

Development of Binary Classification of Structural Chromosome Aberrations for a Diverse Set of Organic Compounds from Molecular Structure

J. R. Serra,[†] E. D. Thompson,[‡] and P. C. Jurs^{*,†}

The Pennsylvania State University, 152 Davey Laboratory, Chemistry Department, University Park, Pennsylvania 16802, and The Procter & Gamble Co., Miami Valley Laboratories, Cincinnati, Ohio 45253-8707

Received August 14, 2002

Classification models are generated to predict in vitro cytogenetic results for a diverse set of 383 organic compounds. Both k-nearest neighbor and support vector machine models are developed. They are based on calculated molecular structure descriptors. Endpoints used are the labels clastogenic or nonclastogenic according to an in vitro chromosomal aberration assay with Chinese hamster lung cells. Compounds that were tested with both a 24 and 48 h exposure are included. Each compound is represented by calculated molecular structure descriptors encoding the topological, electronic, geometrical, or polar surface area aspects of the structure. Subsets of informative descriptors are identified with genetic algorithm feature selection coupled to the appropriate classification algorithm. The overall classification success rate for a k-nearest neighbor classifier built with just six topological descriptors is 81.2% for the training set and 86.5% for an external prediction set. The overall classification success rate for a three-descriptor support vector machine model is 99.7% for the training set, 92.1% for the cross-validation set, and 83.8% for an external prediction set.

Introduction

Every consumer compound, industrial solvent, and byproduct must be tested for adverse effects to people, animals, plants, and the environment. Chemical toxicity can be characterized as acute, where effects develop shortly after a single exposure, or chronic, where effects develop from small dosages over time (1). During the safety assessment process, the ability of a chemical to induce genotoxic effects, including mutations and chromosome aberrations, is assessed. Experimental techniques to detect chromosome aberrations in vitro are reliable, but can be costly and time-consuming when multiple compounds need to be screened. On the basis of this, we have investigated computational models to predict which chemicals may induce structural chromosomal aberrations in vitro.

The development of classification models using genotoxicity data have found application in high throughput screening. The classification model is therefore assigned the duty of a first line of defense in toxicity testing. The idea behind these classification models is to identify compounds that may be toxic based on molecular structure, thereby aiding the experimental toxicologist.

An advantage of developing inductive classification methods in silico that rely only on molecular structure is that no prior knowledge of mechanism of action is needed. Computational classification models can lend support to a particular mechanism of action, but in general they cannot propose or contradict a mechanism. In this study, no prior information regarding mechanism

of action of the compounds involved was assumed nor were any metabolites, such as any due to alteration by S-9 liver homogenase, considered.

The data for 901 compounds were obtained from *Compilation of Chromosomal Mutation Test Data* which represents 20 years of testing carried out by the National Drug and Food Safety Laboratory and the First Laboratory of the Mutation Genetics Department of the Safety and Biotesting Research Center in Japan (2). In these studies, Chinese hamster lung cells cultured in vitro were exposed to test chemicals. Two types of exposures were used in the analysis described in this report. In the first type, cells were exposed to the test chemical for 24 h, and in the second type, cells were exposed to the test chemical for 48 h. A subset of 383 compounds was selected that had been tested with both a 24–48 h exposure. The experimental endpoint to be predicted is given as positive (10% or greater aberrant cells), equivocal (5%–10% aberrant cells), and negative (5% or less aberrant cells) (2). This paper addresses the binary classification of positive and negative endpoints; therefore, equivocal compounds are omitted from this study. Usable equivocal compounds comprised less than 5% of the data, ruling out a three-class problem. For compounds with multiple trials, the results for all trials must match for a compound to be included in this data set. The compounds include known carcinogens, drugs, food additives, agrochemicals, cosmetic materials, medicinal products, and household materials.

Methodology

Experimental. A detailed description of the experimental methods is provided in Ojima (2). Briefly, test substances were dissolved in physiological saline or distilled water. Insoluble

* To whom correspondence should be addressed.

[†] The Pennsylvania State University

[‡] The Procter & Gamble Co.

compounds were dissolved in DMSO or ethanol. Some insoluble compounds were suspended in CMC (carboxymethylcellulose sodium salt). Cultured Chinese hamster lung cells were then exposed to each test substance for 24 or 48 h. After exposure, cells were processed by standard methods, and chromosomal aberrations were identified.

Data Sets. The overall data set of 383 compounds (Table 1) contained 271 nonclastogenic compounds (71%) and 112 clastogenic compounds (29%). The molecular weights ranged from 30 to 660 amu. The data set was broken down into training sets, cross-validation sets, and a prediction set. The 71:29% ratio was maintained in all subsets. Two training sets were created as part of this work. Training set 1 contains 346 compounds (245 nonclastogenic and 101 clastogenic) which is 90% of the compounds. The remaining 10% of the compounds, 37 compounds (26 nonclastogenic and 11 clastogenic) comprised the external prediction set. These 37 compounds were never used during model formation. They were only used for predictive ability estimation by completed models. The selection of compounds to form the training set and prediction set was done randomly, but with the restriction that the fraction of compounds in the nonclastogenic and clastogenic class must conform to the 71:29% overall distribution. Training set 2 was derived from training set 1 by removing a randomly selected 38 compounds to form a cross-validation set, which is necessary for development of the support vector machine models. Thus, training set 2 contains 308 compounds (218 nonclastogens and 90 clastogens). The cross-validation set contains 38 compounds (27 nonclastogens and 11 clastogens). The prediction set remains as it was with the same 38 compounds (27 nonclastogens and 11 clastogens). The compositions of these subsets are summarized in Table 2.

Model Development. Numerical descriptors that encode topological (3, 4), geometrical (5, 6), and electronic (7, 8) properties of the molecules were used to create classification models. A classification study begins with a digital representation of the molecular structure. Then, descriptors are generated based on those structures. Next, only information rich descriptors are desirable, so descriptor reduction or feature selection is performed. Then, classification models are constructed based on the smallest best subset of descriptors found. Finally, the classification models are validated using prediction sets, previously unknown to the classifier, and tested for chance correlations.

Structure Entry and Optimization. Structures for the test chemicals in *Compilation of Chromosomal Mutation Test Data* (2) were put into an sdf file (MDL ISIS sdf file) by Procter & Gamble. These structures were then checked in HyperChem (Hypercube, Inc. Waterloo, ON) on a PC. Structures were geometry-optimized with the PM3 Hamiltonian (9) using the commercially available software package MOPAC¹ (10). Where accurate charge information was required, a single-point energy calculation was performed using the AM1 Hamiltonian on the PM3 optimized structures using MOPAC (11). The suitability of these Hamiltonians for these purposes is described in the literature (12).

Descriptor Generation and Feature Selection. Descriptors in this work were created using ADAPT (Automated Data Analysis and Pattern Recognition Toolkit) (13, 14) developed by the Jurs research group. The ADAPT software package has been shown to provide highly predictive models for various pharmaceutical and toxic properties (15–17). Descriptors encode molecular structure by calculating numerical values for topological, geometric, and electronic features. Topological descriptors use only the connection table of a molecule and therefore do not require accurate 3-D optimized structures. These descriptors encode simple counts of atom types, bond types, connectivity indices (18), and interatomic distances (19). Topological descrip-

tors have been shown to be correlated with molecular size, shape, and degree of branching (20). Geometric descriptors encode information on the overall size and shape of a molecule, and they therefore require accurate 3-D geometries. Here, the PM3 geometry optimized structures are used. Examples of geometric descriptors include length-to-breadth ratios (21), 2-D shadow projection areas (6), and solvent accessible surface areas (5). Electronic descriptors encode charge information of the molecule. As with geometric descriptors, electronic descriptors require accurate 3-D geometries. The electronic descriptors use single-point AM1 charges from the PM3 geometry optimized structures. Examples of electronic descriptors include atomic partial charges (22), dipole moment, and electron–core repulsion energies. The surface areas of the geometric descriptors are combined with the partial charges of the electronic descriptors to form a hybrid set of descriptors, charged partial surface areas (CPSAs) (8). These descriptors provide information on atomic charges relative to the whole molecule, weighted partial charges relative to surface areas, and fractional partial charges relative to surface areas. CPSA descriptors are closely related to Polar Surface Area Descriptors that are widely used in QSAR applications (23). Selective CPSA descriptors can be formed to create hydrogen bonding (24) descriptors which encode information on proton acceptor and donor sites.

Approximately 250 descriptors were calculated for each compound. Many of those descriptors contained redundant information, highly correlated information, or very little useful information. Objective feature selection attempts to reduce the total descriptor pool by eliminating descriptors that are redundant or contain no new information without the use of the dependent variable (toxicity value). Descriptors were deemed not useful if they contained over 80% redundant information or if they were 80% correlated with another descriptor from the training sets. This reduced the number of descriptors by two-thirds.

Finally, the best smallest subsets of the reduced pool were found using subjective feature selection. Here, the toxicity values were taken into account for the training sets. Very small user-defined subsets of descriptors, usually 3–10 descriptors per model, are identified using a genetic algorithm (25, 26).

Classification Models. Many classification schemes were attempted in this study. They include k-nearest neighbor, linear discriminants (27), probabilistic neural networks (28, 29), and support vector machines (30).

The k-nearest neighbor (k-NN) classification algorithm coupled to a genetic algorithm for feature selection was used initially. A leave-one-out method of validation is used to compare each member of the training set to each other in descriptor space. The compound's class is chosen based on the Euclidian distance of the compound to its k-closest neighbors. In this study, several odd values of k were used, and $k = 3$ provided the results to be discussed.

While the k-NN classification technique is simple, it is also quite powerful. However, k-NN classifiers tend to break down with highly skewed data sets. Support vector machines (SVM), a neural network approach, have been shown to perform well on skewed data sets. Briefly, SVM classification is based on the optimal separation of classes from one another (30). This is achieved by finding a hyperplane between classes such that the distance from the boundary compounds of each class to the hyperplane is maximized. These boundary compounds, in descriptor space, define the hyperplane and are called support vectors. Linearly inseparable data are transformed via kernel functions before the hyperplane is found. An SVM coupled to a genetic algorithm was used in this study.

All descriptor calculations and geometry optimizations for this work were performed on a DEC 300 AXP Model 500 workstation. All classification routines were performed on a 3-Node Linux cluster with 1.0 GHz AMD Athlon CPUs and the Red Hat 7.2 operating system.

¹ Abbreviations: MOPAC, molecular orbital package; ADAPT, automated data analysis and pattern recognition toolkit; CPSA, charged partial surface area; k-NN, k nearest neighbor; SVM, support vector machine; GA, genetic algorithm.

Table 1. Experimental and Observed Results for the 24/48 Hour Structural Aberration Studies Containing Compound Identification Number, CAS Number, Set, Experimental Result, k-Nearest Neighbor Calculated Result, and Support Vector Machine (L1-Norm kernel) Calculated Result

ID	CAS	set ^a	obsd ^b	calcd k-NN ^b	calcd SVM ^b	ID	CAS	set ^a	obsd ^b	calcd k-NN ^b	calcd SVM ^b
1	50-21-5	1,3	NonCL	CL	NonCL	69	120-83-2	1,2	NonCL	NonCL	NonCL
2	5103-71-9	1,3	NonCL	NonCL	NonCL	70	98-95-3	1,2	NonCL	NonCL	NonCL
3	54-11-5	1,3	NonCL	NonCL	NonCL	71	115-32-2	1,2	NonCL	NonCL	NonCL
4	484-78-6	1,3	NonCL	CL	NonCL	72	612-60-2	1,2	NonCL	NonCL	NonCL
5	37415-56-8	1,3	NonCL	NonCL	NonCL	73	71-55-6	1,2	NonCL	NonCL	NonCL
6	101-14-4	1,3	NonCL	CL	NonCL	74	93-58-3	1,2	NonCL	NonCL	NonCL
7	363-03-1	1,3	NonCL	NonCL	NonCL	75	465-42-9	1,2	NonCL	NonCL	NonCL
8	94-75-7	1,3	NonCL	NonCL	NonCL	76	100-41-4	1,2	NonCL	NonCL	NonCL
9	9003-39-8	1,3	NonCL	CL	NonCL	77	76824-35-6	1,2	NonCL	NonCL	NonCL
10	105-54-4	1,3	NonCL	NonCL	NonCL	78	2216-51-5	1,2	NonCL	NonCL	NonCL
11	142-47-2	1,3	NonCL	NonCL	NonCL	79	60-00-4	1,2	NonCL	NonCL	NonCL
12	224-42-0	1,3	NonCL	NonCL	NonCL	80	100-51-6	1,2	NonCL	NonCL	NonCL
13	6494-88-8	1,3	NonCL	CL	NonCL	81	61347-09-9	1,2	NonCL	NonCL	NonCL
14	70497-14-2	1,3	NonCL	NonCL	NonCL	82	25104-18-1	1,2	NonCL	NonCL	NonCL
15	77-73-6	1,3	NonCL	NonCL	NonCL	83	61-50-7	1,2	NonCL	NonCL	NonCL
16	135-88-6	1,3	NonCL	NonCL	NonCL	84	3081-61-6	1,2	NonCL	NonCL	NonCL
17	68061-82-5	1,3	NonCL	NonCL	NonCL	85	59-02-9	1,2	NonCL	NonCL	NonCL
18	17673-25-5	1,3	NonCL	NonCL	NonCL	86	148-79-8	1,2	NonCL	CL	NonCL
19	3817-11-6	1,3	NonCL	NonCL	NonCL	87	94-74-6	1,2	NonCL	NonCL	NonCL
20	84-74-2	1,3	NonCL	NonCL	NonCL	88	59665-06-4	1,2	NonCL	CL	NonCL
21	328-50-7	1,3	NonCL	NonCL	NonCL	89	122-39-4	1,2	NonCL	NonCL	NonCL
22	50-14-6	1,3	NonCL	NonCL	NonCL	90	50-70-4	1,2	NonCL	NonCL	NonCL
23	57-83-0	1,3	NonCL	NonCL	NonCL	91	127-47-9	1,2	NonCL	NonCL	NonCL
24	90-30-2	1,3	NonCL	NonCL	NonCL	92	64005-59-0	1,2	NonCL	NonCL	NonCL
25	84-66-2	1,3	NonCL	NonCL	NonCL	93	129-00-0	1,2	NonCL	NonCL	NonCL
26	57912-86-4	1,3	NonCL	NonCL	NonCL	94	110-54-3	1,2	NonCL	NonCL	NonCL
27	120-82-1	1,3	NonCL	NonCL	NonCL	95	108-30-5	1,2	NonCL	CL	NonCL
28	107-92-6	1,2	NonCL	NonCL	NonCL	96	1014-70-6	1,2	NonCL	NonCL	NonCL
29	90-80-2	1,2	NonCL	NonCL	NonCL	97	77-92-9	1,2	NonCL	NonCL	NonCL
30	144-62-7	1,2	NonCL	NonCL	NonCL	98	58-94-6	1,2	NonCL	CL	NonCL
31	439-14-5	1,2	NonCL	NonCL	NonCL	99	2921-88-2	1,2	NonCL	NonCL	NonCL
32	15950-66-0	1,2	NonCL	NonCL	NonCL	100	123-11-5	1,2	NonCL	NonCL	NonCL
33	14929-11-4	1,2	NonCL	NonCL	NonCL	101	97-56-3	1,2	NonCL	NonCL	NonCL
34	38869-91-9	1,2	NonCL	NonCL	NonCL	102	4247-02-3	1,2	NonCL	NonCL	NonCL
35	59-87-0	1,2	NonCL	NonCL	NonCL	103	70699-77-3	1,2	NonCL	CL	NonCL
36	92-52-4	1,2	NonCL	NonCL	NonCL	104	659-70-1	1,2	NonCL	NonCL	NonCL
37	5566-34-7	1,2	NonCL	NonCL	NonCL	105	611-32-5	1,2	NonCL	NonCL	NonCL
38	50-06-6	1,2	NonCL	CL	NonCL	106	78-70-6	1,2	NonCL	NonCL	NonCL
39	58-86-6	1,2	NonCL	NonCL	NonCL	107	64-19-7	1,2	NonCL	NonCL	NonCL
40	142-04-1	1,2	NonCL	CL	NonCL	108	539-89-9	1,2	NonCL	NonCL	NonCL
41	127-18-4	1,2	NonCL	NonCL	NonCL	109	10097-16-2	1,2	NonCL	NonCL	NonCL
42	54-12-6	1,2	NonCL	CL	NonCL	110	62-75-9	1,2	NonCL	NonCL	NonCL
43	7451-46-9	1,2	NonCL	NonCL	NonCL	111	1107-26-2	1,2	NonCL	NonCL	NonCL
44	106-46-7	1,2	NonCL	NonCL	NonCL	112	87-69-4	1,2	NonCL	NonCL	NonCL
45	117-81-7	1,2	NonCL	NonCL	NonCL	113	50-55-5	1,2	NonCL	NonCL	NonCL
46	81-15-2	1,2	NonCL	CL	NonCL	114	105-46-4	1,2	NonCL	CL	NonCL
47	935-95-5	1,2	NonCL	NonCL	NonCL	115	81-07-2	1,2	NonCL	CL	NonCL
48	121-14-2	1,2	NonCL	NonCL	NonCL	116	153-18-4	1,2	NonCL	CL	NonCL
49	88-06-2	1,2	NonCL	NonCL	NonCL	117	87-29-6	1,2	NonCL	NonCL	NonCL
50	64005-58-9	1,2	NonCL	NonCL	NonCL	118	156-59-2	1,2	NonCL	NonCL	NonCL
51	123-86-4	1,2	NonCL	NonCL	NonCL	119	542-18-7	1,2	NonCL	NonCL	NonCL
52	464-49-3	1,2	NonCL	NonCL	NonCL	120	609-19-8	1,2	NonCL	NonCL	NonCL
53	514-78-3	1,2	NonCL	NonCL	NonCL	121	1079-21-6	1,2	NonCL	NonCL	NonCL
54	104-67-6	1,2	NonCL	NonCL	NonCL	122	69-93-2	1,2	NonCL	NonCL	NonCL
55	101-97-3	1,2	NonCL	CL	NonCL	123	110-17-8	1,2	NonCL	NonCL	NonCL
56	99-08-1	1,2	NonCL	NonCL	NonCL	124	128-37-0	1,2	NonCL	CL	NonCL
57	19666-30-9	1,2	NonCL	NonCL	NonCL	125	54897-62-0	1,2	NonCL	CL	NonCL
58	94-26-8	1,2	NonCL	NonCL	NonCL	126	95-51-2	1,2	NonCL	NonCL	NonCL
59	25013-16-5	1,2	NonCL	NonCL	NonCL	127	123-92-2	1,2	NonCL	NonCL	NonCL
60	422-05-9	1,2	NonCL	NonCL	NonCL	128	860-22-0	1,2	NonCL	CL	NonCL
61	79-01-6	1,2	NonCL	NonCL	NonCL	129	5324-12-9	1,2	NonCL	NonCL	NonCL
62	107-06-2	1,2	NonCL	NonCL	NonCL	130	7235-40-7	1,2	NonCL	NonCL	NonCL
63	26087-47-8	1,2	NonCL	NonCL	NonCL	131	35089-66-8	1,2	NonCL	NonCL	NonCL
64	87084-52-4	1,2	NonCL	NonCL	NonCL	132	121-32-4	1,2	NonCL	NonCL	NonCL
65	110-44-1	1,2	NonCL	NonCL	NonCL	133	71-00-1	1,2	NonCL	NonCL	NonCL
66	57-63-6	1,2	NonCL	NonCL	NonCL	134	333-41-5	1,2	NonCL	NonCL	NonCL
67	4191-73-5	1,2	NonCL	NonCL	NonCL	135	10236-47-2	1,2	NonCL	CL	NonCL
68	126-73-8	1,2	NonCL	NonCL	NonCL	136	73-22-3	1,2	NonCL	CL	NonCL

Table 1 (Continued)

ID	CAS	set ^a	obsd ^b	calcd k-NN ^b	calcd SVM ^b	ID	CAS	set ^a	obsd ^b	calcd k-NN ^b	calcd SVM ^b
137	101-67-7	1,2	NonCL	NonCL	NonCL	205	3322-93-8	1,2	NonCL	NonCL	NonCL
138	598-50-5	1,2	NonCL	NonCL	NonCL	206	226-36-8	1,2	NonCL	NonCL	NonCL
139	91-22-5	1,2	NonCL	NonCL	NonCL	207	123-68-2	1,2	NonCL	NonCL	NonCL
140	108-70-3	1,2	NonCL	NonCL	NonCL	208	99-99-0	1,2	NonCL	NonCL	NonCL
141	62-53-3	1,2	NonCL	CL	NonCL	209	834-12-8	1,2	NonCL	NonCL	NonCL
142	52-86-8	1,2	NonCL	NonCL	NonCL	210	115-77-5	1,2	NonCL	NonCL	NonCL
143	75-09-2	1,2	NonCL	CL	NonCL	211	592-17-6	1,2	NonCL	NonCL	NonCL
144	67-42-5	1,2	NonCL	NonCL	NonCL	212	4543-95-7	1,2	NonCL	NonCL	NonCL
145	118-74-1	1,2	NonCL	NonCL	NonCL	213	1582-09-8	1,2	NonCL	NonCL	NonCL
146	7450-62-6	1,2	NonCL	NonCL	NonCL	214	1792-17-2	1,2	NonCL	NonCL	NonCL
147	69-65-8	1,2	NonCL	NonCL	NonCL	215	299-88-7	1,2	NonCL	NonCL	NonCL
148	110-86-1	1,2	NonCL	NonCL	NonCL	216	52423-28-6	1,2	NonCL	CL	NonCL
149	106-27-4	1,2	NonCL	NonCL	NonCL	217	554-00-7	1,2	NonCL	NonCL	NonCL
150	4901-51-3	1,2	NonCL	NonCL	NonCL	218	140-03-4	1,2	NonCL	NonCL	NonCL
151	623-78-9	1,2	NonCL	NonCL	NonCL	219	7491-76-1	1,2	NonCL	NonCL	NonCL
152	64-77-7	1,2	NonCL	NonCL	NonCL	220	1912-24-9	1,2	NonCL	NonCL	NonCL
153	112-31-2	1,2	NonCL	NonCL	NonCL	221	121-33-5	1,2	NonCL	NonCL	NonCL
154	80-05-7	1,2	NonCL	NonCL	NonCL	222	2371-42-8	1,2	NonCL	NonCL	NonCL
155	110-45-2	1,2	NonCL	NonCL	NonCL	223	5392-40-5	1,2	NonCL	NonCL	NonCL
156	61-68-7	1,2	NonCL	NonCL	NonCL	224	591-62-8	1,2	NonCL	NonCL	NonCL
157	59-67-6	1,2	NonCL	NonCL	NonCL	225	141-97-9	1,2	NonCL	NonCL	NonCL
158	58-15-1	1,2	NonCL	CL	NonCL	226	132-27-4	1,2	NonCL	NonCL	NonCL
159	110-82-7	1,2	NonCL	NonCL	NonCL	227	59665-03-1	1,2	NonCL	NonCL	NonCL
160	75-35-4	1,2	NonCL	NonCL	NonCL	228	1156-19-0	1,2	NonCL	CL	NonCL
161	1393-63-1	1,2	NonCL	NonCL	NonCL	229	535-87-5	1,2	NonCL	CL	NonCL
162	120-61-6	1,2	NonCL	NonCL	NonCL	230	72-18-4	1,2	NonCL	NonCL	NonCL
163	50-81-7	1,2	NonCL	NonCL	NonCL	231	105-40-8	1,2	NonCL	NonCL	NonCL
164	105-37-3	1,2	NonCL	CL	NonCL	232	103-84-4	1,2	NonCL	CL	NonCL
165	94-36-0	1,2	NonCL	NonCL	NonCL	233	87-86-5	1,2	NonCL	NonCL	NonCL
166	58-89-9	1,2	NonCL	NonCL	NonCL	234	548-93-6	1,2	NonCL	NonCL	NonCL
167	105-60-2	1,2	NonCL	NonCL	NonCL	235	139-40-2	1,2	NonCL	NonCL	NonCL
168	72-43-5	1,2	NonCL	NonCL	NonCL	236	140-11-4	1,2	NonCL	NonCL	NonCL
169	88-72-2	1,2	NonCL	NonCL	NonCL	237	7512-17-6	1,2	NonCL	NonCL	NonCL
170	56-40-6	1,2	NonCL	NonCL	NonCL	238	5458-83-3	1,2	NonCL	NonCL	NonCL
171	54897-63-1	1,2	NonCL	NonCL	NonCL	239	75-34-3	1,2	NonCL	NonCL	NonCL
172	62-56-6	1,2	NonCL	NonCL	NonCL	240	99-09-2	1,2	NonCL	NonCL	NonCL
173	64005-60-3	1,2	NonCL	NonCL	NonCL	241	53-86-1	1,2	NonCL	NonCL	NonCL
174	16561-29-8	1,2	NonCL	NonCL	NonCL	242	108-90-7	1,2	NonCL	NonCL	NonCL
175	319-84-6	1,2	NonCL	NonCL	NonCL	243	627-06-5	1,2	NonCL	NonCL	NonCL
176	520-45-6	1,2	NonCL	NonCL	NonCL	244	3648-21-3	1,2	NonCL	NonCL	NonCL
177	95-95-4	1,2	NonCL	NonCL	NonCL	245	56-84-8	1,2	NonCL	NonCL	NonCL
178	56-53-1	1,2	NonCL	NonCL	NonCL	246	1401-55-4	1,3	CL	CL	CL
179	50-60-2	1,2	NonCL	NonCL	NonCL	247	10589-74-9	1,3	CL	CL	CL
180	54-88-6	1,2	NonCL	NonCL	NonCL	248	58-55-9	1,3	CL	CL	CL
181	100-42-5	1,2	NonCL	NonCL	NonCL	249	2052-01-9	1,3	CL	NonCL	CL
182	67-21-0	1,2	NonCL	NonCL	NonCL	250	50-00-0	1,3	CL	CL	NonCL
183	23333-91-7	1,2	NonCL	NonCL	NonCL	251	56986-35-7	1,3	CL	CL	CL
184	631-27-6	1,2	NonCL	NonCL	NonCL	252	2783-94-0	1,3	CL	CL	CL
185	9004-67-5	1,2	NonCL	CL	NonCL	253	106-50-3	1,3	CL	CL	CL
186	108-88-3	1,2	NonCL	NonCL	NonCL	254	61-25-6	1,3	CL	NonCL	CL
187	119-36-8	1,2	NonCL	NonCL	NonCL	255	2111-75-3	1,3	CL	NonCL	NonCL
188	101-68-8	1,2	NonCL	NonCL	NonCL	256	59-92-7	1,3	CL	NonCL	NonCL
189	95-50-1	1,2	NonCL	NonCL	NonCL	257	51-21-8	1,2	CL	NonCL	CL
190	584-79-2	1,2	NonCL	NonCL	NonCL	258	501-30-4	1,2	CL	NonCL	NonCL
191	56-23-5	1,2	NonCL	NonCL	NonCL	259	67-20-9	1,2	CL	CL	CL
192	625-52-5	1,2	NonCL	NonCL	NonCL	260	28895-91-2	1,2	CL	CL	CL
193	107-35-7	1,2	NonCL	CL	NonCL	261	494-03-1	1,2	CL	NonCL	CL
194	83-86-3	1,2	NonCL	NonCL	NonCL	262	615-53-2	1,2	CL	CL	CL
195	6915-15-7	1,2	NonCL	NonCL	NonCL	263	57-14-7	1,2	CL	CL	CL
196	59-00-7	1,2	NonCL	CL	NonCL	264	58139-33-6	1,2	CL	CL	CL
197	924-16-3	1,2	NonCL	NonCL	NonCL	265	57-50-1	1,2	CL	CL	CL
198	35089-69-1	1,2	NonCL	NonCL	NonCL	266	62-73-7	1,2	CL	CL	CL
199	50-37-3	1,2	NonCL	NonCL	NonCL	267	1343-78-8	1,2	CL	CL	CL
200	56-86-0	1,2	NonCL	NonCL	NonCL	268	57-13-6	1,2	CL	NonCL	NonCL
201	352-97-6	1,2	NonCL	NonCL	NonCL	269	64005-62-5	1,2	CL	CL	CL
202	73-32-5	1,2	NonCL	NonCL	NonCL	270	60391-92-6	1,2	CL	CL	CL
203	110-15-6	1,2	NonCL	NonCL	NonCL	271	396-01-0	1,2	CL	CL	CL
204	85-01-8	1,2	NonCL	NonCL	NonCL	272	122-14-5	1,2	CL	CL	CL

Table 1 (Continued)

ID	CAS	set ^a	obsd ^b	calcd k-NN ^b	calcd SVM ^b	ID	CAS	set ^a	obsd ^b	calcd k-NN ^b	calcd SVM ^b
273	90-04-0	1,2	CL	CL	CL	329	55726-47-1	1,2	CL	NonCL	CL
274	154-93-8	1,2	CL	CL	CL	330	66-27-3	1,2	CL	CL	CL
275	75321-20-9	1,2	CL	CL	CL	331	3688-53-7	1,2	CL	CL	CL
276	100-22-1	1,2	CL	NonCL	CL	332	67-64-1	1,2	CL	NonCL	CL
277	598-72-1	1,2	CL	NonCL	NonCL	333	1116-54-7	1,2	CL	NonCL	CL
278	156-43-4	1,2	CL	CL	CL	334	121-88-0	1,2	CL	NonCL	CL
279	42397-65-9	1,2	CL	CL	CL	335	50-18-0	1,2	CL	CL	CL
280	81-88-9	1,2	CL	NonCL	CL	336	62450-07-1	1,2	CL	CL	CL
281	19935-86-5	1,2	CL	CL	CL	337	2451-62-9	1,2	CL	CL	CL
282	56525-09-8	1,2	CL	CL	CL	338	67977-01-9	1,2	CL	CL	CL
283	83-88-5	1,2	CL	CL	CL	339	62-50-0	1,2	CL	CL	CL
284	1934-21-0	1,2	CL	NonCL	CL	340	133-06-2	1,2	CL	NonCL	CL
285	680-31-9	1,2	CL	CL	CL	341	230-27-3	1,2	CL	NonCL	CL
286	96-13-9	1,2	CL	CL	CL	342	133-67-5	1,2	CL	NonCL	CL
287	63-25-2	1,2	CL	CL	CL	343	154-23-4	1,2	CL	NonCL	CL
288	15972-60-8	1,2	CL	CL	NonCL	344	50-07-7	1,2	CL	CL	CL
289	13010-08-7	1,2	CL	CL	CL	345	79-10-7	1,2	CL	CL	CL
290	55-98-1	1,2	CL	NonCL	CL	346	79-06-1	1,2	CL	CL	CL
291	260-94-6	1,2	CL	CL	CL	347	120-12-7	4	NonCL	NonCL	NonCL
292	104-55-2	1,2	CL	CL	CL	348	80-68-2	4	NonCL	NonCL	NonCL
293	760-56-5	1,2	CL	CL	CL	349	87-61-6	4	NonCL	NonCL	NonCL
294	52-24-4	1,2	CL	CL	CL	350	78-43-3	4	NonCL	NonCL	NonCL
295	58-08-2	1,2	CL	CL	CL	351	123-66-0	4	NonCL	NonCL	NonCL
296	58139-35-8	1,2	CL	CL	CL	352	24019-05-4	4	NonCL	NonCL	NonCL
297	121-75-5	1,2	CL	CL	CL	353	86-30-6	4	NonCL	NonCL	NonCL
298	339-44-6	1,2	CL	NonCL	CL	354	105-68-0	4	NonCL	NonCL	NonCL
299	93-46-9	1,2	CL	CL	CL	355	89-65-6	4	NonCL	NonCL	NonCL
300	75-07-0	1,2	CL	CL	CL	356	59-30-3	4	NonCL	NonCL	CL
301	106-89-8	1,2	CL	CL	CL	357	471-80-7	4	NonCL	NonCL	NonCL
302	616-23-9	1,2	CL	CL	CL	358	106-24-1	4	NonCL	NonCL	NonCL
303	17902-23-7	1,2	CL	NonCL	CL	359	7287-19-6	4	NonCL	NonCL	NonCL
304	42397-64-8	1,2	CL	CL	CL	360	526-95-4	4	NonCL	NonCL	NonCL
305	74-31-7	1,2	CL	CL	CL	361	72-19-5	4	NonCL	NonCL	NonCL
306	54-31-9	1,2	CL	CL	CL	362	56-81-5	4	NonCL	NonCL	NonCL
307	62450-06-0	1,2	CL	CL	CL	363	106-32-1	4	NonCL	NonCL	NonCL
308	684-93-5	1,2	CL	CL	CL	364	90-43-7	4	NonCL	NonCL	NonCL
309	147-94-4	1,2	CL	CL	CL	365	1897-45-6	4	NonCL	NonCL	NonCL
310	6558-78-7	1,2	CL	CL	CL	366	101-25-7	4	NonCL	CL	NonCL
311	458-37-7	1,2	CL	NonCL	CL	367	108-64-5	4	NonCL	NonCL	NonCL
312	70-25-7	1,2	CL	CL	CL	368	62-55-5	4	NonCL	NonCL	NonCL
313	121-79-9	1,2	CL	NonCL	CL	369	103-36-6	4	NonCL	NonCL	NonCL
314	1129-41-5	1,2	CL	CL	CL	370	106-47-8	4	NonCL	NonCL	NonCL
315	24423-85-6	1,2	CL	CL	CL	371	34522-32-2	4	NonCL	CL	NonCL
316	306-37-6	1,2	CL	NonCL	CL	372	5522-43-0	4	NonCL	NonCL	CL
317	140-88-5	1,2	CL	CL	CL	373	614-95-9	4	CL	CL	CL
318	63885-23-4	1,2	CL	CL	CL	374	968-81-0	4	CL	NonCL	CL
319	6494-81-1	1,2	CL	CL	CL	375	7090-25-7	4	CL	CL	CL
320	1024-57-3	1,2	CL	NonCL	CL	376	38604-70-5	4	CL	CL	NonCL
321	60-27-5	1,2	CL	CL	CL	377	60-51-5	4	CL	CL	CL
322	869-01-2	1,2	CL	CL	CL	378	118-71-8	4	CL	NonCL	NonCL
323	122-60-1	1,2	CL	NonCL	CL	379	54-85-3	4	CL	CL	CL
324	128-44-9	1,2	CL	NonCL	NonCL	380	59-98-3	4	CL	NonCL	NonCL
325	98-92-0	1,2	CL	NonCL	CL	381	96-09-3	4	CL	CL	CL
326	95-54-5	1,2	CL	CL	CL	382	25956-17-6	4	CL	CL	CL
327	49606-40-8	1,2	CL	CL	CL	383	56986-37-9	4	CL	CL	CL
328	59665-11-1	1,2	CL	NonCL	CL						

^a Training set for k-nearest neighbor model = 1, training set for support vector machine model = 2, cross-validation set for the support vector machine model = 3, prediction set common to both the k-nearest neighbor and support vector machine models = 4. ^b CL = clastogenic and NonCL = nonclastogenic.

Results and Discussion

Two classification techniques provide the best results for this study. The first is the k-nearest neighbor (k-NN) classifier, and the second is the more complex support vector machine (SVM) classifier. The k-NN model developed in this work classifies based only on topological descriptors. This can be advantageous for the data sets

that contain very large and complex compounds whose geometry optimization may not be feasible. All descriptors, topological, geometric, electronic, and hybrids are used to create the SVM model. Both classifiers perform adequately, and the results are discussed.

k-Nearest Neighbor (k-NN) Model. The k-NN part of this study was done with training set 1 (upper section

Table 2

Training Set 1. Used for k-NN Model Development			
	nonclastogenic	clastogenic	total
training set	245	101	346
prediction set	26	11	37
Training Set 2. Used for SVM Model Development			
	nonclastogenic	clastogenic	total
training set	218	90	308
cross-validation set	27	11	38
prediction set	26	11	37

Table 3. Confusion Matrix for the Training and Prediction Set Compounds Using the k-Nearest Neighbor Classification Method^a

Training Set (81.2% overall correct)			
	nonclastogenic	clastogenic	percent correct
nonclastogenic	211	34	86.1%
clastogenic	31	70	69.3%
Prediction Set (86.5% overall correct)			
	nonclastogenic	clastogenic	percent correct
nonclastogenic	24	2	92.3%
clastogenic	3	8	72.7%

^a Bold values indicate correct classification.

of Table 2). After objective feature selection, 64 descriptors remained in the reduced pool. A genetic algorithm coupled to the k-NN fitness evaluator determined the best subset of descriptors from the reduced pool. Models ranging from 3–20 descriptors were formed. A six-descriptor k-NN classification model with $k = 3$ gave the best results balancing model accuracy and predictive power while using few descriptors.

The best k-NN model was determined based on number of false negatives in the training set. This criterion was chosen over the total overall classification rate of the training set, because the data set is heavily skewed toward nonclastogenic compounds and the number of clastogenic compounds misclassified as nonclastogenic (false negatives) is generally less desirable than the number of nonclastogenic compounds misclassified as clastogenic (false positives).

The classification results for the k-NN model are shown in Table 3. This 3-*nn* model has an overall training set classification rate of 81.2%, correctly classifying 281 compounds out of the 346 compound training set. This model correctly classified 211 out of 245 nonclastogenic training set compounds (86.1%) and 70 of 101 clastogenic training set compounds (69.3%). This k-NN model was then tested on the 37 compounds in the external prediction set. The overall prediction set classification rate was 86.5% (32 of 37 correct). This 3-*nn* model correctly classified 24 of 26 nonclastogenic prediction set compounds at a rate of 92.3%. The model also correctly classified 8 of 11 clastogenic prediction set compounds (72.7%). The lower success rates for the clastogenic

compound class is the usual outcome for k-NN classifiers when the data set distribution is skewed as it is for this study.

Only topological descriptors were used in this six-descriptor k-NN model. They are EAVE-2 (31), MOLC-9 (32, 33), NLP-19, MDEC-11 (19, 34), S5CH-17 (18, 35), and KAPA-6 (3, 36), as shown in Table 4. The average correlation coefficient for the six descriptors is $8.40 \times 10^{-2} \pm 0.225$ with the maximum value of 0.646 occurring between MDEC-11 and NLP-19. The descriptor EAVE-2 calculates the average E-state value over all heteroatoms in the molecule. The E-state attempts to encode valence information by calculating the ratio of the number of electrons involved in σ bonds, π bonds, and lone pairs to the number of electrons involved in sigma bonds for each atom in the molecule. These values are then appropriately summed to produce an E-state value for the compound. This descriptor gives information on the reactivity of each atom of the compound. The EAVE-2 descriptor value ranges from 0.00 to 13.49 with a mean of 7.200. The molecular connectivity descriptor, MOLC-9, calculates the topological index J and ranges from 1.268 to 4.808 with a mean of 2.797 for the training set. The topological index J , first described by Alexandru Balaban in the early 1980s (32), encodes information about degree of branching in a compound by summing the average distance connectivity. This is achieved by first calculating the distance sums per atom of each molecule, then normalizing with respect to the total number of bonds in the compound. Finally, these values are summed as one over the square root. The descriptor NLP-17 encodes the number of lone pairs of electrons in the compound and ranges from 0 to 48 with a mean of 6.5. The descriptor MDEC-11 encodes the molecular distance edge between primary carbon atoms. A molecular distance edge is defined as the through bond distance from atom A to atom B. This descriptor encodes molecular size and branching information of the molecule by taking into account only sp^3 hybridized carbon atoms. The values of MDEC-11 range from 0.00 to 27.95 with a mean of 0.4951. The descriptor S5CH-17 is the simple chi index for fifth order path chains. A simple χ index does not distinguish between single, double, triple, and aromatic bonds. A fifth order path chain contains five atoms in contiguous order. This can be in a ring of five atoms, a ring of four atoms with one branch point, or a ring of three atoms with two branch points. The S5CH-17 descriptor encodes information on molecular size and branching and ranges from 0.00 to 1.281 with a mean of 3.815×10^{-2} . The descriptor KAPA-6 encodes the third order κ index, as described by Kier and Hall, corrected for the number of atoms in the compound. The KAPA-6 descriptor describes the size and degree of branching of a compound by calculating the number of third order paths in the straight chained compound with the same number of atoms multiplied by the number of third order paths in the theoretical star compound and dividing that

Table 4. Six Topological Descriptors Used in the k-Nearest Neighbor Model

descriptor ID	type	Tset range	Tset mean	description
EAVE-2	topological	0.00–13.49	7.200	avg E-state value over all heteroatoms (34)
MOLC-9	topological	1.268–4.808	2.797	topological index J (35, 36)
NLP-19	topological	0–48	6.5	no. of lone pair electrons
MDEC-11	topological	0.00–27.95	0.4951	molecular distance edge between primary carbons (22, 37)
S5CH-17	topological	0.00–1.281	3.815×10^{-2}	fifth order path chain (21, 38)
KAPA-6	topological	0.00–18.90	4.154	third order kappa index corrected for no. of atoms (6, 39)

Table 5. Misclassified Nonclastogenic Prediction Set Compounds for the Topological k-Nearest Neighbor Model

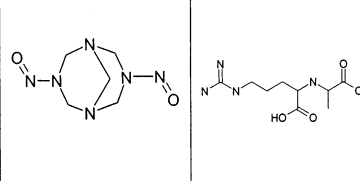
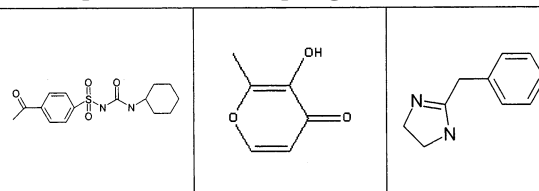
Descriptor Name	Overall Tset Range mean \pm st. dev.	Nonclastogenic Tset Range		
			ID: 366 101-25-7	ID: 371 34522-32-2
EAVE-2	0 – 13.49 7.2 \pm 2.6	2.097 – 12.58 6.5 \pm 1.8	4.11	6.987
MOLC-9	1.268 – 4.808 2.80 \pm 0.69	1.325 – 4.808 2.8 \pm 0.77	2.00	4.03
NLP-19	0 – 48 6.5 \pm 5.7	1 – 36 7.9 \pm 5.5	10	13
MDEC-11	0 – 27.95 0.49 \pm 2.4	0 – 16.9 0.55 \pm 2.4	0.00	0.1667
S5CH-17	0 – 1.281 0.038 \pm 0.14	0 – 0.933 0.048 \pm 0.14	0.00	0.00
KAPA-6	0 – 18.9 4.15 \pm 3.1	0 – 13.8 3.60 \pm 2.3	2.209	8.859

Table 6. Misclassified Clastogenic Prediction Set Compounds for the Topological k-Nearest Neighbor Model

Descriptor Name	Overall Tset Range mean \pm st. dev.	Clastogenic Tset Range			
			ID: 374 968-81-0	ID: 378 118-71-8	ID: 380 59-98-3
EAVE-2	0 – 13.49 7.2 \pm 2.6	0 – 13.49 7.5 \pm 2.8	7.087	8.06	3.796
MOLC-9	1.268 – 4.808 2.80 \pm 0.69	1.268 – 4.67 2.8 \pm 0.66	2.00	2.92	2.04
NLP-19	0 – 48 6.5 \pm 5.7	0 – 48 6.0 \pm 5.7	10	6	2
MDEC-11	0 – 27.95 0.49 \pm 2.4	0 – 27.95 0.473 \pm 2.4	0.00	0.00	0.00
S5CH-17	0 – 1.281 0.038 \pm 0.14	0 – 1.281 0.034 \pm 0.15	0.00	0.00	0.1443
KAPA-6	0 – 18.9 4.15 \pm 3.1	0 – 18.9 4.38 \pm 3.3	5.412	1.429	2.331

by the square of the actual number of third order paths in the compound. The KAPA-6 descriptor ranges from 0.00 to 18.90 with a mean of 4.154.

With such a structurally diverse data set, understanding why compounds are not classified properly is a daunting task. Of the 26 nonclastogenic prediction set compounds, 2 were misclassified (false positives). These are compounds 366 and 371, see Table 5. Compound 366 lies outside the first standard deviation about the mean of the nonclastogenic training set for descriptors EAVE-2 and MOLC-9. The pentamethylenetetramine structure is not represented in either the nonclastogenic or the clastogenic members of the training set. The only bicyclics are compounds 2, 15, 37, and 52, but none of those contain nitrogen. Compound 371 lies outside the first standard deviation about the mean of the nonclastogenic training set for descriptors MOLC-9, S5CH-17, and KAPA-6. No nonclastogenic training set compound has an overall structure similar to that of 371. However, the clastogenic compound 271 does share similar functionality. The misclassification of this compound could be due to half of the descriptors falling outside the range of the training set.

Of the 11 clastogenic compounds of the prediction set, 3 were misclassified as nonclastogenic. They are compounds 374, 378, and 380, see Table 6. All descriptor values for compound 374 lay within one standard deviation about the mean for the clastogenic training set descriptors. There are no compounds in the clastogenic training set that contains a benzene ring connected to a cyclohexane via a nitrogen-sulfate bridge. All descriptor values for compound 378 also lay within one standard deviation about the mean for each descriptor value. The pyrone backbone is found in the clastogenic training set compound 258, which is misclassified. A similar backbone is found in compound 257 of the clastogenic training set and is also misclassified. The descriptors EAVE-2 and MOLC-9 lay outside the first standard deviation about the mean for the clastogenic training set for compound 380. The imidazole ring is not represented in the clastogenic training set and, therefore, could be a cause for misclassification.

Support Vector Machine (SVM) Model. The SVM part of this study was done with training set 2 (lower section of Table 2). A cross-validation set is needed for SVM training to avoid overtraining. The same prediction

Table 7. Confusion Matrix for the Training, Cross-Validation and Prediction Set Compounds Using the Support Vector Machine Classification Method with an L1-Norm Kernel^a

Training Set (99.7% overall correct)			
	nonclastogenic	clastogenic	percent correct
nonclastogenic	218	0	100%
clastogenic	1	89	98.9
Cross-Validation Set (92.1% overall correct)			
	nonclastogenic	clastogenic	percent correct
nonclastogenic	27	0	100%
clastogenic	3	8	72.7%
Prediction Set (83.8% overall correct)			
	nonclastogenic	clastogenic	correct
nonclastogenic	23	3	88.5%
clastogenic	3	8	72.7%

^a Bold values indicate correct classification.

set that was used for the k-NN model was used for the SVM model. It consisted of 37 compounds, of which 26 were nonclastogenic and 11 were clastogenic. After objective feature selection, 81 descriptors remained in the reduced pool. Subsets of descriptors ranging from 3 to 15 were created using a genetic algorithm coupled to an SVM using the linear, polynomial ($n = 2$), chi-square, Gaussian radial basis function, and L1-Norm kernel. The best model found contained only three descriptors using the L1-Norm kernel. Once again, model accuracy and predictive power were achieved using only a minimal number of molecular descriptors. All available types of descriptors were calculated for this SVM model.

The classification results for the SVM model are shown in Table 7. The L1-Norm SVM model correctly classified 307 of 308 training set compounds (99.7%). This model correctly classified all 218 nonclastogenic training set compounds, and it correctly classified 89 of 90 clastogenic training set compounds. The classification rate for clastogenic training set compounds was 98.9%. The model correctly classified 35 of 38 cross-validation set members (92.1%). The model correctly classified all 27 nonclastogenic cross-validation set compounds and correctly classified 8 of 11 clastogenic compounds (72.7%). This SVM

model was then tested on the 37 compounds in the external prediction set. The overall prediction set classification rate was 83.8% (31 of 37 correct). Of the nonclastogenic prediction set compounds, 23 were correctly classified and 3 were misclassified (88.5%). Of the 11 clastogenic prediction set compounds, 8 were correctly classified (72.3%), while 3 were misclassified. The classification results achieved by this SVM model are the best we have found for this very structurally diverse set of compounds.

The L1-Norm SVM model needed only three descriptors to give good classification results, see Table 8. Of these descriptors, two were topological, and one was a hybrid electronic-geometric. The average correlation coefficient is $9.07 \times 10^{-2} \pm 6.94 \times 10^{-2}$ with the maximum of 0.141 between descriptors EAVE-2 and ELEC-0. The first topological descriptor, N6CH-6 (18, 35), encodes branching and molecular size information by calculating the number of rings with six constituents. The values of N6CH-6 range from 0 to 47 with a mean of 2.01 for the training set. This includes the number of six member rings, the number of five member rings, with one substituent, four member rings with two substituents, and three-member rings with three substituents. The other topological descriptor, EAVE-2 (31, 37), calculates the average E-state value over all heteroatoms. The E-state is calculated as described previously. The range of EAVE-2 for the training set is 0.00–13.45 with a mean of 7.17. The final descriptor, ELEC-0, is a hybrid electronic-geometric descriptors with values ranging from 3.167 to 6.682 with a mean of 4.92 for the training set. ELEC-0 is the electronic energy of the MOPAC PM3 geometry optimized structure. The electronic energy is determined by performing a single point MOPAC AM1 energy calculation on the MOPAC PM3 geometry optimized structure.

Compounds that were misclassified using the SVM L1-Norm approach were analyzed as the k-NN data was. Of the 26 nonclastogenic prediction set compounds, 3 were misclassified as clastogenic (false positives). As Table 9 shows, compound 356 was misclassified as clastogenic, but all descriptor values lay within one standard deviation about the mean for the nonclastogenic

Table 8. Descriptors Used in the Support Vector Machine Model

descriptor ID	type	range	mean	description
N6CH-6	topological	0–47	2.01	no. of rings with 6 constituents (21, 38)
EAVE-2	topological	0.00–13.45	7.17	avg E-state value over all heteroatoms (34)
ELEC-0	hybrid electronic-geometric	3.167–6.682	4.92	minimum electronic energy from MOPAC

Table 9. Misclassified Nonclastogenic Prediction Set Compounds for the Support Vector Machine Model

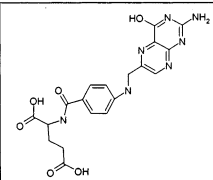
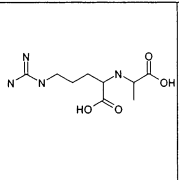
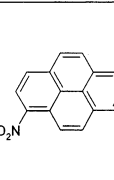
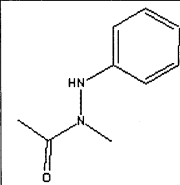
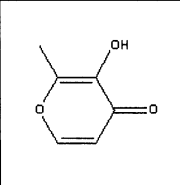
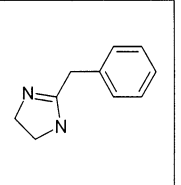
Descriptor Name	Overall Tset Range mean \pm st. dev.	Nonclastogenic Tset Range	Chemical Structure 1	Chemical Structure 2	Chemical Structure 3
					
			ID: 356 59-30-3	ID: 360 526-95-4	ID: 372 5522-43-0
N6CH-6	0–47 2.01 \pm 4.3	0–47 1.8 \pm 4.1	3	0	4
EAVE-2	0–13.45 7.2 \pm 2.5	0–13.45 7.5 \pm 2.7	6.862	8.844	7.331
ELEC-0	3.167–6.682 4.9 \pm 0.65	3.282–6.682 4.8 \pm 0.61	5.155	4.984	5.251

Table 10. Misclassified Clastogenic Prediction Set Compounds for the Support Vector Machine Model

Descriptor Name	Overall Tset Range mean \pm st. dev.	Clastogenic Tset Range			
			ID: 376 38604-70-5	ID: 378 118-71-8	ID: 380 59-98-3
N6CH-6	0 - 47 2.01 \pm 4.3	0 - 30 2.6 \pm 4.8	1	5.059	4.301
EAVE-2	0 - 13.45 7.2 \pm 2.5	2.097 - 12.58 6.4 \pm 1.8	1	8.006	4.758
ELEC-0	3.167 - 6.682 4.9 \pm 0.65	3.167 - 6.658 5.1 \pm 0.70	2	3.796	4.437

training set. However, no other compound in the non-clastogenic training set contains the pteridine ring system of compound 356, and this may contribute to the misclassification. Compound 360 was misclassified as clastogenic. All three of the descriptors for this compound fell within the first standard deviation about the mean for the nonclastogenic training set. Two similar nonclastogenic training set compounds, 1 and 90, share similar structural attributes and were correctly classified. Compound 1 has a carboxylic acid group attached to a hydroxylated carbon, and compound 90, which is also correctly classified, has the same six-member hydroxylated carbon backbone but does not contain the carboxylic acid group. Since none of the three descriptors of this model explicitly characterize the carboxyl group, that could be the reason for misclassification. Compound 372 is also misclassified as clastogenic. All descriptor values lie within one standard deviation about the mean for each descriptor in the SVM model. Several compounds with the pyrene base structure exist in the training set. Pyrene, compound 93, is correctly classified as nonclastogenic in the training set. However, the correctly classified clastogenic training set compounds 275, 279, and 304 differ from compound 372 by a second NO₂ group. Therefore, the classification algorithm may be confused, since no explicit information about nitrogen is encoded in these descriptors.

Some common characteristics of the descriptors chosen by both the k-NN and SVM models can provide some insight into the toxicity of these compounds to Chinese hamster lung cells. As shown in Table 4, four of the six descriptors that were selected by the k-NN classifier deal with size and degree of branching (MOLC-9, MDEC-11, S5CH-17, and KAPA-6). From Table 8, one (N6CH-6) of the three descriptors selected by the SVM classifier was topological, which also deals with molecular size and degree of branching. Selection of these descriptors may suggest that molecular size and steric hindrance plays a role in the aberration of lung cells from molecular structure. Both the k-NN and SVM classification models contain the average E-state over all heteroatoms, EAVE-2. This may suggest that electronic availability plays a role in the activity of these compounds to Chinese hamster lung cells.

Of the 11 clastogenic prediction set compounds, three were misclassified as nonclastogenic (false negatives), as shown in Table 10. Two of the three descriptors of compound 376 lie outside the first standard deviation about the mean for the clastogenic training set. Descrip-

tors EAVE-2 and ELEC-0 lay outside this range. Compounds 378 and 380 were also misclassified in the SVM model. Descriptor ELEC-0 of compound 378 lies outside the first standard deviation about the mean of the clastogenic training set data. Possible structural reasons for misclassification were given previously.

Testing for Chance Correlations. The two classification models used in this study were tested for chance correlations using perturbation testing and scrambling experiments. For each of the classification algorithms, five unique sets of descriptors were randomly chosen and evaluated. The overall average classification rate for the training set of the k-NN model was 67.7% and the clastogenic classification rate was 35.6%. The overall average classification rate for the prediction set of the k-NN model was 73.7% and the clastogenic classification rate was 40.0%. The overall average classification rate for the training set of the SVM model was 77.6% and the average clastogenic classification rate was 23.6%. The overall average cross-validation set classification rate was 72.6% and the average clastogenic classification rate was only 7.3%. The overall average classification rate of the prediction set was 73.2% while the average clastogenic classification rate was 7.3%. These results show that both the k-NN and SVM algorithms select relevant, information rich descriptor subsets.

In the scramble calculations, the dependent variable (nonclastogenic/clastogenic) was randomly scrambled. Then the classification algorithms coupled with the GA were run with the same training set, cross-validation set, and prediction set distributions as in the real experiments. For each model, the scrambling experiment was redone five times. The overall training set classification rate for the k-NN model was 77.2% and the clastogenic classification rate was 34.0%. The overall prediction set classification rate for this model was 42.1% and the clastogenic classification rate was only 11.1%. The extremely low prediction rate shows that the results achieved with the real experiments are very unlikely to have been influenced by chance. These lopsided values are expected due to the nature of the dataset. Roughly 70% of the compounds are nonclastogenic, so the scrambled models classify most compounds as nonclastogenic. This accounted for the rather high overall classification rate of the training set and the very poor classification rate of the prediction set clastogenic compounds.

The overall training set classification for the SVM (L1-Norm) model was 98.4% and the classification rate for the clastogenic compounds was 94.0%. The unusually

high training set classification rates can be of some concern. Ideally, all classification rates should be near random, which is 59% for this unevenly distributed data set. However, the classification rate of the training set was not considered in the creation of the best SVM model reported previously. The cross-validation set classification rates are of primary concern here. The overall classification rate of the cross-validation set was 81.6%, and the classification rate of the clastogenic compounds was 46.2%. This is what is expected from the 70% nonclastogenic compound distribution. Basically, most compounds were classified as nonclastogenic. This is shown in the prediction set, where the overall classification rate is just 47.4% and none of the clastogenic compounds are predicted correctly. The random prediction set classification success rate shows that the SVM has found no connection between the molecular structure descriptors and the biological activity class, just as is proper, since there is no connection in this scrambled data set.

Overall, the results from the perturbation tests and scramble calculations show that the classification rates for the real experiments were very unlikely to have been influenced by chance effects.

Summary and Conclusions

Two classification schemes were presented based on molecular descriptors that encode structural information for a diverse set of 383 industrial, household, cosmetic, and pharmaceutical compounds. The first, a six-descriptor k-nearest neighbor classification model using only topological descriptors with a training set of 245 nonclastogenic compounds and 101 clastogenic compounds was evaluated using a prediction set of 26 clastogenic compounds and 11 nonclastogenic compounds. The overall classification rate for the 346 member training set was 81.2%. The classification rate for the 245 member nonclastogenic training set was 86.1%, and the classification rate for the 101 member clastogenic training set was 69.3%. The overall classification rate of the 37 member prediction set was 86.5%. The classification rate for the 26 member nonclastogenic prediction set was 92.3%, and the classification rate for the 11 member clastogenic prediction set was 72.7%. This k-NN model, based only on topological descriptors, could be used to quickly screen a large and structurally diverse data set for chromosomal aberrations.

The more sophisticated support vector machine (SVM) model using an L1-Norm kernel function produced a classification model using topological, geometric, electronic, and hybrid descriptors. The three descriptors of the SVM model included the topological descriptor N6CH-6, which gives information on branching, the topological descriptor EAVE-2, which encodes information about valence, and the hybrid geometric-electronic descriptor ELEC-0, which encodes the ground-state electronic energy. The overall classification rate for the 308 compound training set is 99.7%. The classification rate of the 218 member nonclastogenic training set is 100%, and the classification rate of the 90 member clastogenic training set is 98.9%. The overall classification rate of the 38 member cross-validation set was 92.1%. The classification rate of the 27 member nonclastogenic cross-validation set was 100%, and the classification rate of the 11 member clastogenic cross-validation set was 72.7%. The overall classification rate of the 37 member prediction set was

83.8%. The classification rate of the 26 member nonclastogenic prediction set was 88.5%, and the classification rate of the 11 member clastogenic prediction set was 72.7%. This SVM model requires more computational effort because accurate molecular geometries are required for the descriptor ELEC-0.

Scrambling experiments were performed to show that these models were not built by chance correlation. Overall, these classification models adequately describe the acute structural chromosome aberrations for Chinese hamster lung cells as represented by the compounds from ref 2 used in this study.

The descriptors chosen by both the k-nearest neighbor and the support vector machine models suggest that molecular size and degree of branching combined with the electronic accessibility may play a role in the structural aberration of Chinese hamster lung cells for these compounds. These models were formed with no a priori knowledge of mechanism of action. These models could be used as filters in high throughput screening applications.

Acknowledgment. The authors are grateful to Dr. David Stanton from Procter and Gamble for providing molecular structures and toxicity values and Dr. Marilyn Aadema for adding valuable input and proof reading the paper. This project was supported by Procter and Gamble.

References

- (1) Trimbrell, J. (2000) *Principle of biochemical toxicology*, Vol. 3, Taylor and Francis, London.
- (2) Ojima, T., Hayashi, S., and Matsuoka, A., Eds. (1998) *Compilation of Chromosomal Mutation Test Data*, Life Science Information Center, Japan.
- (3) Kier, L. B. (1985) A shape index from molecular graphs. *Quant. Struct.-Act. Relat.* **4**, 109–116.
- (4) Randic, M., Brisse, G. M., Spencer, R. B., and Wilkins, C. L. (1979) Search for all self-avoiding paths for molecular graphs. *Comput. Chem.* **3**, 5–13.
- (5) Pearlman, R. S. (1980) Molecular surface areas and volumes and their use in structure/activity relationships. In *Physical chemical properties of drugs* (Valvani, S. C., Ed.) Vol. 10, p 361, Marcel Dekker, New York.
- (6) Stouch, T. R., and Jurs, P. C. (1986) A simple method for the representation, quantification, and comparison of the volumes and shapes of chemical compounds. *J. Chem. Inf. Comput. Sci.* **26**, 4–12.
- (7) Sutter, J. M., Dixon, S. L., and Jurs, P. C. (1995) Automated descriptor selection for quantitative structure–activity relationships using generalized simulated annealing. *J. Chem. Inf. Comput. Sci.* **35**, 77–84.
- (8) Stanton, D. T., and Jurs, P. C. (1990) Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure–property relationship studies. *Anal. Chem.* **62**, 2323–2329.
- (9) Stewart, J. P. P. (1990) MOPAC: A semiempirical molecular orbital program. *J. Comput.-Aid. Mol. Des.* **4**, 1–105.
- (10) Stewart, J. P. P. Quantum Chemistry Program Exchange, Indiana University, Version 6.
- (11) Dewar, M. J. S., Zoebisch, E. G., Healy, E. F., and Stewart, J. P. P. (1985) AM1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **107**, 3902–3909.
- (12) Aleman, C., Luque, F. J., and Orozco, M. (1993) Suitability of the PM3-derived molecular electrostatic potentials. *J. Comput. Chem.* **14**, 799–808.
- (13) Jurs, P. C., Chou, J. T., and Yuan, M., Eds. (1979) *Computer Assisted Drug Design*, American Chemical Society, Washington, DC.
- (14) Stuper, A. J., Brugger, W. E., and Jurs, P. C. (1979) *Computer-assisted studies of chemical structure and biological function*, John Wiley & Sons, New York.
- (15) Bakkan, G. A., and Jurs, P. C. (2000) Classification of multidrug-resistance reversal agents using structure-based descriptors and linear discriminant analysis. *J. Med. Chem.* **43**, 4534–4541.

- (16) Mattioni, B. E., and Jurs, P. C. (2002) Development of quantitative structure–activity relationship and classification models for a set of carbonic anhydrase inhibitors. *J. Chem. Inf. Comput. Sci.* **42**, 94–102.
- (17) Serra, J. R., Jurs, P. C., and Kaiser, K. L. E. (2001) Linear regression and computational neural network prediction of Tetrahymena acute toxicity for aromatic compounds from molecular structure. *Chem. Res. Toxicol.* **14**, 1535–1545.
- (18) Kier, L. B., Hall, L. H., and Murray, W. J. (1975) Molecular connectivity I: Relationship to nonspecific local anesthesia. *J. Pharm. Sci.* **64**, 1971–1974.
- (19) Muller, W. R., Szymanski, K., and Knop, J. V. (1987) An algorithm for construction of the molecular distance matrix. *J. Comput. Chem.* **8**, 170–173.
- (20) Todeschini, R., and Consonni, V. (2000) *Handbook of molecular descriptors*, John Wiley & Sons, New York.
- (21) Rohrbaugh, R. H., and Jurs, P. C. (1987) Molecular shape and the prediction of high-performance liquid chromatographic retention indexes of polycyclic aromatic hydrocarbons. *Anal. Chem.* **59**, 1048–1054.
- (22) Dixon, S. L., and Jurs, P. C. (1992) Atomic charge calculations for quantitative structure–property relationships. *J. Comput. Chem.* **13**, 492–504.
- (23) Kelder, J., Grootenhuis, P. D. J., Bayada, D. M., Delbressine, L. P. C., and Ploemen, J. P. (1999) Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs. *Pharm. Res.* **16**, 1514–1519.
- (24) Pimentel, G. C., and McClellan, A. L. (1960) *The Hydrogen Bond*, Reinhold Pub. Corp., New York.
- (25) Kimura, T., Hasegawa, K., and Funatsu, K. (1998) Strategy for variable selection in QSAR studies: GA-Based region selection for CoMFA modeling. *J. Chem. Inf. Comput. Sci.* **38**, 276–282.
- (26) Luke, B. T. (1994) Evolutionary programming applied to the development of quantitative structure–activity relationships and quantitative structure–property relationships. *J. Chem. Inf. Comput. Sci.* **34**, 1279–1287.
- (27) Huberty, C. J. (1994) *Applied Discriminant Analysis*, John Wiley & Sons, New York.
- (28) Masters, T. (1995) *Advanced algorithms for neural networks: A C++ sourcebook*, Vol. 431, John Wiley & Sons, New York.
- (29) Mosier, P. D., and Jurs, P. C. (2003) QSAR/QSPR studies using probabilistic neural networks and generalized regression neural networks. *J. Chem. Inf. Comput. Sci.* (in press).
- (30) Cristianini, N., and Shawe-Taylor, J. (2000) *Support Vector Machines and other kernel-based learning methods*, p 189, Cambridge University Press, Cambridge, U.K.
- (31) Kier, L. B., and Hall, L. H. (1990) An electropological-state index for atoms in molecules. *Pharm. Res.* **7**, 801–807.
- (32) Balaban, A. T. (1982) A highly discriminating distance-based topological index. *Chem. Phys. Lett.* **89**, 399–404.
- (33) Randic, M. (1975) On characterization of molecular branching. *J. Am. Chem. Soc.* **97**, 6609–6615.
- (34) Liu, S., Cao, C., and Zhiliang, L. (1998) Approach to estimation and prediction for normal boiling point (NBP) of alkanes based on a novel molecular distance-edge (MDE) vector, λ . *J. Chem. Inf. Comput. Sci.* **38**, 387–394.
- (35) Kier, L. B., and Hall, L. H. (1976) Molecular connectivity VII: Specific treatment to heteroatoms. *J. Pharm. Sci.* **65**, 1806–1809.
- (36) Kier, L. B. (1986) Shape indexes of orders one and three from molecular structure. *Quant. Struct.-Act. Relat.* **5**, 1–7.
- (37) Kier, L. B., and Hall, L. H. (1997) The E-state as an extended free valance. *J. Chem. Inf. Comput. Sci.* **37**, 548–552.

TX020077W