Prediction of Chemical Carcinogenicity from Molecular Structure

Hongmao Sun*

Discovery Chemistry, Hoffmann-La Roche Inc., 340 Kingsland Street, Nutley, New Jersey 07110

Received March 2, 2004

Carcinogens represent a serious threat to human health. In vivo determination of carcinogenicity is time-consuming and expensive, thus in silico models to predict chemical carcinogenicity are highly desirable for virtual screening of compound libraries of both pharmaceutically and other commercially interesting molecules. In the present study, a PLS-DA (partial least squares discriminant analysis) model was developed to predict carcinogenicities in each of four rodent models: male *mouse* (MM), female *mouse* (FM), male *rat* (MR), and female *rat* (FR). The data set that was used contained over 520 compounds from both the NTP and the FDA databases. All the models were built from the same molecular descriptor system, which is based on atom typing [Sun, H. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 748−757], enabling the comparison of atomic contributions to carcinogenicity with respect to species and gender. Using four components, the models were able to achieve excellent fitting and prediction, with $r^2 = 0.987$ and $q^2 = 0.944$ for MM, $r^2 = 0.985$ and $q^2 = 0.950$ for FM, $r^2 = 0.989$ and $q^2 = 0.962$ for MR, and $r^2 = 0.990$ and $q^2 = 0.965$ for FR. The models were further validated by response permutation testing and external validation, and the results indicated that the models were both statistically significant and predictive. Variable influence on projection (VIP) analysis identified the key atom types and fragments that contributed to carcinogenicities and response differences across species and gender.

INTRODUCTION

Carcinogens represent a serious threat to human health. There are more than 80 000 chemicals registered for use in commerce in the United States, and an estimated 2000 new ones are introduced annually for use in everyday items such as foods, personal care products, prescription drugs, household cleaners, and lawn care products.[2] It is desirable to carefully assess the carcinogenicity of all the chemical substances to which people may be exposed, but accurate measurements of this property are laborious, time-consuming, and costly. According to regulatory guidelines of the Food and Drug Administration (FDA), carcinogenicity is required to be evaluated against multiple biological models.[3] Various shorter-term, less costly screening methods[4] have been applied to predict chemical carcinogenicity. The most common of these is the Ames assay,[5] an in vitro bacterial assay method using several strains of *Salmonella typhimurium*, which is based on the principle that mutation is an essential effect for cancer induction.[6,7] However, the concordance of *Salmonella* mutagenicity with rodent carcinogenicity has been found to be as low as 60%.[8]

Chemical carcinogenesis is an important subject not only for environmental hazard assessment but also for drug discovery. Drug discovery is an optimization process, during which a large number of compounds are synthesized and tested. Most of the compounds in this process are ultimately filtered out because of their undesirable properties. To quickly focus on the molecules most likely to succeed, compounds with adverse properties, such as carcinogenicity, should be removed from the pipeline at an early stage. New technologies adopted by the pharmaceutical industry, such as high throughput screening (HTS) and combinatorial chemistry, have dramatically increased the number of compounds entering the drug discovery pipeline and significantly speeded up the preclinical drug discovery process. This has made it impractical to experimentally evaluate the carcinogenicity of every compound of interest. Thus, a reliable tool for predicting carcinogenicity would be highly desirable. Also, such a prediction would aid in prioritizing compounds for possible synthesis that have been proposed via design or virtual screening.

To date, toxicology modeling is mainly rule-based or structure-based. Rule-based approaches, such as the program DEREK (deductive estimate of risk from existing knowledge),[9] provide flags indicating whether a specific toxic response may occur, on the basis of rules derived by experts from the study of a large experimental data set. DEREK also gives alerts concerning the molecular substructures associated with toxic end points, implying an implicit relationship between molecular structure and toxicity. QSAR approaches,[10,11] on the other hand, quantify molecular descriptors that depict the electronic and topological properties of a molecule, attempting to directly construct a quantitative relationship between molecular structure and activity. Cariello and co-workers[12] compared the performance of the rule-based program DEREK and the QSAR program TOPKAT[13] in their ability to predict of the Ames bacterial mutagenicity of 400 drug-like compounds. TOPKAT outperformed DEREK marginally, but the overall concordance of both predictions with the Ames assay results was around 70%.

Compared with the rule-based methods, QSAR approaches are easier to implement. QSAR models[14,15] have been built to predict Ames mutagenicity from data sets of different size

* Corresponding author phone: (973)562 3870; e-mail: hongmao.sun@roche.com.

10.1021/ci049917y CCC: $27.50
Published on Web 05/20/2004

and structural diversity. Alternatively, other approaches[16,17] have attempted to predict rodent carcinogenicity from short-term mutagenicity assay results, despite the observation that not all cancers are due to DNA damaging events. Using an inductive learning program RL, Lee et al.[16] correctly predicted the carcinogenicity of 70−80% of nongenotoxic chemicals. Taningher and co-workers[17] did a molecular connectivity analysis on a data set of 145 chemicals with both rodent carcinogenicity and Ames test results, and they identified the most significant fragments, or biophores, responsible for carcinogenesis. Recently, Klopman et al.[18] developed a new "hybrid" method, Expert System Prediction (ESP), to predict the rodent carcinogenicity of 113 new chemicals with improved concordance and sensitivity. ESP is a machine learning program, using an artificial neural network employing genetic algorithms (GA-ANN), to learn the relationship between structure and carcinogenicity from a training set of around 1000 compounds.

An atom type classification approach using a universal molecular descriptor system has proven to be quite powerful in predicting ADME related properties.[1] In the current study, this approach has been extended to construct a simple predictive model for rodent carcinogenicity.

DATA SETS

The data sets used in the present study were the National Toxicology Program (NTP) data set and the Food and Drug Administration (FDA) data set, downloaded from the public domain.[19,20] Both the NTP and the FDA carried out rodent carcinogenicity studies on male and female rats and mice. The models were based on a single 2-year carcinogenicity study to identify trans-species tumorigens. Although the NTP and the FDA used different rodent strains, the relative proportion of positive response compounds and the overall concordance ratio between rats and mice were similar across both the NTP and the FDA databases.[3] The data sets contain chemical carcinogenicities from experimental measurements on both mice and rats. The NTP data set contains 487 compounds, with the carcinogenic activity of each classified as CE (clear evidence of carcinogenic activity), SE (some evidence of carcinogenic activity), EE (equivocal evidence of carcinogenic activity), NE (no evidence of carcinogenic activity), IS (inadequate study of carcinogenic activity), P (positive), E (equivocal), or N (negative). Those data labeled as SE, EE, IS, or E were not included in the final training set, to avoid introducing less confident data. The FDA data set contains 223 pharmaceutically related compounds, with the carcinogenicity of each labeled "+" for positive or "−" for negative. The final combined data set contained 526 compounds, with 150 (28.5%) positives for MM (male mouse), 542 compounds (173, or 31.9% positives) for FM (female mouse), 520 compounds (172 or 33.1% positives) for MR (male rat), and 542 compounds (143, or 26.4% positives) for FR (female rat). Molecules in both data sets are structurally diverse, laying a solid foundation for a robust predictive model.

METHOD

Atom Type Classification. Atom types and correction factors were used as the general molecular descriptors. For each atom, the type was determined by its own chemical properties and the neighboring atoms and bonds, reflecting its chemical environment. The primary classification tree, which was constructed on the basis of experience and chemical intuition, was trained by optimizing the logP predictions of the compounds in Starlist. Through analysis of the structures of the outliers, the "variable importance in projection" (VIP), and the standard errors of the coefficients, the atom type classification tree was optimized in terms of where to split and where to stop splitting the tree. Details on the method of atom type classification are described in ref 1.

Partial Least Square Discriminant Analysis (PLS-DA). PLS-DA is an extension of PLS, also known as "projections to latent structures", a powerful multivariate analysis technique.[21] By projecting intercorrelated data of poor quality from high-dimensional space into low-dimensional orthogonal space, the newly formed variables, which are linear combinations of the original variables, become orthogonal to each other. Through finding the "discriminant plane" to effectively separate data into different classes, PLS-DA is capable of separating "tight" classes of observations on the basis of their X variables. The Y vector encoding the class membership is a set of "dummy" variables, denoted "positive" and "negative" in this study. X variables were mean-centered and scaled to unit variance (UV) before the PLS discriminant analysis. The scaling weight employed in this study was $1/S_k$, where S_k represents the standard deviation of variable k. A default 7-fold cross-validation was used for all four models, where the data set was randomly split into seven even groups—while keeping one-seventh of the data out of model development, models were built on the basis of the rest of the data points, then the activities of the one-seventh compounds left-out were predicted and compared with the actual values. Each data point was left out only once. PLS-DA analysis was performed by using SIMCA-P10.[22]

RESULTS AND DISCUSSION

Atom Type Classification. The atom type classification tree, after being trained by logP,[1] identified 218 atom types, including 88 types of carbon, 7 types of hydrogen, 55 types of nitrogen, 31 types of oxygen, 8 types of halides, 23 types of sulfur, and 6 types of phosphorus. Many atom types carried fragment information. For example, the 34th type of nitrogen, N34, is a nitrogen atom in a nitro group bonded to an aromatic ring, thus, atom type N34 was associated with the fragment of aromatic nitro group. In some cases, two or more atom types are needed to define a specific fragment. For example, C66, a carbon atom in an amide group bonded to aromatic ring, together with N22, a nitrogen atom in an amide group linked with a hydrogen and an aliphatic carbon, defined an amide, which was bonded to an aromatic ring on the carbonyl side and to an aliphatic carbon on the amine side. On the other hand, some atom types do not explicitly relate to any molecular fragments, but they indicate particular chemical features. C13, for example, is an aromatic carbon linked to other three aromatic carbon atoms, indicating a fused aromatic system in the molecule of interest.

The correction factors were slightly different from those used in ref 1. An additional factor was introduced, i.e. M27, the longest unbranched hydrocarbon chain in a molecule.

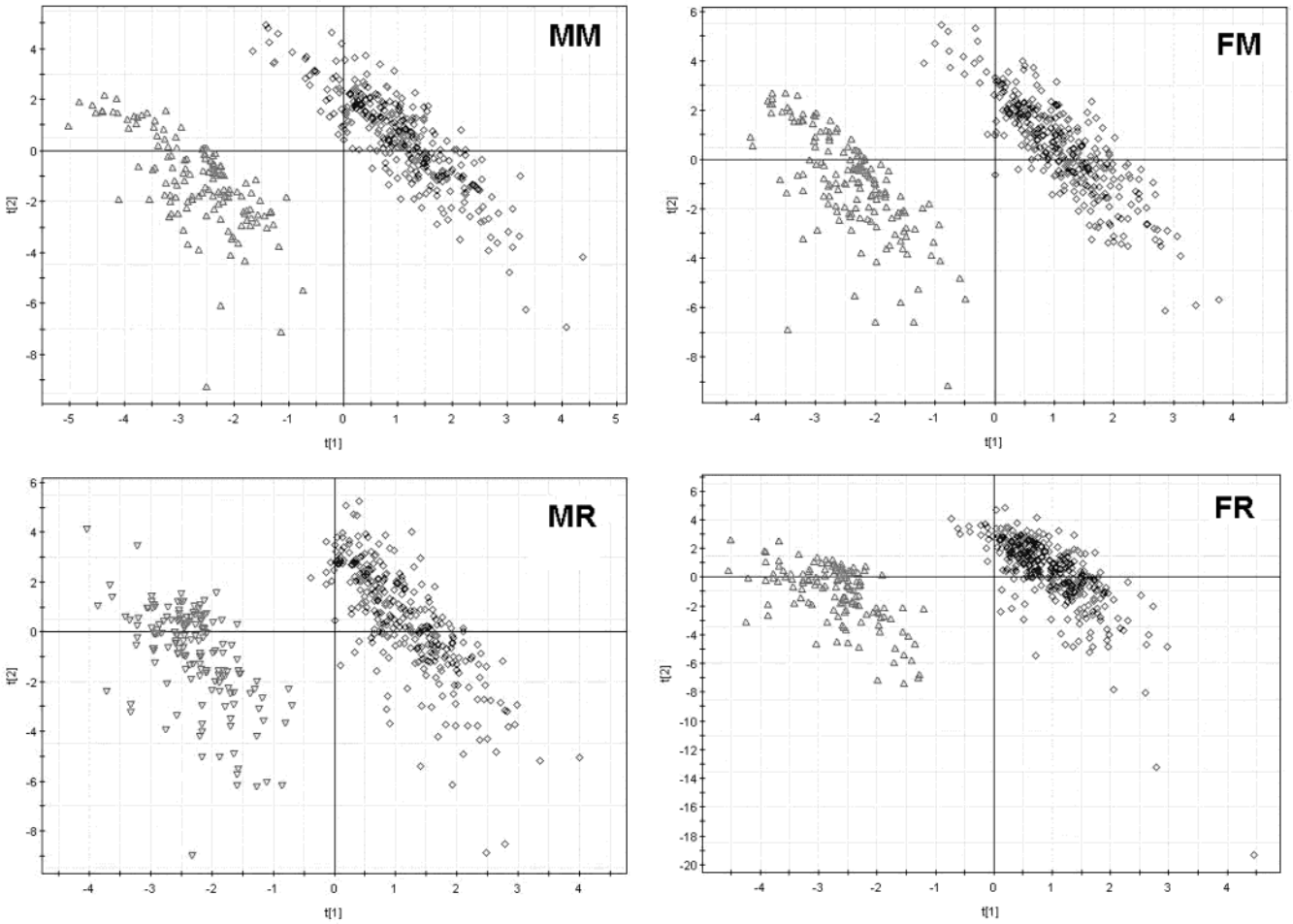

Figure 1. PLS-DA t1/t2 score plots for (a) MM, (b) FM, (c) MR, and (d) FR. Open triangles (△) are carcinogenic positives, and open diamonds (◇) are negatives.

Also, M9, total rotatable bonds in the molecule divided by four, replaced the number of fused atoms (Table 1).

PLS-DA Models for MM, FM, MR, and FR. Chemical toxicities, especially those from in vivo measurements, are often expressed qualitatively, such as those used in this study. In cases where the observations combined into discrete classes, regression methods that are commonly used for continuous numerical responses, such as multiple regression (MR) and PLS, are not the best choice for model construction. Instead, recursive partitioning (RP)[23] and PLS-DA are more effective in these situations. RP works very well on low-dimensional variable space,[24] while PLS-DA can handle high-dimensional variables with autocorrelation problems.

Using the same molecular descriptors, highly predictive PLS-DA models were built for MM, FM, MR, and FR. For MM, a 4-component PLS-DA model predicted carcinogenicity of the 526-compound data set with $r^2 = 0.987$ and $q^2 = 0.944$. With 4 components, PLS-DA models were also deduced for FM with $r^2 = 0.985$ and $q^2 = 0.950$, MR with $r^2 = 0.989$ and $q^2 = 0.962$, and FR with $r^2 = 0.990$ and $q^2 = 0.965$. Figure 1 illustrated the score plots t1/t2 for MM, FM, MR, and FR, where clear separations of carcinogenic positives and negatives were achieved on the first two principal components for all four models.

Validation of the Models. Judging from the values of q^2, the PLS-DA models were highly predictive. There is, however, one limitation of cross-validation, i.e. it only assesses the predictive power of a model but does not address the question of statistical significance. To estimate the significance of q^2 values, a permutation method, called

Table 1. List of Correction Factors

CF ID	correction factor
M1	(int) (molecular weight/100)
M2	intramolecular HB
M3	adjacent halides
M4	(C15 > 4) ? (C15−4) / 2 : 0
M5	(C43 > 5) ? (C43−5) / 2 : 0
M6	fraction of rotatable bonds
M7	alpha amino acid
M8	amide
M9	(int) (rotatable bonds / 4)
M10	salic acid
M11	1,4-dioxane
M12	acetyl urea
M13	number of aromatic rings
M14	number of aliphatic rings
M15	multiple oxygen atoms (> 8)
M16	zwitter ions
M17	multiple hydroxyl groups
M18	multiple acids
M19	linear zwitter ion
M20	ring number = 0
M21	number of fused bonds (fb)
M22	2 * fb − number of fused atoms
M23	pyrazine
M24	ortho functional groups
M25	meta functional groups
M26	para functional groups
M27	longest chain

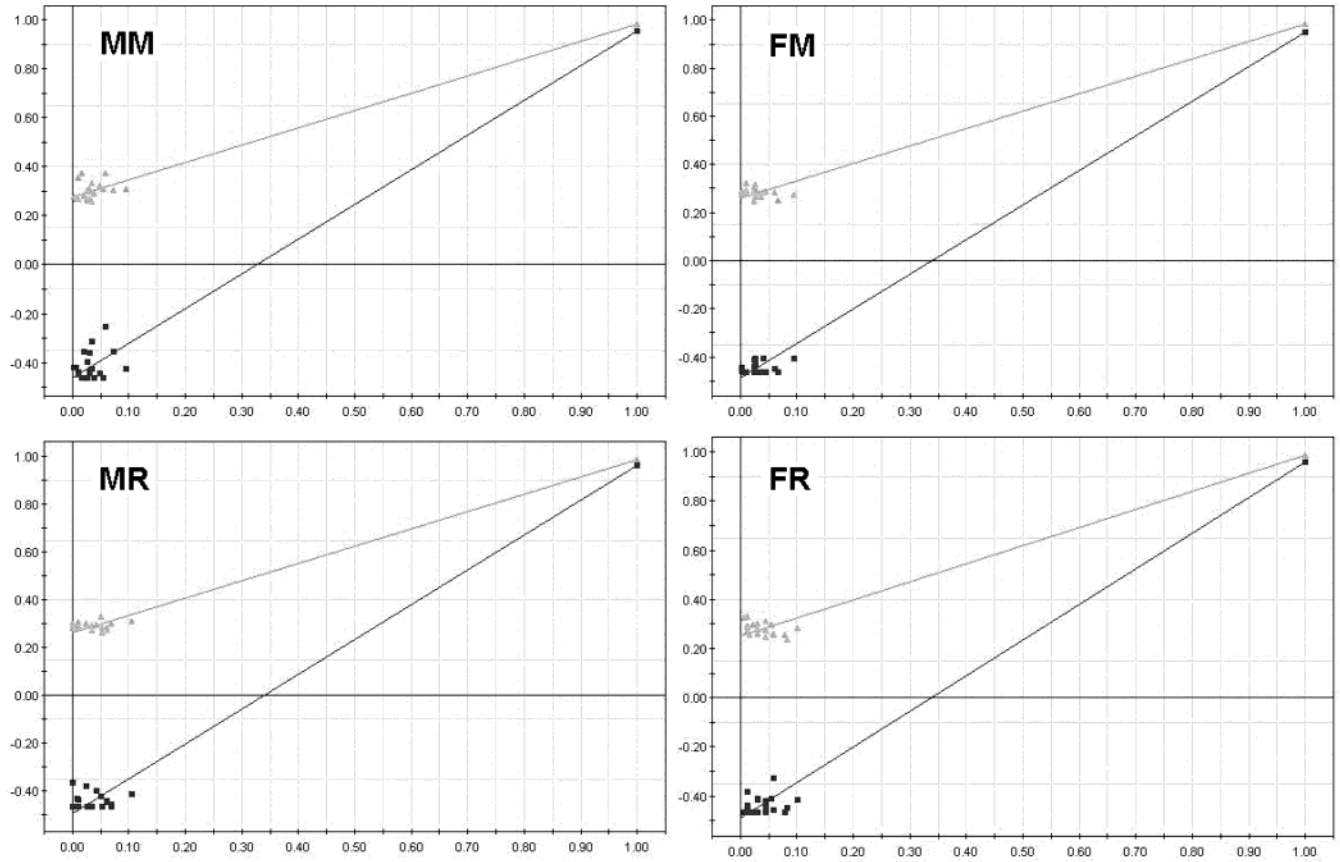

Figure 2. Validate plots of (a) MM, (b) FM, (c) MR, and (d) FR. The *Y* axis represents r^2 (triangles) and q^2 (squares) for every model, and the *X* axis designates the correlation coefficient between original and permuted response data.

validate, in SIMCA-P was employed.[21] To run *validate*, the X matrix remains untouched, while the Y data are randomly shuffled. A PLS-DA model was then reconstructed with the permuted Y data, and r^2 and q^2 were computed. Comparing the two values of r^2 and q^2 of the "permuted" model with those of the "real" model gives an indication of the significance of the latter values. By repeating the permutation procedure many times, it is possible to achieve reference distributions of r^2 and q^2 based on random response data. As shown in Figure 2, the original r^2 and q^2 values of the four models were substantially higher than the corresponding "permuted" ones, indicating that all four models were statistically significant. Another read-out of the *validate* testing is the model complexity: the more scattered the distribution of the permuted r^2 and q^2 data points, the higher the complexity of the model. The overall complexities of the four PLS-DA models were low, which authenticate good predictivity of the models.

Predictive validation by means of cross-validation and the response permutation test provides a reasonable first approximation of the predictive ability of a PLS model, yet a more demanding and rigorous way of testing predictive performance is to predict an independent external data set.[21] In the current study, the original data sets were randomly split into training sets and testing sets, by placing every third compound into the testing set. For example, in the case of MR data set, the 347-compound training set had 118, or 34.0% positives, which was close to the ratio of the original set (33.1%). The 4-component PLS-DA model was built for the training set with $r^2 = 0.984$ and $q^2 = 0.927$. The model was then used to predict the carcinogenicities of the 173 compounds in the test set, and the results were classified as the "probability" of a certain compound being "positive" or "negative". It turned out that the activities of all 173 compounds were correctly predicted (Table 2), among which 148 compounds were predicted with a high confidence of over 0.9, 23 compounds over 0.8, and only 2 compounds between 0.75 and 0.8. The results of all three validation methods—cross-validation, response permutation, and external validation—indicated the PLS-DA models were excellent in robustness and predictive power.

Carcinogenicities of the compounds not included in MR model development, which were labeled as "IS", "E", "EE", and "SE", were predicted (Table 3). The predicted values differed from external validation mainly in two ways: first, 22 out of 90 compounds were predicted with low confidence, whose probability was between 0.4 and 0.6; second, 46 out of 90 compounds were positive, a ratio significantly higher than the original data set (33.1%). Compounds labeled as "SE" had 20 out of 34 (58.8%) positives; compounds labeled "EE" or "E" had 19 out of 44 (43.2%) positives; compounds labeled "IS" had 7 out of 12 (58.3%) positives.

Analysis of the PLS-DA Models. Chemical carcinogenicity is related to molecular structure, and the relationship could potentially be extracted either by an expert system or through regression. As discussed earlier, atom typing is a method to depict the molecular structure by analyzing each atom in the context of its neighbors. The PLS-DA models in this study differed from previous rule-based or structure-based methods in that there were no fragments predefined, reducing the possibility of introducing bias at the beginning of the model construction. Since the atom types implicitly carried frag-

Table 2. Experimental and Predicted Carcinogenicities of the Test Set of MR, Containing Every Third Compound of the Original Data Set

compd ID	compd name	carcinogenicity[a]	DA1[b,c]	DA2[c]	compd ID	compd name	carcinogenicity[a]	DA1[b,c]	DA2[c]
3	TR002	N	**0.9936**	0.0064	228	TR351	P	0.0604	0.9396
6	TR007	N	**0.8470**	0.1530	231	TR354	N	**0.9256**	0.0744
9	TR010	N	**0.9326**	0.0674	234	TR359	P	−0.0358	1.0358
12	TR014	N	**0.9621**	0.0379	237	TR368	P	0.0579	0.9421
15	TR019	P	−0.0764	1.0764	240	TR372	P	−0.0090	1.0090
18	TR022	N	**0.9880**	0.0120	243	TR377	N	**1.1042**	−0.1042
21	TR025	N	**1.0282**	−0.0282	246	TR383	P	−0.3722	1.3722
24	TR030	N	**1.0557**	−0.0557	249	TR387	N	**0.9968**	0.0032
27	TR033	N	**0.9441**	0.0559	252	TR390	P	0.0002	0.9999
30	TR036	N	**0.9691**	0.0309	255	TR395	N	**0.9976**	0.0024
33	TR041	P	0.0322	0.9678	258	TR398	P	−0.8412	1.8412
36	TR047	P	0.0218	0.9782	261	TR402	P	0.1342	0.8658
39	TR050	N	**0.9326**	0.0674	264	TR406	N	**0.9001**	0.0999
42	TR054	P	0.1084	0.8916	267	TR424	N	**0.9982**	0.0019
45	TR058	P	0.0276	0.9724	270	TR431	N	**0.9785**	0.0215
48	TR061	N	**1.0821**	−0.0821	273	TR437	N	**1.1202**	−0.1202
51	TR066	N	**0.9449**	0.0551	276	TR442	N	**1.0329**	−0.0329
54	TR072	P	0.0528	0.9472	279	TR448	P	0.0139	0.9861
57	TR076	P	−0.4761	1.4761	282	TR455	N	**1.0581**	−0.0581
60	TR080	P	−0.0130	1.0130	285	TR464	N	**0.9624**	0.0376
63	TR083	N	**0.9243**	0.0757	288	TR467	P	0.1241	0.8759
66	TR086	P	−0.0643	1.0643	291	TR477	N	**0.9955**	0.0045
69	TR091	N	**1.0702**	−0.0702	294	TR481	N	**0.9213**	0.0787
72	TR094	P	0.1005	0.8995	297	TR490	N	**0.9819**	0.0181
75	TR098	N	**0.9871**	0.0129	300	TR496	P	0.0034	0.9966
78	TR101	N	**0.9177**	0.0823	303	hydroxyquinoline, 8-	N	**1.0005**	−0.0005
81	TR104	N	**1.1922**	−0.1922	306	acetohexamide	N	**0.9326**	0.0674
84	TR109	N	**1.0036**	−0.0036	309	allopurinol	N	**0.9687**	0.0313
87	TR112	N	**0.9941**	0.0059	312	amlodipine	N	**1.1463**	−0.1463
90	TR115	P	0.1179	0.8821	315	amrinone	N	**0.9992**	0.0008
93	TR120	N	**1.1031**	−0.1031	318	benazepril	N	**1.0995**	−0.0995
96	TR123	N	**1.0817**	−0.0817	321	bisoprolol	N	**1.0329**	−0.0329
99	TR126	N	**0.8927**	0.1073	324	budesonide	P	−0.0337	1.0337
102	TR129	N	**0.8879**	0.1121	327	buspirone	N	**1.0391**	−0.0391
105	TR131b	N	**1.0060**	−0.0060	330	carteolol	N	**0.9647**	0.0353
108	TR134	N	**0.8535**	0.1465	333	chlorpheniramine	N	**1.1058**	−0.1058
111	TR137	N	**1.0696**	−0.0696	336	ciprofloxacin	N	**0.8673**	0.1327
114	TR141	N	**0.8585**	0.1415	339	clozapine	N	**1.0044**	−0.0044
117	TR144	P	−0.0534	1.0534	342	cyclobenzaprine	N	**1.1416**	−0.1416
120	TR147	N	**0.9349**	0.0651	345	dantrolene	N	**0.9219**	0.0781
123	TR150	N	**1.0120**	−0.0120	348	dexfenfluramine	N	**0.9460**	0.0540
126	TR155	P	0.1644	0.8356	351	didanosine	N	**0.8965**	0.1035
129	TR158	N	**0.9660**	0.0340	354	diphenhydramine	N	**0.9931**	0.0069
132	TR161	N	**0.8777**	0.1223	357	doxylamine	P	0.1454	0.8546
135	TR165	N	**0.9326**	0.0674	360	ephedrine	N	**0.9971**	0.0029
138	TR169	N	**0.9508**	0.0492	363	estradiol mustard	N	**0.9651**	0.0349
141	TR174	N	**0.9173**	0.0827	366	etretinate	N	**0.7840**	0.2160
144	TR178	N	**1.0081**	−0.0081	369	famotidine	N	**1.1897**	−0.1897
147	TR181	P	−0.0128	1.0128	372	flecainide	N	**0.8268**	0.1732
150	TR186	P	0.0401	0.9599	375	fluoxetine	N	**0.8955**	0.1045
153	TR191	N	**0.9796**	0.0204	378	fluvastatin	P	0.1871	0.8129
156	TR195	N	**0.9148**	0.0852	381	furazolidone	N	**0.8494**	0.1506
159	TR203	N	**0.9988**	0.0012	384	gemfibrozil	P	0.0614	0.9386
162	TR206	P	−0.0927	1.0927	387	granisetron	P	0.0795	0.9205
165	TR209	P	0.2032	0.7968	390	hydrochlorothiazide	N	**0.9090**	0.0910
168	TR212	N	**0.9416**	0.0584	393	iodinated glycerol	P	0.0521	0.9479
171	TR217	P	0.0374	0.9626	396	isosorbide	N	**0.9005**	0.0995
174	TR222	P	0.0334	0.9666	399	ketoconazole	N	**0.7549**	0.2451
177	TR226	P	0.0163	0.9837	402	labetalol	N	**0.9162**	0.0838
180	TR234	P	0.0016	0.9984	405	levamisole	N	**0.9210**	0.0790
183	TR245	P	0.2065	0.7935	408	lorazepam	N	**1.0671**	−0.0671
186	TR253	P	0.0978	0.9022	411	mebendazole	N	**1.0406**	−0.0406
189	TR259	P	0.1165	0.8835	414	metaproterenol	N	**1.0577**	−0.0577
192	TR269	P	0.0828	0.9172	417	metoprolol	N	**1.0113**	−0.0113
195	TR276	N	**1.0005**	−0.0005	420	midazolam	P	0.1428	0.8572
198	TR282	N	**0.7816**	0.2184	423	misoprostol	N	**0.9721**	0.0279
201	TR287	P	0.0236	0.9764	426	nabumetone	N	**1.0262**	−0.0262
204	TR299	P	−0.1686	1.1686	429	nalidixic acid	P	0.0579	0.9421
207	TR305	N	**0.8409**	0.1591	432	netilmicin	N	**1.1067**	−0.1067
210	TR311	P	0.1264	0.8737	435	nisoldipine	N	**0.9124**	0.0876
213	TR314	N	**0.9386**	0.0614	438	nitrofurazone	N	**0.9127**	0.0873
216	TR321	P	−0.0337	1.0337	441	omeprazole	P	−0.1289	1.1289
219	TR329	P	0.0459	0.9541	444	oxazepam	P	0.1586	0.8414
222	TR335	N	**1.0071**	−0.0071	447	paroxetine	N	**0.9975**	0.0025
225	TR344	N	**0.8564**	0.1436	450	pentaerythritol	N	**1.0357**	−0.0357

Table 2 (Continued)

compd ID	compd name	carcinogenicity[a]	DA1[b,c]	DA2[c]	compd ID	compd name	carcinogenicity[a]	DA1[b,c]	DA2[c]
453	perindopril	N	**1.0430**	−0.0430	489	sotalol	N	**1.0271**	−0.0271
456	phenformin	N	**0.8470**	0.1530	492	sulfisoxazole	N	**0.9081**	0.0919
459	phenylbutazone	N	**0.9130**	0.0870	495	temazepam	N	**1.0557**	−0.0557
462	pimozide	N	**1.0700**	−0.0700	498	terbutaline	N	**1.0668**	−0.0668
465	piroxicam	N	**1.0838**	−0.0838	501	theophylline	N	**1.0587**	−0.0587
468	probenecid	N	**0.9976**	0.0024	504	timolol	P	−0.0068	1.0068
471	propafenone	N	**0.9673**	0.0327	507	tolbutamide	N	**0.9567**	0.0433
474	pyrilamine	N	**0.9457**	0.0543	510	tramadol	N	**0.9433**	0.0567
477	ramipril	N	**1.0159**	−0.0159	513	tryptophan	N	**0.9297**	0.0703
480	resorcinol	N	**1.0287**	−0.0287	516	valproic acid	P	0.0935	0.9065
483	rifampin	N	**0.9495**	0.0505	519	zolpidem	P	0.1021	0.8979
486	scopolamine	N	**1.0345**	−0.0345					

[a] "N" refers to carcinogenic negative; "P" is positive. [b] The compounds with a DA1 value of greater than 0.5, indicating carcinogenic negative as predicted, are shown in boldface in the column of DA1. [c] DA1 is the possibility of a compound to be carcinogenic negative, and DA2 is the possibility of being positive. DA1 + DA2 = 1.0; DA1 > 0.5 means the compound is predicted negative, while DA2 > 0.5 is positive.

Table 3. Experimental and Predicted Carcinogenicities of the MR Data Set, Whose Activities Were Labeled as "IS", "E", "EE", and "SE"

compd ID	compd name	carcinogenicity	DA1[a]	DA2	compd ID	compd name	carcinogenicity	DA1[a]	DA2
1	TR003	MR=IS	0.4451	0.5549	46	TR337	MR=EE	**0.9856**	0.0144
2	TR005	MR=E	0.0783	0.9217	47	TR339	MR=SE	0.3794	0.6206
3	TR006b	MR=E	**0.5743**	0.4257	48	TR341	MR=SE	**1.2733**	−0.2733
4	TR013	MR=IS	0.3785	0.6215	49	TR342	MR=SE	0.0864	0.9136
5	TR016	MR=E	0.1778	0.8222	50	TR345	MR=EE	**0.5719**	0.4281
6	TR018	MR=E	**1.3923**	−0.3923	51	TR346	MR=EE	0.4625	0.5375
7	TR021a	MR=E	0.4697	0.5303	52	TR350	MR=SE	−0.1552	1.1552
8	TR026	MR=IS	**0.5684**	0.4316	53	TR355	MR=EE	**0.8964**	0.1036
9	TR027	MR=E	0.4237	0.5763	54	TR356	MR=EE	**1.2616**	−0.2616
10	TR042	MR=IS	**1.0495**	−0.0495	55	TR360	MR=SE	0.4866	0.5134
11	TR043	MR=IS	0.3829	0.6171	56	TR363	MR=SE	0.2747	0.7253
12	TR052	MR=E	0.3044	0.6956	57	TR364	MR=EE	**0.6262**	0.3738
13	TR057	MR=E	0.4374	0.5626	58	TR365	MR=EE	0.2284	0.7716
14	TR062	MR=IS	**1.5800**	−0.5800	59	TR366	MR=SE	**0.7078**	0.2922
15	TR065	MR=IS	0.0817	0.9183	60	TR367	MR=EE	**1.0387**	−0.0387
16	TR069	MR=E	**1.4504**	−0.4504	61	TR369	MR=SE	**0.6882**	0.3118
17	TR070	MR=E	**1.5027**	−0.5027	62	TR376	MR=EE	0.3809	0.6191
18	TR075	MR=E	**0.7592**	0.2408	63	TR380	MR=IS	**1.1730**	−0.1730
19	TR088	MR=E	0.1428	0.8572	64	TR382	MR=SE	**0.9265**	0.0735
20	TR106	MR=IS	0.3171	0.6829	65	TR404	MR=EE	**1.0163**	−0.0163
21	TR116	MR=E	0.1861	0.8139	66	TR407	MR=SE	0.2396	0.7604
22	TR131c	MR=E	0.4175	0.5825	67	TR409	MR=SE	**1.0775**	−0.0775
23	TR152	MR=E	**2.6139**	−1.6139	68	TR411	MR=EE	0.4027	0.5973
24	TR189	MR=E	0.3439	0.6561	69	TR414	MR=SE	0.4051	0.5949
25	TR213	MR=IS	**0.6236**	0.3764	70	TR419	MR=EE	**1.0125**	−0.0125
26	TR215	MR=E	**0.9166**	0.0834	71	TR420	MR=EE	0.3767	0.6233
27	TR232	MR=E	0.4286	0.5714	72	TR422	MR=SE	0.2322	0.7678
28	TR237	MR=E	0.4503	0.5497	73	TR423	MR=SE	0.3367	0.6633
29	TR240	MR=E	**1.0263**	−0.0263	74	TR430	MR=SE	−0.3677	1.3677
30	TR243	MR=IS	0.4640	0.5360	75	TR436	MR=SE	**0.5254**	0.4746
31	TR261	MR=E	0.4784	0.5216	76	TR447	MR=EE	**1.0193**	−0.0193
32	TR267	MR=SE	0.4330	0.5670	77	TR449	MR=EE	**0.9900**	0.0100
33	TR271	MR=EE	**1.0077**	−0.0077	78	TR456	MR=SE	−0.2147	1.2147
34	TR274	MR=EE	**0.5879**	0.4121	79	TR457	MR=SE	**0.8023**	0.1977
35	TR291	MR=SE	**0.6275**	0.3725	80	TR458	MR=SE	**0.6236**	0.3764
36	TR298	MR=SE	−0.0186	1.0186	81	TR463	MR=SE	**0.5369**	0.4631
37	TR303	MR=IS	0.2345	0.7655	82	TR468	MR=EE	**1.2112**	−0.2112
38	TR306	MR=SE	**0.7271**	0.2729	83	TR470	MR=SE	0.4995	0.5005
39	TR309	MR=SE	−1.3934	2.3934	84	TR475	MR=SE	**0.5410**	0.4590
40	TR315	MR=EE	**1.1919**	−0.1919	85	TR476	MR=EE	**1.6505**	−0.6505
41	TR318	MR=EE	**1.8655**	−0.8655	86	TR482	MR=SE	**0.6901**	0.3099
42	TR320	MR=EE	0.2928	0.7072	87	TR483	MR=SE	**0.6629**	0.3371
43	TR323	MR=SE	0.2928	0.7072	88	TR485	MR=EE	**1.1150**	−0.1150
44	TR332	MR=SE	0.2392	0.7608	89	TR487	MR=SE	0.0804	0.9196
45	TR334	MR=SE	0.3794	0.6206	90	TR494	MR=SE	−0.0081	1.0081

[a] The compounds with a DA1 value of greater than 0.5, indicating carcinogenic negative as predicted, are shown in boldface in the column of DA1.

mental information, it was straightforward for PLS-DA models based on atom types to identify the chemical moieties that related to carcinogenicity. Variable influence on projection (VIP) analysis of PLS-DA model gives quantitative estimates of the discrimination power of each atom type or correction factor. The VIP plots of PLS-DA models, as

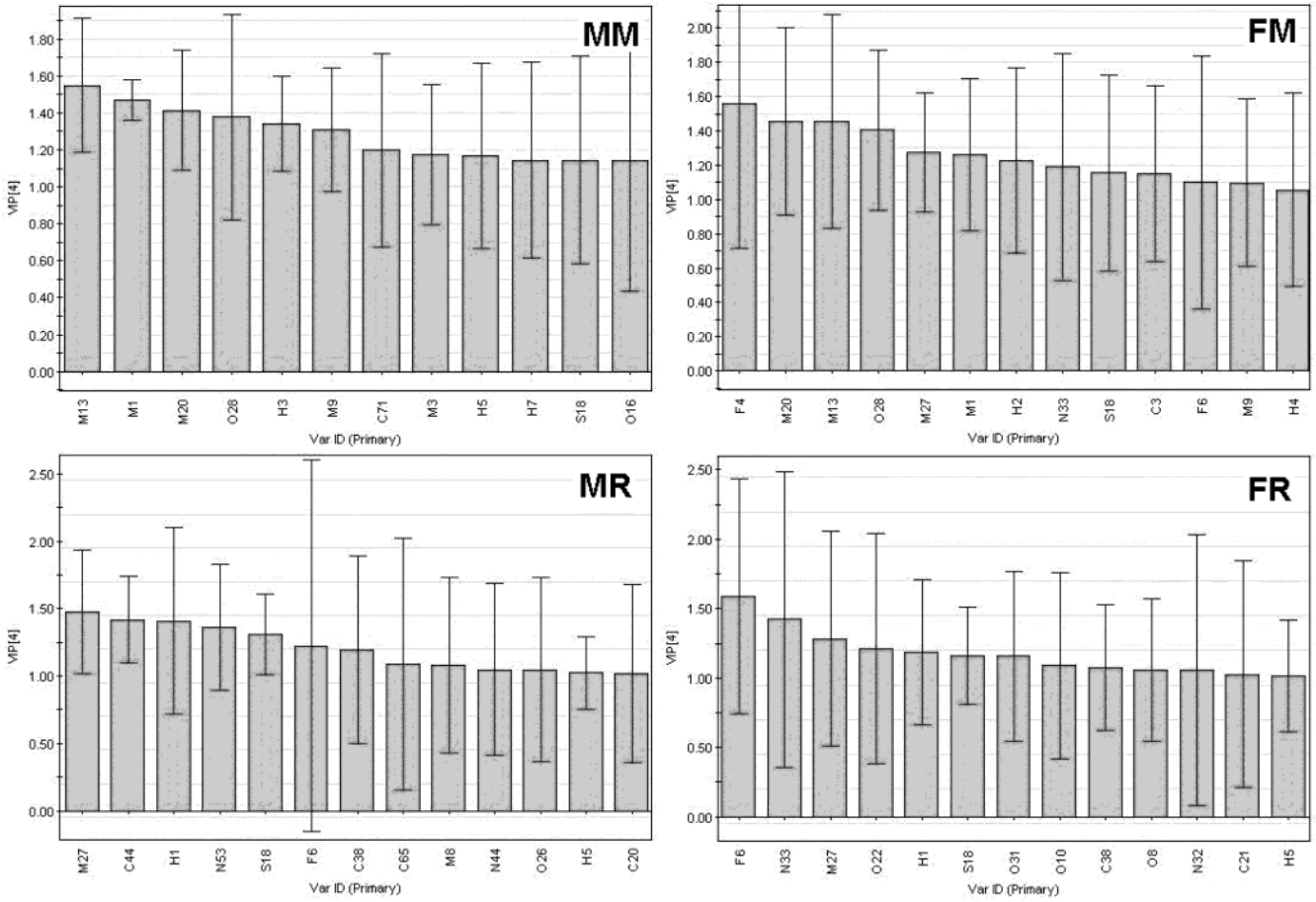

Figure 3. VIP plot of PLS-DA models for MM, FM, MR, and FR.

shown in Figure 3, described the atom types and correction factors with the strongest discrimination power. For the MM and FM models, molecular weight (M1), number of aromatic rings (M13), and molecule without any ring (M20) were important factors to separate positives from negatives. For the MM, oxygen atom in sulfone or phosphate group (O28) was the atom type with the highest discrimination power. The atom type O28 occurred in 56 molecules, and only 6 of them were positive, with a positive ratio of 10.7% (6/56), significantly lower than 28.5% of positives in the original data set. The positive ratio dropped to 5.0% (2/40) for those compounds containing more than one O28 atom. The same atom type, O28, also appeared less frequently in FM positives. For molecules containing O28 atom, only 9 out of 57 (15.8%) molecules were positives, and only 3 out of 39 (7.7%) were positives, when a molecule contained more than one O28 atom. The results implied that compounds with a sulfone or a phosphate group were less likely to be carcinogenic in both the male and female mouse. However, the conclusion was less solid for male or female rat, where the positive ratios of the molecules with the atom type O28 were 27.3% (15/55) and 21.1% (12/57) for MR and FR, respectively. Another interesting observation from analyzing VIPs across species and gender was that the differences across gender were less than those across species. Comparing the 12 variables with the highest discrimination power, male and female mice had 6 in common, male and female rats had 5 in common, while male mice and rats had only 2 in common, and female mice and rats had 4 in common. There was one atom type, S18, a sulfur atom double bonded to a non-carbon atom, that played an important role in all four models, suggesting that compounds containing sulfone, sulfonamide, or thiophosphate were least likely to be carcinogenic.

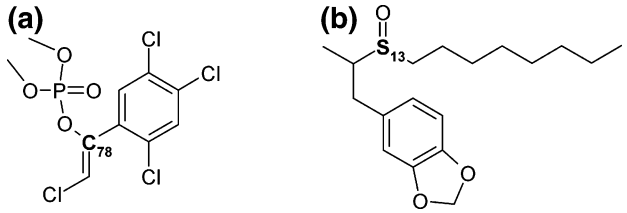

Figure 4. Structures of compounds (a) TR033 and (b) TR124.

Atom types or correction factors that contributed the most to differentiating the carcinogenic activities across species and gender can be identified by coefficient analysis of the four PLS-DA models. As listed in Table 4, compounds with atom type C78, a carbon atom double bonded to another carbon and single bonded to an aromatic carbon and a non-carbon, tend to be carcinogenic positive for male mice and negative for male rats, such as in compound TR033. On the other hand, compounds with atom type S13, a sulfur atom double bonded to an oxygen and single bonded to two aliphatic atoms, are most likely to be positive for male mice and negative for female mice, such as in compound TR124 (Figure 4). Ultimately, it is desired to be able to predict carcinogenicity across a range of different animal species and across genders; therefore, it is necessary to build a global single model. To build such a global model, reliable submodels, such as those described herein, are a first, but

Table 4. Coefficients of the Molecular Descriptors in Four PLS-DA Models

Var ID	MM	FM	MR	FR	Var ID	MM	FM	MR	FR	Var ID	MM	FM	MR	FR
constant	0.5085	0.4953	0.5057	0.5040	C79	0.0226	−0.0945	0.0154	0.0275	O11	0.0046	0.0205	0.0018	0.0035
C1	0.0101	0.0158	−0.0010	0.0057	C80	0.0032	0.0051	0.0619	0.0301	O12	0.0153	0.0179	0.0360	0.0412
C2	−0.0077	−0.0183	0.0010	0.0064	C81	−0.0083	0.0130	−0.0114	0.0069	O13	0.0093	0.0168	0.0008	0.0220
C3	−0.0003	−0.0003	−0.0007	0.0003	C82	0.0430	0.0151	−0.0431	−0.0192	O15	0.0177	−0.1058	0.0018	0.0003
C4	−0.0026	0.0070	0.0075	0.0052	C83	0.0233	−0.0107	−0.0428	−0.0432	O16	−0.1645	0.0078	0.0402	−0.0591
C5	0.0048	−0.0121	0.0099	−0.0003	C84	0.0161	0.0328	−0.0056	−0.0341	O17	−0.0049	0.0158	−0.0105	−0.0029
C6	0.0070	0.0134	−0.0159	0.0069	C85	0.0749	−0.0913	0.1155	0.0478	O18	0.0222	0.0127	0.0215	0.0211
C7	−0.0173	0.0380	−0.0088	−0.0195	C87	−0.0622	−0.0973	−0.0774	−0.1314	O19	−0.0016	−0.0021	−0.0059	−0.0042
C8	0.0314	0.0472	0.0177	0.0302	C88	−0.0715	0.0003	−0.0776	0.0295	O21	0.0144	−0.0213	0.0147	0.0019
C9	−0.0117	−0.0214	0.0091	−0.0197	H1	0.0003	−0.0004	0.0000	−0.0002	O22	−0.0489	0.0154	−0.0161	−0.0351
C10	0.0014	0.0024	−0.0026	−0.0015	H2	−0.0004	−0.0054	−0.0007	0.0004	O23	0.0357	0.0097	0.0106	0.0273
C11	0.0001	0.0020	0.0198	0.0192	H3	−0.0045	0.0100	−0.0002	0.0017	O24	0.0025	0.0028	0.0034	0.0106
C12	−0.0004	−0.0043	−0.0012	−0.0020	H4	0.0113	0.0073	0.0107	0.0077	O26	−0.0005	0.0019	0.0068	0.0028
C13	−0.0066	0.0016	0.0081	0.0083	H5	0.0038	−0.0051	0.0071	0.0072	O27	0.0014	0.0047	0.0006	−0.0017
C14	−0.0042	−0.0038	−0.0102	0.0071	H6	−0.0048	−0.0053	0.0378	0.0115	O28	0.0069	−0.0002	0.0015	0.0007
C15	−0.0005	−0.0018	−0.0018	0.0003	H7	−0.0023	0.0063	−0.0014	−0.0020	O29	0.0049	0.0024	0.0071	0.0048
C16	−0.0216	−0.0082	−0.0117	−0.0383	N2	0.0138	0.0208	0.0148	0.0034	O30	0.0004	0.0142	−0.0086	−0.0070
C17	−0.0013	−0.0016	−0.0021	−0.0047	N3	0.0025	0.0156	−0.0104	−0.0184	O31	−0.0451	−0.0191	−0.0610	−0.0183
C18	−0.0033	−0.0015	0.0047	0.0025	N4	0.0328	0.0081	−0.0310	−0.0276	F1	0.0416	−0.0032	−0.0073	0.0220
C19	0.0069	−0.0219	−0.0009	0.0225	N5	0.0034	−0.0125	0.0100	0.0142	F2	0.0051	0.0000	−0.0065	−0.0173
C20	0.0008	0.0004	−0.0006	−0.0004	N6	0.0022	−0.0983	−0.0076	0.0028	F3	0.0002	−0.0100	0.0093	0.0086
C21	−0.0563	−0.0598	−0.0258	−0.0674	N7	0.0539	0.0287	0.0130	−0.0521	F4	−0.0135	0.0023	−0.0007	−0.0009
C22	0.0046	0.0218	−0.0304	−0.0098	N8	0.0058	−0.0005	0.0628	0.0610	F5	−0.0301	−0.0307	−0.0203	−0.0314
C23	0.0196	0.0059	0.0788	0.0627	N10	−0.0105	−0.0690	−0.0001	0.0028	F6	−0.0393	0.0539	−0.0339	−0.0475
C24	−0.0036	−0.0025	−0.0108	−0.0126	N11	0.0301	0.0117	−0.0306	−0.0507	F8	0.0558	−0.0521	0.0224	0.0218
C25	−0.0587	−0.0631	0.0135	0.0526	N12	−0.0190	−0.0608	−0.0089	0.0196	P1	0.0214	0.0374	0.0037	−0.0224
C26	−0.0315	−0.0374	−0.0130	−0.0296	N13	−0.0587	0.0657	−0.0008	0.0330	P3	0.0557	0.0775	−0.0146	0.0326
C28	0.0081	0.0256	0.0194	0.0022	N14	0.0389	0.0030	0.0344	0.0562	P4	0.0749	−0.0239	0.0746	0.0629
C29	−0.0155	−0.0297	−0.0333	−0.0175	N15	−0.0131	0.0129	−0.0139	0.0019	P6	−0.0196	0.0582	−0.0132	−0.0151
C30	−0.0025	0.0126	−0.0092	−0.0177	N16	0.0347	0.0229	0.0132	−0.0070	S2	0.0027	−0.1118	−0.0526	−0.0243
C31	0.0278	0.0445	0.0490	0.0460	N17	0.0130	0.0049	0.0228	0.0164	S3	0.0198	0.0097	−0.0152	−0.0455
C32	0.0844	0.0397	0.1005	−0.0003	N18	−0.0157	0.0204	0.0372	0.0161	S4	0.0301	−0.0283	−0.0306	0.0036
C34	0.1347	0.1000	0.1115	0.1000	N20	0.0180	−0.0055	−0.0313	0.0251	S5	0.0821	0.0367	−0.0366	−0.0205
C35	−0.0022	−0.0098	0.0215	0.0172	N21	0.0270	0.0235	−0.0087	−0.0048	S6	0.0609	0.1190	0.0391	0.0700
C36	0.0061	−0.0078	0.0299	−0.0231	N22	−0.0193	0.0151	0.0022	0.0015	S7	0.1518	0.0785	0.0637	−0.1353
C37	−0.2497	−0.2286	−0.1624	−0.1813	N23	−0.0384	−0.0146	0.0358	0.0250	S8	0.0893	−0.0713	0.0786	0.0665
C38	0.0048	0.0051	0.0014	0.0010	N24	0.0071	0.0640	0.0107	−0.0315	S9	0.0878	0.0029	−0.0401	−0.1492
C39	0.0152	0.0114	−0.0081	−0.0154	N25	0.0732	0.0536	−0.0453	0.0014	S10	−0.0422	0.0172	0.0424	−0.0503
C40	0.0093	0.0075	0.0012	0.0114	N27	−0.0372	−0.0513	−0.0274	0.0081	S12	0.0129	−0.0874	−0.0581	0.0511
C41	0.0026	0.0116	−0.0036	−0.0076	N28	0.0214	0.0249	−0.0279	−0.0300	S13	−0.2070	0.1070	−0.1391	−0.0285
C42	0.0855	0.0182	0.0294	0.0151	N29	−0.0535	−0.0856	0.0252	0.0238	S14	−0.0716	0.1297	−0.0375	−0.2338
C43	0.0004	0.0001	0.0005	−0.0010	N30	0.0182	−0.0550	0.0231	0.0198	S15	0.1028	−0.0134	0.1193	−0.0081
C44	0.0172	0.0141	0.0231	0.0110	N31	0.0142	−0.0106	0.1092	0.0288	S16	0.1238	0.0142	0.0004	0.0884
C45	−0.0711	−0.0557	0.0742	−0.0312	N32	0.0072	−0.0870	−0.0322	−0.0196	S18	−0.0016	−0.0084	0.0104	0.0083
C46	−0.0073	−0.0062	0.0028	−0.0011	N33	−0.0579	0.0035	−0.0520	−0.0669	S19	0.0039	0.0373	−0.0011	−0.0045
C47	0.0036	0.0061	0.0003	−0.0114	N34	0.0079	0.0082	0.0011	0.0058	S22	0.0068	−0.0239	−0.0307	0.0196
C48	−0.0151	−0.0537	−0.0122	0.0123	N36	0.0023	−0.0263	0.0183	0.0105	S23	0.0467	−0.0045	−0.0132	−0.0225
C49	−0.0040	−0.0059	−0.0034	−0.0032	N38	−0.0297	0.0104	−0.0237	−0.0455	S24	−0.0196	−0.0049	−0.0030	−0.0151
C50	−0.0006	0.0160	−0.0303	−0.0124	N39	−0.0065	−0.0558	0.0181	0.0012	M1	−0.0043	0.0081	−0.0066	−0.0036
C51	−0.0024	−0.0190	−0.0206	−0.0116	N40	−0.0347	0.0173	−0.0130	−0.0041	M2	−0.0055	0.0109	0.0024	−0.0050
C52	0.0528	0.0411	0.0529	0.0484	N42	0.0854	0.0113	0.0502	0.0907	M3	0.0065	−0.0160	0.0253	0.0041
C53	0.0424	0.0368	−0.0111	−0.0144	N43	0.0156	−0.0134	−0.0227	−0.0225	M4	0.0065	−0.0010	0.0210	0.0015
C54	−0.0127	−0.0053	0.0035	0.0249	N44	0.0090	0.0136	0.0210	−0.0150	M5	−0.0013	0.0554	−0.0177	0.0229
C55	0.0489	0.0162	−0.0001	0.0180	N45	0.0223	−0.0521	0.0152	−0.0092	M6	−0.0294	−0.0175	−0.0696	−0.0135
C57	−0.0158	0.0170	−0.0069	0.0122	N46	0.0214	0.0205	0.0037	−0.0224	M7	0.0174	−0.0063	0.0057	−0.0801
C58	0.0179	0.0197	0.0223	0.0162	N47	0.0257	0.0285	−0.0107	0.0191	M8	−0.0024	0.0043	0.0007	−0.0171
C59	−0.1509	0.0728	0.1430	0.1179	N48	0.0242	0.1190	0.0396	−0.0333	M9	−0.0017	−0.0010	−0.0150	0.0010
C60	0.0711	0.0205	0.0211	0.0481	N49	0.0534	0.0151	−0.0048	−0.0094	M10	−0.0434	0.0032	−0.0041	−0.0014
C61	0.0454	0.0433	0.0327	0.0464	N50	0.0233	0.0038	−0.0428	−0.0432	M11	−0.0041	−0.0013	0.0238	0.0032
C62	0.0430	0.0097	0.0061	0.0002	N51	0.0357	−0.0044	0.0187	−0.0068	M12	−0.0125	0.0002	−0.0023	0.0360
C63	0.0025	0.0081	0.0034	0.0106	N52	−0.0007	−0.0140	−0.0073	−0.0043	M13	−0.0009	−0.0440	−0.0041	−0.0005
C65	0.0023	0.0127	0.0063	0.0049	N53	−0.0070	−0.0557	−0.0090	−0.0096	M14	−0.0015	−0.0258	−0.0171	−0.0027
C66	0.0029	−0.0021	0.0029	0.0017	N54	−0.1658	0.0217	−0.1477	−0.1007	M15	−0.0259	−0.0136	0.0138	0.0042
C67	−0.0015	−0.0141	0.0055	0.0010	N55	0.0160	0.0242	0.0480	0.0351	M16	0.0071	−0.0568	−0.0252	0.0139
C68	−0.0138	−0.0106	0.0029	0.0007	N56	0.0074	−0.0569	0.0023	0.0216	M17	−0.0037	0.0359	−0.0352	−0.0174
C69	−0.0047	0.0353	−0.0147	−0.0109	O1	−0.0492	0.0250	−0.0718	−0.0167	M18	−0.0761	0.0262	−0.2116	−0.0303
C70	−0.0107	−0.0018	0.0207	0.0607	O2	0.0214	−0.0353	0.0037	0.0187	M20	0.0266	0.0006	0.0113	−0.0135
C71	−0.0114	−0.0240	0.0074	0.0030	O3	−0.0204	0.0118	−0.0278	−0.0186	M21	0.0019	0.0064	−0.0011	0.0145
C72	−0.0284	0.0318	0.0263	−0.0021	O4	0.0156	0.0088	0.0196	0.0066	M22	0.0093	0.0035	0.0012	0.0006
C73	−0.0003	0.0223	−0.0238	0.0014	O5	0.0077	0.0052	0.0239	−0.0015	M23	0.0062	0.0051	0.0396	0.0026
C74	0.0186	−0.0534	0.0107	−0.0469	O6	0.0269	0.0112	0.0313	0.0156	M24	0.0049	0.0033	−0.0045	0.0334
C75	−0.0288	−0.0177	−0.1333	−0.0427	O7	0.0053	0.0176	−0.0018	0.0028	M25	0.0014	0.0106	0.0060	−0.0024
C76	−0.0142	0.0085	0.0058	0.0390	O8	0.0127	0.0052	0.0190	0.0167	M26	−0.0005	0.0045	0.0004	0.0069
C77	−0.0017	−0.1634	−0.0081	−0.0179	O9	0.0023	−0.0102	0.0029	0.0052	M27	0.0011	0.4484	0.0025	0.0107
C78	−0.1966	0.0115	0.1157	−0.2154	O10	−0.0094	0.0022	0.0129	−0.0191					

essential, step. More extensive data sets will be needed to bridge the gap between rodent carcinogenesis and human carcinogenesis.

SUMMARY

The application of a university molecular descriptor system has been extended to the field of toxicology modeling. Four PLS-DA models were built to predict carcinogenicities in each of four rodent models: MM, FM, MR, and FR. The models were highly predictive, with $r^2 = 0.987$ and $q^2 = 0.944$ for MM, $r^2 = 0.985$ and $q^2 = 0.950$ for FM, $r^2 = 0.989$ and $q^2 = 0.962$ for MR, and $r^2 = 0.990$ and $q^2 = 0.965$ for FR. Model validations were carried out by response permutation testing and external validation, and the results indicated that the models were both statistically significant and robust. Coefficient analysis and VIP analysis identified the specific atom types and fragments that contributed most significantly to carcinogenicity and response differences across species and gender.

ACKNOWLEDGMENT

The author would like to acknowledge Drs. Sung-Sau So and David Fry for critical reading of the manuscript and insightful suggestions.

REFERENCES AND NOTES

(1) Sun, H. A Universal Molecular Descriptor System for Prediction of logP, logS, logBB and Absorption. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 748−757.
(2) National Toxicology Program. http://ntp-server.niehs.nih.gov/main_pages/about_NTP.html.
(3) Contrera, J. F.; Jacobs, A. C.; DeGorge, J. J. Carcinogenicity testing and the evaluation of regulatory requirements for pharmaceuticals. *Regul. Toxicol. Pharmacol.* **1997**, *25*, 130−145.
(4) Sato, S.; Tomita, I. Short-Term Screening Method for the Prediction of Carcinogenicity of Chemical Substances: Current Status and Problems of an in vivo Rodent Micronucleus Assay. *J. Health Sci.* **2001**, *47*, 1−8.
(5) Ames, B. N.; Durston, W. E.; Yamasaki, E.; Lee, F. D. Carcinogens Are Mutagens: A Simple Test System Combining Liver Homogenates for Activation and Bacteria for Detection. *Proc. Natl. Acad. Sci. U.S.A.* **1973**, *70*, 2281−2285.
(6) Maron, D. M.; Ames, B. N. Revised Methods for the Salmonella Mutagenicity Test. *Mutat. Res.* **1983**, *113*, 173−215.
(7) Yuspa, S. H.; Poirier, M. C. Chemical Carcinogenesis: From Animal Models to Molecular Models in One Decade. *Adv. Cancer Res.* **1988**, *50*, 25−70.
(8) Lee, Y.; Buchanan, B. G.; Klopman, G.; Dimayuga, M.; Rosenkranz, H. S. The Potential of Organ Specific Toxicity for Predicting Rodent Carcinogenicity. *Mutat. Res.* **1996**, *358*, 37−62.
(9) Sanderson, D. M.; Earnshaw, C. G. Compuer Prediction of Possible Toxic Action from Chemical Structure; the DEREK System. *Human Exp. Toxicol.* **1991**, *10*, 261−273.
(10) Franke, R.; Gruska, A.; Giuliani, A.; Benigni, R. Prediction of Rodent Carcinogenicity of Aromatic Amines: A Quantitative Structure−Activity Relationships Model. *Carcinogenisis* **2001**, *22*, 1561−1571.
(11) Benigni, R.; Passerini, L. Carcinogenicity of the Aromatic Amines: From Structure−Activity Relationships to Mechanisms of Action and Risk Assessment. *Mutat. Res.* **2002**, *511*, 191−206.
(12) Cariello, N. F.; Wilson, J. D.; Britt, B. H.; Wedd, D. J. Burlinson, B.; Gombar, V. Comparison of the Computer Programs DEREK and TOPKAT to Predict Bacterial Mutagenicity. *Mutagenesis* **2002**, *14*, 321−329.
(13) Enslein, K.; Gombar, V. K.; Blake, B. W. Use of SAR in Computer-Assisted Prediction of Carcinogenicity and Mutagenicity of Chemicals by the TOPKAT Program. *Mutat. Res.* **1994**, *305*, 47−61.
(14) Debnath, A. K.; Lopez de Compadre, R. L.; Debnath, G.; Shusterman, A. J.; Hansch, C. Structure−Activity Relationship of Mutagenic Aromatic and Heteroaromatic Nitro Compounds. *J. Med. Chem.* **1991**, *34*, 786−797.
(15) Young, S. S.; Gombar, V. K.; Emptage, M. R.; Cariello, N. F.; Lambert, C. Mixture Deconvolution and Analysis of Ames Mutagenicity Data. *Chem. Intell. Lab. Sys.* **2002**, *60*, 5−11.
(16) Lee, Y.; Buchanan, B. G.; Mattison, D. M.; Klopman, G.; Dimayuga, M.; Rosenkranz, H. S. Learning Rules to Predict Rodent Carcinogenicity of Non-Genotoxic Chemicals. *Mutat. Res.* **1995**, *328*, 127−149.
(17) Taningher, M.; Malacarne, D.; Perrotta, A.; Parodi, S. Computer-Aided Analysis of Mutagenicity and Cell Transformation Data for Assessing Their Relationship With Carcinogenicity. *Environ. Mol. Mutagen.* **1999**, *33*, 226−239.
(18) Klopman, G.; Chakravarti, S. K.; Zhu, H.; Ivanov, J. M.; Saiakhov, R. D. ESP: A Method To Predict Toxicity and Pharmacological Properties of Chemicals Using Multiple MCASE Databases. *J. Chem. Inf. Comput. Sci.* **2004**, *41*, 671−678.
(19) http://www.predictive-toxicology.org/data/ntp/.
(20) http://www.predictive-toxicology.org/data/fda/.
(21) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Multi- and Megavariate Data Analysis Principles and Applications*; 2001; Umetrics Academy: Kinnelon, NJ.
(22) SIMCA-P 10.0, Umetric Inc., Kinnelon, NJ. http://www.umetrics.com.
(23) ChemTree, Golden Helix, Inc. Bozeman, MT. http://www.goldenhelix.com.
(24) Rao, S. N.; Stockfisch, T. P. Partially Unified Multiple Property Recursive Partitioning (PUMP−RP) Analyses of Cyclooxygenase (COX) Inhibitors. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1614−1622.

CI049917Y