Prediction of the Rodent Carcinogenicity of 60 Pesticides by the DEREKfW Expert System

Pierre Crettaz[†,*] and Romualdo Benigni[‡]

Swiss Federal Office of Public Health, 3003 Bern, Switzerland, and Istituto Superiore di Sanita', Environment and Health Department, Experimental and Computational Carcinogenesis Unit, Viale Regina Elena 299, 00161 Rome, Italy

Received April 26, 2005

The two-year rodent bioassay represents the golden standard for evaluating the carcinogenicity of chemicals. Because of practical and ethical reasons, alternative approaches have been investigated for many years. Among these approaches, the (quantitative) structure−activity relationships [(Q)SARs] offer promising perspectives for quickly screening a large number of chemicals. To increase the acceptance of (Q)SARs among the regulators, their predictive power needs to be scientifically validated. In this article, we tested the capacity of the DEREKfW expert system to qualitatively predict the rodent carcinogenicity and the genotoxic potential of 60 pesticides recently registered in Switzerland. The percentage of false negatives was found to be 31% for carcinogenicity. The associated sensitivity of 69% indicates that most of the pesticides with positive rodent bioassay results were detected by DEREKfW. On the other hand, the low specificity of 47% indicates that many pesticides may be flagged as carcinogenic while rodent bioassays would not confirm this potential. This may lead to unnecessary testing or the unnecessary restriction of a chemical.

1. INTRODUCTION

Health effects of chemical compounds are mainly evaluated by conducting animal studies. In vivo toxicological studies can be complex, time-consuming, and expensive. This is especially true for rodent bioassays carried out to assess carcinogenic potential. This potential is conventionally tested in a two-year rodent bioassay with two species (rat and mouse) and two sexes. These bioassays represent the gold standard for evaluating carcinogenicity. Because of practical (time and cost of these tests) and ethical (animal welfare) reasons, only a small number of chemicals on the market have been tested for their carcinogenic potential, and it is unlikely that this situation will change in the foreseeable future. Alternative approaches have, therefore, been investigated for many years. Rhomberg [1] discussed three types of alternative approaches to replacing the two-year rodent carcinogenesis bioassay. The first possible approach is to rely on short-term assays. Such tests have been developed to assess the genotoxic potential and different epigenetic modes of action that can lead to the development of tumors. The principle of these short-term tests is that chemical carcinogenesis consists of some fundamental biological mechanisms. The demonstration of any of these mechanisms raises the presumption that the chemical can be a carcinogen. A second alternative approach consists of developing accelerated in vivo methods, for instance, by using transgenic mice or rats. The hope is that the increased sensitivity of these rodents will overcome some of the limitations of classical two-year rodent bioassays, such as the length, expense, and need for high doses to ensure adequate sensitivity.

A third approach to replacing the rodent bioassay involves waiving any long-term animal testing and predicting the carcinogenic potential on the basis of chemical structure. The assumption that biological activity is implicit from chemical structure has been around for over 100 years.[2] The prediction of effects from chemical structure encompasses a broad range of methodologies, generally referred to as structure−activity relationships (SARs). These methodologies are simplified representations of complex chemical−biological interactions and are consequently more uncertain than the underlying test data. They can be divided into two major types, qualitative SARs and quantitative SARs (QSARs). QSARs are quantitative tools approximating the complex relationships between chemical properties and toxic activities of compounds, while SARs are qualitative relationships which can include, for example, structural alerts incorporating molecular substructures related to the presence or absence of toxic activity.[3] Different (Q)SARs have been developed for hazard identification. Carcinogenicity has been the target of more efforts than any other toxicological endpoint. This is mainly due to the high cost of the two-year rodent bioassay, the pressure to reduce animal testing, the prominent role played by a positive carcinogenicity result in regulatory toxicology, and the relatively large knowledge on carcinogenicity.[4,5]

(Q)SARs offer promising alternatives for quickly screening the toxicological outcome of a large number of chemicals. This screening would enable the reduction of animal testing, the saving of money and time, and the speeding-up of the number of evaluated chemicals. This is especially important with regard to the new European Union chemical policy, with 30 000 existing chemicals to be evaluated within the next 12 years and ca. 2500 chemicals requiring an evaluation

* Author to whom correspondence should be addressed. Tel.: +41 (0)31 322 96 31. Fax: +41 (0)31 322 97 00. E-mail: Pierre.Crettaz@bag.admin.ch.
† Swiss Federal Office of Public Health.
‡ Istituto Superiore di Santita'.

10.1021/ci050150z CCC: $30.25
Published on Web 09/29/2005

of carcinogenic potential.[6] However, there is currently resistance to the application of (Q)SARs in the regulatory arena. The lack of validated and appropriate (Q)SARs, the lack of confidence in (Q)SAR predictions, and the lack of (Q)SAR expertise within regulatory authorities are key considerations for this resistance. Regulators often fear that (Q)SARs do not offer a sufficient level of protection for human health and are, thus, often reluctant to use them. To increase the acceptability of (Q)SARs by decision makers and to minimize their misuse, the predictive power of (Q)-SARs needs to be scientifically validated. This important acceptability criterion is presently only partially fulfilled for most of the (Q)SARs. The validation of (Q)SARs within the regulatory environment is, therefore, urgently needed. The purpose of this article is to test the capacity of the DEREKfW expert system (Lhasa Limited, version 7.0) to qualitatively predict the rodent carcinogenicity of a range of pesticides recently registered in Switzerland. Prediction of the genotoxic potential will also be discussed, since it can help to evaluate the carcinogenic potency. DEREKfW is evaluated in this paper, because it is a promising qualitative SAR model which is user-friendly and provides peer-reviewed arguments for each prediction, which makes them understandable and relatively transparent.

2. METHODOLOGY

DEREKfW for Windows, Version 7.0. DEREKfW for Windows (DEREK = Deductive Estimation of Risk from Existing Knowledge) is developed and supplied by LHASA Limited, an independent not-for-profit company located at the University of Leeds. It is a knowledge-based expert system originating from Sanderson and Earnshaw[7] at Schering Agrochemicals. It is designed to assist toxicologists to predict toxicological hazards on the basis of an analysis of chemical structure and physicochemical properties. Two-dimensional chemical structures can be input into DEREK-fW, and molecules with up to 999 atoms can be processed through the software.[8]

DEREKfW uses a knowledge base, which contains structural alerts describing structure−toxicity relationships and rules that describe the relationship between physicochemical properties and biological activities. The alerts are derived from the expertise of toxicologists from industry, academia, and government. The development of alerts is a dynamic process and allows for continuous refinements. The users can incorporate their own in-house alerts to adapt DEREKfW to their specific needs. DEREKfW identifies the toxicophore or substructure associated with an adverse effect and highlights this to the user with a brief statement about the hazard it represents. Additional information, including literature references and supporting example compounds with known toxicological bioassay results, are provided when available. Since a comprehensive list of references used to develop an alert is not given, the quality of the data employed to derive an alert cannot be fully verified. DEREKfW provides the domain of an alert. Structural alerts found in structures outside the domain of the alert are predicted as negative. Adjacent atoms to an alert are generally taken into account, but more remote fragments are not considered. Details on the internal performance are not known, although LHASA Limited does a lot of testing internally to check that alerts are firing exactly as designed (Laraway, E.; personal communication).

DEREKfW indicates whether a specific toxic response may occur but does not provide a quantitative estimate of the prediction. DEREKfW is, therefore, not suitable for risk assessment, where a measure of the potency is required and a no-observed-effect level has to be derived. Nine levels are used to characterize the likelihood of an occurrence of an adverse effect: certain, probable, plausible, equivocal, doubted, improbable, impossible, open, and contradicted. Since these terms have specific meanings, they are defined by LHASA Limited in the DEREKfW user guide. A plausible prediction means, for instance, that the weight of evidence supports the adverse effect, while an equivocal prediction means that there is an equal weight of evidence for and against the occurrence of an adverse effect.[9]

DEREKfW includes toxicity alerts for a broad range of toxicological endpoints, including carcinogenicity, genotoxicity, skin sensitization, respiratory sensitization, and irritation. Its main strengths lie in the areas of carcinogenicity, genotoxicity, and skin sensitization,[10] the other adverse effects being characterized by a limited number of alerts. Among the 312 structural alerts defined in DEREKfW version 7.0, 47 and 96 alerts are available for carcinogenicity and genotoxicity, respectively. More than 75% of the genotoxicity alerts refer to mutagenicity, which indicates that the incidence of chromosomal aberrations is not well-covered. Since we are interested in predicting the carcinogenic potential of chemical compounds, our analysis is limited, in this article, to the DEREKfW predictions for carcinogenicity and genotoxicity. From that perspective, it should be stressed that DEREKfW does not explicitly identify metabolites of a query compound. It only partly accounts for some of the metabolic pathways that a chemical can undergo; for instance, it is stated in the comments of the alerts "aromatic amide" and "aromatic nitro compounds" that cellular metabolism is required for carcinogenic activity of these compounds. This may have an impact on the carcinogenicity predictions, since many carcinogenic chemicals are active only after biotransformation to electrophilic species that can bind to nucleophilic DNA sites. Such compounds may be missed by DEREKfW and appear as false negatives.

Selection of Compounds. The predictive capabilities of DEREKfW for assessing the potential carcinogenicity and genotoxicity of pesticides have been evaluated in this article by comparing test and prediction results for a set of 60 pesticides. These compounds have been evaluated in Switzerland since 1998 and are listed in Tables 1 and 2. They offer the advantage to present a full set of reliable and high-quality toxicological data, which is vital for a validation exercise. The data set includes the two-year rodent carcinogenesis bioassay with rats and mice (male and female) and a battery of in vitro and in vivo genotoxicity tests. Unfortunately, the structures of the pesticides cannot be provided, since a confidentiality issue applies.

The selected pesticides have not been found in the illustrating examples provided with the alerts and have probably not been used to develop the DEREKfW alerts, although we were not able to check it since details of the training set are not provided. Only a subset of the references used to derive an alert is provided in DEREKfW. This is

Table 1. DEREKfW Predictions for Carcinogenicity of 60 Selected Pesticides

	DEREKfW			
	alert for carcinogenicity	likelihood, species	prediction[a]	bioassays results[b]
compound 1			−	neg.
compound 2	072 epoxide	plausible, mammal	+	neg.
compound 3	113 polycyclic aromatic hydrocarbon or hetero-analogue	plausible, mammal	+	neg.
compound 4	107 aromatic amide 121 substituted pyrimidine or purine	plausible, mammal	+	neg.
compound 5	073 alkylating agent	plausible, mammal	+	neg.
compound 6	107 aromatic amide 123 halogenated alkene	plausible, mammal	+	liver (Rmf) + parathyroid (Rm)
compound 7	073 alkylating agent 107 aromatic amide	plausible, mammal	+	nasopharynx (Rm)
compound 8			−	neg. (testicle tumors in Rm, control range)
compound 9	107 aromatic amide 116 polyhalogenated aromatic	plausible, mammal	+	neg.
compound 10	121 substituted pyrimidine or purine	plausible, mammal	+	neg.
compound 11	121 substituted pyrimidine or purine	plausible, mammal	+	neg.
compound 12	107 aromatic amide	plausible, mammal	+	neg.
compound 13	121 substituted pyrimidine or purine	plausible, mammal	+	liver (Mmf)
compound 14	107 aromatic amide 116 polyhalogenated aromatic	plausible, mammal	+	thyroid (Rmf) + liver (Rf +Mmf)
compound 15	105 aromatic nitro compound	plausible, mammal	+	neg.
compound 16			−	neg.
compound 17	no alert, but rule 162	plausible, mammal	+	neg.
compound 18			−	thyroid (Rm) + liver (RMmf)
compound 19	105 aromatic nitro compound	plausible, mammal	+	neg. [weak thyroid (Rf)]
compound 20			−	neg.
compound 21			−	neg.
compound 22	113 polycyclic aromatic hydrocarbon or hetero-analogue 116 polyhalogenated aromatic	plausible, mammal	+	neg.
compound 23			−	neg.
compound 24	123 halogenated alkene	plausible, mammal	+	liver (Rf) + liver (Mf)
compound 25			−	thyroid (Rm) + uterus (Rf) + ovary (Mf)
compound 26			−	thyroid (Rmf)
compound 27	no alert, but rule 161	plausible, rat/mouse	+	neg.
compound 28			−	neg.
compound 29			−	neg.
compound 30	no alert, but rule 161	plausible, rat/mouse	+	neg.
compound 31	113 polycyclic aromatic hydrocarbon or hetero-analogue	plausible, mammal	+	uterus (Rf) + liver (Rm + Mmf) + thyroid (Mm)
compound 32	070 n-nitro or n-nitroso compound	plausible, mammal	+	neg.
compound 33			−	neg.
compound 34			−	neg.
compound 35			−	thyroid (Rf)
compound 36			−	neg. (testicle tumors in Rm, control range)
compound 37			−	neg.
compound 38	121 substituted pyrimidine or purine	plausible, mammal	+	neg. (uterus tumors Rf, control range)
compound 39	107 aromatic amide	plausible, mammal	+	neg.
compound 40	107 aromatic amide 121 substituted pyrimidine or purine	plausible, mammal	+	neg.
compound 41			−	neg.
compound 42			−	uterus (Rf) + 5 other sites (Rf)
compound 43			−	neg.
compound 44	121 substituted pyrimidine or purine	plausible, mammal	+	neg.
compound 45	074 mono- or di-alkylhydrazine	plausible, mammal	+	neg.
compound 46	116 polyhalogenated aromatic 121 substituted pyrimidine or purine	plausible, mammal	+	kidney (Rmf)
compound 47	107 aromatic amide	plausible, mammal	+	thyroid (Rmf)
compound 48			−	neg.
compound 49	073 alkylating agent	plausible, mammal	+	thyroid (Rm) + liver (Mm)
compound 50			−	neg. (leucemia Rm, control range)
compound 51	no alert, but rule 162	plausible, mammal	+	thyroid (Mf + Rm) + liver (Mmf +Rf)
compound 52	112 thioamide or thiourea	plausible, mammal	+	neg.
compound 53			−	neg.
compound 54	106 aromatic hydroxylamine	plausible, mammal	+	neg.
compound 55	no alert, but rule 161	plausible, rat/mouse	+	liver (Mmf)
compound 56			−	neg.
compound 57	116 polyhalogenated aromatic	plausible, mammal	+	uterus (Rf) + Leydig cells (Rm) + liver
compound 58			−	neg.
compound 59	070 n-nitro or n-nitroso compound	plausible, mammal	+	liver (Mmf)
compound 60			−	thyroid (Rmf)

[a] Predictions: −, no alerts found; +, alerts found. [b] Bioassay results: neg. = negative; M, mouse; R, rat; m, male; f, female

Table 2. DEREKfW Predictions for Genotoxicity of 60 Selected Pesticides

	DEREKfW				
	specific endpoint	alert for the specific endpoint	likelihood, species	prediction[a]	test results[b]
compound 1				−	+ in vitro (CHO)
compound 2	mutagenicity chromosome damage	019 epoxide 361 α,β-unsaturated ester or thioester−class II or class III	plausible, bacterium plausible, mammal	+	neg.
compound 3				−	neg.
compound 4				−	+ in vitro (HPRT)
compound 5	mutagenicity	027 alkylating agent	plausible, bacterium	+ (only AT)	+ in vitro (CHO)
compound 6	mutagenicity	331 halogenated alkene	plausible, bacterium	+ (only AT)	neg.
compound 7	mutagenicity chromosome damage	027 alkylating agent 061 haloacetanilide or analogue	plausible, bacterium plausible, mammal	+	neg.
compound 8				−	+ in vitro (HLT + UDS)
compound 9	mutagenicity	352 aromatic amine or amide	plausible, bacterium	+ (only AT)	neg.
compound 10				−	neg.
compound 11				−	neg.
compound 12				−	neg.
compound 13				−	neg.
compound 14				−	neg.
compound 15	mutagenicity	329 aromatic nitro compound	plausible, bacterium	+ (only AT)	neg.
compound 16				−	neg.
compound 17				−	neg.
compound 18				−	neg.
compound 19	chromosome damage mutagenicity	309 substituted vinyl ketone 329 aromatic nitro compound	plausible, mammal plausible, bacterium	+	neg.
compound 20	chromosome damage	309 substituted vinyl ketone	plausible, mammal	+	neg.
compound 21				−	neg.
compound 22				−	neg.
compound 23				−	neg.
compound 24	chromosome damage mutagenicity	309 substituted vinyl ketone 331 halogenated alkene	plausible, mammal plausible, bacterium	+	neg.
compound 25				−	neg.
compound 26	mutagenicity	004 potentially labile halogen	plausible, bacterium	+ (only AT)	+ in vitro (V79)
compound 27				−	neg.
compound 28				−	neg.
compound 29				−	neg.
compound 30				−	neg.
compound 31	chromosome damage	308 alkyl carbamate	plausible, mammal	+	+ in vitro (AT)
compound 32	mutagenicity	007 N-nitro or N-nitroso compound	plausible, bacterium	+ (only AT)	+ in vitro (AT + MLT + CHL)
compound 33				−	neg.
compound 34				−	+ in vitro (MLT)
compound 35				−	neg.
compound 36				−	+ in vitro (MLT)
compound 37				−	+ in vitro (HLT + MLT)
compound 38				−	neg.
compound 39				−	neg.
compound 40	mutagenicity	352 aromatic amine or amide	plausible, bacterium	+ (only AT)	+ in vitro (HTL)
compound 41				−	neg.
compound 42	chromosome damage	308 alkyl carbamate	plausible, mammal	+	neg.
compound 43				−	+ in vitro (CHL)
compound 44				−	neg.
compound 45	mutagenicity	028 mono- or dialkylhydrazine	equivocal, bacterium	+ (only AT)	neg.
compound 46				−	neg.
compound 47	mutagenicity	004 potentially labile halogen	plausible, bacterium	+ (only AT)	neg.
compound 48				−	neg.
compound 49	mutagenicity	027 alkylating agent	plausible, bacterium	+ (only AT)	+ in vitro (HLT)
compound 50				−	+ in vitro (HLT + MLT)
compound 51				−	neg.
compound 52				−	+ in vitro (V79 + UDS)
compound 53				−	neg.
compound 54				−	neg.
compound 55				−	+ in vitro (MLT)
compound 56	mutagenicity genotoxicity chromosome damage	306 alkyl aldehyde or precursor 306 alkyl aldehyde or precursor 309 substituted vinyl ketone	plausible, mammal	+	+ in vitro (AT)
compound 57				−	neg.
compound 58				−	+ in vitro (RL + MLT)
compound 59	mutagenicity	007 N-nitro or N-nitroso compound	plausible, bacterium	+ (only AT)	neg.
compound 60				−	+ in vitro (AT)

[a] Predictions: −, no alerts found; +, alerts found. AT: Ames test. [b] Test results: neg. = negative; + in vitro, positive in vitro genotoxicity test; AT, Ames Test; HLT, human lymphocyte test; MLT, mouse lymphoma test; UDS, unschedulded DNA synthesis; HPRT, HPRT (hypoxanthine phosphoribosyl transferase) test; RL, rat lymphocyte; CHO, Chinese hamster ovary cells; CHL, Chinese hamster lung cells; V79, V79 cells.

Table 3. Comparison of DEREKfW Predictions with Experimental Results of Carcinogenicity and Genotoxicity and Derivation of the Sensitivity, Specificity, and Correctness

	DEREK positive[a]	DEREK negative[a]	% false negatives with DEREK	sensitivity	% false positives with DEREK	specificity = selectivity	correctness = concordance = accuracy	positive predictive value	negative predictive value
Carcinogenicity ($n = 60$)									
positive test (19)	13 (TP)	6 (FN)	31%	69%	53%	47%	53%	37%	76%
negative test (41)	22 (FP)	19 (TN)							
Genotoxicity ($n = 60$)									
positive test in vitro (19)	7 (TP)	12 (FN)	63%	37%	29%	71%	60%	37%	71%
negative test (41)	12 (FP)	29 (TN)							

[a] FN: false negative; TN: true negative; FP: false positive; TP: true positive.

sometimes supported with a list of example compounds that have positive or negative results. The nature of DEREKfW is that it stores expert knowledge rather than pieces of information. This is a limitation to the transparency of DEREKfW.

3. RESULTS

Detailed Predictions. The two-dimensional structure of the 60 pesticides was submitted to DEREKfW. DEREKfW was able to process all these compounds, which are all within its domain of applicability. DEREKfW predictions are presented in detail in Table 1 for carcinogenicity and in Table 2 for genotoxicity. The name of the alerts found and the likelihood level of the qualitative prediction are also given. A total of 58% (35/60) and 32% (19/60) of the pesticides present result in at least one alert for carcinogenicity and for genotoxicity, respectively. About one-quarter of the pesticides predicted as carcinogenic or genotoxic present more than one structural alert. A plausible prediction for mammals or bacteria is obtained for all pesticides with one or more alerts, with the exception of one compound for which an equivocal prediction is found. All these predictions have been evaluated as positive, even the equivocal prediction. In addition, five compounds have been predicted as plausible carcinogens by DEREKfW reasoning rules 161 or 162, although no structural alert for carcinogenicity has been triggered. Rules 161 and 162 predict a plausible carcinogenic effect in rodents on the basis of the prediction of peroxisome proliferation and thyroid toxicity, respectively.[8] These predictions have been interpreted as positive.

Predictions Versus Experimental Results. The experimental results gained from two-year rodent bioassays are presented in Table 1, together with the tumor sites and the species/sex in which tumors have been found. The liver and the kidney were the most frequent sites of tumors. The tumor incidence was within the range of historical control data for four pesticides. Their bioassay results were, therefore, evaluated as negative, since the incidence of tumors was spontaneous and not substance-related. The types of in vitro genotoxicity tests leading to positive results are shown in Table 2. A total of 32% (19/60) of the pesticides are carcinogenic in at least one rodent species, and 32% (19/60) are genotoxic in at least one in vitro genotoxicity test. This proportion of chemicals testing positive is important for a validation exercise. In the extreme example that there are, for instance, only carcinogens in the tested compounds, every negative prediction will be a false negative. On the other hand, every positive prediction will be a false positive if there are no carcinogens in the tested compounds.

4. DISCUSSION

Predictive Performance. The predictive ability can be assessed by determining whether DEREKfW is missing many carcinogenic or genotoxic pesticides and whether DEREKfW tends to overpredict these adverse effects. An insufficient level of protection for humans is provided in the first case, while unnecessary animal testing or regulatory restrictions are the consequence in the second case.

The number of false predictions (both false positives and false negatives) is presented in Table 3, together with the number of positive and negative test results. Statistical values such as the rates of false negatives (FN) and false positives (FP), as well as the sensitivity, specificity, and correctness of the prediction, have been derived by comparing predictions with test results. These statistical terms are conventionally used to assess the predictivity of a qualitative expert system. Prediction correctness is defined as the ratio of correctly predicted compounds (true positives and true negatives) to the total number of compounds. It represents the overall concordance with test results. Sensitivity is defined as the percentage of rodent carcinogens which are predicted by DEREKfW as carcinogens, while specificity is the percentage of rodent noncarcinogens predicted as noncarcinogens.

The rate of false negatives and the corresponding sensitivity are key indicators for a regulatory agency. Authorities in charge of the registration of chemicals aim at keeping the rate of false negatives as low as possible, to avoid missing toxic compounds as much as possible. A negative prediction has, therefore, a lower regulatory acceptability than a positive prediction (precautionary application of predictive systems). From that perspective, it should be mentioned that (Q)SARs developers have the option of either (a) writing very restrictive alerts that would tend to cause the program to make false negative predictions but would give users a high level of confidence for positive predictions or (b) making the alerts less restrictive and more likely to fire, thus creating a program that gives more false positive predictions but gives users a high level of confidence in negative predictions. This latter option is preferable from a regulatory point of view. Obviously, the most desirable (Q)SAR software would be one that identifies correctly both the positive and negative compounds.

Carcinogenicity. DEREKfW predictions for carcinogenic activity are correct for 53% of the pesticides. More false positive predictions are found than false negative predictions.

The rate of false negatives is 31%, which is relatively low for SAR prediction. The corresponding sensitivity is 69%. These values indicate that DEREKfW was able to detect more than two-thirds (69%) of the pesticides causing tumors in rodents. Less than one-third (31%) of the pesticides inducing tumors in rodents had no structural alerts and were not detected by DEREKfW. The structure of the six false negatives could be investigated by LHASA Limited to find out if structural alerts not yet available in DEREKfW could be derived and incorporated in the expert system. These new toxicophores could help to improve DEREKfW and to lower the number of false negatives. Furthermore, the underlying biological mechanisms and carcinogenic potencies of the missed carcinogens have been considered by analyzing the bioassay results of the six false negatives. Thyroid tumors were reported for five of the six false negatives, indicating that this type of tumor may be difficult to predict by DEREKfW. Furthermore, the tumors reported for five of these false negatives are of little significance since they were, for instance, reported only at the maximum tolerated dose. This may indicate that these false negatives have only a weak carcinogenic potential in rodents.

The rate of false positives is 53%. The specificity is, therefore, only 47%. These rates indicate that about half of the pesticides which do not induce tumors in rodents are predicted by DEREKfW as plausible rodent carcinogens. The prediction, therefore, tends to be quite conservative. This is acceptable for a regulatory agency but unfavorable to industry, which can, nevertheless, always carry out a rodent carcinogenesis bioassay to find out if a positive prediction is correct or not. Alerts that are particularly conservative can be identified, and their refinement is recommended (see below).

Genotoxicity. DEREKfW predictions for genotoxicity are correct for 60% of the pesticides. The rate of false negatives is 63%. The sensitivity is, thus, only 37%. This indicates that more than half (63%) of the pesticides which are genotoxic in at least one in vitro test are not detected by DEREKfW. DEREKfW is, therefore, missing a large number of in vitro genotoxic pesticides. This indicates that there is an insufficient coverage of the genotoxic structures of pesticides by DEREKfW. Since a high proportion of in vitro genotoxic pesticides are erroneously predicted to be nongenotoxic, a negative DEREKfW prediction must be interpreted with caution and cannot be regarded as a reason for waiving genotoxicity testing. It should, however, be mentioned that genotoxicity testing is relatively rapid and inexpensive compared to a two-year rodent bioassay. The need for reliable predictions in this field is, therefore, less urgent than for carcinogenicity. Furthermore, none of the pesticides examined in this study have been found to be genotoxic in vivo. This may indicate that these pesticides have a low probability of being genotoxic to humans.

The rate of false positives is 29%. The corresponding specificity is 71%. This indicates that about one-third (29%) of the pesticides which do not induce genotoxicity in a battery of bioassays are predicted by DEREKfW as plausible genotoxic compounds. The percentage of false positives is lower than the percentage of false negatives, which is exactly the contrary of what has been found for carcinogenicity. This may partly be due to the fact that DEREKfW alerts are defined more specifically for genotoxicity than for carcinogenicity. Another reason for the high percentage of false negatives may be that the coverage of assays resulting in chromosome damage is not yet complete in DEREKfW, mainly because of the reduced availability of data. Further work on chromosome damage is, therefore, required to improve the sensitivity of the genotoxicity prediction.

Comparison with Other Prediction Exercises. The predictive performance reported in this paper can be compared with a number of other external validations of DEREKfW which have been published in the literature.

Regarding carcinogenicity, Hulzebos and Posthumus [11] evaluated DEREKfW for 29 chemicals discussed in the Environmental Health Criteria (EHC) Monographs or recently notified in The Netherlands. A better predictive capacity was reported for these chemicals (sensitivity = 88% and specificity = 62%). Another evaluation of DEREKfW regarded 28 Bayer compounds (21 agrochemicals, 6 pharmaceutical chemicals, and 1 organic chemical).[12] The prediction was characterized by a sensitivity of 60% and a specificity of 26%, which is worse than our values. Little credit can, however, be paid to this sensitivity, since only 5 of the 28 compounds tested positive in the rodent bioassays. In an exercise on 61 chemicals from the IUCLID database, the sensitivity and specificity were 68% and 62%, respectively.[12] Pearl et al.[13] derived a sensitivity of 75% and a specificity of 40% for 142 drugs, which is comparable with our values, although the chemical universe of drugs is different from that of pesticides. At the beginning of the 1990s, the developers of some predictive methods were invited to challenge their tools by predicting, in advance, the carcinogenic potential for rodents of 44 chemicals that were about to be tested by the National Toxicology Program (NTP). Sanderson and Earnshaw[7] presented the DEREKfW predictions. Parry[14] and Benigni[15] summarized the performance of the different predictive methods. A total of 59% of the predictions made by DEREKfW were correct. This is in line with the levels of correct predictions that we have found. The results of a second similar NTP comparative exercise were presented by Benigni and Zito[16] for another set of 30 NTP chemicals. A lower correctness of 43% was derived for DEREKfW, on the basis of the blind prediction presented by Marchant.[17] This value may represent a lower limit of the predictive capacity of DEREKfW, since this expert system was substantially improved in the past eight years. In addition, it should be noted that the 30 NTP chemicals included several nongenotoxic carcinogens, thus presenting a considerable challenge for the prediction systems.

In the literature, there are also a number of external validations of DEREKfW for genotoxicity. Hulzebos and Posthumus[11] found for 44 chemicals a sensitivity of 90% and a specificity of 74%, which is much better than the sensitivity found in our analysis. Three validation studies were reported in an ECETOC report.[12] One of the studies regards 27 chemicals tested by Bayer Toxicology, with a much better resulting sensitivity (sensitivity = 100%, specificity = 61%). The perfect sensitivity is, however, of little significance, since only 4 of the 28 compounds were found to be mutagenic in experimental testing. The second study reported by ECETOC [12] regarded 44 simple aromatic amines, still with high sensitivity (86%), and the third study included 169 new proprietary Novartis pharmaceutical

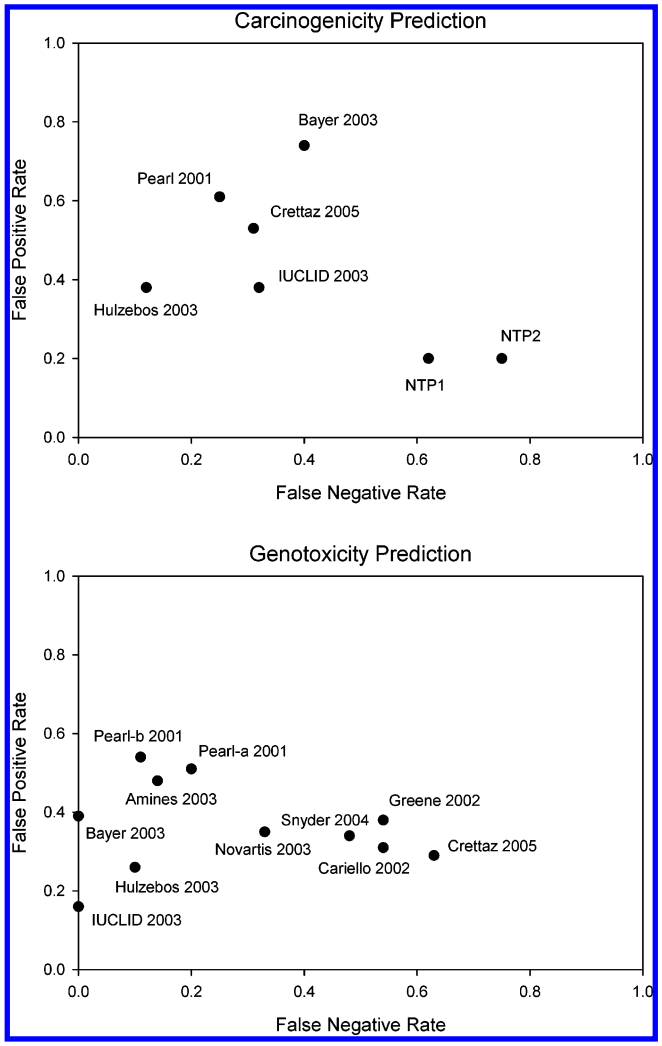

Figure 1. Performance of DEREKfW in external validation exercises, for carcinogenity and genotoxicity.

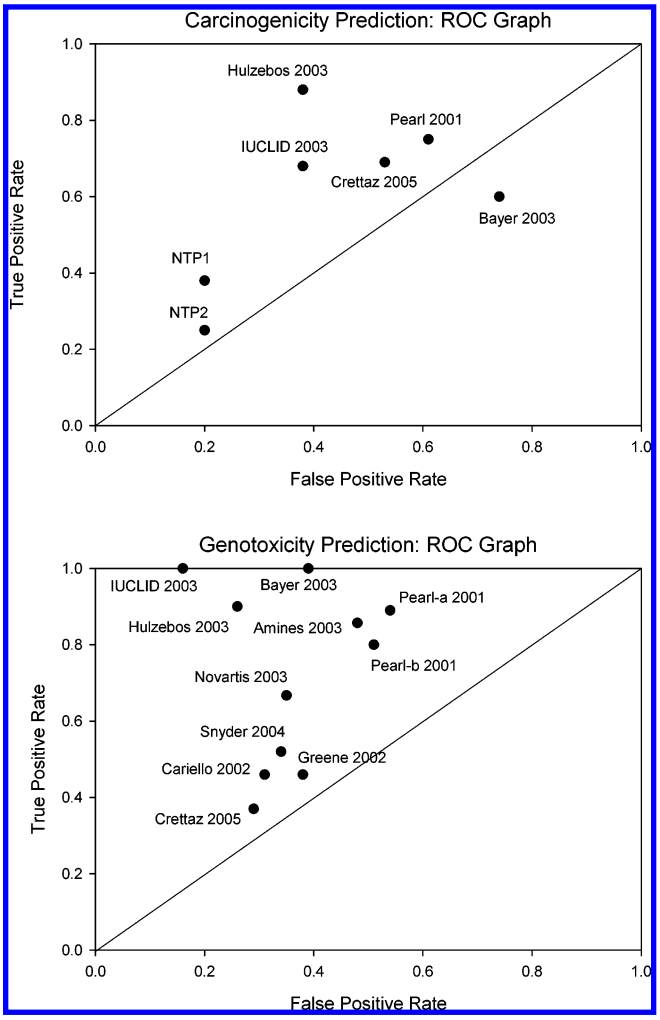

Figure 2. ROC graphs of carcinogenity and genotoxicity prediction in external validation.

candidates, with lower sensitivity (67%). In the third study on 54 chemicals from the IUCLID database, both sensitivity and specificity were particularly high (100% and 84%, respectively).[12] Cariello et al.[18] presented predictions of somewhat similar performance for 409 pharmaceuticals tested by GlaxoSmithKline (sensitivity = 46% and specificity = 69% for the Ames test). A predictive capacity in agreement with the one presented in the present paper was also found by Greene[10] for 972 proprietary Pfizer structures (sensitivity = 46% and specificity = 62% for the Ames test) and by Snyder et al.,[19] who studied 394 marketed pharmaceuticals. Finally, Pearl et al.[13] studied, separately, a set of 123 pharmaceutical drugs and another set of 516 nondrugs, in both cases with a resulting sensitivity higher than 80% and a specificity around 50%.

A graphical overview of the results of the validation studies with DEREKfW is provided in Figures 1 and 2. Figure 1 displays the percentage of false negatives versus false positives. These are conventionally used measures of the goodness of prediction. Figure 2 displays the Receiver Operating Characteristic (ROC) graphs, which have the advantage of summarizing and comparing, simultaneously, the performance of several systems. For example, the accuracy index alone does not distinguish between positive and negative predictions and is influenced by the performance on the most numerous class, whereas the axes of the ROC graphs display, independently, the information relative to the prediction of positive and negative chemicals. In a ROC plot, the true positive rate (or sensitivity) is plotted against the false positive rate (1 − specificity). According to the ROC curve theory, the diagonal line represents random responses; the top left corner is obviously the ideal performance, and the bottom right corner is the completely wrong performance. Thus, the most finely tuned systems are those in the left upper triangle, as close as possible to the corner.[20]

The combined inspection of Figures 1 and 2 shows that the predictive performance of DEREKfW, as resulting from our analysis, is average for carcinogenicity, whereas it is somewhat disappointing for its sensitivity to genotoxins. In particular, the prediction of DEREKfW for the genotoxicity of the present set of chemicals was characterized by the highest false negative rate (Figure 1) and, obviously, the lowest true positive rate (which is the complementary measure) (Figure 2).

The comparison of Figures 1 and 2 also indicates that a "general" performance value cannot be determined. The performance varies greatly from one data set to another. There is no way of knowing in advance the accuracy in a new data set, and this makes it impossible to decide the reliability of the predictions. The latter issue is clearly related to the definition of the applicability domain of the prediction

Table 4. Carcinogenicity and Genotoxicity Alerts Triggered in the Selected Pesticides, Together with the Number of Times that They Were Fired

	carcinogenicity: name of the alert (number of the alert)	number of times the alert was fired	number of times the alert was correctly fired	number of times the alert was incorrectly fired	number of references[a]	number of examples
1	aromatic amide (107)	9	4	5	1 (FDA)	0
2	substituted pyrimidine or purine (121)	8	3	5	15	4
3	polyhalogenated aromatic (116)	5	3	2	1 (FDA)	0
4	alkylating agent (73)	3	2	1	2	0
5	polycyclic aromatic hydrocarbon or heteroanalogue (113)	3	1	2	1 (FDA)	0
6	N-nitro or N-nitroso compound (70)	2	1	1	1	0
7	aromatic nitro compound (105)	2	0	2	12	4
8	halogenated alkene (123)	2	2	0	1 (FDA)	0
9	epoxide (72)	1	0	1	3	0
10	mono- or dialkylhydrazine (74)	1	0	1	3	3
11	aromatic hydroxylamine (106)	1	0	1	12	1
12	thioamide or thiourea (112)	1	0	1	1 (FDA)	0

	genotoxicity: name of the alert (number of the alert)[b]	number of times the alert was fired	number of times the alert was correctly fired	number of times the alert was incorrectly fired	number of references	number of examples
1	*substituted vinyl ketone (309)*	4	1	3	26	2
2	alkylating agent (27)	3	2	1	2	0
3	potentially labile halogen (4)	2	1	1	0	0
4	N-nitro or N-nitroso compound (7)	2	1	1	1	0
5	alkyl aldehyde or precursor (306)	2	2	0	3	0
6	*alkyl carbamate (308)*	2	1	1	5	0
7	aromatic nitro compound (329)	2	0	2	3	0
8	halogenated alkene (331)	2	0	2	5	0
9	aromatic amine or amide (352)	2	1	1	17	4
10	epoxide (19)	1	0	1	2	0
11	*haloacetanilide or analogue (61)*	1	0	1	2	3
12	mono- or dialkylhydrazine (28)	1	0	1	7	4
13	*α,β-unsaturated ester or thioester—class II or class III (361)*	1	0	1	5	0

[a] FDA: alerts derived from a FDA publication (ref 21). [b] Alerts for chromosome damage are noted in italics.

system, for example, to which types of chemicals it can be applied with satisfactory reliability. As explained in the methodology section, DEREKfW provides the domain of its alerts; the results presented here indicate that this aspect should be improved further. However, it should be noted that this is a general problem of the prediction software and is not limited to DEREKfW only.[5]

Frequency of the Alerts. Table 4 presents the 12 carcinogenicity alerts and 13 genotoxicity alerts triggered in the selected pesticides. These numbers indicate that the majority of the alerts defined in DEREKfW for carcinogenicity and genotoxicity have not been found in our study. They are, therefore, not considered by this validation exercise, and their adequacy for predictions cannot be judged. The analysis of more compounds would enable us to consider more alerts and, thereby, to carry out a more complete validation exercise. Interestingly, only 4 of the 13 genotoxicity alerts presented in Table 4 refer to chromosome damage (alert numbers 61, 308, 309, and 361). All the other alerts are gene mutation alerts.

The total number of times that an alert is fired is indicated in Table 4, together with the number of times that an alert is correctly or incorrectly fired. Eight out of the 12 alerts triggered for carcinogenicity and 7 out of the 12 alerts triggered for genotoxicity were fired more often incorrectly than correctly. The alerts "aromatic amide" and "substituted pyrimidine or purine" are the two most commonly triggered alerts for carcinogenicity. They are present in 26% (9/35) and 23% (8/35), respectively, of the pesticides presenting at least one carcinogenicity alert. The alerts "substituted vinyl ketone" and "alkylating agent" are the most commonly triggered for genotoxicity. They are present in 21% (4/19) and 16% (3/19) of the pesticides presenting at least one alert for this effect.

The carcinogenic alert numbers 72, 74, 105, 106, and 112 (the epoxide, mono- or dialkylhydrazine, aromatic nitro compound, aromatic hydroxylamine, thioamide, or thiourea alerts) appear to be particularly conservative, since all pesticides with these alerts are false positives. However, the number of times they appear in the chemicals is too low (one or two occurrences) to draw reliable conclusions. Interestingly, the alert for "aromatic amide" was the most frequently triggered alert in both the true positives and the false positives. This shows the difficulty of applying a rule-based approach, which can point to the presence of alerting chemical functionalities but has difficulty in modulating the carcinogenic potential within each of the toxicophore classes. In other words, the simple alerts only classify a compound into a chemical class that previous research has shown to be statistically and mechanistically related to toxic effects. However, the effect of a chemical functionality (here, alert) strongly depends on the rest of the molecule. For example, chemicals with a very high molecular weight and large size have little chance of being absorbed in significant amounts, and highly hydrophilic compounds are poorly absorbed and, if absorbed, are readily excreted; in this way, the effect of the potentially toxic alert is counterbalanced, and the chemical is detoxified. It can be envisaged that

the (quantitative) consideration of the modulating factors (e.g., neighboring atoms, physical chemical properties) can substantially improve the performance of the system.

Table 4 shows that, while some alerts are well-referenced and illustrated with different examples, only a few references and no examples are provided for the majority of the alerts. There is even a genotoxicity alert without reference and without any illustrating examples, making it difficult to understand its significance. The transparency of the triggered alerts is, thus, unequal and may be limited. Almost half of the carcinogenicity alerts originate from a report of the Food and Drug Administration (FDA) published in 1986.[21] These FDA alerts are generally poorly described and are not illustrated by examples. Furthermore, the broad definition of some of these alerts may lead to more false positive predictions of the carcinogenic potential in comparison to other DEREKfW alerts.[17]

Priority Setting. A priority setting based on DEREKfW predictions found in our validation exercise can be suggested. Two different starting points can be distinguished. First, if a compound presents no structural alert for carcinogenicity, it can be assumed that the compound has a low carcinogenic potential in rodents. The level of concern is, therefore, relatively low for such a compound but is not zero since false negatives can be found for compounds without carcinogenicity alerts (six false negatives in our analysis, five of them with a low potency). A compound predicted negative for carcinogenicity but presenting genotoxicity alerts should receive attention, since it may be a false negative. In our validation exercise, four pesticides presented such characteristics and two of them were false negatives (compounds 26 and 42 in Table 1). Looking at the genotoxic prediction for compounds without structural alerts for carcinogenicity may, therefore, help to decrease the number of false negatives (just luckily for epigenetic carcinogens) but can also decrease the specificity of the analysis. Second, if a compound presents one or more structural alerts for carcinogenicity, it can be suspected to induce tumors in rodents. Since DEREKfW has been found in the present evaluation to be characterized by quite a high rate of false positives for carcinogenicity, this suspicion may not be confirmed in rodent bioassays or may be downgraded by looking at additional data such as subchronic toxicity data or carcinogenicity testing results on structural analogues. It was hoped that knowledge of the genotoxicity potential could help to specify the level of concern, on the basis of the knowledge that genotoxic carcinogens are usually assumed to present no threshold and often affect many organs in many species, thereby causing more concern than epigenetic carcinogens. However, not much information can be gained from the genotoxicity predictions since the genotoxicity alerts of DEREKfW mainly refer to in vitro genotoxicity and not to in vivo genotoxicity. Finally, it is important to stress that this priority setting has been established on the basis of the examination of a set of pesticides. Differences in the mode of action between pesticides and other chemicals are significant. This may have consequences when trying to prioritize compounds other than pesticides.

5. CONCLUSION

We investigated the application of the rule-based system DEREKfW to qualitatively predict the rodent carcinogenic potential of a set of 60 pesticides. Predictions were compared with experimental bioassay data to evaluate the performance of DEREKfW. The percentage of false negatives was found to be 31% for carcinogenicity. The associated sensitivity of 69% indicates that most of the pesticides with positive rodent bioassay results were detected by DEREKfW. The low specificity of 47%, however, indicates that many pesticides may be flagged as carcinogenic while rodent bioassays would not confirm this potential. DEREKfW is, thus, reasonably successful at detecting carcinogens, at the cost of a high percentage of false positives. While providing a high margin of safety for human health is a must for a regulatory agency, it may lead to unnecessary testing with high extra costs for industry or to the unnecessary restriction of a chemical.

Our results indicate that DEREKfW offers interesting perspectives and could, for instance, be used for hazard screening, priority setting, or to get insight into the mode of action. This technology, however, has not yet reached a level that would enable us to make regulatory decisions about the carcinogenic hazard of individual chemicals only on the basis of an expert system like DEREKfW. The sensitivity and the selectivity of the prediction must be improved in order to guarantee a sufficient protection level for human health and to prevent unnecessary animal testing. From that perspective, industry and regulatory agencies should define when a model is sufficiently predictive (how good is good enough) and agree on acceptable values for both the sensitivity and the specificity. Animal testing will still be required as part of the registration process if the false negative results are not low enough. Finally, further refinements and optimizations of DEREKfW are needed to enhance its performance and, thus, its acceptance by decision makers. Additional validations by independent bodies are required to better characterize the predictive performance. With this aim, we intend to perform a more comprehensive validation of DEREKfW in the future, by examining a larger number of pesticides. Other types of compounds may also be selected in this future exercise, since pesticides are not representative of the chemical universe. A comparison of DEREKfW with other (Q)SARs is also planned and should help to assess the benefits of using a battery of predictive models.

ACKNOWLEDGMENT

The author would like to thank Edith Laraway, Carol Marchant, and Karen Young from LHASA Limited for their full support and valuable advice. Also, thanks to Etje Hulzebos, Iain Purchase, Ann Richard, and Julian Preston for their valuable comments on an earlier draft of this paper. Cecilia Bossa is acknowledged for her contribution to the revision of the manuscript.

REFERENCES AND NOTES

(1) Rhomberg, L. Risk assessment and the use of information on underlying biologic mechanisms: a perspective. *Mutat. Res.* **1996**, *365*, 175−89.

(2) Cronin, M. T. D.; Jaworska, J. S.; Walker, J. D.; Comber, M. H. I.; Watts, J. D.; Worth, A. P. Use of QSARs in International Decision-

Making Frameworks to Predict Health Effects of Chemical Substances. *Environ. Health Perspect.* **2003**, *111*, 1391−401.
(3) Jaworska, J. S.; Comber, M. H. I.; Auer, C. M.; van Leeuwen, C. J. Summary of a Workshop on Regulatory Acceptance of (Q)SARs for Human Health and Environmental Endpoints. *Environ. Health Perspect.* **2003**, *111*, 1358−60.
(4) Richard, A. M.; Benigni, R. AI and SAR approaches for predicting chemical carcinogenicity: survey and status report. *SAR QSAR Environ. Res.* **2002**, *13*, 1−19.
(5) Benigni, R. Structure−activity relationship studies of chemical mutagens and carcinogens: mechanistic investigations and prediction approaches. *Chem. Rev.* **2005**, *105*, 1767−1800.
(6) BgVV fordert Nachbesserungen bei der kuenftingen EU−Chemikalienpolitik; Bundesinstitut für gesundheitlichen Verbraucherschutz und Veterinärmedizin (BgVV): Berlin, 2001, Pressedienst.
(7) Sanderson. D. M.; Earnshaw, C. G. Computer prediction of possible toxic action from chemical structure: the DEREKFW system. *Hum. Exp. Toxicol.* **1991**, *10*, 261−71.
(8) *DEREKfW for Windows, Version 7.0, User Guide*; Lhasa Ltd.: Leeds, U. K., 2003.
(9) Judson P. N.; Marchant, C. A.; Vessey, J. D. Using argumentation for absolute reasoning about the potential toxicity of chemicals. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1364−70.
(10) Greene, N. Computer systems for the prediction of toxicity: an update. *Adv. Drug Delivery Rev.* **2002**, *54*, 417−31.
(11) Hulzebos, E. M. Posthumus, R. (Q)SARs: Gatekeepers against risk on chemicals? *SAR QSAR Environ. Res.* **2003**, *14*, 285−316.
(12) (Q)SARs: Evaluation of the commercially available software for human health and environmental endpoints with respect to chemical management applications; Technical Report No. 89.; ECETOC: Brussels, Belgium, 2003.
(13) Pearl, G. M. Livingstone-Carr, S.; Durham, S. K. Integration of computational analysis as a sentinel tool in toxicologic assessments. *Curr. Top. Med. Chem.* **2001**, *1*, 247−55.
(14) Parry, J. M. Detecting and predicting the activity of rodent carcinogens. *Mutagenesis* **1994**, *9*, 3−5.
(15) Benigni, R. The first US National Toxicology Program exercise on the prediction of rodent carcinogenicity: definitive results. *Mutat. Res.* **1997**, *387*, 35−45.
(16) Benigni, R.; Zito, R. The second National Toxicology Program comparative exercise on the prediction of rodent carcinogenicity: definitive results. *Mutat. Res. Rev.* **2004**, *566*, 49−63.
(17) Marchant, C. A. Prediction of Rodent Carcinogenicity Using the DEREKFW system for 30 chemicals Currently Being Tested by the National Toxicology Program. *Environ. Health Perspect.* **1996**, *104*, 1065−74.
(18) Cariello, N. F.; Wilson, J. D.; Britt, B. H.; Wedd, D. J.; Burlinson, B.; Gombar, V. K. Comparison of the computer programs DEREKFW and Topkat to predict bacterial mutagenicity. *Mutagenesis* **2002**, *17*, 321−9.
(19) Snyder, R. D.; Pearl, G. M.; Mandakas, G.; Choy, W. N.; Goodsaid, F.; Rosenblum, I. Y. Assessment of the sensitivity of the computational programs DEREKFW, TOPKAT, and MCASE in the prediction of the genotoxicity of pharmaceutical molecules. *Environ. Mol. Mutagen.* **2004**, *43*, 143−58.
(20) Provost, F.; Fawcett, T. Robust classification for imprecise environment. *Machine Learn. J.* **2001**, *42*, 5−11.
(21) General principles for evaluating the safety of compounds used in food-producing animals, Appendix 1: Carcinogen structure guide; U.S. Food and Drug Adminstration (FDA): Washington, DC, 1986.

CI050150Z