-congeneric database (Database No. 1, Table 1), we compared predictions for the same chemicals. Thus, based upon this enriched model (Database 2), 2,4,5-trimethylaniline is again predicted to be a rodent carcinogen (Figure 6). However, the

probability of carcinogenicity, based upon Toxicophore 4 (Table 4), increased to 82% (compared to 75%, Figure 1). This is due to the enrichment in the subset of carcinogens that contain that toxicophore. The predicted potency has also been increased significantly to 69 SAR units corresponding to a TD_{50} value of 0.06 mmol/kg/day (Database 2, Table 1). In fact, this is closer to the reported TD_{50} value (0.05 mmol/kg/day) [22] than the projected TD_{50} value (0.62 mg/mmol/kg/day) found with the model based upon Database 1 (Table 5).

The prediction for 2-toluidine based upon the congeneric model (Database 2, Table 1) (Figure 5) when compared to the corresponding one based upon the non-congeneric (Database 1, Table 1) data base (Figure 2) is also instructive. It is to be noted that the new prediction, which is based upon Toxicophore 3 (Table 4), is also associated with a greater probability (82%) of carcinogenicity than the 75% associated with the toxicophore of Database 1. This increase in probability (Figure 2) was achieved even without the contribution of the second toxicophore derived from non-arylamine carcinogens that was seen with the non-congeneric model (Figure 2). Moreover, the predicted potency (0.24 mmol/kg/day) is identical (Table 5) to the experimentally determined one [22].

While the prediction of the lack of carcinogenicity of methyl anthranilate derived from the parental non-congeneric SAR (Database 1, Table 1) model was based upon the presence of deactivating moieties external to the arylamines (Figure 3), that chemical was also predicted to be non-carcinogenic (Figure 7) by the congeneric data base (Database 2, Table 2). However, in this instance the prediction was based entirely upon the presence of "internal" inactivating modulators associated with the modified arylamine toxicophore 3 (Table 4).

Finally, the SAR model based upon the "congeneric" database (Database 2) could not identify any relevant toxicophores in pABA and hence, by default, this chemical is predicted to be non-carcinogenic (data not shown). This is in contrast to the prediction of carcinogenicity based upon the non-congeneric SAR model (Database 1, Figure 4).

A comparison of the results obtained with the two SAR models (Databases 1 and 2) indicates (Table 5) that the "congeneric" database (Database 2), even without the contribution of toxicophores derived from carcinogens not containing the shared toxicophore that defines the class (*i.e*. NH_2-c=cH), has a performance that appears superior to the model based upon the total parental (non-congeneric) learning set (Database 1). The congeneric model (Database 2) also appears to be more informative as it contains a series of relevant toxicophores (Table 5) and associated modulators *vs*. a single arylamine toxicophore for the non-congeneric set. Moreover, the model based on Database 2 projects potencies that are nearly identical to experimentally determined values (Table 5).

A comparison of the predictions of the two models (Database Nos. 1 and 2) for NH_2-c=cH- containing chemicals indicates a concordance of 82.6% (sensitivity: 79.4%; specificity: 92.0%) for arylamines external to the models. The primary difference derives from the fact that the model derived from the congeneric subset predicts fewer chemicals to be carcinogenic, *e.g*. as for pABA (Table 5).

Previous studies had suggested [11,34-36] that a ratio of actives/inactives of unity in the learning set was optimal for the predictive performance of the resulting SAR model. Database 2 (Table 1) is deficient in that respect. Accordingly, we supplemented the above "congeneric" learning set (Database 2) with non-arylamine non-carcinogens to achieve a ratio of unity (Database 3, Table 1). This resulted in the generation of a single toxicophore that represents a generic arylamine (NH_2-c=), present in 48 carcinogens, 2 marginal carcinogens and 15 non-carcinogens, see Figures 8-10, Toxicophore A). That toxicophore is associated with a 74% probability of carcinogenicity and a basal potency of 44.0 SAR units or a TD_{50} value of 1.38 mmol/kg/day. A series of associated modulators (Table 6) was also identified. Interestingly, however, the log P (P=octanol/water partition coefficient) is a significant activating modulator (Figures 8-10) *i.e*. the greater the lipophilicity of the molecule, the greater the likely carcinogenic

potency. This may reflect the possibility that the effective dose of the putative carcinogen is increased due to accumulation in fatty tissues (see also [16]). The generation of a single toxicophore is in contrast to the multiple toxicophores identified by the SAR model based only on arylamine (Database 2, Table 1) albeit in that data set the ratio of actives/inactives is skewed towards actives (*i.e.* 3.3).

Three replicates of database 3 were generated by the independent triplicate random selection of 31 non-arylamine non-carcinogens. As expected [36], the predictive performance of the models were similar and the individual predictions were identical. Accordingly, the results obtained with only one of the subsets are illustrated herein.

Because this model (*i.e*. Database 3, Table 1) contains a single toxicophore and since the ratio of active/inactive arylamines is 1.0, the probability of activity will not exceed 75%. In that respect, that SAR model is similar to the parental non-congeneric SAR model (Database 1) but different from the above "congeneric" model (Database 2). Thus, the probability of carcinogenicity of 2,4,5-trimethylaniline using the l:l congeneric SAR model (Database 3) is 75% and the major activating modulator is log P (*i.e*. 6.6* log P, Table 6). The log P for this chemical is 2.17. Accordingly, its total contribution is 14.25 SAR units for a total potency of 58.2 SAR units (which corresponds to a TD_{50} of 0.23 mmol/kg/day) which is significantly less potent than the experimentally determined value (Table 5).

Similarly, for 2-toluidine (data not shown) the probability of carcinogenicity is also 75%, the log P is 1.89, and the log P contribution is 9.10 SAR units for a total potency of 53.1 SAR units, *i.e*. a TD_{50} value of 0.44 mmol/kg/day. That potency is also less than the experimentally determined value (Table 5). This model predicted methyl anthranilate (Figure 9) to be devoid of carcinogenicity. That projection is based upon the presence of a series of inactivating modulators (Figure 9). *p*-Aminobenzoic acid is predicted to possess marginal carcinogenicity (Figure 10). That prediction (Figure 10) is dependent upon a deactivating moiety (Figure 10B)

that is external to the arylamines, *i.e.* it is derived from non-arylamine non-carcinogenic molecules and it does not abolish entirely the carcinogenicity potential inherent in the toxicophore. Overall, given these observations, this 1:1 SAR model (Database 3) is inferior (see Table 5) to the congeneric SAR model described above (Database 2). However, the predictivity (0.81) of this model is still very respectable (χ^2 =39.4) (Table 1).

A comparison of the performance of this model (Database 3) with the non-congeneric one (Database 1) indicates a concordance of 89.1%, a sensitivity of 87.9% and a specificity of 92.3% for arylamines external to the models. Similarly, the comparison between the congeneric SAR model (Database 2) and the 1:1 supplemented congeneric one (Database 3) indicates a concordance of 87.0%, a sensitivity of 96.4% and a specificity of 72.2%.

Based upon these considerations, as well as the internal coherence of the toxicophores, it would appear that in this instance the non-supplemented congeneric data base (Database 2) generates the most informative SAR model (Table 5), albeit it is unbalanced with respect to the overall actives/inactives ratio. The model derived from the parental non-congeneric data base (Database 1), is less informative and in the instance of the arylamines predicts lower potencies than determined experimentally (Table 5). However, it may be useful in identifying moieties, external to the arylamine-associated toxicophore, that affect the mechanism or spectrum of the carcinogenic response. As indicated below, a model combining the features of Database Nos. 1 and 2 might be the most informative.

CONCLUSIONS

The present analyses have shed light on a number of the performance characteristics and variables associated with the MULTICASE SAR system (*e.g.* informational content, ratio of active to inactives, probability of activity, and variables influencing potency, *i.e.* modulators). Overall, these analyses have indicated the robustness of the MULTICASE SAR system.

Interestingly, the analyses indicate that apparently no uniform predictive advantage is achieved by intervention in the SAR model building process such as segregating structurally-restricted toxicophore-centered subsets for SAR analysis. This is partially due to the fact that when using non-congeneric data bases, the analyses use information from structural classes that appear to be unrelated to the primary class of the agent under consideration but nevertheless contribute to the model's performance. On the other hand, in some instances (*e.g.* Model 2), the model derived from a subset is more predictive especially with respect to projected potencies as well as more informative. In order to capture these nuances and improve predictivity it would seem appropriate for in depth analysis to include both types of SAR models, *i.e.* congeneric (subset-based) as well as non-congeneric. This recommendation is based upon the realization that the toxicophores derived from the toxicophore-based subsets appear to have greater statistical significance than the modulators associated with the parental (non-congeneric) databases. Such an approach might allow a refinement in the understanding of the structural basis of activity as well as improve predictivity. Moreover, methods to combine the outputs of multiple models into a single prediction are available [28,30]. Based upon this rationale and with knowledge of the predictivities of the individual SAR models generated herein (Table 1), we analyzed a combination of models using a bayesian weight-of-evidence approach [30]. As part of these analyses, we made alternate sets of evaluation for the non-congeneric model (Database 1) based upon (a) its overall predictivity for non-arylamines (sensitivity 0.67; specificity 0.65) designated "Model 1", and (b) its predictivity for arylamines (Table 1), designated "Model lA". The predictivities of the combinations of models are summarized in Table 7.

Combining either Model 1 or lA with Model 2, in all instances results in predictions that agree with experimental findings. This is noteworthy as this concordance occurs even with Model 1. The latter has a significantly lower predictivity than Model 1A. Additionally, even when the predictions based upon Model 1 or 1A are divergent from those based on Model 2 (e.g.

Analyses 13, 14, 19, 20, 25 and 26), the performance of Model 2 overcomes the contributions of Models 1 or 1A, and the predictions agree with the experimental findings. The same occurs with respect to Model 1 or 1A and Model 3, i.e. the latter outweighs the former when they diverge (Analyses 15, 16, 27 and 28). It is noteworthy that the projections of Models 2 and 3 are also dominant when they project carcinogenicity and Model 1 (or 1A) projects non-carcinogenicity (Analyses 25-28).

The results obtained when the combination contains Model 1 (or 1A) and Model 3 are not as reliable (*e.g.* pABA is predicted to be a carcinogen (Analyses 21 and 22) and divergent results can be obtained depending on whether Model 1 or 1A is used (Analyses 26 and 27)). This evaluation of the lack of reliability of Model 3 is reinforced by the findings obtained when the 3 models are combined. When Models 2 and 3 diverge (Analyses 23 and 24), the overall projections become ambiguous and unreliable.

In view of these analyses, we would opt for a consensus based upon Models 1 and 2. Obviously, the same type of analyses would be needed whenever new or updated related SAR models are developed.

Finally, the present findings indicate that the performance of an SAR model is a function of the nature of the database as well as of the model-building paradigm. Interactions between these have been discovered in this study. These appear to be database- and model- specific and, hence, no generalization regarding performance trends can, *a priori*, be made. This indicates that in order to gain a better understanding of the applicability of specific SAR models as well as improve predictivity that the type of analyses demonstrated herein be carried out.

The molecule contains the Toxicophore (nr.occ.= 1):

```
        NH2  -c
(A)           \\
              cH
```

*** 48 out of the known 65 molecules (74%) containing such a toxicophore are rodent carcinogens (conf.level=100%)

*** QSAR Contribution : Constant is 50.28

** Total projected QSAR activity 50.28

*** The probability that this molecule is a Rodent Carcinogen is 75.0% **

** The projected potency is 50.3 SAR units **

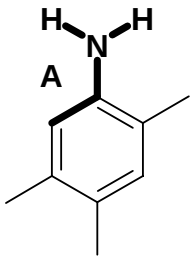


Figure 1: Prediction of the carcinogenicity in rodents of 2,4,5-trimethylaniline. A potency of 50.3 SAR units corresponds to a TD_{50} value of 0.62 mmol/kg/day. This model is based upon Database 1 (Table 1).

The molecule contains the Toxicophore (nr.occ.= 1):

NH2 -c
(A) \\
cH

*** 48 out of the known 65 molecules (74%) containing such a toxicophore are rodent carcinogens (conf.level=100%)

*** QSAR Contribution : Constant is 50.28
** The following Modulator is also present:
(B) cH =cH -c =c <- <3-CH3> Activating 12.09

** Total projected QSAR activity 62.37

The molecule also contains the Toxicophore:
(C) CH3-c =cH -cH =cH -
*** 6 out of the known 7 molecules (86%) containing such a toxicophore are rodent carcinogens
*** QSAR Contribution : Constant is 14.04
** The following Modulators are also present:

(D) CH3-c =cH -cH =cH -cH =c <- Activating 39.44

Log partition coeff.= 1.39 ; LogP**2 contribution is 4.51

** Total projected QSAR activity 57.99

*** The probability that this molecule is a Rodent Carcinogen is increased to 84% due to the presence of the extra Toxicophore

** The projected potency is 62.4 SAR units **

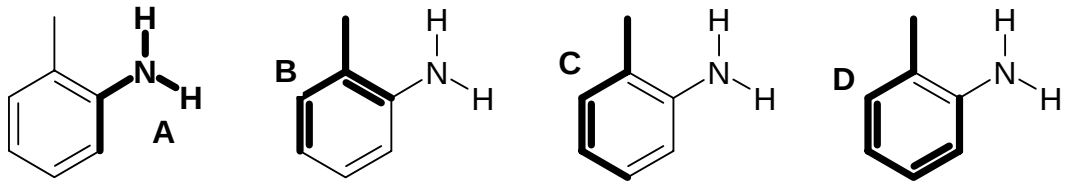


Figure 2: Prediction of the carcinogenicity of 2-toluidine. Toxicophore C was derived from non-arylamine carcinogens. A potency of 62.4 SAR units corresponds to a TD_{50} value of 0.14 mmol/kg/day. This model is based upon Database 1 (Table 1).

The molecule contains the Toxicophore (nr.occ.= 1):

(A) NH2 -c
\\
cH

*** 48 out of the known 65 molecules (74%) containing such a toxicophore are rodent carcinogens (conf.level=100%)

*** QSAR Contribution : Constant is 50.28

** Total projected QSAR activity 50.28

** The molecule contains the following DEACTIVATING Fragment:

(B) CO -c cH
\\ //
c < -cH

*** The probability that this molecule is a Rodent Carcinogen is 62.5% **

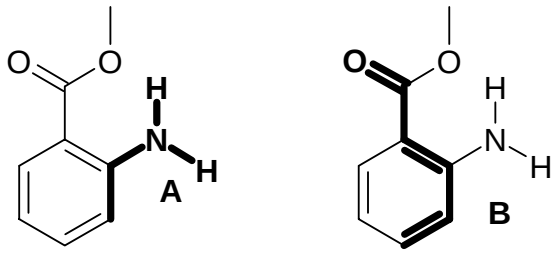


Figure 3: Prediction of the non-carcinogenicity of methyl anthranilate. The deactivating fragment is derived from non-arylamine non-carcinogens in the dataset. This model is based upon Database 1 (Table 1).

The molecule contains the Toxicophore (nr.occ.= 2):

NH2 -c
(A) \\
cH

*** 48 out of the known 65 molecules (74%) containing such a toxicophore are rodent carcinogens (conf.level=100%)

*** QSAR Contribution : Constant is 50.28

** Total projected QSAR activity 50.28

*** The probability that this molecule is a Rodent Carcinogen is 75.0% **

** The projected potency is 50.3 SAR units **

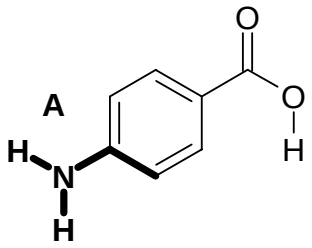


Figure 4: Prediction of the carcinogenicity in rodents of *p*-aminobenzoic acid. This prediction is based upon Database 1 (Table 1).

The molecule contains the Toxicophore (nr.occ.= 1):

(A) NH2 -c cH
\\ //
c -cH

*** 17 out of the known 21 molecules (81%) containing such a Toxicophore are rodent carcinogens (conf.level= 98%)

*** QSAR Contribution : Constant is 36.96
** The following Modulators are also present:

(B) NH2-c =c -cH =cH -cH = <3-CH3> Activating 14.48

Log partition coeff.= 1.39 ; LogP contribution is 6.56

** Total projected QSAR activity 58.00

*** The probability that this molecule is a Rodent Carcinogen is 81.8% **

** The projected potency is 58.0 SAR units **

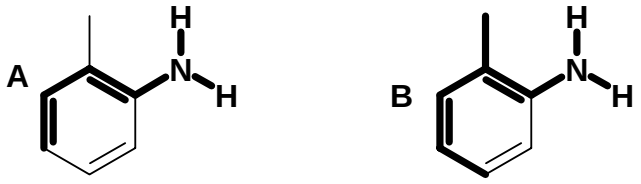


Figure 5: Prediction of the carcinogenicity in rodents of 2-toluidine. A potency of 58 SAR units corresponds to a TD_{50} value of 0.24 mmol/kg/day. This model is based upon Database 2 (Table 1).

The molecule contains the Toxicophore (nr.occ.= 3):

```
        CH3 -c
(D)          \\
             cH  -c"
```

*** 8 out of the known 10 molecules (80%) containing such a Toxicophore are rodent carcinogens

*** QSAR Contribution : Constant is 44.42

** The following Modulators are also present:

(E) CH3-c =cH -c =c <- Activating 12.93

Log partition coeff.= 2.17 ; LogP contribution is 11.70

** Total projected QSAR activity 69.06

*** The probability that this molecule is a Rodent Carcinogen is 81.8% **

** The projected potency is 69.1 SAR units **

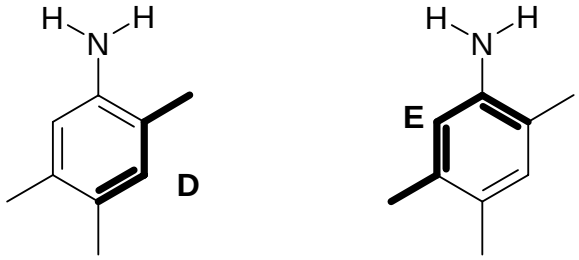


Figure 6: Prediction of the carcinogenicity in rodents of 2,4,5-trimethylaniline. The model is based upon Database 2 (Table 1). A potency of 69.1 SAR units corresponds to a TD_{50} value of 0.06 mmol/kg/day.

The molecule contains the Toxicophore (nr.occ.= 1):

(A) NH2 -c cH
\\ //
c -cH

*** 17 out of the known 21 molecules (81%) containing such a Toxicophore are rodent carcinogens (conf.level= 98%)

*** QSAR Contribution :			Constant is	36.96
** The following Modulators are also present:				
	CO -c =c <-		Inactivating	-1.14
	cH =cH -c >=c <-		Inactivating	-1.14
	CO -c =c -cH =		Inactivating	-1.14
	CO c =c -NH2		Inactivating	-1.14
	CO -c =c <-cH =		Inactivating	-1.14
	CO -c =cH -cH =		Inactivating	-1.14
	cH =c >-c =cH -	<3-CO >	Inactivating	-1.14
(B)	cH =cH -c =c <-	<3-CO >	Inactivating	-1.14
	NH2-c =c -cH =	<3-CO >	Inactivating	-1.14
	CO -c =c -cH =	<3-NH2>	Inactivating	-1.14
	cH =cH -c >=c <-cH =		Inactivating	-1.14
	NH2-c =c >-cH =cH -		Inactivating	-1.14
	cH =cH -c =c <-cH =	<3-CO >	Inactivating	-1.14
	NH2-c =c -cH =cH -	<3-CO >	Inactivating	-14.76
	Log partition coeff.= 0.91 ;		LogP contribution is	4.32
** Total projected QSAR activity				11.75

*** The probability that this molecule is a Rodent Carcinogen is 0.0% **

** The projected potency is 11.8 SAR units **

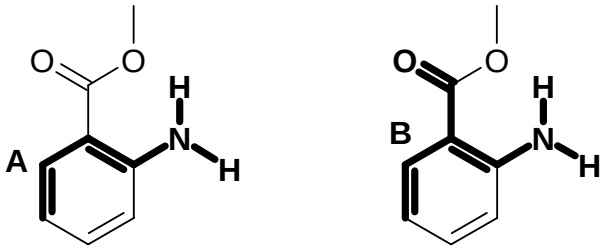


Figure 7: Prediction of the lack of carcinogenicity in rodents of methyl anthranilate. The model is based upon Database 2 (Table 1). A predicted potency of 11.8 SAR units corresponds to a TD_{50} value of 79.2 mmol/kg/day which is considered to indicate lack of carcinogenicity.

The molecule contains the Toxicophore (nr.occ.= 1):

(A) NH2 -c"

*** 48 out of the known 65 molecules (74%) containing such a toxicophore
are rodent carcinogens (conf.level=100%)

*** QSAR Contribution : Constant is 43.98
** The following Modulator is also present:
Log partition coeff.= 2.17 ; LogP contribution is 14.25

** Total projected QSAR activity 58.23

*** The probability that this molecule is a rodent carcinogen is 75.0% **

** The projected potency is 58.2 SAR units **

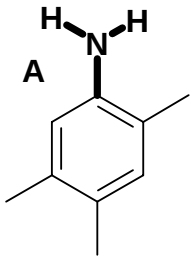


Figure 8: Prediction of the carcinogenicity in rodents of 2,4,5-trimethylaniline.

The activity of the toxicophore is modulated by 6.57 * the logP of the chemical (see Table 6). A potency of 58.2 SAR units corresponds to a TD_{50} value of 0.23 mmol/kg/day. This model is based upon Database 3 (Table 1).

The molecule contains the Toxicophore (nr.occ.= 1):

(A) NH2 -c"

*** 48 out of the known 65 molecules (74%) containing such a Toxicophore
are rodent carcinogens (conf.level=100%)

*** QSAR Contribution :	Constant is	43.98
** The following Modulators are also present:		
(B) cH =cH -c >=c <-	Inactivating	-2.49
CO -c =c -NH2	Inactivating	-2.49
cH =cH -c =c <-	<3-CO > Inactivating	-2.49
(C) NH2-c =c -cH =	<3-CO > Inactivating	-19.93
CO -c =c -cH =	<3-NH2> Inactivating	-2.49
cH =cH -c >=c <-cH =	Inactivating	-2.49
NH2-c =c >-cH =cH -	Inactivating	-2.49
cH =cH -c =c <-cH =	<3-CO > Inactivating	-2.49
NH2-c =c -cH =cH -	<3-CO > Inactivating	-2.49
Log partition coeff.= 0.91 ;	LogP contribution is	5.99
** Total projected QSAR activity		10.10

*** The probability that this molecule is a rodent carcinogen is 0.0% **

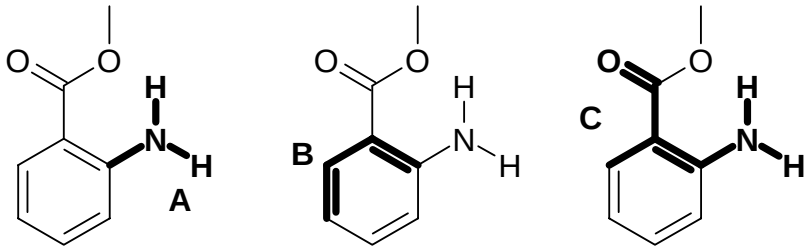


Figure 9: Prediction of the non-carcinogenicity in rodents of methyl anthranilate. This model is based upon Database No. 3 (Table 1).

The molecule contains the Toxicophore (nr.occ.= 1):

(A) NH2 -c"

*** 48 out of the known 65 molecules (74%) containing such a Toxicophore
are rodent carcinogens (conf.level=100%)

*** QSAR Contribution : Constant is 43.98
** The following Modulator is also present:
Log partition coeff.= 0.34 ; LogP contribution is 2.24

** The projected QSAR activity is 46.22

** The molecule also contains the following DEACTIVATING Fragment:

(B) CO -OH
present in 8 molecules, all inactive
The molecule may therefore be inactive

*** The probability that this molecule is a Rodent Carcinogen is 62.5% **

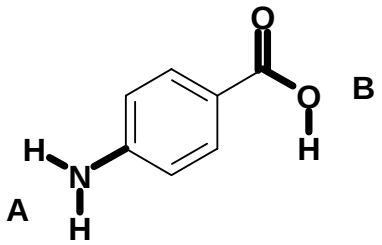


Figure 10: Prediction of the marginal carcinogenicity of *p*-aminobenzoic acid. The deactivating moiety B is derived from non-arylamine non-carcinogens. The model is derived from Database 3 (Table 1).

Table 1: Characteristic of the rodent carcinogenicity databases used for SAR modeling

No.	Description*	N	Actives/ Inactives	Concordance	Sensitivity	Specificity	χ^2
1	Non-congeneric	437	1.1	0.85	0.88	0.73	22.5
1A		437		0.66	0.67	0.65	44.8
2	Congeneric:	65	3.3	0.91	0.92	0.87	37.0
3	Congeneric:	97	1.0	0.81	0.86	0.77	39.4

The construction and composition of each dataset are described in the text and in reference [29], respectively.
Concordance is defined as the ratio of correctly predicted activity/known experimental results for chemicals external to the model.
A χ^2-value of 6.63 corresponds to a 99% confidence level [43]. The values for the present models are significantly greater.
To determine the predictivity of Model 1, it was challenged with either arylamines (Model 1) or by the entire dataset chemicals (Model 1A) external to the model.

Table 2: Some of the major toxicophores associated with rodent carcinogenicity: Database 1

1---2---3---4---5---6---7---8---9---10		Nr. of Fragments	Inactives	Marginals	Actives	Toxicophore No.
NH2-c =cH -		65	15	2	48	1
NH -C =N -		9	1	0	8	2
[Cl -] <-- 4.0A --> [Cl -]		21	2	0	19	3
CH2-N -CH2-		29	7	0	22	4
O -CH =		7	0	0	7	5
N -C =		5	0	0	5	6
O -C =		14	1	0	13	7
O^ -CH2-		6	0	0	6	8
Br -CH2-		5	0	0	5	9
cH =cH -c =cH -cH =	<3-Cl>	14	3	0	11	10
PO –O		11	1	0	10	11
CH3-N -c =cH -	<2-CH3>	6	1	0	5	12
cH =c -cH =cH -c <=	<2-NH>	6	1	1	4	13
Cl -CH2-		26	4	1	21	14
c.”-CO -c. =		7	0	0	7	15
NO2-C =CH -		14	0	0	14	16
cH =cH -c =cH -cH =	<3-C=>	11	1	0	10	17
CH3-c =cH -cH =cH -		7	0	1	6	18

Toxicophore No. 1 is shown in Figures 1-4; No. 18 in Figure 2C.

“c” and “C” refer to aromatic and acyclic atoms, respectively; c. indicates a carbon atom shared by two rings; 0^ indicates an epoxide; c” indicates a carbon atom connected by a double bond to another atom. <3-Cl> indicates a chlorine atom substituted on the 3rd non-hydrogen atom from the left. ←4.0A→ indicates a 2-D 4 Angstrom distance descriptor.

In toxicophore No. 18, the second carbon from the left is shown as unsubstituted. This means that it can be substituted with any atom except a hydrogen. On the other hand, for this toxicophore, the last carbon on the right is shown with an attached hydrogen. This means it cannot be substituted by any other atom but hydrogen. Finally, in toxicophore No. 10 the third non-hydrogen atom from the left is shown as unsubstituted. It can only be substituted by a chlorine atom.

Table 3: Modulators associated with the carcinogenic toxicophore NH2-c =cH - : Database 1

Fragment		Constant= 50.3 QSAR	Modulator No.
1---2---3---4---5---6---7---8---9---10			
2D [N -] <-- 2.6A --> [N =]		29.1	1
CO -NH2		28.6	2
N =CH -C =		-18.9	3
NH -C =CH -		15.4	4
n =c -cH =	<2-NH2>	19.0	5
c =cH -c =c -		-23.8	6
cH =c. -N =C -		-32.3	7
OH -CO -c =c <-		-20.1	8
Cl -c =cH -c <=		-23.2	9
cH =cH -c =c <-	<3-CH3>	12.1	10
cH =c -cH =cH -c <=		17.7	11
NH2-c =cH -cH =c -NH -		-20.1	12
NH2-c =cH -cH =c -NH2		-25.5	13
NH2-c =cH -cH =c -CH =		25.1	14
NH2-c =cH -cH =cH -c =		-35.3	15
NH -c =cH -cH =c -cH =	<5-NH2>	-20.1	16
NH2-c =cH -cH =c -SO2-NH -		-20.1	17
NH2-c =cH -cH =c <-cH =c <-		-20.1	18
NH2-c =cH -cH =c >-cH =c -NH2		-20.1	19

For an explanation of the fragments, see legend to Table 2. "c." indicates a carbon atom shared by two rings.
Modulator No. 10 is shown in Figure 2B.

Table 4: Toxicophores associated with carcinogenic arylamine: Database 2

Fragment 1---2---3---4---5---6---7---8---9---10	Nr. of Fragments	Inactives	Marginals	Actives	Toxicophore No.
c <=cH -cH =c -	20	1	0	19	1
c. =cH -cH =c -	5	0	0	5	2
NH2-c =c -cH =cH -	21	3	1	17	3
CH3-c =cH -c =	10	1	1	8	4

For an explanation of the fragments, see legend to Table 2. "c" indicates a carbon atom shared by two rings.
Toxicophore No. 3 is shown in Figures 5A and 7A and No. 4 in Figure 6D.

Table 5: Summary of predictions of the carcinogenicity in rodents of selected arylamines

Chemical	SAR Model Database	Probability %	Potency* (mmol/kg/day)
2,4,5-Trimethylaniline	1	75	0.62
	2	82	0.06
	3	75	0.23
	Exper.		0.05
2-Toluidine	1	84	0.14
	2	82	0.24
	3	75	0.44
	Exper.		0.24
Methyl anthranilate	1	62	NC
	2	0	80*NC
	3	0	98*NC
Anthranilic Acid	Exper.		NC
p-Aminobenzoic Acid	1	75	0.62
	2	0	NC
	3	62.5	Marginal

*Calculated from SAR units using Eq. 1. Potency >32 mmol/kg/day are considered non-carcinogens. NC indicates lack of carcinogenicity. "Exper." indicates experimentally determined TD_{50} values.

Table 6: Modulators associated with the carcinogenic toxicophore NH2-c = (Database 3)

Fragment 1---2---3---4---5---6---7---8---9---10	Constant= 44.0 QSAR	Modulator No.
cH =cH -c >=c <-	-2.5	1
NH2-c =c -c =	-17.7	2
NH2-c =c -c =	-10.7	3
NH2-c =n -c =	12.0	4
CO -c =c -NH2	-2.5	5
c <=cH -c =c <- <3-CH=>	0.5	6
cH =cH -c =c <- <3-CO >	-2.5	7
NH2-c =c -cH = <3-CO >	-19.9	8
CO -c =c -cH = <3-NH2>	-2.5	9
cH =cH -c >=c <-cH =	-2.5	10
CH3-c =c <-cH =c <-	0.9	11
NH2-c =c -c =cH -	-17.7	12
NH2-c =c >-cH =cH -	-2.5	13
NH2-c =c <-cH =c <-	-20.4	14
cH =c -c <=cH -c <= <2-CH3>	0.9	15
cH =cH -c =c <-cH = <3-CO >	-2.5	16
cH =cH -c =cH -c = <3-NH2>	-38.7	17
NH2-c =c -cH =c >- <3-NH2>	-20.0	18
NH2-c =c -cH =cH - <3-CO >	-2.5	19
cH =cH -c <=cH -c <=cH -	-1.6	20
cH =cH -cH =c <-cH =c <-	-1.6	21
CH3-c =c <-cH =c <-cH =	0.9	22
NH -c =cH -cH =c -cH =	-9.1	23
NH -c =cH -cH =c <-cH =	-9.1	24
NH2-c =cH -c <=c -CH3	16.2	25
NH2-c =cH -cH =c -NH -	-27.4	26
NH2-c =cH -cH =c -CH =	13.7	27

Table 6 continued

Table 6: Modulators associated with the carcinogenic toxicophore NH2-c = (Database 3)

Fragment	QSAR	Modulator No.
NH2-c =c -cH =c >-cH = <3-NH2>	-2.5	28
NH2-c =c -cH =cH -c >= <3-NH2>	-2.5	29
NH2c =c -cH =cH -c <= <3-CH3>	0.9	30
cH =cH -cH =c -cH =c <- <4-NH2>	-1.6	31
NO2-c =cH -c =c -NH2 <4-NH2>	-2.5	32
NH -c =cH -cH =c -cH = <5-NH2>	-9.1	33
CH3-c =c <-cH =c <-cH =cH -	0.9	34
NH2-c =cH -c <=cH -cH =cH -	-18.0	35
NH2-c =cH -cH =c -SO2-NH -	-18.9	36
NH2-c =cH -cH =c -C =CH -	15.2	37
CH3-O -c =cH _cH =c <-cH = <3-cH=>	-20.3	38
NH2-c =cH -c =cH -cH =cH - <4-NH2>	-1.6	39
NH2-c =cH -cH =c -cH =cH - <5-c =>	12.8	40
NH2-c =cH -cH =c -cH =cH - <5-NH2>	-36.1	41
NO2-c =cH -cH =c -c" -NH2 <5-NH2>	-2.5	42
NH2-c =cH -cH =cH -c =cH - <6-NH2>	-1.6	43
CH3-O -c =cH -cH =c -cH =cH -	-6.8	44
CH3-O -c =cH -cH =c <-cH =cH -	-6.8	45
NH2-c =cH -c <=cH -cH =c -CH3	0.9	46
NH2-c =cH -cH =c >-cH =c -NH2	-2.5	47
NH2-c =cH -c =cH -cH =c -CH3 <4-NH2>	0.9	48
NH2-c =cH -cH =c -cH =c -NH2 <5-NO2>	-2.5	49
CH3-O -c =cH -cH =c -cH =cH - <6-NH2>	-6.8	50
Log P	6.6	51

For an explanation of the fragments see legend to Table 2.

Modulators 1 and 8 are seen in Figures 9B and 9C, respectively.

Modulator 51 indicates that the estimated logP of the predicted chemical needs to be multiplied by 6.6; see Figure 8.

Table 7: Predictions of carcinogenicity based upon combinations of SAR models

Analysis	Chemical	Models	Predictions[a]	Probability[b]	Conclusion	Experimental
1	2,4,5-Trimethylaniline	1 & 2	+ +	0.931	CA	CA
2		1A&2	+ +	0.959	CA	
3		1 & 3	+ +	0.874	CA	
4		1A& 3	+ +	0.924	CA	
5		1,2&3	+ + +	0.981	CA	
6		1A,2&3	+ + +	0.989	CA	
7	2-Toluidine	1 & 2	+ +	0.931	CA	CA
8		1A&2	+ +	0.959	CA	
9		1 & 3	+ +	0.874	CA	
10		1A&3	+ +	0.924	CA	
11		1,2&3	+ + +	0.981	CA	
12		1A,2&3	+ + +	0.989	CA	
13	Methyl anthranilate	1 & 2	+ -	0.150	NC	(NC)
14		1A&2	+ -	0.231	NC	
15		1 & 3	+ -	0.258	NC	
16		1A&3	+ -	0.372	NC	
17		1,2&3	+ - -	0.032	NC	
18		1A,2&3	+ - -	0.052	NC	
19	p-Aminobenzoic Acid	1 & 2	+ -	0.150	NC	(NC)
20		1A&2	+ -	0.231	NC	
21		1 & 3	+ +	0.878	CA	
22		1A&3	+ +	0.924	CA	
23		1,2&3	+ - +	0.405	(NC)	
24		1A,2&3	+ - +	0.528	Marginal	
25	Example A	1 & 2	- +	0.783	CA	
26		1A&2	- +	0.883	CA	
27	Example B	1 & 3	- +	0.656	CA	
28		1A&3	- +	0.383	NC	
29	Example C	1 & 2	- -	0.045	NC	
30		1A&2	- -	0.013	NC	
31	Example D	1 & 3	- -	0.085	NC	
32		1A&3	- -	0.029	NC	

a) The predictions are shown in Table 5. A designation such as "+ +" (see Analysis No. 1) indicates positive predictions with SAR models 1 and 2.

b) Indicates the posterior probability based on a prior probability of 0.5 which reflects the ratio of unity of actives/inactives in model 1.

Abbreviations: CA and NC carcinogenicity and non-carcinogenicity, respectively. (NC) indicates putative non-carcinogenicity.

Acknowledgments

This research was supported by the Vira Heinz Endowment and the Department of Defense Breast Cancer Research Program under award number DAMD17-01-0376. Views and opinions of, and endorsements by the author(s) do not reflect those of the US Army or the Department of Defense.

References

[1] I.D. McKinney, A. Richard, C. Waller, M.C. Newman and F. Gerberick, The practice of structure-activity relationships (SAR) in toxicology, Toxicol. Sci. 56, (2000) 8-17.

[2] Commission of the European Communities. White Paper: Strategy for a Future Chemicals Policy. 2001. http://europa.eu.int./comm/environment/chemicals/whitepaper.htm

[3] M.T.D. Cronin, J.S. Jaworska, J.D. Walker, M.H.I. Comber, C.D. Watts and A.P. Worth, Use of QSARs in international decision-making frameworks to predict health effects of chemical substances, Environmental Health Perspectives 111, (2003) 1391-1401.

[4] ECOTOC Workshop on regulatory acceptance of (Q)SARS for human health and environmental endpoints. European Centre for Ecotoxicology and Toxicology of Chemicals. March 4-6, 2002. www.ecetoc.org, Setubal, Portugal, 2002.

[5] ATSDR Chemical-specific health consultation toxicological information on substances identified by the State of New Jersey, Dept. of Health and Human Services, Agency for Toxic Substances and Disease Registry, Div. of Toxicology, Atlanta, GA, 2000, pp. 83.

[6] E.M. Hulzebos and R. Posthumus, (Q)SARS: Gatekeepers against risk on chemicals?, SAR and QSAR Environ. Res. 14, (2003) 285-316.

[7] J.S. Jaworska, M. Comber, C. Auer and C.J. Van Leeuwen, Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints, Environ. Health Pers. 111, (2003) 1358-1360.

[8] M.T.D. Cronin, J.D. Walker, J.S. Jaworska, M.H.I. Comber, C.D. Watts and A.P. Worth, Use of quantitative structure-activity relationships in international decision-making

frameworks to predict ecological effects and environmental fate of chemical substances, Environ. Health Pers. 111, (2003) 1376-1390.

[9] J.D. Walker, L. Carlsen and J. Jaworska, Improving opportunities for regulatory acceptance of QSARS: The importance of model domain, uncertainty, validity and predictability, Quant. Struct.-Act. Rela. 22, (2003) 346-350.

[10] L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Cronin, R.M. McDowell and P. Gramatica, Methods for reliability, uncertainty assessment, and for applicability evaluations of classification and regression based QSARs, Environ. Health Pers. 111, (2003) 1361-1375.

[11] H.S. Rosenkranz, A.R. Cunningham, Y.P. Zhang, H.G. Claycamp, O.T. Macina, N.B. Sussman, S.G. Grant and G. Klopman, Development, characterization and application of predictive-toxicology models, SAR QSAR Environ. Res. 10, (1999) 277-298.

[12] H.S. Rosenkranz, Structural concepts in the prediction of the toxicity of therapeutical agents, in: Burger's Medicinal Chemistry and Drug Discovery, John Wiley & Sons, New York, 2003, pp. 827-847.

[13] H.S. Rosenkranz and B.P. Thampatty, Applications of substructure-based SAR in toxicology, in: C. Helma (Ed.), Predictive Toxicology, Marcel Dekker, Inc., New York, 2003, pp. in press.

[14] R. Benigni and R. Zito, The second National Toxicology Program comparative exercise on the prediction of rodent carcinogenicity: definitive results, Mutat. Res. 566, (2004) 49-63.

[15] E.J. Matthews and J.F. Contrera, A new highly specific method for predicting the carcinogenic potential of pharmaceuticals in rodents using enhanced MCASE QSAR-ES software, Regul. Toxicol. Pharmacol. 28, (1998) 242-264.

[16] H.S. Rosenkranz, SAR in the Assessment of Carcinogenesis: The MULTICASE approach, in: R. Benigni (Ed.), Quantitative Structure-Activity Relationship (QSAR) Models of Mutagens and Carcinogens, CRC Press, LLC, Boca Raton, FL, 2003, pp. 175-206.

[17] G. Klopman and H.S. Rosenkranz, Prediction of carcinogenicity/mutagenicity using MULTICASE, Mutat. Res. 305, (1994) 33-46.

[18] G. Klopman and H.S. Rosenkranz, Toxicity estimation by chemical substructure analysis: The Tox II Program, Toxicol. Lett. 79, (1995) 145-155.

[19] H.S. Rosenkranz, Practical aspects of SAR modeling of allergic contact dermatitis in humans: Congeneric *vs.* non-congeneric learning sets. 2004. http://bioser.biomed.fau.edu/biomedical/research/rosenkranz/refdB.html

[20] Y.P. Zhang, N. Sussman, G. Klopman and H.S. Rosenkranz, Development of methods to ascertain the predictivity and consistency of SAR models: application to the US National Toxicology Program rodent carcinogenicity bioassays, Quant. Struct.-Act. Rela. 16, (1997) 290-295.

[21] L. Eriksson, E. Johansson and S. Wold, QSAR model validation, in: F. Chen and G. Schüürmann (Eds.), Quantitative structure-activity relationships in environmental sciences. Proceedings of the 7th International Workshop on QSAR in Environmental Sciences, June 24-28, Elsinore, Denmark, VII. SETAC Press, Pensacola, FL, 1997, pp. 381-397.

[22] L.S. Gold, T.H. Slone and B.N. Ames, Overview and update of analyses of the carcinogenic potency database, in: L.S. Gold and E. Zeiger (Eds.), Handbook of Carcinogenic Potency and Genotoxicity Databases, CRC Press, Boca Raton, FL, 1997, pp. 661-685.

[23] L.S. Gold, C.B. Sawyer, R. Magaw, G.M. Backman, M. deVeciana, R. Levinson, N.K. Hooper, W.R. Havender, L. Bernstein, R. Peto, M.C. Pike and B.N. Ames, A carcinogenic potency database of the standardized results of animal bioassays, Environ. Health Perspect. 58, (1984) 9-319.

[24] R. Peto, M.C. Pike, L. Bernstein, L.S. Gold and B.N. Ames, The TD50: A proposed general convention for the numerical description of the carcinogenic potency of chemicals in chronic-exposure animal experiments, Environ. Health Perspect. 58, (1984) 1-8.

[25] A.R. Cunningham, G. Klopman and H.S. Rosenkranz, The carcinogenicity of diethylstilbestrol: structural evidence for a non-genotoxic mechanism, Arch. Toxicol. 70, (1996) 356-361.

[26] A.R. Cunningham, G. Klopman and H.S. Rosenkranz, A study of the structural basis of the carcinogenicity of tamoxifen, toremifene and their metabolites, Mutat. Res. 349, (1996) 85-94.

[27] H.S. Rosenkranz, Y.P. Zhang and G. Klopman, Risk identification using structural concepts: The potential carcinogenicity of praziquantel., Reg. Toxicol. Pharmacol. 22, (1995) 152-161.

[28] Y.P. Zhang, N. Sussman, O.T. Macina, H.S. Rosenkranz and G. Klopman, Prediction of the carcinogenicity of a second group of chemicals undergoing carcinogenicity testing, Environ. Health Pers.104, (Suppl. 5) (1996) 1045-1050.

[29] H.S. Rosenkranz, Practical aspects of MULTICASE SAR modeling of toxicological phenomena: Congeneric vs. non-congeneric learning sets. III. Nature of the learning sets. 2004. http://bioser.biomed.fau.edu/biomedical/research/rosenkranz/refdB.html

[30] V. Chankong, Y.Y. Haimes, H.S. Rosenkranz and J. Pet-Edwards, The carcinogenicity prediction and battery selection (CPBS) method: A Bayesian approach, Mutat. Res. 153, (1985) 135-166.

[31] F.K. Ennever and H.S. Rosenkranz, Predicting the carcinogenicity of the aromatic amine derivatives tested in the second UKEMS Collaborative Study, Mutagenesis 1, (1986) 119-123.

[32] F.K. Ennever and H.S. Rosenkranz, Evaluating batteries of short-term genotoxicity tests, Mutagenesis 1, (1986) 293-298.

[33] A.M. Richard and R. Benigni, AI and SAR approaches for predicting chemical carcinogenicity: survey and status report, SAR QSAR Environ. Res. 13, (2002) 1-19.

[34] M. Liu, N. Sussman, G. Klopman and H.S. Rosenkranz, Estimation of the optimal data base size for structure-activity analyses: The *Salmonella* mutagenicity data base, Mutat. Res. 358, (1996) 63-72.

[35] H.S. Rosenkranz and A.R. Cunningham, SAR Modeling of Unbalanced Data Sets, SAR QSAR Environ. Res. 12, (2001) 267-274.

[36] H.S. Rosenkranz and A.R. Cunningham, SAR modeling of genotoxic phenomena: the effect of supplementation with physiological chemicals, Mutat. Res. 476, (2001) 133-137.

[37] G. Klopman, Artificial intelligence approach to structure-activity studies. The computer automated structure evaluation of biological activity of organic molecules, J. Amer. Chem. Soc. 106, (1984) 7315-7321.

[38] H.S. Rosenkranz, SAR modeling of genotoxic phenomena: The consequence on predictive performance of deviation from a unity ratio of genotoxicants/non-genotoxicants, Mutat. Res. 559, (2004) 67-71.

[39] H.S. Rosenkranz and G. Klopman, An expert system approach to the prediction and elucidation of the structural basis of toxicological activities, in: A.M. Goldberg (Ed.), Alternative Methods in Toxicology. Vol. 8, InVitro Toxicology: New Technology, 8, In Vitro Toxicology: New Technology. Mary Ann Liebert, Inc., New York, 1991, pp. 145-162.

[40] J. Ashby and R.W. Tennant, Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the U.S. NTP, Mutat. Res. 257, (1991) 229-306.

[41] L.B. Lave, F.K. Ennever, H.S. Rosenkranz and G.S. Omenn, Information value of the rodent bioassay, Nature 336, (1988) 631-633.

[42] NTP, National Toxicology Program-Long Term Study Abstracts. 2004. http://ntp-server.niehs.nih.gov/htdocs/pub.html

[43] G. Klopman and H.S. Rosenkranz, Quantification of the predictivity of some short-term assays for carcinogenicity in rodents, Mutat. Res. 253, (1991) 237-240.

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†
Ethylenediamine	+	1.00	1.00
Fenthion	+	0.91	0.96
Hexamethylene diisocyanate	+	1.00	0.38†
Isononanoyl oxybenzene sulfonate	+	0.98	0.82
Isophorone diisocyanate	+	0.22†	0.17†
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†	0.14†
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†	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†	0.35

Table 2 – *continued*

Chemical	*Experimental Activity*	*Model 3-7/3/0.90*	
		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

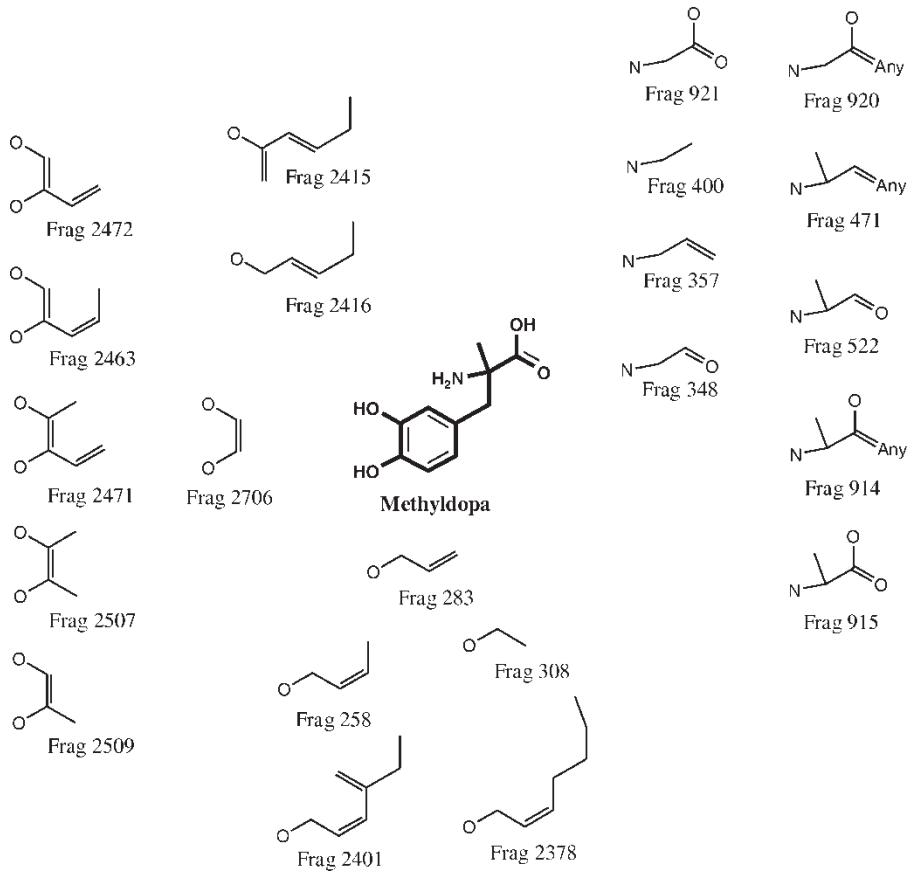


Figure 1. Illustration of the 22 significant fragments contributing to the active validation prediction of methyldopa.

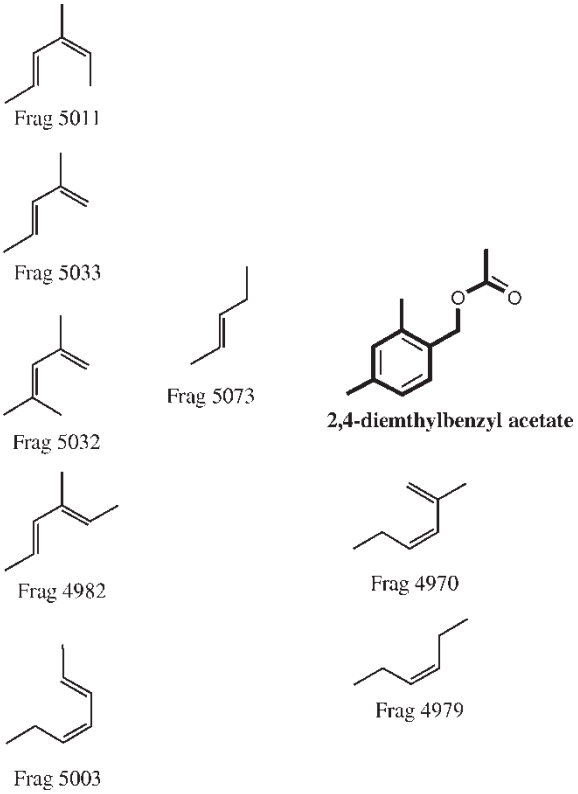


Figure 2. Illustration of the eight significant fragments contributing to the inactive validation prediction of 2,4-dimethylbenzyl acetate.

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).

SAR and QSAR in Environmental Research, VOL. 15 (1), February 2004, pp. 55–67

Taylor & Francis
Taylor & Francis Group

STRUCTURE–ACTIVITY APPROACH TO THE IDENTIFICATION OF ENVIRONMENTAL ESTROGENS: THE MCASE APPROACH

A.R. CUNNINGHAM[a,*], S.L. CUNNINGHAM[a] and H.S. ROSENKRANZ[b]

[a]*Department of Environmental Studies, Louisiana State University, 1285 Energy, Coast & Environment Building, Baton Rouge, LA 70803 USA;* [b]*Department of Biomedical Sciences, Florida Atlantic University, Boca Raton, FL 33431 USA*

(Received 24 April 2003; In final form 25 October 2003)

A sizable number of environmental contaminants and natural products have been found to possess hormonal activity and have been termed endocrine-disrupting chemicals. Due to the vast number (estimated at about 58,000) of environmental contaminants, their potential to adversely affect the endocrine system, and the paucity of health effects data associated with them, the U.S. Congress was led to mandate testing of these compounds for endocrine-disrupting ability. Here we provide evidence that a computational structure–activity relationship (SAR) approach has the potential to rapidly and cost effectively screen and prioritize these compounds for further testing. Our models were based on data for 122 compounds assayed for estrogenicity in the ESCREEN assay. We produced two models, one for relative proliferative effect (RPE) and one for relative proliferative potency (RPP) for chemicals as compared to the effects and potency of 17β-estradiol. The RPE and RPP models achieved an 88 and 72% accurate prediction rate, respectively, for compounds not in the learning sets. The good predictive ability of these models and their basis on simple to understand 2-D molecular fragments indicates their potential usefulness in computational screening methods for environmental estrogens.

Keywords: Environmental estrogens; Xenoestrogens; Structure–activity relationship (SAR); Computational modeling

INTRODUCTION

Compounds that mimic the activity of 17β-estradiol are of interest and concern for several reasons. First, many environmental contaminants have been found to possess estrogenic activity. These xenoestrogens are more generally known as endocrine disruptors. Another group of estrogenically active agents are of medicinal value. These are the selective estrogen receptor modulators (i.e. SERMs) that are actively being investigated as breast cancer therapies. The widely used tamoxifen and to a lesser extent rolaxifene are two such examples. Additionally, interest is focusing on plant derived estrogens (i.e. phytoestrogens) as chemopreventative agents [1] as well as alternative therapies for postmenopausal hormone replacement therapies [2,3].

*Corresponding author. E-mail: arc@lsu.edu

ISSN 1062-936X print/ISSN 1029-046X online © 2004 Taylor & Francis Ltd
DOI: 10.1080/1062936032000169679

With the obvious usefulness of SERMs, medicinal chemistry has added a great deal of understanding to the phenomena of estrogenicity and some of the health effects associated with these compounds. Although exceptionally useful, the investigation of SERMs does not cover the entire plethora of environmental concerns regarding endocrine active agents. The consequence of exposure to estrogen mimics can cause a vast array of toxicological and pharmacological responses including cancer [4–6], cancer therapy, developmental abnormalities and altered sexual differentiation [7,8], immune disturbances [9] as well as no observable adverse effects or even beneficial responses [1]. It has also been observed that the timing (e.g. fetal vs. adult), hormonal status, and level and duration of exposure affect the biological consequences associated with exposure to these agents. Moreover, apart from diversity in biological response, the estrogen mimics, as a group, display minimal structural homology [10]. This presents a challenge to structure–activity relationship (SAR) approaches aimed at their identification (i.e. predicted activity), activity and understanding mechanisms of action.

The United States Environmental Protection Agency (EPA) was mandated under the 1996 Food Quality Protection Act by the United States Congress to develop a screening and testing strategy to determine whether exogenous substances may have an effect in humans similar to those of natural hormones [11]. The EPA considers 87,000 chemicals as potentially requiring analysis for endocrine activity [12]. To facilitate this, a stated key goal of the EPA is to pursue computational methods for their analysis [13]. Computational SAR have gained recent acceptance in the regulatory community for both human health [14] and ecological endpoints [15].

Waller and others [16–19] have demonstrated the ability of comparative molecular field analysis (CoMFA) to accurately predict the relative binding affinities (RBA) of several series of compounds for the estrogen receptor. However, due to the limitations on CoMFA, these analyses had to rely on congeneric series of compounds for the training sets. However, with this limitation, these models are quite capable of predicting the activity of compounds that fit this model space. Additionally, the National Center for Toxicological Research has published a set of rat uterine cytosol RBA data [20]. Shi *et al.* [21] successfully analyzed this dataset and produced predictive CoMFA and holographic quantitative structure–activity relationship (HQSAR) models. Moreover, this same group has recently demonstrated the use of structural alerts for estrogen activity in a logical tree-based method to prioritize upwards of 58,000 compounds that are of environmental concern [22].

The ESCREEN dataset was chosen for several reasons. Basically, the ESCREEN assay measures estrogen-induced growth of human MCF-7 breast cancer cells [23,24]. Given the broad spectrum of biological assays for estrogenicity, the ESCREEN assays fall somewhere in the middle of the biological complexity scale (i.e. above *in vitro* receptor binding and below *in vivo* whole animal assays). This assay is well characterized, and the investigators report estrogenic response of chemicals using two unique parameters (i.e. relative proliferative potency (RPP) and relative proliferative effect (RPE)). RPP is the ratio between the least amount of 17β-estradiol needed to produce maximum proliferation and the least amount of the test chemical needed to produce a comparable effect [25]. That is, RPP compares the estrogenic potency of a compound to the potency of the standard estrogen 17β-estradiol. On the other hand, it is realized that many estrogenic compounds, no matter how high the dose, will never produce cell proliferation at the rate of 17β-estradiol. The RPE measures this effect. The PRE is 100 times the ratio of the greatest cell yield obtained with a test chemical and that obtained by 17β-estradiol [25].

The present investigation uses the MCASE algorithm (MultiCASE, Inc., Beechwood, OH) to predict estrogenic activity as measured in the ESCREEN assay [25]. The advantage of this approach is its ability to deal with non-congeneric datasets, as does HQSAR.

However, unlike HQSAR and CoMFA approaches that require continuous-type data, MCASE works by identifying molecular attributes associated with biological activity by comparing attributes of active (i.e. estrogenic) to inactive (i.e. non-estrogenic) compounds (i.e. a binary-type response). Although the MCASE program uses binary information to discriminate among structural features associated with active and inactive compounds, the program in this setting also took into account potency values for the active compounds. The models and subsequent predictions based on this dichotomy can then be used to examine structural features associated with estrogenicity and predicted the potential estrogenic activity of unknown compounds respectively.

The present report demonstrates the ability of MCASE to adequately assess compounds for their ability to induce an estrogenic response in MCF-7 cells. With these promising results, we are currently assessing the method's applicability to assess estrogen receptor binding ability as well as uterine growth stimulation and inhibition. Overall, considering the work from the National Center for Toxicological Research and the preliminary SAR modeling approach discussed here for environmental estrogens, it seems plausible that computational methods singly or in combination will be able to provide a reliable method to prioritize compounds for further testing and for regulatory classification. These methods, could therefore drastically reduce the tremendous financial cost, time and use of animals associated with meeting the mandate to assess these compounds for endocrine disrupting ability.

MATERIALS AND METHODS

Database

Two learning sets of 122 chemicals were created from publications of Soto and colleagues [23,25]. Both sets contain the same chemicals. The RPP learning set consisted of 50 active (i.e. estrogenic) and 72 inactive (i.e. non-estrogenic) chemicals. The RPE learning set consisted of 73 active and 49 inactive chemicals. Potency values for each endpoint were scaled to conform to MCASE requirements that SAR potency units range between 10 and 99 activity units. In this scale, inactive compounds were less than 30 and actives were greater than or equal to 30. Inactive RPE compounds were assigned 10 units and active compounds were scaled using the conversion equation, $\text{SAR units} = 0.62\,(\text{RPE}) + 29.38$. Inactive RPP compounds were assigned 10 SAR activity units and the active compounds were scaled using the conversion equation, $\text{SAR units} = 9.57 \log\,(\text{RPP}) + 70.29$.

The potency values obviously differed between the RPP and RPE models. However, the overall designation of compounds as estrogenic or non-estrogenic also differed between the two. Twenty three chemicals designated as inactive in the RPP set were listed as active in the RPE model (compounds 2,2′,3,3′,5,5′-hexachlorobiphenyl through 6-bromonaphthol-2, Table II). All the 23 compounds in question had very low RPE values. Although the original authors of the studies chose to call these compounds non-estrogens, we chose to call them active since activity (although minimal) was reported.

MCASE Methodology

The MCASE methodologies have been described [26–28]. Basically, MCASE selects its own descriptors automatically from a learning set composed of active and inactive molecules. The descriptors are readily recognizable single, continuous structural fragments that are embedded in the complete molecule. The descriptors consist of either activating or inactivating fragments termed as biophores and biophobes, respectively. Each of

the fragments is associated with a confidence level and a probability of activity that is derived from the distribution of these biophores and biophobes among active and inactive molecules. MCASE then selects the most important of these fragments as a biophore (i.e. the functionality that is associated with the largest number of active molecules and fewest number of inactive molecules). A biophore may also be a 2-D distance descriptor based upon the presence of lipophilic centers or heteroatoms in the molecule [29]. At this point, a congeneric series of chemicals has been identified with the biophore being the unifying feature. MCASE then performs a series of defined chemical substitutions of the atoms in the first biophore (e.g. halogen for halogen or nitrogen for carbon in aromatic systems) and then searches for similar biophores in the pool of fragments significantly related to active chemicals. All chemicals containing these related structural features are grouped together under a single biophore designation. Thus, a biophore may consist of a single feature or a family of chemically similar features. Using the molecules contained in this family as a learning set, MCASE derives a local QSAR equation for this series of chemicals. The regression variables may be chemical properties (e.g. structural fragments), physicochemical (e.g. $\log P$, water solubility), or quantum chemical parameters such as the energy of the highest occupied molecular orbital (HOMO) and the energy of the lowest unoccupied molecular orbital (LUMO). These features ("modulators") thus augment or decrease the basal activity associated with the biophore. The identified biophore and modulators will then be used to derive a local QSAR equation for chemicals within this subset. If the data set is congeneric, then the single biophore and associated modulators may explain the activity of the entire training set; this usually does not occur and there is a residue of molecules not explained by the single biophore and modulators. When this happens, the program will remove from consideration the molecules already explained by this biophore and will search for the next biophore and associated modulators. The process is iterated until all of the active molecules in the learning set have been explained or until no further biophores are identified.

The MCASE SAR program yields two numerical parameters when challenged with unknown chemicals. These are a predicted probability of activity and a predicted potency value. We have found that the ability to identify active or inactive compounds can be optimized by separate analyses of each of the two parameters to define optimal cutoff values for each that best separate predicted active from predicted inactive chemicals and therefore yield the best concordance between predictions and experimental results. Bayes' Theorem was used to combine the two individual parameters to yield an indication of the model's overall sensitivity, specificity and concordance [30,31] (Table I). Briefly, Bayes' Theorem

TABLE I Predictive performance summary for RPP and RPE MCASE models

Model	*Concordance*	*Sensitivity*	*Specificity*
RPP			
SAR Units	0.72	0.72	0.72
Probability	0.74	0.70	0.76
Overall	0.72	0.72	0.72
RPE			
SAR Units	0.87	0.88	0.86
Probability	0.86	0.92	0.78
Overall	0.88	0.86	0.89

Notes:
Concordance: number of correct predictions / total number of predictions.
Sensitivity: number of correct positive predictions / total number of positives.
Specificity: number of correct negative predictions / total number of negatives.
Overall: combined SAR models derived from Bayes' Theorem.

states that the joint probability of two events is the product of the probability of one of the events and the conditional probability of the second event, given that the first event occurs. The system employed here starts with a prior probability for the first event, which is set at 0.5. This reflects the fact that SAR models are constructed wherein the ratio of active to inactive chemicals is unity. Using Bayes' Theorem, this prior probability (i.e. 0.5) is updated with the specificity and sensitivity of the first SAR submodel. This posterior probability serves as the new prior probability to which the sensitivity and specificity of the second submodel is incorporated. This process is iterated for the two MCASE parameters to derive an overall probability that a chemical is active based upon the combined information of the SAR submodels [30–32].

To examine the predictivity of the MCASE SAR models, 10-fold cross-validation tests were conducted [33]. From the RPE and RPP learning sets, 10 mutually exclusive test sets were prepared. These sets were created by the random removal of approximately 10% of the chemicals in the database. The activity of each chemical in the test set was predicted from models developed with the remaining 90% of the database as a learning set. This allowed the determination of sensitivity, specificity and concordance between experimental and predicted results.

To analyze the potential of chemicals demonstrating estrogenic activity in the ESCREEN assay to induce other toxicological phenomena including cancer and developmental toxicity we used the "Chemical Diversity Approach". This is a method based on comparisons of the predicted toxicological profiles of a group of 10,000 chemicals chosen to represent a random assortment of all chemicals and chemical features [34]. These chemicals were derived from chemical structure libraries and from a random sample of chemical structures from the National Cancer Institute Repository of potential cancer chemotherapeutic agents. The various toxicological properties of these chemicals are predicted using validated SAR models including the models for RPE and RPP. The prevalence of chemicals predicted to possess two toxicological properties simultaneously is then quantified and compared to the expected prevalence. If the two effects are assumed to be independent of one another (i.e. null hypothesis), then the observed and expected values should be nearly equal. A significantly greater observed than expected prevalence indicates a similarity in mechanism among the toxicological effects that are being studied. Likewise, a significantly lower observed than expected prevalence suggests a possible antagonism between the phenomena under investigation. The applicability of the methodology to the study of diverse toxicological phenomena has been demonstrated by successfully estimating the number of potential *Salmonella* mutagens in the environment [35]. The inhibition of gap junctional intercellular communication is related to rodent carcinogenesis through cellular and systemic toxicity but not genotoxicity [34].

RESULTS AND DISCUSSION

Examination of the performance of the RPE and RPP models indicates that both have acceptable predictive performances to identify estrogenic and non-estrogenic compounds (Tables I and II). Overall, the RPP model correctly assessed the estrogenic activity of 72% of the compounds not included in the learning sets, while the RPE model correctly assessed 88% of the compounds. Interestingly, using the same 122 compounds, the RPE model outperformed the RPP models by 16%. This was achieved by both an increase in sensitivity and specificity. The change of 23 activity designations (see above) could have altered the structural components of the models to the point that they were over-weighted with either active or inactive chemicals and thus possibly, over-predict either group. However, this

was not the case as each overall model nearly equally maintained values for sensitivity and specificity (Table I). Therefore, the results suggest that these exceedingly weak chemicals are nonetheless true estrogens and contribute information to the model.

As mentioned, 23 compounds (compounds 2,2′,3,3′,5,5′-hexachlorobiphenyl through 6-bromonaphthol-2, Table II) have disparate activity in the two assays. These were mostly weak RPE compounds and negative RPP ones which were in some instances categorized as negative. Interestingly, 22 of the 23 are accurately predicted for RPE activity (Table II). However, the RPP model "erroneously" predicted these RPP inactive compounds as active. The predictivity of a model has been used as an acceptable measure for assessing the "meaningfulness" of a model [36]. Moreover, we have consistently observed that good predictivity is based on mechanistically sound and interpretable models [37,38]. Therefore, we consider that the RPE model, which includes the very weak estrogens, is a more informative model than that based on the RPP data set. This finding is significant with respect to applying the model to environmental estrogens and phytoestrogens, many of which are exceedingly weak compared to 17β-estradiol.

The major structural attributes that composed both models are listed in Tables III and IV and shown in Figs. 1 and 2. The sets of biophores designated with the letter are expanded biophores. Once the primary biophore (i.e. version a) is identified, the program searches for similar structural fragments to include in the expanded biophore family. The MCASE model for RPE consisted of five biophores (Table III and Fig. 1) and the model for RPP consisted of nine biophores (Table IV and Fig. 2). The RPE biophores divided the data into three basic groups: a phenolic ring with varying substitution patterns, the chlorinated nonaromatic compounds, and moieties that depict the keto and hydroxy substituents of 17β-estradiol derivatives. The RPP model was made up of biophores that were similar in nature to the RPE biophores. The major biophore in each model was the phenolic A-ring. It should be noted that this major biophore although not specific for a hydroxyl substitution which is commonly associated with estrogenicity was derived predominately from phenols.

Interestingly, the RPE biophores were more robust, each typically being derived from more chemicals than those in the RPP model. For example, what is explained with biophores 1–4 in the RPP model is explained with only two biophores in the RPE model. It is noteworthy that the RPE model outperformed the RPP model with fewer structural moieties being associated with activity. Therefore, designating the 23 weak estrogens as active compounds facilitated a refinement of features associated with estrogenicity. That is to say, the model contained more robust structural features and thus also indicates the superiority of the RPE over the RPP model.

Since the RPE model is both simpler and more predictive, we used it in the "Chemical Diversity Approach" to investigate the possible role of estrogens in other toxicological phenomena. Essentially, based on the greater than expected prevalence of chemicals possessing two toxicological properties simultaneously (one being estrogenicity in this exercise) we can hypothesize on the underlying mechanisms of action being related. The first analysis consisted of comparing the RPP and RPE models. As expected, there was a high degree of similarity verifying that, although they both are measuring different estrogenic endpoints (i.e. proliferative potency and effect relative to 17β-estradiol), these endpoints are related (Table V, Analysis 1). The two major health concerns related to environmental estrogens are their potential to induce cancer and developmental effects. We found that generally the RPE model did not significantly overlap with chemicals that have the potential to induce mutagenicity, unscheduled DNA synthesis, and chromosomal aberrations (Table V, Analyses 2–4). This is not unexpected as estrogens are not genotoxic *per se*. Only the SOS Chromotest showed significant commonality with estrogens (Table V, Analysis 5). This may be a reflection of the fact that this assay, unlike the others, responds

TABLE II Experimental results and MCASE predictions for RPP and RPE

	RPP		*RPE*	
Chemical	*Experimental*	*Prediction*	*Experimental*	*Prediction*
1,2-Dichloropropane	-	-	-	-
1-Naphthol	-	+	-	-
2,3,7,8-TCDD	-	-	-	-
2,4-DB Acid	-	-	-	-
2,4-Dichlorophenoxyacetic acid	-	-	-	-
2-Naphthol	-	+	-	+
4-Butooxyphenol	-	+	-	-
4-Hexyloxyphenol	-	+	-	-
5,6,7,8-Tetrahydronaphthol-2	-	+	-	-
Alachlor	-	-	-	-
Atrazine	-	-	-	-
Bendiocarb	-	-	-	-
Butylate	-	-	-	-
Butylated hydroxytoluene	-	-	-	-
Carbaryl	-	-	-	-
Carbofuran	-	-	-	-
Chlordimeform	-	-	-	+
Chlorothalonil	-	-	-	-
Chlorpyrifos	-	-	-	-
Cyanazine	-	-	-	-
Dacthal	-	-	-	-
Diamyl phthalate	-	-	-	-
Diazinon	-	-	-	-
Dibutyl phthalate	-	-	-	-
Dimethyl isophthalate	-	-	-	-
Dimethyl terephthalate	-	-	-	-
Dinonyl phthalate	-	-	-	-
Dinoseb	-	-	-	-
Hexachlorobenzene	-	-	-	-
Hexazinone	-	-	-	-
Kelthane	-	-	-	+
Lindane	-	-	-	-
Malathion	-	-	-	-
Maneb or zineb	-	-	-	-
Methoprene	-	-	-	-
Metalochlor	-	-	-	-
Mirex	-	+	-	+
Octachlorostyrene	-	-	-	-
Parathion	-	+	-	-
Phenol	-	+	-	-
Picloram	-	-	-	-
Propazin	-	-	-	-
Rotenone	-	-	-	-
Simazine	-	-	-	-
Styrene	-	-	-	+
Tetrachloroethylene	-	-	-	-
Thiram	-	-	-	-
Trifluralin	-	-	-	-
Ziram	-	-	-	-
2,2′,3,3′,5,5′-Hexachlorobiphenyl	-	-	1	+
2,3,3′,4,5-Pentachlorobiphenyl	-	+	1	+
3,5-Dichloro-4-hydroxybiphenyl	-	-	1.5	+
4-Monochlorobiphenyl	-	-	2.1	+
2,3′,5-Trichlorobiphenyl	-	-	2.2	+
3,5-Dichlorobiphenyl	-	-	2.7	+
2,3,5,6-Tetrachlorobiphenyl	-	-	3.1	+
2,6-Dichlorobiphenyl	-	-	3.4	+
Decachlorobiphenyl	-	-	3.5	+
2,5-Dichlorobiphenyl	-	-	3.7	+
Chlordene	-	+	4	+

TABLE II – *continued*

	RPP		*RPE*	
Chemical	*Experimental*	*Prediction*	*Experimental*	*Prediction*
Gibberellic acid	-	+	4	-
2,3,4,5,6-Pentachlorobiphenyl	-	+	4.4	+
2-Monochlorobiphenyl	-	-	4.4	+
2,3,4,4′-Tetrachlorobiphenyl	-	+	4.7	+
2′,3′,4′,5,'5-Pentachloro-2-hydroxybiphenyl	-	+	4.8	+
4-Ethylphenol	-	+	5	+
Chlordane	-	+	5	+
3,5-Dichloro-2-hydroxybiphenyl	-	+	5.4	+
2,3,6-Trichlorobiphenyl	-	-	5.8	+
Heptachlor	-	+	8	+
4-Propylphenol	-	+	17	+
6-Bromonaphthol-2	-	+	38	+
t-Butylhydroxyanisol	0.00006	+	30	+
2′,5′-Dichloro-2-hydroxybiphenyl	0.0001	-	13	+
2′,3′,4′,5′-Tetrachloro-3-hydroxybiphenyl	0.0001	+	35.3	+
2,3,4,5-Tetrachlorobiphenyl	0.0001	-	39.2	+
1-Hydroxychlordene	0.0001	+	40	+
Toxaphene	0.0001	-	51.9	-
Dieldrin	0.0001	-	54.89	+
Methoxychlor	0.0001	+	57	+
2,2′,3,3′,6,6′-Hexachlorobiphenyl	0.0001	-	61.6	+
2,2′,4,5-Tetrachlorobiphenyl	0.0001	+	61.6	+
2′,5′-Dichloro-3-hydroxybiphenyl	0.0001	+	69.9	+
p,*p*′-DDT	0.0001	+	71	+
2,4,4′,6-Tetrachlorobiphenyl	0.0001	-	75.7	+
2,3,4-Trichlorobiphenyl	0.0001	-	77	+
Endosulfan	0.0001	-	81.25	-
Kepone	0.0001	-	84	-
o,*p*′-DDD	0.0001	-	84	+
o,*p*′-DDT	0.0001	+	86.14	+
4-*tert*-Butylphenol	0.0003	+	71	+
4-*sec*-Butylphenol	0.0003	+	76	+
Bisphenol A	0.0003	+	82	+
4,4′-Dihydroxyiphenyl	0.0003	+	84	+
4-Hydroxybiphenyl	0.0003	+	87	+
Butylbenzylphthalate	0.0003	-	90	-
4-*iso*-Pentylphenol	0.0003	-	93	-
4-*tert*-Pentylphenol	0.0003	+	105	+
2,2′,5-Trichloro-4-hydroxybiphenyl	0.001	+	37.8	+
2′,5′-Dichloro-4-hydroxybiphenyl	0.001	+	71.2	+
Tamoxifen	0.001	+	75*	+
2′,3′,4′,5′-Tetrachloro-4-hydroxybiphenyl	0.001	+	92	+
Coumestrol	0.001	+	93	-
Bisphenol A dimethacrylate	0.003	-	84	+
4-Nonylphenol	0.003	+	100	+
2′,4′,6′-Trichloro4-hydroxybiphenyl	0.01	+	99.8	+
4-Octylphenol	0.03	+	100	+
5-Octylphenol	0.03	+	100	+
Pseudo diethylstilbestrol	0.1	+	100	+
16-Hydroxyestrone	0.1	+	-	+
Zearalenone	1	+	88	-
Zearalenol	1	+	93	-
Equilenin	1	+	100	+
Estrone	1	+	100	+
Allenolic acid	1	-	105	-
Estriol	10	+	100	+
Indenestrol	10	+	100	+
17β-estradiol	100	+	100	+
Ethynylestradiol	100	+	100	+
11β-chloromethylestradiol	1000	+	110	+

TABLE II – *continued*

	RPP		*RPE*	
Chemical	*Experimental*	*Prediction*	*Experimental*	*Prediction*
Moxestrole	1000	+	110	+
Diethylstilbestrol	1000	+	112	+

Note:
* RPE potency value estimated from compounds with similar RPP values.

to oxidative mutagens [39] of the type that are derived from estrogens [40–43]. Interestingly, there is antagonism between estrogenicity and the induction of micronuclei as indicated by the significantly less than expected overlap (Table V, Analysis 6). This could reflect that there are two mechanisms to induce micronuclei: genotoxic vs. non-genotoxic (e.g. via inhibition of tubulin polymerization). However, analysis of estrogens and carcinogens

TABLE III MCASE biophores associated with estrogenic activity measured by RPE in the ESCREEN assay

Biophore	*Total*	*Inactive*	*Active*
1a. cH ═cH —c ═cH —	70	13	57
1b. c ⟨ ═cH —c ═c ⟩ —	2	0	2
1c. cH ═ c —c ═ c—	24	0	24
1d. cH ═ cH —c ═c —	26	0	26
2. Cl —c ═c —c ═	24	0	24
3. Cl —C —C ═	7	0	7
4. OH —CH —	7	0	7
5. CO —C —	5	0	5

Each biophore is accompanied by the number of compounds contributing to it, the number of active and inactive compounds, and their average activity.
Notes:
Biophore interpretation:
c: aromatic carbon.
⟨⟩: attachment of electron withdrawing or electron donating group.
ˆ: epoxide.
⟨#-atom⟩: biophore branch at atom # with substituent.
See Fig. 1 for illustration of biophores.

TABLE IV MCASE biophores associated with estrogenic activity measured by RPP in the ESCREEN assay

Biophore		*Total*	*Inactive*	*Active*
1a. cH ═cH —c ⟨ ═cH —		41	9	32
1b. c ═cH —c ⟨ ═c —		1	0	1
1c. c ⟨═cH —c ═c ⟩—		2	0	2
2a. cH ═c —c ═c —c ═cH —	⟨2-Cl ⟩	4	0	4
2b. cH ═c —cH ═c —c ═c —	⟨2-Cl ⟩	2	0	2
3a. cH ═cH —c ═cH —cH ═c —CH —		6	0	6
3b. CH2 —c ═cH —cH ═cH —cH ═cH —		1	0	1
5. Cl —c ═c —c ═c —c ═cH —cH ═cH —cH ═ ⟨5-cH═ ⟩		2	0	2
6. C —C —C —C —	⟨2-Cl ⟩	3	1	2
7. OH —CH —CH —		2	0	2
8. Ô —CH —		1	0	1
9. O —SO —O —CH2—CH —C —C —	⟨6-Cl ⟩	1	0	1

Each biophore is accompanied by the number of compounds contributing to it, the number of active and inactive compounds, and their average activity.
Notes: see Table III.
See Fig. 2 for illustration of biophores.

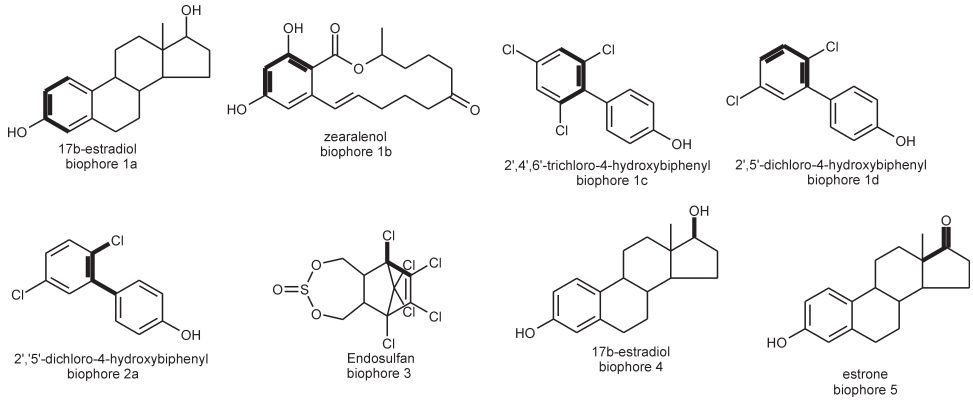


FIGURE 1 Illustration of MCASE biophores associated with estrogenic activity measured by RPE in the ESCREEN assay.

indicate that they may significantly share a common underlying mechanism (Table V, Analyses 7–10). Only the CPDB rat model did not significantly overlap with estrogenicity (Table V, Analysis 8). Analyses 11 and 12 of Table V show a significant overlap between estrogenicity and developmental toxicity in both humans and hamsters. Overall, these findings provide credibility to the mechanistic basis of the ESCREEN models.

CONCLUSIONS

The present analysis of estrogenicity with the MCASE program clearly indicates the utility of the program in assessing unknown compounds for estrogenicity. Given the complex structural nature of estrogenic compounds, it is imperative that any computational method

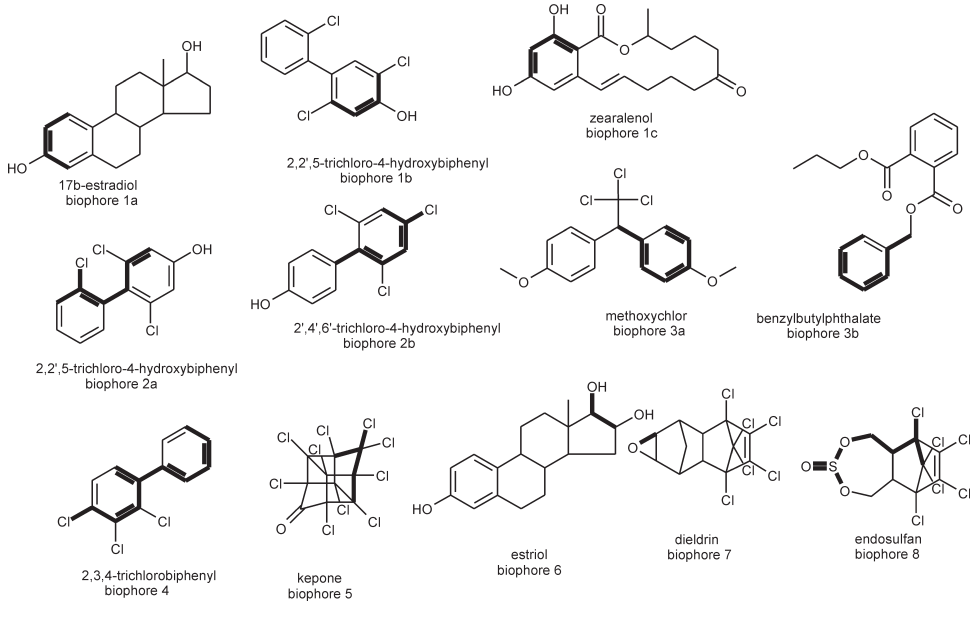


FIGURE 2 MCASE biophores associated with estrogenic activity measured by RPP in the ESCREEN assay.

TABLE V Mechanistic relationships of the ESCREEN RPE assay to other toxicological endpoints including genotoxicity, developmental effects and carcinogenesis

Analysis (References)	*Observed*	*Expected*	*p-value*	*Δ**	*100Δ/Expected*
1. ESCREEN relative proliferative potency	776	236	< 0.0001	540	228.8
2. Salmonella mutagenicity [44,45]	470	461	0.763	9	1.9
3. Unscheduled DNA synthesis [46]	485	445	0.179	40	9.0
4. Chromosomal aberrations [47]	364	400	0.184	− 36	− 9
5. SOS chromotest [48,49]	338	273	< 0.0001	65	23.8
6. Induction of micronuclei [50]	16	125	< 0.0001	− 109	− 87.2
7. CPDB mouse [37]	561	466	0.002	95	20.4
8. CPDB rat [38]	531	490	0.188	41	8.4
9. NTP mouse [47]	843	555	< 0.0001	288	51.9
10. NTP rat [47]	407	271	< 0.0001	136	50.2
11. Hamster developmental toxicity [51]	491	416	0.011	75	18.0
12. Human developmental toxicity [52]	338	274	0.009	64	23.4

Notes:
Observed: Number of compounds simultaneously identified to be estrogens using the RPE model and the row-listed endpoint.
Expected: The product of the individual prevalences of compounds identified to be estrogens using the RPE model and the row-listed endpoint.
p-value: Difference of two means test.
Δ: Difference of observed from expected.
100Δ/Expected: Percent difference from expected.

applied to their analysis is capable of coping with noncongeneric datasets. As evidenced by MCASE's predictive performance, it seems likely that this program has the potential to be a useful tool for screening and prioritizing environmental agents for subsequent testing. Moreover, we are applying this method in the development of models depicting relative binding ability to the estrogen receptor and for uterotropic and antiuterotropic activity. We speculate that although the program could be used as a stand-alone entity for screening potentially endocrine active compounds, it seems more likely and prudent that it could contribute as part of a battery of computational tools aimed at prioritizing compounds.

Acknowledgements

We gratefully acknowledge support for this work from the Congressionally Directed Medical Research Program for Breast Cancer Idea Award DAMD17-01-0376.

References

[1] Adlercreutz, H. (1993) "Phytoestrogens: Epidemiology and a possible role in cancer protection", *Environ. Health Perspect.* **103**(Suppl 7), 103–112.
[2] Clarkson, T.B., Anthony, M.S., Williams, J.K., Honore, E.K. and Cline, J.M. (1998) "The potential of soybean phytoestrogens for postmenopausal hormone replacement therapy", *Proc. Soc. Exp. Biol. Med.* **217**, 365–368.
[3] Arjmandi, B.H. (2001) "The role of phytoestrogens in the prevention and treatment of osteoporosis in ovarian hormone deficiency", *J. Am. Coll. Nutr.* **20**, 398s–402s.
[4] Marselos, M. and Tomatiz, L. (1992) "Diethylstilbestrol: I, Pharmacology, toxicology and carcinogenicity in humans", *Eur. J. Cancer* **28A**, 1182–1189.
[5] Marselos, M. and Tomatiz, L. (1993) "Diethylstilbestrol: II, Pharmacology, toxicology and carcinogenicity in experimental animals", *Eur. J. Cancer* **29A**, 149–155.
[6] IARC (1979) Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans, Sex Hormones (II) (International Agency for Research on Cancer, Lyon) Vol. **21**.
[7] vom Saal, F.S., Montano, M.M. and Wang, M.H. (1992) "Sexual differentiation in mammals", In: Colborn, T. and Clement, C., eds, Chemically-Induced Alterations in Sexual Development: The Wildlife/Human Connection (Princeton Scientific Publishing, Princeton, NJ), pp 17–84.
[8] Gray, J.L.E. (1992) "Chemical-induced alterations of sexual differentiation: a review of effects in humans and rodents", In: Colborn, T. and Clement, C., eds, Chemically-Induced Alterations in Sexual Development: The Wildlife/Human Connection (Princeton Scientific Publishing, Princeton, NJ), pp 203–230.

[9] Blair, B.B. (1992) "Immunologic studies of women exposed *in utero* to diethylstilbestrol", In: Colborn, T. and Clement, C., eds, Chemically-Induced Alterations in Sexual Development: The Wildlife/Human Connection (Princeton Scientific Publishing, Princeton, NJ), pp 289–294.
[10] Katzenellenbogen, J.A. (1995) "The structural pervasiveness of estrogenic activity", *Environ. Health Perspect. Suppl.* **103**(Suppl 7), 99–101.
[11] EPA (1998) "Endocrine Disruptor Screening and Testing Advisory Committee (EDSTAC) Final Report", [www.http://www.epa.gov/oscpmont/oscpendo/history/finalrpt.htm].
[12] EPA (2002) "Priority-Setting in the Endocrine Disruptor Screening Program (EDSP)-Background. Washington, DC: Environmental Protection Agency", [www.http://www.epa.gov/oscpmont/oscpendo/prioritysetting/background.htm].
[13] Timm, G (2002) "EPA update on the validation and standardization, In: TestSmart Endocrine Disruptors", [www.http://caat.jhsph.edu/programs/workshops/testsmart/endo02-proc.htm].
[14] Cronin, M.T.D., Jaworska, J.S., Walker, J.D., Comber, M.H.I., Watts, C.D. and Worth, A.P. (2003) "Use of quantitative structure–activity relationships in international decision-making frameworks to predict health effects of chemical substances", *Environ. Health Perspect.* **111**, 1376–1390.
[15] Cronin, M.T.D., Walker, J.D., Jaworska, J.S., Comber, M.H.I., Watts, C.D. and Worth, A.P. (2003) "Use of quantitative structure–activity relationships in international decision-making frameworks to predict ecological effects and environmental fate of chemical substances", *Environ. Health Perspect.* **111**, 1376–1390.
[16] Waller, C.L., Oprea, T.I., Chae, K., Park, H.-K., Korach, K.S., Laws, S.C., Wiese, T.E., Kelce, W.R. and Gray, J.L.E. (1996) "Ligand-based identification of environmental estrogens", *Chem. Res. Toxicol.* **9**, 1240–1248.
[17] Tong, W., Perkins, R., Strelitz, R., Collantes, E.R., Keenan, S., Welsh, W.J., Branham, W.S. and Sheehan, D.M. (1997) "Quantitative structure–activity relationships (QSARs) for estrogen binding to the estrogen receptor: predictions across species", *Environ. Health Perspect.* **105**, 1116–1124.
[18] Tong, W., Perkins, R., Xing, L., Welsh, W.J. and Sheehan, D.M. (1997) "QSAR models for binding of estrogenic compounds to estrogen receptor alpha and beta subtypes", *Endocrinology* **138**, 4022–4025.
[19] Tong, W., Lowis, D.R., Perkins, R., Chen, Y., Welsh, W.J., Goddette, D.W., Heritage, T.W. and Sheehan, D.M. (1998) "Evaluation of quantitative structure–activity relationship methods for large-scale prediction of chemicals binding to the estrogen receptor", *J. Chem. Inf. Comput. Sci.* **38**, 669–677.
[20] Blair, R.M., Fang, H., Branham, W.S., Hass, B.S., Dial, S.L., Moland, C.L., Tong, W., Shi, L., Perkins, R. and Sheehan, D.M. (2000) "The estrogen receptor relative binding affinities of 188 natural and xenochemicals: Structural diversity of ligands", *Toxicol. Sci.* **54**, 138–153.
[21] Shi, L.M., Fang, H., Tong, W., Wu, J., Perkins, R., Blair, R.M., Branham, W.S., Dial, S.L., Moland, C.L. and Sheehan, D.M. (2001) "QSAR models using a large diverse set of estrogens", *J. Chem. Inf. Comput. Sci.* **41**, 186–195.
[22] Hong, H., Tong, W., Fang, H., Shi, L., Xie, W., Wu, J., Perkins, R., Walker, J.D., Branham, W. and Sheehan, D.M. (2002) "Prediction of estrogen receptor binding for 58,000 chemicals using an integrated system of a tree-based model with structural alerts", *Environ. Health Perspect.* **110**, 29–36.
[23] Soto, A.M., Lin, T.-M., Justicia, H., Silvia, R.M. and Sonnenschein, C. (1992) "An "in culture" bioassay to assess the estrogenicity of xenobiotics (E-SCREEN)", In: Colborn, T. and Clement, C., eds, Chemically-Induced Alterations in Sexual Development: The Wildlife/Human Connection (Princeton Scientific Publishing, Princeton, NJ), pp 295–309.
[24] Soto, A.M., Sonnenschein, C., Chung, K.L., Fernandez, M.F., Olea, N. and Serrano, F.O. (1995) "The E-SCREEN assay as a tool to identify estrogens: An update on estrogenic environmental pollutants", *Environ. Health Perspect.* **103**(Suppl 7), 113–122.
[25] Sonnenschein, C., Soto, A.M., Fernandez, M.F., Olea, N., Olea-Serrano, M.F. and Ruiz-Lopez, M.D. (1995) "Development of a marker of estrogenic exposure in human serum", *Clin. Chem.* **41**, 1888–1895.
[26] Klopman, G. (1984) "Artificial intelligence approach to structure–activity studies. Computer automated structure evaluation of biological activity of organic molecules", *J. Am. Chem. Soc.* **106**, 7315–7321.
[27] Klopman, G. (1992) "MULTICASE 1. A hierarchical computer automated structure evaluation program", *Quant. Struct. Act. Relat.* **11**, 176–184.
[28] Klopman, G. and Rosenkranz, H.S. (1994) "Approaches to SAR in carcinogenesis and mutagenesis. Prediction of carcinogenicity / mutagenicity using MULTI-CASE", *Mutat. Res.* **305**, 33–46.
[29] Cunningham, A.R., Klopman, G. and Rosenkranz, H.S. (1996) "The carcinogenicity of diethylstilbestrol: structural evidence for a non-genotoxic mechanism", *Arch. Toxicol.* **70**, 356–361.
[30] Murrill, W.B., Brown, N.M., Zhang, J.-X., Manzolillo, P.A., Barnes, S. and Lamartiniere, C.A. (1996) "Prepubertal genistein exposure suppresses mammary cancer and enhances gland differentiation in rats", *Carcinogenesis* **17**, 1451–1457.
[31] Macina, O.T., Zhang, Y.P. and Rosenkranz, H.S. (1998) "Improved predictivity of carcinogens: the use of a battery of SAR models", In: Kitchin, K., ed, Testing, Predicting and Integrating Carcinogenicity (Marcel Dekker, New York), pp 227–250.
[32] Chankong, V., Haimes, Y.Y., Rosenkranz, H.S. and Pet-Edwards, J. (1985) "The carcinogenicity prediction and battery selection (CPBS) method: a Bayesian approach", *Mutat. Res.* **153**, 135–166.
[33] Zhang, Y.P., Sussman, N., Klopman, G. and Rosenkranz, H.S. (1997) "Development of methods to ascertain the predictivity and consistency of SAR models: Application to the U.S. National Toxicology Program rodent carcinogenicity bioassays", *Quant. Struct. Act. Relat.* **16**, 290–295.
[34] Pollack, N., Cunningham, A.R., Klopman, G. and Rosenkranz, H.S. (1999) "Chemical diversity approach for evaluating mechanistic relatedness among toxicological phenomena", *SAR QSAR Environ. Res.* **10**, 533–543.

[35] Rosenkranz, H.S. and Cunningham, A.R. (2001) "Prevalence of mutagens in the environment: experimental data vs. simulation", *Mutat. Res.* **484**, 49–51.
[36] Shi, L.M., Fan, Y., Myers, T.G., O'Connor, P.M., Paull, K.D., Friend, S.H. and Weinstein, J.N. (1998) "Mining the NCI anticancer drug discovery database: genetic function approximation for the QSAR of anticancer ellipticine analogues", *J. Chem. Inf. Comput. Sci.* **38**, 189–199.
[37] Cunningham, A.R., Rosenkranz, H.S., Zhang, Y.P. and Klopman, G. (1998) "Identification of "genotoxic" and "non-genotoxic" alerts for cancer in mice: the carcinogenic potency database", *Mutat. Res.* **398**, 1–17.
[38] Cunningham, A.R., Rosenkranz, H.S. and Klopman, G. (1998) "Identification of structural features and associated mechanisms of action for carcinogens in rats", *Mutat. Res.* **405**, 9–28.
[39] Rosenkranz, H.R. (2002) "A paradigm for determing the relevance of short-term assays: Application to oxidative mutagenesis", *Mutat. Res.* **508**, 21–27.
[40] Sipe, H.J., Jr., Jordan, S.J., Hanna, P.M. and Mason, R. (1994) "The metabolism of 17b-estradiol by lactoperoxidase: a possible source of oxidative stress in breast cancer", *Carcinogenesis* **15**, 2637–2643.
[41] Liehr, J.G., DaGua, B.B. and Ballatore, A.M. (1985) "Reactivity of 4′,4-diethystilbestrol quinone, a metabolic intermediate of diethylstilbestrol", *Carcinogenesis* **6**, 829–836.
[42] Roy, D., Floyd, R.A. and Liehr, J.G. (1991) "Elevated 8-hydroxyguanosine levels in DNA of diethylstilbestrol-treated Syrian Hamsters: Covalet DNA damage by free radicals generated by redox cycling of diethylstilbestrol", *Cancer Res.* **51**, 3882–3885.
[43] Liehr, J.G. (1990) "Genotoxic effects of estrogens", *Mutat. Res.* **238**, 269–276.
[44] Zeiger, E., Ashby, J., Bakale, G., Enslein, K., Klopman, G. and Rosenkranz, H.S. (1996) "Prediction of Salmonella mutagenicity", *Mutagenesis* **11**, 471–484.
[45] Lui, M., Sussman, N., Klopman, G. and Rosenkranz, H.S. (1996) "Estimation of the optimal database size for structure-activity analyses: the Salmonella mutagenicity database", *Mutat. Res.* **358**, 63–72.
[46] Rosenkranz, H.S., Zhang, Y.P. and Klopman, G. (1994) "Evidence that cell toxicity may contribute to the genotoxic response", *Regul. Toxicol. Pharmacol.* **19**, 176–182.
[47] Ennever, F.K., Rosenkranz, H.S., Lave, L.B. and Omenn, G.S. (1990) "Value-of-information analysis of testing strategies: estimating the effect of uncertainty about the proportion of chemicals that are true human carcinogens", In: Mendelsohn, M.L. and Albertini, R.J., eds, Mutation and the Environment, Part D: Carcinogenesis (Wiley-Liss, Hoboken, NJ), pp 23–48.
[48] Mersch-Sundermann, V., Klopman, G. and Rosenkranz, H.S. (1996) "Chemical structure and genotoxicity: studies of the SOS Chromotest", *Mutat. Res.* **340**, 81–91.
[49] Mersch-Sundermann, V., Schneider, U., Klopman, G. and Rosenkranz, H.S. (1994) "SOS-Induction in *E. coli* and *Salmonella* mutagenicity: a comparison using 330 compounds", *Mutagenesis* **9**, 205–224.
[50] Yang, W.-L., Klopman, G. and Rosenkranz, H.S. (1992) "Structural basis of the *in vivo* induction of micronuclei", *Mutat. Res.* **272**, 111–124.
[51] Gómez, J., Macina, O.T., Mattison, D.R., Zhang, Y.P., Klopman, G. and Rosenkranz, H.S. (1999) "Structural determinants of developmental toxicity in hamsters", *Teratology* **60**, 190–205.
[52] Ghanooni, M., Mattison, D.R., Zhang, Y.P., Macina, O.T., Rosenkranz, H.S. and Klopman, G. (1997) "Structural determinants associated with risk of human developmental toxicity", *Am. J. Obstet. Gynecol.* **176**, 799–806.