Research Article

Testing computational toxicology models with phytochemicals

Luis G. Valerio, Jr.[1], *Kirk B. Arvidson*[2], *Emily Busta*[2], *Barbara L. Minnier*[1,3], *Naomi L. Kruhlak*[1] *and R. Daniel Benz*[1]

[1] Informatics and Computational Safety Analysis Staff, Science and Research Staff, Office of Pharmaceutical Science, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, USA
[2] Division of Food Contact Notifications, US Food and Drug Administration, Center for Food Safety and Applied Nutrition, Office of Food Additive Safety, College Park, MD, USA
[3] GlobalNet Services, Inc., Rockville, MD, USA

Computational toxicology employing quantitative structure–activity relationship (QSAR) modeling is an evidence-based predictive method being evaluated by regulatory agencies for risk assessment and scientific decision support for toxicological endpoints of interest such as rodent carcinogenicity. Computational toxicology is being tested for its usefulness to support the safety assessment of drug-related substances (*e.g.* active pharmaceutical ingredients, metabolites, impurities), indirect food additives, and other applied uses of value for protecting public health including safety assessment of environmental chemicals. The specific use of QSAR as a chemoinformatic tool for estimating the rodent carcinogenic potential of phytochemicals present in botanicals, herbs, and natural dietary sources is investigated here by an external validation study, which is the most stringent scientific method of measuring predictive performance. The external validation statistics for predicting rodent carcinogenicity of 43 phytochemicals, using two computational software programs evaluated at the FDA, are discussed. One software program showed very good performance for predicting non-carcinogens (high specificity), but both exhibited poor performance in predicting carcinogens (sensitivity), which is consistent with the design of the models. When predictions were considered in combination with each other rather than based on any one software, the performance for sensitivity was enhanced, However, Chi-square values indicated that the overall predictive performance decreases when using the two computational programs with this particular data set. This study suggests that complementary multiple computational toxicology software need to be carefully selected to improve global QSAR predictions for this complex toxicological endpoint.

Received: June 4, 2009
Revised: August 6, 2009
Accepted: August 8, 2009

Keywords:
Computational toxicology / Phytochemicals / Quantitative structure–activity relationship / Rodent carcinogenicity / Safety assessment

1 Introduction

Phytochemicals are widely prevalent in normal dietary sources and can be found in many regulated consumer products. Phytochemicals are natural products and are present as constituents in conventional foods, components of natural mixtures (*e.g.* flavoring agents, botanicals) used as food ingredients, and botanical extracts used as ingredients

Correspondence: Dr. Luis G. Valerio, Jr., Informatics and Computational Safety Analysis Staff, Science and Research Staff, Office of Pharmaceutical Science, Center for Drug Evaluation and Research, US Food and Drug Administration, White Oak 51 Room 4128, 10903 New Hampshire Ave., Silver Spring, MD 20993-0002, USA

E-mail: Luis.Valerio@fda.hhs.gov
Fax: +1-301-796-9997

Abbreviations: χ^2, Chi-square; **LMA,** Leadscope Model Applier; **QSAR,** quantitative structure–activity relationship in dietary supplements and botanical drug products. Unfortunately, a common problem with these substances is the lack of toxicology data that are useful for evaluating the safety of chronic human exposure. Chronic toxicity of a chemical is often pivotal evidence for regulatory decision-making on the safety of the product in which the chemical is present. The carcinogenicity endpoint is among the most important chronic toxicities used to assess risk for human exposure to chemicals and in safety evaluations of regulated products. Regulatory guidance recommends the use of 2-year rodent carcinogenicity studies in two species and sexes to support the safety of US Food and Drug Administration (FDA) regulated products (http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm078924.pdf [1], and there is internationally harmonized guidance (http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm074911.pdf) [2]. Evidence that a chemical is a rodent carcinogen can adversely impact the regulatory approval of products and knowledge of carcinogenic potential is of great importance to protecting public health [3]. Although the need for rodent carcinogenicity study data for a chemical is great, relatively few substances, especially phytochemicals, have been tested for carcinogenicity. There are several practical reasons for the lack of carcinogenicity test data, including the exorbitant financial cost ($3 million for species/sex standardized study), intensive resources (experts and review) and the long period of time required to conduct the study according to standardized protocols such as those described in US FDA guidance documents [1]. These challenges not withstanding, assessing cancer risk for chemicals humans are exposed to in regulated products is important in the context of hazard and risk characterization, which may lead to regulatory action, and for prioritization of these substances for further study in the interest of protecting public health [3–5].

In silico methods have been proposed as a way to predict both efficiently and accurately the outcome of rodent carcinogenicity studies on the scientific basis of structure–activity relationships [6]. The use of *in silico* methods is now supported in the EU by enacted legislation in reaction to public desire to reduce the use of animals in testing [7]. Moreover, these methods have been recommended by the US National Research Council [8], and are considered to be useful in setting testing priorities [8, 9]. *In silico* models of rodent carcinogenicity using quantitative structure–activity relationship (QSAR) analyses of phytochemicals have been previously reported to be a predictive tool indicating some degree of promise for predicting naturally occurring carcinogens derived from plants [6, 10]. At the US FDA Center for Drug Evaluation and Research (CDER), Office of Pharmaceutical Science (OPS), Informatics and Computational Safety Analysis Staff (ICSAS), and the Center for Food Safety and Applied Nutrition (CFSAN), Office of Food Additive Safety (OFAS), Division of Food Contact Notifications (DFCN), the use of computational toxicology software programs is being employed to help support regulatory decision-making in the safety evaluation of human pharmaceuticals, their metabolites, and impurities, and indirect food additives [6, 9–14]. Given the aforementioned rationale for regulatory use of *in silico* approaches, its potential application with diverse sets of chemicals, and the data poor situation for phytochemicals in terms of available chronic toxicology study data, the use of *in silico* modeling for predicting rodent carcinogenicity of these substances is of significant interest. The purpose of the present study was to examine the utility of *in silico* QSAR-based tools in current use at the FDA (ICSAS and DFCN) for predicting rodent carcinogenicity by employing two computational software programs to screen a data set of 43 phytochemicals. Predictive performance was validated by comparing computational predictions with empirical data. This analysis is considered to be an external validation study because the phytochemicals that were screened *in silico* were not used to construct the predictive QSAR models used by the computational software.

2 Materials and methods

2.1 Computer hardware and computational software programs

The computer hardware used for the computational toxicology software programs in this study was a PC with Microsoft Windows XP Professional version 2002. The computational software programs used were Leadscope Model Applier (LMA) version 1.0, available from Leadscope (www.leadscope.com) [15], and MC4PC version 2.1.0.11, available from MultiCASE (www.multicase.com) [16]. The following seven QSAR rodent carcinogenicity models were employed with the LMA in the Rodent Carcinogenicity suite: carc mouse, carc mouse female, carc mouse male, carc rat, carc rat female, carc rat male, and in the Miscellaneous Toxicity Endpoint Suite: carc rodent. The QSAR prediction paradigm used in the LMA software has been described previously [17]. The following seven QSAR rodent carcinogenicity models were employed with the MC4PC software: carcinogenicity – rodents (+proprietary) AGU, carcinogenicity – rats (+proprietary) AGV, carcinogenicity – male rats (+proprietary) AG1, carcinogenicity – female rats (+proprietary) AG2, carcinogenicity – mice (+proprietary) AGW, carcinogenicity – male mice (+proprietary) AG3, carcinogenicity – female mice (+proprietary) AG4.

Details on how the carcinogenicity models were built (including the acceptance criteria for rodent carcinogenicity data, classification, stratification, and scoring of rodent tumor findings) have been described previously [17, 18]. The predictive paradigm for the QSAR computational software has also been described previously [17].

2.2 Data set

Forty-three phytochemicals with rodent carcinogenicity studies were identified from sources in the public domain as described previously [6]. Briefly, these substances are organic molecules with a known activity or inactivity for rodent carcinogenicity. In addition, a data set of ten synthetic chemicals with known rodent carcinogenic activity was added to the test to increase the robustness of the study by adding statistical power to the analysis of predictive performance. The synthetic chemicals were a group of constituents obtained from CFSAN files that are known carcinogenic chemicals [6]. All molecules described in this study were non-proprietary and two-dimensional molecular structures may be found in the public domain at the US National Library of Medicine's PubChem web site (http://pubchem.ncbi.nlm.nih.gov/).

2.3 Assessment of experimental evidence for rodent carcinogenicity and external validation experiment

The assessment of experimental evidence for rodent carcinogenicity was performed using published chronic bioassay data and conclusions published in the US National Toxicology Program, International Agency for Research on Cancer, Priority-based Assessment of Food Additives, and Gold Carcinogenicity Potency Database resources. Because of the small number of phytochemicals with rodent carcinogenicity study data, the carcinogenicity of the phytochemicals and synthetic chemicals was evaluated irrespective of target organs for tumor formation. Any positive or negative finding from the published literature of carcinogenicity studies was based on the author's opinion in the published paper. Phytochemicals lacking chronic carcinogenicity studies were excluded from this study.

In order to perform the external validation study, the predictive performance of the rodent carcinogenicity QSAR models was tested with a set of 43 phytochemicals, comprising 24 active (high carcinogenic potential) and 19 inactive (low carcinogenic potential) molecules, and a set of ten synthetic chemicals to increase the robustness of the statistical analysis. None of the 53 external validation chemicals were ever part of the QSAR model training data sets.

2.4 Statistical analysis of predictive performance

The QSAR model performance for predicting the carcinogenicity endpoint was calculated according to the method of Cooper *et al.* [19]. The parameters include sensitivity, specificity, positive predictivity, negative predictivity, false positives, false negatives, concordance, and Chi-square (χ^2). The sensitivity index is defined as the percentage of correctly predicted carcinogens from the total number of carcinogens. Specificity is defined as the percentage of correctly predicted non-carcinogens from the total number of non-carcinogens. Positive predictivity is defined as the percentage of correctly predicted carcinogens from the total number of positive predictions from the test, and negative predictivity is defined as the percentage of correctly predicted non-carcinogens from the total number of negative predictions from the test. The false positives represent the percentage of incorrectly classified non-carcinogens from the total number of non-carcinogens. The false negative parameter represents the percentage of incorrectly classified carcinogens from the total number of carcinogens. Concordance is defined as the percentage of correctly predicted non-carcinogens and carcinogens from the total number of chemicals tested. The χ^2 test is an overall measure of the association between the predicted and experimental results for the test chemicals, and is used to express the degree of association relative to a confidence level (*p*-value). Coverage is a measure of the percentage of chemicals in the test set for which the model is able to make a prediction, and is a reflection of the relationship of the molecular diversity of the training data set with the test chemical being screened by the computational software program. Test chemicals containing structural features not well represented in the training data set are considered to be uncovered and, consequently, no prediction call can be made for them and these outputs by the computational software are referred to as "no call".

2.5 Combining predictions from two software programs

Predictions from both MC4PC and LMA were combined to assess whether the overall predictive performance improved as compared with that of any single program. The following protocol was used: if at least one of the two software programs predicted a test chemical (*i.e.* phytochemical) to be positive then the molecule was given an overall positive call. The only situation leading to an overall negative call was consequently when both software programs gave negative (consensus) predictions. In cases where a test chemical was covered in only one of the two software programs, the prediction from the single software program became the overall call. In the single case where a compound was uncovered by both software, no overall call was made (*i.e.* "no call"). The χ^2 test was used to determine whether the degree of association between the combined predictions and experimental calls was higher or lower than that of each software alone.

3 Results

The 43 phytochemicals used for this validation study are presented in Table 1, along with their respective natural

Table 1. Forty-three phytochemicals tested with computational toxicology software programs in external validation

Phytochemical	Natural occurrence
1′-Hydroxyestragole	Basil, *Ocimum basilicum;* metabolite of estragole
2-Ethyl-1-hexanol	Sassafras, *Sassafras albidum;* metabolite of safrole
5-Methoxypsoralen	Parsley, *Petroselinum sativum*
6-Methylcoumarin	Oregano, *Origanum vulgare*
Capsaicin	Hot peppers, *Capsicum annum*
Dehydromonocrotaline	Russian comfrey, *Symphytum uplandicum*
Estragole	Basil, *Ocimum basilicum*
Heliotrine	Russian comfrey, *Symphytum uplandicum*
Vanillin	Vanilla, *Vanilla planifolia*
Ptaquilosin	Bracken fern, *Pteridum aquilinum*
Ptaquilosin (APT) dienone	Bracken fern, *Pteridum aquilinum*
Hydroxysenkirkine	Medicinal herb, *Crotalaria laburnifolia*
4-Methylphenylhydrazine	Edible mushroom, *Agaricus bisporus*; metabolite
Allyl hexanoate	Tea tree oil, *Melaleuca alternifolia*; flavor
Anethole	Fennel, *Foeniculum vulgare*
β-Apo-8′-carotenal	Carrot, *Paucus carota*
Citrate	Lemon, *Citrus limon*
Crotonaldehyde	Potato, *Solanum tuberosum*
Curcumin	Turmeric, *Curcuma longa*
Epicatechin	Green tea, *Camellia sinensis*
Formic acid	Carrot, *Paucus carota*
Gallic acid	Mango, *Mangifera indica*
Indole	Corn, *Zea mays*
Indole-3-acetic acid	Strawberry fruit, *Fragaria vesca*
Linalool	Apricots, *Prunus armeniaca*
Lipoic acid	Spinach, *Spinacia oleracea*
Maltol	Roasted coffee, *Coffea Arabica*
Piperonal	Vanilla, *Vanilla planifolia*
Piperine	Black pepper, *Piper nigrum*
Intermedine	Comfrey, *Symphytum Officinale*
Isosafrole	Oil of Sassafras, *Sassafras albidum*
Jacobine	Ragwort herb, *Senecio jacobaea*
Lycopsamine	Comfrey, *Symphytum Officinale*
Methylglyoxal	Roasted coffee, *Coffea arabica*
Parasorbic acid	Rowan berry, *Sorbus aucubaria*
Propionic acid	Tomato, *Lycopersicon esculentum*
Retronecine	Medicinal herb, *Crotalaria laburnifolia*
Senecionine	Ragwort herb, *Senecio jacobaea*
Seneciphyllinine	Ragwort herb, *Senecio jacobaea*
Tannic acid	Tea, *Camellia sinensis*
Hydroxymethylphenylhydrazine	Edible mushroom, *Agaricus bisporus*
1-Octacosanol	Perilla seeds, *Perilla frutescens*
Protocatechuic acid	Shallot onions, *Allium cepa*

sources, and the set of ten synthetic chemicals, which are known rodent carcinogens, are listed in Table 2. Together these 53 compounds were run as a test set against the MC4PC and LMA computational software models in order to assess each program's respective performance in predicting rodent carcinogenicity.

The predictive performance statistics for MC4PC on the 53 natural and synthetic compounds are summarized in Table 3. MC4PC made correct predictions for 94% of the phytochemicals with experimentally determined low risk for rodent carcinogenicity. This specificity value is considered to be of high performance. For phytochemicals with experimentally determined high risk for inducing carcinogenicity in rodents, the program predicted only 47% of the compounds correctly. This sensitivity value is considered of low or poor performance. Concordance for accurate predictions was 63%, a marginal value overall since it is only slightly above 50%. However, a low false positive rate and high positive predictivity (94%) was observed. A total of four phytochemicals were not in the domain of applicability of the MC4PC QSAR training set, and thus a prediction could not be made for these molecules (*i.e.* no call). The χ^2 test gave a value of 8.4835 with $p = 0.0036$, showing the overall predictive performance of this model to be highly statistically significant.

Table 2. Ten synthetic chemicals known to be rodent carcinogens tested with computational toxicology software programs in external validation

Synthetic chemicals
1,3-Dichloropropanol
3-Chloro-1,2-propanediol
4-Aminoazobenzene
4-Hydroxyphenylacetamide
Ammonium perfluorooctanoic acid
Diazoaminobenzene
Dibutyltin diacetate
Permanent orange
Quinoline
Tri-*n*-butyl phosphate

Table 3. External validation statistics for the MC4PC computational software program for predicting rodent carcinogenicity of 43 phytochemicals and ten synthetic dietary chemicals based on QSAR analysis of seven models

Performance parameters: MC4PC	Value
Coverage	92%
Specificity	94%
Sensitivity	47%
Concordance	63%
False positives	6%
False negatives	53%
Positive predictivity	94%
Negative predictivity	48%
χ^2	8.4835 ($p = 0.0036$)

Table 4. External validation statistics for the Leadscope Model Applier computational software program for predicting rodent carcinogenicity of 43 phytochemicals and ten synthetic dietary chemicals based on QSAR analysis of seven models

Performance parameters: Leadscope Model Applier	Value
Coverage	92%
Specificity	59%
Sensitivity	50%
Concordance	53%
False positives	41%
False negatives	50%
Positive predictivity	70%
Negative predictivity	38%
χ^2	0.3470 ($p = 0.5558$)

Table 4 summarizes the performance parameters for the LMA when assessing the 53 natural and synthetic compounds. The LMA was able to correctly predict 59% of the phytochemicals with low risk for rodent carcinogenicity and only 50% of the high-risk rodent carcinogens, leading to an overall concordance of 53%. Positive predictivity was higher at nearly 70% suggesting that when a positive prediction is made it will be made with a higher level of confidence for this class of chemicals; however, the negative predictivity was low at 38% suggesting that the opposite is true of negative predictions. A total of four phytochemicals were not in the domain of applicability of the LMA training set, and thus a prediction could not be made (*i.e.* no call). The χ^2 test resulted in a value of only 0.3470 with a p value $= 0.5558$, indicating a correlation between predicted and experimental values that is not statistically significant.

A comparison of the predictions made by MC4PC and LMA software is provided in Table 5. Comparing the predictions between MC4PC and LMA, 51% of the predictions were concordant between the two programs while 49% of the predictions were non-concordant between the two programs. Of the concordant predictions, 63% corresponded to low-risk predictions made by both of the programs. Within the non-concordant predictions, MC4PC made a higher percentage of correct low-risk rodent carcinogenicity predictions. However, LMA correctly predicted a greater number of high-risk calls (54%). An analysis of the predictive performance when combining the results was performed and is presented in Table 6. The protocol used for combining the results favored sensitivity by accepting any single positive as the basis for an overall positive call. This led to more positive calls and an increase in sensitivity (to 68%) as compared with the use of any one software program alone. Specificity slightly increased compared with one software program but by comparison decreased with the other program. However, the level of negative predictivity was maintained upon combining results due the consensus-derived negative calls. False positive and negative rates were comparable. The increase in sensitivity while maintaining negative predictivity when results from two different computational toxicology software are combined supports the use of multiple (two or more) computational software packages in applications where tolerance for risk is low. The association between experimental carcinogenicity and the combined predictions from the two software programs gave a χ^2 index of $\chi^2 = 3.9878$ with a p value of 0.0458. This indicates a statistically significant association, but the association is lower than that of MC4PC used alone.

4 Discussion

Human exposure to phytochemicals occurs on a daily basis, mainly through dietary sources. However, few chronic toxicity studies have been conducted for this class of chemicals, and this information is needed to help understand potential toxicological risks. Therefore, validated tools for risk assessment of these substances would be useful to help assess their safety. In the present study, the objective was to evaluate the

Table 5. Concordance between the MC4PC overall rodent call and the LMA maximum positive prediction for rodent carcinogenic potential for the 43 phytochemicals and ten synthetic chemicals

Phytochemical	Experimental evidence based carcinogenic risk	MC4PC overall rodent call	LMA overall rodent call	Concordance (√) or non-concordance (x) of MC4PC prediction with LMA prediction
1′-Hydroxyestragole	High	–	–	√
1-Octacosanol	Low	–	–	√
2-Ethyl-1-hexanol	High	+	–	x
4-Methylphenylhydrazine	High	+	+	√
5-Methoxypsoralen	High	+	+	√
6-Methylcoumarin	High	+	–	x
Allyl hexanoate	Low	–	–	√
Anethole	Low	–	+	x
β-Apo-8′carotenal	Low	No call	–	x
Capsaicin	High	+	–	x
Citrate	Low	–	–	√
Crotonaldehyde	High	–	No call	x
Curcumin	Low	+	+	√
Dehydromonocrotaline	High	+	No call	x
Epicatechin	Low	–	–	√
Estragole	High	+	–	x
Formic acid	Low	–	+	x
Gallic acid	Low	–	+	x
Heliotrine	High	–	+	x
Hydroxymethylphenylhydrazine	High	–	+	x
Hydroxysenkirkine	High	+	+	√
Indole	Low	–	No call	x
Indole-3-acetic acid	Low	–	–	√
Intermedine	High	–	+	x
Isosafrole	High	–	+	x
Jacobine	High	+	+	√
Linalool	Low	–	–	√
Lipoic acid	Low	–	–	√
Lycopsamine	High	–	+	x
Maltol	Low	–	–	√
Methylglyoxal	High	–	–	√
Parasorbic acid	High	–	–	√
Piperine	Low	–	+	x
Piperonal	Low	–	–	√
Propionic acid	High	–	–	√
Protocatechuic acid	Low	–	+	x
Ptaquilosin	High	–	–	√
Ptaquilosin (APT) dienone	High	No call	–	x
Retronecine	High	–	+	x
Senecionine	High	+	+	√
Seneciphyllinine	High	+	+	√
Tannic acid	Low	No call	No call	√
Vanillin	Low	–	+	x
1,3-Dichloropropanol	High	+	–	x
3-Chloro-1,2-propanediol	High	+	–	x
4-Aminoazobenzene	High	+	+	√
4-Hydroxyphenylacetamide	High	–	–	√
Ammonium perfluorooctanoic acid	High	–	+	x
Diazoaminobenzene	High	+	+	√
Dibutyltin diacetate	High	No call	–	x
Permanent orange	High	–	+	x
Quinoline	High	–	–	√
Tri-*n*-butyl phosphate	High	–	–	√

(–) indicates a negative prediction, low-rodent carcinogenic risk, and (+) indicates a positive prediction, high-rodent carcinogenic risk, while ''No call'' means the chemical is not in the domain of applicability of the training set data.

utility of *in silico* predictive global QSAR-based tools currently used at the FDA for predicting rodent carcinogenicity of a set of phytochemicals. Rodent carcinogenicity is the most commonly computationally modeled endpoint for several reasons, including cost- and time-efficiency as compared with that of a traditional 2-year rodent bioassay, and its prominence as pivotal evidence for regulatory risk assessors and safety evaluators. In this study, two different *in silico* QSAR-based tools were employed to screen an external validation set of phytochemicals and synthetic dietary compounds in the interest of assessing the predictive performance of the software programs and determining their usefulness as a substitute for *in vivo* testing. The results found excellent predictive performance with MC4PC in the context of predicting non-carcinogens (Table 3); however, the performance for predicting carcinogens was poor for both MC4PC and LMA, with sensitivity of approximately 50% (Tables 3 and 4). Examination of the internal validation statistics used to establish the models clearly demonstrates that performance was optimized for predicting non-carcinogens: the models were constructed to provide higher specificity with less concern for sensitivity or negative predictivity [17]. The best approach(s) in terms of model performance for sensitivity or specificity is still being evaluated at FDA/CDER and CFSAN. Thus, because of this fact and larger relative number of inactives in the QSAR training sets of these models, the expectation in predictive performance would be for high specificity not high sensitivity. The results of this study are consistent with other validation studies for these models [17, 18, 20]. Overall, concordance was not optimal for both software programs for predicting the known rodent carcinogenicity of the test set, but when results were combined between the two computational software programs, the analysis yielded higher sensitivity at the expense of specificity (Table 6). It could be argued that a larger training set of chemicals is needed in order to improve performance, or that there is a lack of appropriate descriptors in the QSAR programs that are needed to identify SAR relevant oncogenic activity of these molecules. In addition, lack of natural product representation in the training data set could be another point to consider; however, the training sets are considered “multipurpose” in that they contain pharmaceuticals, food additives, industrial chemicals, pesticides, and some natural products, so a lack of phytochemical representation in the model may not be an issue. Other considerations are that the software evaluated in this study use global QSARs, pooling knowledge of carcinogenicity by multiple mechanisms of action, in contrast to mechanism-specific local QSARs. The latter often show good predictive performance for test compounds within a narrow applicability domain but are limited when predictions are needed for a broad range of molecules.

Table 6. External validation statistics for the consensus predictions made by the LMA and MC4PC software programs combined for predicting rodent carcinogenicity of 43 phytochemicals and ten synthetic dietary chemicals based on QSAR analysis

Performance parameters: consensus	Value
Coverage	98%
Specificity	61%
Sensitivity	68%
Concordance	65%
False positives	39%
False negatives	32%
Positive predictivity	77%
Negative predictivity	50%
χ^2	3.9878 ($p = 0.0458$)

With respect to objectively assessing performance, the real dilemma in conducting external validation studies is that when new carcinogenicity studies are published, they are rapidly incorporated into predictive models rather than used in an external validation study. Thus, it is difficult to find a large data set of chemicals with toxicology study data that were never used to construct the QSAR model.

The calculated association between the carcinogenicity results and the predictions in this study, as measured by χ^2 values, indicated that one software program had statistically significant discriminatory power over the other. When the predictions from the two software programs were combined, the degree of association between experimental and predicted findings was still considered statistically significant but at a lower confidence level than when one of the software programs was used alone. While this may suggest that there is no benefit to combining predictions from the two software programs, the significant increase in sensitivity observed by applying this methodology supports its use in some situations, particularly in a regulatory environment where the goal is to minimize the risk to human health

A comparison of the predictions from the two software programs one to another found a 50% concordance for the test set, and within the set of concordant predictions, specificity remained high at 90% (Table 5). Over the entire test set, it was also notable that a high positive predictivity value was observed with both computational toxicology software programs either when assessing the results from the perspective of each individual program (Tables 3 and 4) or combined (Table 6), giving the user a high degree of confidence in a positive prediction. These results taken together with previous external validation studies with natural products using other computational software suggest the desirability to use multiple computational platforms for high or low throughput screening. This notion is supported by the outcome of a previous external validation study with phytochemicals where high sensitivity (97%) was achieved for predicting rodent carcinogenicity but poor specificity (53%) was observed [6]. One logical approach given the limitations of computational predictive modeling is to combine techniques that perform well for specificity with those computational methods that have demonstrated high sensitivity, and then perform a weight of evidence approach for using multiple predictions from different software programs.

To take this one step further, added confidence can be assigned when predictions from different software programs are in complete agreement, since the software are each interpreting molecular structure using different parameters and arriving at the same conclusion. Some studies have tested this possibility with *in silico* methods showing some success; however, only a few phytochemicals were examined [10], or the chemicals were not natural products [17]. The problem arises in interpreting conflicting predictions and there is no clear level of confidence to assign to the weight of evidence. In such a scenario, the solution may be that it is simply not solvable by the computational predictive software and based on current technology there is no way to address this conundrum.

From the perspective of regulating food additives, the 1958 amendment to the U.S. Food, Drug, and CosmeticAct states that "...no additive shall be deemed to be safe if it is found to induce cancer when ingested by man or animal, or if it is found, after tests which are appropriate for the evaluation of the safety of food additives, to induce cancer in man or animal,..." (Delaney Clause). Therefore, from a regulatory standpoint, it would be more desirable to better predict possible carcinogens with the aim of identifying them and keeping them out of the food supply. Under this paradigm, the poor ability to predict potential rodent carcinogens by an *in silico* method would not be acceptable. Thus, the low sensitivity (poor ability for predicting potential rodent carcinogens) of both MC4PC and the LMA when used by themselves is not ideal for regulatory purposes. However, given the software programs' higher sensitivity with combined or consensus predictions, and the higher specificity when used alone (better predictions for rodent non-carcinogens), use of these programs in drug discovery might be envisioned as a desirable predictive tool for filtering a high volume of molecules. This notion has also been recommended by a recent industry paper on the prediction of genotoxicity of human pharmaceuticals [21]. The study also asserted, however, that for predicting the potential of genotoxic impurities in the final drug product, a high sensitivity would be desired [21]. The whole issue of whether it is best to have a predictive model with higher specificity or sensitivity is paradoxical since arguments can be made for either, and it is likely most dependent upon the intended use of the *in silico* method and potential regulatory framework where it would be applied.

One feasible solution suggested by this study is to employ multiple computational platforms that are complementary in the sense that each takes different approaches to molecular structure interpretation. One advantage of such an approach is if one falls short in covering an area of chemical space another can assist in giving a needed prediction for that area, as seen between MC4PC and the LMA where only a single test chemical was uncovered by both programs (Table 5). The net result in this study was that combined coverage increased by 6%. Thus, the ability of one software package to complement another is an important consideration when screening chemicals.

The low predictive performance for sensitivity with both computational software programs can be related to the way the training sets were constructed for each model and so expectations may be tempered by model construction techniques. Perhaps with an expanded training set of chemicals the statistical parameter of specificity for the LMA might improve. The results of this study also indicate a need for the development of additional training sets emphasizing sensitivity, with follow-up studies to analyze their predictive potential within the regulatory framework. Depending on the anticipated application, screening of drug candidates or use in regulatory review for safety, one could select the most appropriate set of models, or combination of models, for a particular need. Computational toxicology methods could potentially serve as part of an *in silico* toolbox of techniques for prioritizing further testing of phytochemicals when used in combination with other evidences (*e.g.* structural alert classification schemes and empirical data). Further research needs to be done with larger training and test data sets containing different classes of compounds, like industrial chemicals, pharmaceuticals, *etc.* to further investigate and establish computational toxicology software-predictive performance. In addition, this study also suggests the need for construction of additional training sets emphasizing sensitivity in order to further evaluate their predictive potential within the regulatory frame for cases when consensus predictions are not adopted.

In conclusion, the application of QSAR-predictive computational models for rodent carcinogenic activity of phytochemicals is found to be useful for predicting non-carcinogens and, if a consensus approach is adopted, an improvement in sensitivity for predicting carcinogens is achieved. The deployment of multiple computational platforms may well be the optimal way of utilizing this type of predictive information in order to best fit into the safety and risk assessment paradigms in which these substances are encountered; however, the software used for this approach must be carefully selected to ensure complementarity in predictive performance. Further research is needed to help make this determination.

The authors acknowledge Dr. Edwin Matthews for building the QSAR models.

This research report is not an official US Food and Drug Administration guidance or policy statement. No official support or endorsement by the US Food and Drug Administration is intended or should be inferred.The authors have declared no conflicts of interest.

5 References

[1] FDA, Guidance for industry: Carcinogenicity study protocol submissions. FDA Center for Drug Evaluation and Research (CDER), *Department of Health and Human Services* 2002.

[2] ICH, Guidance for industry: the need for long-term carcinogenicity studies of pharmaceuticals, ICH1A, 1996.

[3] Jacobs, A., Jacobson-Kram, D., Human carcinogenic risk evaluation, Part III: assessing cancer hazard and risk in human drug development. *Toxicol. Sci.* 2004, *81*, 260–262.

[4] Contrera, J. F., Jacobs, A. C., DeGeorge, J. J., Carcinogenicity testing and the evaluation of regulatory requirements for pharmaceuticals. *Regul. Toxicol. Pharmacol.* 1997, *25*, 130–145.

[5] Jacobs, A., Prediction of 2-year carcinogenicity study results for pharmaceutical products: how are we doing? *Toxicol. Sci.* 2005, *88*, 18–23.

[6] Valerio, L. G., Jr., Arvidson, K. B., Chanderbhan, R. F., Contrera, J. F., Prediction of rodent carcinogenic potential of naturally occurring chemicals in the human diet using high-throughput QSAR predictive modeling. *Toxicol. Appl. Pharmacol.* 2007, *222*, 1–16.

[7] EU, Regulation (EC) No 1907/2006 of The European Parliment and of the Council concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. *Official Journal of the European Union* 2006, *L396*, 1–843.

[8] NRC, *Toxicity Testing in the 21st Century: A Vision and a Strategy*, National Academy Press, Washington, D.C. 2007.

[9] Bailey, A. B., Chanderbhan, R., Collazo-Braier, N., Cheeseman, M. A., Twaroski, M. L., The use of structure-activity relationship analysis in the food contact notification program. *Regul. Toxicol. Pharmacol.* 2005, *42*, 225–235.

[10] Arvidson, K. B., Valerio, L. G., Diaz, M., Chanderbhan, R. F., In Silico toxicological screening of natural products. *Toxicol. Mech. Methods* 2008, *18*, 229–242.

[11] Yang, C., Benz, R. D., Cheeseman, M. A., Landscape of current toxicity databases and database standards. *Curr. Opin. Drug Discov. Dev.* 2006, *9*, 124–133.

[12] Matthews, E. J., Contrera, J. F., In silico approaches to explore toxicity end points: issues and concerns for estimating human health effects. *Exp. Opin. Drug Metab. Toxicol.* 2007, *3*, 125–134.

[13] Kruhlak, N. L., Contrera, J. F., Benz, R. D., Matthews, E. J., Progress in QSAR toxicity screening of pharmaceutical impurities and other FDA regulated products. *Adv. Drug Deliv. Rev.* 2007, *59*, 43–55.

[14] Mayer, J., Cheeseman, M. A., Twaroski, M. L., Structure-activity relationship analysis tools: validation and applicability in predicting carcinogens. *Regul. Toxicol. Pharmacol.* 2008, *50*, 50–58.

[15] Yang, C., Hasselgren, C. H., Boyer, S., Arvidson, K. *et al.*, Understanding genetic toxicity through data mining: the process of building knowledge by integrating multiple genetic toxicity databases. *Toxicol. Mech. Methods* 2008, *18*, 277–295.

[16] Saiakhov, R. D., Klopman, G., MultiCASE Expert Systems and the REACH Initiative. *Toxicol. Mech. Methods* 2008, *18*, 159–175.

[17] Matthews, E. J., Kruhlak, N. L., Benz, R. D., Contrera, J. F. *et al.*, Combined use of MC4PC, MDL-QSAR, BioEpisteme, Leadscope PDM, and Derek for Windows Software to achieve high-performance, high-confidence, mode of action-based predictions of chemical carcinogenesis in rodents. *Toxicol. Mech. Methods* 2008, *18*, 189–206.

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

[19] Cooper 2nd, J. A., Saracci, R., Cole, P., Describing the validity of carcinogen screening tests. *Br. J. Cancer* 1979, *39*, 87–89.

[20] Contrera, J. F., Kruhlak, N. L., Matthews, E. J., Benz, R. D., Comparison of MC4PC and MDL-QSAR rodent carcinogenicity predictions and the enhancement of predictive performance by combining QSAR models. *Regul. Toxicol. Pharmacol.* 2007, *49*, 172–182.

[21] Snyder, R. D., An update on the genotoxicity and carcinogenicity of marketed pharmaceuticals with reference to in silico predictivity. *Environ. Mol. Mutagen.* 2009, *50*, 435–450.