

Prediction of the Carcinogenicity of a Second Group of Organic Chemicals Undergoing Carcinogenicity Testing

Y.P. Zhang,¹ N. Sussman,¹ O.T. Macina,¹ H.S. Rosenkranz,¹ and G. Klopman²

¹Department of Environmental and Occupational Health, University of Pittsburgh, Pittsburgh, Pennsylvania; ²Department of Chemistry, Case Western Reserve University, Cleveland, Ohio

Twenty-four organic compounds currently undergoing testing within cancer bioassays under the aegis of the U.S. National Toxicology Program (NTP) were submitted to the computer automated structure evaluation (CASE) and multiple computer automated structure evaluation (MULTICASE) system for predictions of activity. Individual predictions resulting from the NTP combined rodent, NTP mouse, Carcinogenic Potency Database (CPDB) combined rodent, and CPDB mouse databases were combined using Bayes' theorem to yield an overall probability of rodent carcinogenicity. Based upon an arbitrary probability cut-off of 0.50, nine compounds were predicted to be rodent carcinogens. The predicted carcinogens are chloroprene, 1-chloro-2-propanol, codeine, emodin, furfuryl alcohol, isobutyraldehyde, primaclone, sodium xylenesulfonate, and *t*-butylhydroquinone. — Environ Health Perspect 104(Suppl 5):1045–1050 (1996)

Key words: CASE, MULTICASE, Bayes' Theorem, NTP predictions

Introduction

This article presents the results of the predictions of carcinogenicity in rodents of a group of organic molecules currently undergoing cancer bioassays under the aegis of the U.S. National Toxicology Program (NTP). We apply the computer automated structure evaluation (CASE) and the multiple computer automated structure evaluation (MULTICASE) structure-activity relational expert systems (SAR) (1,2) using several carcinogenicity SAR models (e.g., NTP carcinogenicity, Carcinogenic Potency Data Base [CPDB]). The results obtained with the different models are combined into a single prediction by the

application of Bayes' theorem (3–5) to the sensitivities and specificities determined for the individual SAR models.

Methodology

CASE/MULTICASE

Detailed descriptions of the algorithms employed in CASE and MULTICASE have been published previously (1,2). The necessary input to the CASE/MULTICASE system consists of a database comprised of the chemical structures of biologically active compounds as well as experimentally measured values of the activity. The experimental values are expressed in terms of CASE and MULTICASE units ranging from 10 (inactive) to 99 (extremely active). In addition to this quantitative scale, compounds can be simply entered as inactive or active. The CASE/MULTICASE system automatically fragments each molecular structure into units of 2 to 10 heavy atoms together with their associated hydrogens. Accordingly, to be predictable by CASE/MULTICASE, a molecule must consist of at least 3 nonhydrogen atoms. The CASE/MULTICASE system accommodates fragments with branching at one position along the linear atomic chain. The resulting fragments are

cataloged according to their origins (parent compounds). Fragments arising from active compounds are labeled as activating, while inactive compounds give rise to inactivating fragments.

The CASE/MULTICASE system performs a statistical analysis to identify those molecular fragments that are relevant to the observed activity. A binomial distribution is assumed, and any considerable deviation from a random distribution of a fragment among the active and inactive classes of compounds indicates potential significance to biological activity.

CASE utilizes the set of statistically significant fragments in predictions of the activity or inactivity of compounds. Predictions of activity/inactivity are expressed as probabilities and are based on the observed fragment distributions encountered within the learning set. Quantitative estimation of the potency (quantitative structure-activity relationship [QSAR]) for all the compounds within a database is achieved by a multivariate linear regression analysis based on the stepwise selection of a subset of molecular fragments. In addition, calculated values of the logarithm of the partition coefficient and its squared value ($\log P$, $\log P^2$) are included as potential variables. The coefficient of each of the molecular fragments within the QSAR is a measure of the activating/inactivating contribution made to the activity by the presence of the fragment.

MULTICASE, on the other hand, utilizes the set of statistically significant molecular fragments to find a descriptor (biophore) accounting for the majority of the active class of compounds. In addition to molecular fragments, MULTICASE considers two-dimensional distance descriptors. Compounds containing the primary biophore (fragment or distance) are removed from further consideration, and subsequent biophores are selected that explain the activity of the remaining compounds. This iterative process of selection is continued until either all the active compounds are accounted for or no significant descriptors remain. The presence or absence of relevant biophores determines the prediction of activity/inactivity.

Unlike CASE, MULTICASE attempts to derive a QSAR only within each group of compounds sharing a common biophore. These local QSARs identify molecular features that modulate the activity/potency of compounds containing the biophore. The

This paper is part of the NIEHS Predictive-Toxicology Evaluation Project. Manuscript received 6 January 1996; manuscript accepted 19 April 1996.

This investigation was supported by the U.S. Department of Defense and the Center for Alternatives to Animal Testing.

Address correspondence to Dr. H.S. Rosenkranz, Department of Environmental and Occupational Health, University of Pittsburgh, 260 Kappa Drive - RIDC Park, Pittsburgh, PA 15238. Telephone: (412) 967-6510. Fax: (412) 624-1289. E-mail: rsnkranz@vms.cis.pitt.edu

Abbreviations used: CASE, computer automated structure evaluation; MULTICASE, multiple computer automated structure evaluation; MCu, MULTICASE potency units; MC%, MULTICASE probability; Cu, CASE potency units; C% CASE probability.

modulators are selected from the associated pool of molecular fragments, distance descriptors, calculated electronic indices (lowest unoccupied molecular orbital, highest occupied molecular orbital, etc.), and calculated transport parameters (log *P*, water solubility, molecular weight). The contribution (positive or negative) of the modulators to the activity is valid only within the context of compounds containing the relevant biophore.

At this point, the analyzed CASE/MULTICASE database can be queried regarding predictions of the activity of chemicals not present in the learning data set or experimentally untested agents. For each compound both CASE and MULTICASE predictions are made according to the respective SAR model. Compounds containing structural fragments not encountered in the learning set are flagged as "warnings" since the contribution of these unknown functionalities to biological activity is uncertain.

A typical example of a CASE/MULTICASE prediction is shown in Figure 1. CASE/MULTICASE predictions are shown for furfuryl alcohol tested against the CPDB mouse carcinogen database. The first part of the output deals with MULTICASE predictions. Furfuryl alcohol contains the biophore $O=C=CH-CH=$ present in 13 carcinogenic compounds within the learning database. The compound also contains an extra biophore ($O-CH=$). The overall conclusion from MULTICASE is that furfuryl alcohol has a 97% chance of being a carcinogen with an activity calculated to be 55 units, corresponding to a TD_{50} value of 0.17 mmol/kg/day. The CASE results are listed in the second half of the output. The overall conclusion from CASE is a probability of 89% with a calculated activity of 54 (0.20 mmol/kg/day).

Carcinogenicity Databases

Carcinogenicity databases used in the present study to predict the activity of chemicals currently undergoing experimental evaluation consist of results obtained from previous NTP evaluations (6) and the CPDB data assembled by Gold and associates (7-11). Data for both the rat and the mouse are available in both compilations. Furthermore, the rat and mouse data have also been combined into a rodent database whereby a chemical carcinogenic to either rat or mouse is classified as a rodent carcinogen. In the present study, only the NTP and CPDB rodent and mouse databases were used.

Because of the nature of the CASE/MULTICASE system, only organic molecules can be analyzed. Furthermore, these molecules must consist of more than two heavy atoms (e.g., formaldehyde would not be suitable for analysis). Neither inorganics nor mixtures of organic compounds can be evaluated by CASE/MULTICASE.

Presently our NTP rodent database contains 313 compounds (173 active, 39 equivocal, 101 inactive). The NTP mouse and rat compilations consist of 319 (131 active, 20 equivocal, 168 inactive) and 316 (131 active, 28 equivocal, 157 inactive) chemicals, respectively. The CPDB rodent database consists of 433 compounds (261 active, 10 marginally active, 162 inactive), while the separate mouse and rat databases consist of 636 (291 active, 11 marginal, 334 inactive) and 745 (383 active, 14 marginal, 348 inactive) compounds, respectively.

CASE/MULTICASE Predictions

Submission of an unknown molecule to the CASE/MULTICASE system yields predictions based on both CASE and MULTICASE SAR models (Figure 1). Specifically the predictions generate MULTICASE potency units (MCu), MULTICASE probabilities (MC%), CASE potency units (Cu), and CASE probabilities (C%). To maximize the use of the information within these four predictive indices, each was evaluated individually as to its predictivity. This was accomplished by performing *n*-fold cross-validation studies using a range of threshold values (corresponding to CASE/MULTICASE potency units and probabilities) serving as demarcations between actives and inactives. Indices that consistently resulted in poor sensitivities or specificities were eliminated from consideration. As a result of these studies, we were able to assign optimum threshold values to the acceptable CASE/MULTICASE potency units and probabilities. These values served to classify compounds as active (+) or inactive (-) and were used to determine the respective sensitivities and specificities of each individual index (MCu, MC%, Cu, C%). The individual activity classifications (+ or -) corresponding to the four individual prediction indices can be combined according to Bayes' theorem to provide an overall predicted probability of activity within each SAR model.

For the NTP rodent, mouse, and rat databases, complete leave-one-out validation studies were performed. These involved removing one compound at a time from the complete database and rederiving a CASE/MULTICASE SAR model for the remaining compounds. The compound originally removed was subsequently submitted for a prediction of its activity. This procedure was performed *n* times, where *n* represented the total number of compounds within each of the respective databases. The predicted values of Cu, C%, MCu, and MC% for each of the chemicals were tabulated for comparison with the actual experimental activity. Because of the large size of the CPDB set, a complete leave-one-out cross-validation was not attempted. Alternatively, only 57 randomly selected compounds were removed one at a time from the rodent, mouse, and rat databases.

The optimal threshold values for classifying a compound as active (carcinogenic) or inactive (noncarcinogenic) were based on the Cu, C%, MCu, and MC% obtained from the above validation studies.

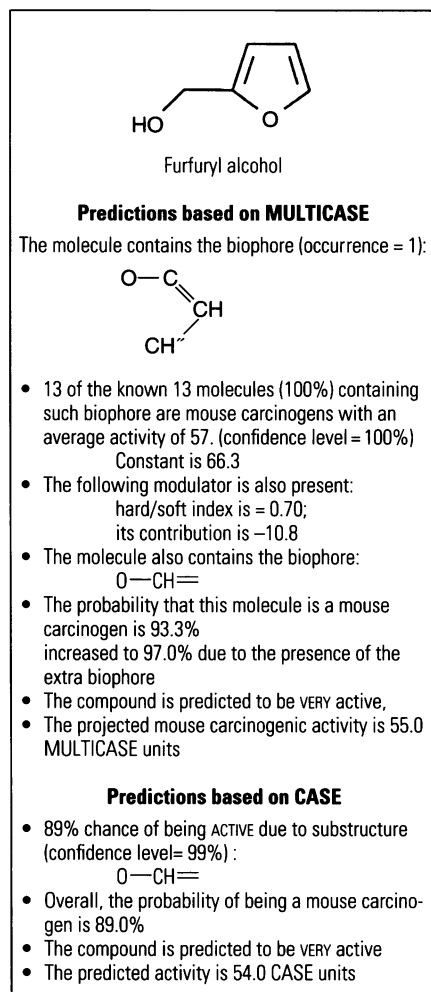


Figure 1. Example of a CASE/MULTICASE prediction.

The selection of optimal threshold values for each of the prediction indices was performed in isolation; e.g., the value chosen for MCu was independent of the values selected for MC%, Cu, and C%. The chemicals for which predictions were made during the cross-validation studies were used as the calibration set for determining predictions of activity/inactivity on the basis of the threshold value being considered. Comparison with the experimentally observed activities yielded the respective concordances, sensitivities, and specificities for each value being considered as a cut-off. The optimal threshold values as well as the corresponding concordances, sensitivities, and specificities are listed in Table 1 for the NTP and CPDB rodent and mouse databases. The NTP and CPDB rat models were found to lack sufficient concordance to be included in the battery of SAR models used for prediction. The NTP rodent and mouse SAR models employ only the MCu, MC%, and C% as predictive indices to arrive at a conclusion of activity. The CPDB rodent and mouse SAR models, on the other hand, use all four indices. This is not unexpected; for the NTP-derived models, the assigned potencies reflect a spectrum of carcinogenic activities, while for the CPDB models, they reflect TD₅₀ values expressed as mmole/kg/day (12).

Application of Bayes' Theorem

To combine all four predictions into a single overall conclusion, we sequentially

applied Bayes' theorem (5). Bayes' theorem is based on the fact that the joint probability of two events can be written as the product of the probability of one of the events and the conditional probability of the second event, given the first event. If we designate the two events as "A" (or "not A") and "+" (or "-"), the form of Bayes' theorem is as follows:

$$P(A/+)=\frac{P(+/A)*P(A)}{P(+/A)*P(A)+P(+/notA)*P(notA)}$$

The left side of Bayes' theorem is the probability that we are dealing with the state "A," given that we have observed the data "+." The probability P(+/A) represents the probability of the observed data "+," given that we are dealing with "A." Finally, P(A) is our current belief concerning the probability of "A". In the above form, P(A/+) is termed the posterior probability, i.e., the updated probability, given that we have observed a "+". P(A) and P(not A) are termed prior probabilities because they are decided on before any new data are known to us.

When we are in the diagnostic or predictive setting, Bayes' theorem takes on the following form:

$$P(Active/+)=\frac{Sensitivity * P(Active)}{Sensitivity * P(Active) + (1 - Specificity) * P(not Active)}$$

Using this form of Bayes' theorem, estimates of sensitivity and specificity associated with the prediction models and an estimate of the probability of activity of the chemical, we can estimate the new probability that a chemical is active, given that we have obtained a positive prediction when we apply our model. For the purpose of the present analysis, we have assumed a prior probability of 0.50, which reflects the prevalence of carcinogens in the NTP database. Similar expressions are available for P(Active/-), P(not Active/+), and P(not Active/-) (13).

Each additional application of Bayes' theorem uses the previously estimated posterior probability of activity as the new prior probability and the relevant estimates of sensitivity and specificity from the new source. By sequentially assimilating the data through Bayes' theorem, we obtain an increasingly updated estimate of the probability of activity for the chemical under consideration. This probability is based on the consideration of all four metrics (probabilities and units) obtained from the CASE/MULTICASE SAR models. Table 2 lists the calculated probabilities for the NTP/CPDB rodent/mouse SAR models that result from combining the activity classifications (+/-) with respect to MCu, MC%, Cu, and C%. The overall probabilities for each possible activity pattern were calculated by Bayes' theorem with the appropriate values of sensitivity and specificity (Table 1).

The Bayesian approach can further be used in situations in which more than one

Table 1. Optimal threshold values.

Database	MULTICASE unit				MULTICASE probability				CASE unit				CASE probability			
	Cut-off	Conc.	Sens.	Spec.	Cut-off	Conc.	Sens.	Spec.	Cut-off	Conc.	Sens.	Spec.	Cut-off	Conc.	Sens.	Spec.
N_rod.	22	0.61	0.61	0.62	72%	0.61	0.55	0.67	Unusable				58%	0.56	0.65	0.45
N_mice	30	0.64	0.33	0.85	57%	0.65	0.40	0.82	Unusable				53%	0.58	0.46	0.66
N_rat	Unusable															
G_rod.	30	0.54	0.53	0.56	68%	0.58	0.69	0.56	20	0.60	0.66	0.52	58%	0.58	0.59	0.56
G_mice	29	0.67	0.42	0.85	50%	0.68	0.46	0.85	20	0.74	0.54	0.88	66%	0.70	0.67	0.73
G_rat	Unusable															

N_rod., NTP rodent carcinogenicity database; N_mice, NTP mouse carcinogenicity database; N_rat, NTP rat carcinogenicity database; G_rod., CPDB rodent carcinogenicity database; G_mice, CPDB mouse carcinogenicity database; G_rat, CPDB rat carcinogenicity database; conc., concordance; sens., sensitivity; spec., specificity.

Table 2. Probabilities of possible combinations of predictions.

NTP rodent		NTP mouse		CPDB rodent				CPDB mouse			
Pattern	Probability	Pattern	Probability	Pattern	Probability	Pattern	Probability	Pattern	Probability	Pattern	Probability
++*+	0.7597	++*+	0.8687	++++	0.7769	----	0.1819	++++	0.9897	----	0.0929
++*-	0.6754	++*-	0.8000	++--	0.6554	----+	0.2894	+++-	0.9459	----+	0.3599
+-*+	0.5603	+-*+	0.6853	++*+	0.6235	---+	0.3187	++*+	0.9176	---+	0.4686
+*-+	0.4561	+*-+	0.5684	++*-	0.4749	---++	0.4614	++*-	0.6699	---++	0.8288
--*-	0.2473	--*-	0.3206	++*+	0.5515	---+-	0.3865	++*+	0.9521	---+-	0.3309
--*+	0.3330	--*+	0.4383	++*-	0.4017	---++	0.5357	++*-	0.7835	---++	0.7308
+*-+	0.4492	+*-+	0.5890	++*+	0.3689	---+-	0.5699	++*+	0.6976	---+-	0.8098
-+*+	0.5534	-+*+	0.7033	++--	0.2420	---++	0.7082	++--	0.2959	---++	0.9590

The pattern is ordered as MCu, MC%, Cu, C%; asterisk represents a discarded index. Probability > 0.5 is scored as "+."

SAR model is available. This is especially relevant to the prediction of carcinogenicity, wherein several databases can be used in making assessments of carcinogenic activity. Given the complexity of carcinogenicity, we do not expect a single SAR model to be as effective as the use of a battery of carcinogenic SAR models. Indeed, we expect that using such a battery for an overall prediction will overcome some of the limitations associated with individual SAR models.

In the present study, the carcinogenicity of the target compounds was predicted based on four carcinogenicity databases for which predictive SAR models had been derived: NTP rodent, NTP mouse, CPDB rodent, and CPDB mouse. To make an overall prediction of carcinogenicity, we applied Bayes' theorem using the predictions from the four models and their respective sensitivities and specificities (calculated from the *n*-fold cross validation studies). Table 3 lists the calculated probabilities resulting from combining all the possible predictions using the four carcinogenic SAR models. Elevated overall probabilities can be assumed to be indicative of an increased potential for rodent carcinogenicity. In the present analysis we arbitrarily selected a probability cut-off of 0.50 to classify compounds as carcinogens.

Results and Discussion

Our previously established CASE/MULTI-CASE SAR models (14,15) pertaining to NTP rodent, NTP mouse, CPDB rodent, and CPDB mouse data were used to rank the potential for carcinogenic activity of 24 organic compounds currently undergoing testing (Table 4). Organic salts were analyzed as the free bases/acids. The four inorganic compounds (cobalt sulfate heptahydrate, gallium arsenide, molybdenum trioxide, vanadium pentoxide) are not amenable to CASE/MULTICASE analyses. Predictions were also not made for nitromethane and sodium nitrite because these compounds are too small for CASE/MULTICASE analyses (above). All possible isomers of sodium xylenesulfonate were analyzed individually since commercial samples of xylene are a mixture of the *ortho*-, *meta*- and *para*-isomers.

Table 4 lists the results of our analysis of the carcinogenic potential exhibited by the chemicals of interest. Predictions of carcinogenic activity within each individual CASE/MULTICASE SAR model used in this study (Table 1) are shown, as well as the overall calculated probability resulting

from the sequential application of Bayes' theorem as outlined above. Individual SAR model predictions may differ from one another because the SAR models differ (the fragments identified within the mouse model are not necessarily identical to the rodent fragments). The chemicals and their overall predictions are listed in terms of the increasing probability of exhibiting carcinogenicity. If we arbitrarily assign a probability value of ≥ 0.50 to indicate a potential for carcinogenicity (+), nine compounds would be classified as such. Our SAR models, however, are based on oral administration and hence the predictions are based on the same assumption.

Five chemical compounds are predicted as carcinogens in at least two of the CASE/MULTICASE SAR models (1-chloro-2-

propanol, emodin, furfuryl alcohol, primaclone, and *t*-butylhydroquinone). Furfuryl alcohol and *t*-butylhydroquinone are predicted to be carcinogenic in three and four, respectively, of the SAR models. The remaining compounds (chloroprene, codeine, isobutyraldehyde, and sodium xylenesulfonate) are concluded to be carcinogenic by virtue of being so predicted by the CPDB mouse SAR model.

In the case of sodium xylenesulfonate, only the 1,3-dimethyl-4-sulfonate and 1,4-dimethyl-2-sulfonate isomers yielded positive predictions. Assuming that sodium xylenesulfonate is a mixture, the isomers responsible for carcinogenicity may not reach critical levels to induce carcinogenicity under experimental conditions.

Table 3. Predictions based upon combining carcinogenicity SAR models.

Pattern	Probability	Pattern	Probability	Pattern	Probability	Pattern	Probability
++++	0.9429	+++	0.8447	----	0.1535	-+--	0.3553
+++-	0.7204	++-	0.4633	---+	0.5332	-++-	0.7764
+++	0.9029	+++	0.7536	---+	0.2238	-++-	0.4949
++--	0.5906	++--	0.3268	---+	0.6701	----	0.8606

The order is NTP rodent, NTP mouse, CPDB rodent, and CPDB mouse. Probability > 0.5 is scored as "+."

Table 4. Predictions of carcinogenic activity resulting from oral administration.^{a,b,c,d}

Compound	NTP ^e		CPDB ^f		Overall probability ^g
	Rodent	Mouse	Rodent	Mouse	
Cinnamaldehyde	-	-	-	-	0.15(-)
Citral	-	-	-	-	0.15(-)
D&C Yellow 11	-	-	-	-	0.15(-)
Diethanolamine	-	-	-	-	0.15(-)
1,2-Dihydro-2,2,4-trimethylquinoline	-	-	-	-	0.15(-)
Ethylbenzene	-	-	-	-	0.15(-)
Ethylene glycol monobutyl ether	-	-	-	-	0.15(-)
Isobutene	-	-	-	-	0.15(-)
Oxymethalone	-	-	-	-	0.15(-)
Pyridine	-	-	-	-	0.15(-)
Scopolamine hydrobromide trihydrate	-	-	-	-	0.15(-)
Tetrahydrofuran	-	-	+	-	0.22(-)
Phenolphthalein	+	-	-	-	0.33(-)
Anthraquinone	-	+	-	-	0.36(-)
Methyl eugenol	-	+	-	-	0.36(-)
Chloroprene	-	-	-	+	0.53(+)
Codeine	-	-	-	+	0.53(+)
Isobutyraldehyde	-	-	-	+	0.53(+)
Sodium xylenesulfonate ^h	-	-	-	+	0.53(+)
1-Chloro-2-propanol	-	+	-	+	0.78(+)
Emodin	-	+	-	+	0.78(+)
Primaclone	-	+	-	+	0.78(+)
Furfuryl alcohol	-	+	+	+	0.86(+)
<i>t</i> -Butylhydroquinone	+	+	+	+	0.94(+)

^aChemicals are listed in ascending probability of carcinogenicity. ^bPredictions were not made for the inorganic compounds cobalt sulfate heptahydrate, gallium arsenide, molybdenum trioxide, and vanadium pentoxide; predictions were also not made for nitromethane and sodium nitrite, since the compounds are too small for CASE/MULTICASE analysis. ^cSalts were tested as the free bases/acids. ^dThese predictions are based on SAR models using data restricted to oral administration. ^eNTP refers to the National Toxicology Program. ^fCPDB refers to the Carcinogenicity Potency Database. ^gPositives (+) have a calculated probability of ≥ 0.50 ; negatives (-) have a calculated probability of < 0.50 . ^hAll possible isomers were tested; the 1,3-dimethyl-4-sulfonate and 1,4-dimethyl-2-sulfonate isomers were predicted as positives.

Figure 2 contains the chemical structures of the compounds predicted to be rodent carcinogens. In addition, the molecular regions identified as associated with carcinogenicity are indicated. These molecular regions consist of fragments identified by the CASE and/or MULTICASE SAR models of CPDB mouse data. Table 5 lists the putative carcinogens predicted in this study, together with known rodent carcinogens sharing carcinogenic CASE/MULTICASE fragments used within the CPDB mouse SAR model. It is presumed that compounds containing identical fragments in the same molecular environment will

exhibit similar modes of action and consequently exhibit similar biological activities.

Our approach entails the rational evaluation of a chemical's potential to exhibit carcinogenic activity on the basis of chemical structure alone. The quality of predictions depends on the learning set used to derive the resulting SAR model as well as on the complexity of the biological phenomenon modeled. Thus, in a previous study using a *Salmonella* mutagenicity SAR model, the concordance between experimental and predicted CASE/MULTICASE results for 100 chemicals was 76% (16). Obviously, given the multiple mechanisms

associated with carcinogenicity, we do not expect the SAR model for carcinogenicity to be as good as that for mutagenicity. However, we anticipate that the use of a battery of carcinogenic SAR models for an overall prediction will overcome some of the limitations associated with individual SAR models. Indeed, we have compared the respective concordances, sensitivities, and specificities of individual SAR model predictions with the carcinogenic SAR model battery predictions for an independent test set of compounds (17). The SAR model battery possessed the highest overall concordance with a significantly enhanced sensitivity and a comparable specificity. Furthermore, the SAR model battery predictions were significantly more accurate than using *Salmonella* results alone or in combination. Finally, the SAR data used to derive our models are based on the oral route of administration and may not be applicable to the prediction of carcinogenicity by other routes.

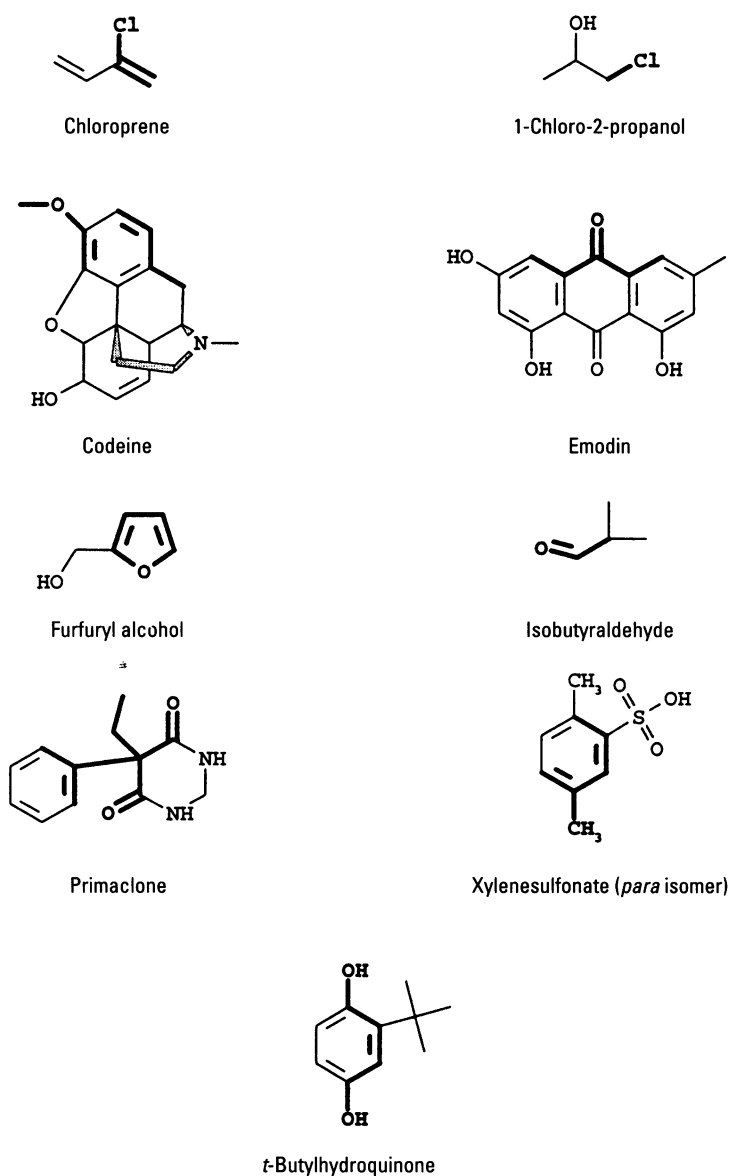


Table 5. Some representative rodent carcinogens containing fragments identified within the predicted carcinogens.^a

Predicted carcinogen	Example of known rodent carcinogen
Chloroprene	Vinylidene chloride Sulfallate
1-Chloro-2-propanol	Melphalan Chlorambucil Phenesterin Estradiol mustard Phenoxybenzamine Aramite
Codeine	Sterigamatcystin Aflatoxin 2-Acetylaminofluorene
Emodin	2-Methyl-1-nitroanthraquinone 1-Amino-2-methylantraquinone Chrysazin
Furfuryl alcohol	Furosemide Nitrofurazone 2-Amino-4-(5-nitro-2-furyl)-thiazole
Isobutyraldehyde	Leupeptin 2-Chloropropanal
Primaclone	Phenobarbital
Sodium xylenesulfonate	2,4-Xylidine <i>p</i> -Cresidine 2,4,6-Trimethylaniline
<i>t</i> -Butylhydroquinone	Luteoskyrin

^aCASE/MULTICASE fragments within the CPDB mouse SAR model were used.

Figure 2. Predicted carcinogens containing carcinogenic fragments. Regions identified by CASE and/or MULTICASE SAR models of CPDB mouse carcinogenicity are shown in bold.

REFERENCES

1. Klopman G. Computer automated structure evaluation of organic molecules. *J Am Chem Soc* 106:7315-7324 (1984).
2. Klopman G. MULTICASE 1. A hierarchical computer automated structure evaluation program. *Quant Struct Act Rel* 11:176-184 (1992).
3. Rosenkranz HS, Ennever FK, Chankong V, Pet-Edwards J, Haimes YY. An objective approach to the deployment of short-term tests predictive of carcinogenicity. *Cell Biol Toxicol* 2:425-440 (1986).
4. Pet-Edwards J, Haimes YY, Chankong V, Rosenkranz HS, Ennever FK. Risk Assessment and Decision Making Using Test Results: The Carcinogenicity Prediction and Battery Selection Approach. New York:Plenum Press, 1989.
5. Iversen GR. Bayesian Statistical Inference. Newbury Park, CA: Sage Publications, 1984.
6. Ashby J, Tennant RW. Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the U.S. NTP. *Mutat Res* 257:229-306 (1991).
7. Gold LS, Sawyer CB, Magaw R, Backman GM, deVeciana M, Levinson R, Hooper NK, Havender WR, Berstein L, Peto R, Pike MC, Ames BA. A Carcinogenic Potency Database of the Standardized Results of Animal Bioassays. *Environ Health Perspect* 58: 9-319 (1984).
8. Gold LS, deVeciana M, Backman GM, Lopipero M, Smith M, Blumenthal R, Levinson R, Berstein L, Ames BN. Chronological Supplement to the Carcinogenic Potency Database: Standardized Results of Animal Bioassays Published through December 1982. *Environ Health Perspect* 67:161-200 (1986).
9. Gold LS, Slone TH, Backman GM, Magaw R, DaCosta M, Lopipero P, Blumenthal M, Ames BN. 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 (1987).
10. Gold LS, Slone TH, Backman GM, Eisenberg S, DaCosta M, Wong M, Manley, NB, Rohrbach L, Ames BN. 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 (1990).
11. Gold LS, Manley NB, Slone TH, Garfinkle TH, Rohrbach L, Ames BN. Fifth plot of the carcinogenic potency database: Results of animal bioassays published in the general literature through 1986 and by the National Toxicology Program through 1989. *Environ Health Perspect* 100:86-135 (1993).
12. Cunningham A, Klopman G, Rosenkranz HS. A study of the structural basis of the carcinogenicity of tamoxifen, toremifene and their metabolites. *Mutat Res* 349:85-94 (1996).
13. Chankong V, Haimes YY, Rosenkranz HS, Pet-Edwards J. The Carcinogenicity Prediction and Battery Selection (CPBS) method: a Bayesian approach. *Mutat Res* 153:135-166 (1985).
14. Rosenkranz HS, Klopman G. Structural basis of carcinogenicity in rodents of genotoxicants and nongenotoxicants. *Mutat Res* 228:105-124 (1990).
15. Rosenkranz HS, Klopman G. The application of structural concepts to the prediction of the carcinogenicity of therapeutic agents. In: Burger's Medicinal Chemistry and Drug Discovery. 5th ed. Vol 1: Principles and Practice (Wolff ME, ed). New York:John Wiley & Sons, 1995;223-249.
16. Zeiger E, Ashby J, Bakale G, Enslein K, Klopman G, Rosenkranz HS. Prediction of *Salmonella* mutagenicity. *Mutagenesis* (in press).
17. Macina OT, Zhang Y, Rosenkranz HS. Improved predictivity of carcinogens: the use of a battery of SAR models. In: Testing, Predicting and Interpreting Carcinogenicity (Kitchin KT, ed). New York:Marcel Dekker, in press.