Classification of Diverse Organic Compounds That Induce Chromosomal Aberrations in Chinese Hamster Cells

Nathan R. McElroy,[†] E. D. Thompson,[‡] and Peter C. Jurs*[,†]

152 Davey Laboratory, Department of Chemistry, The Pennsylvania State University, University Park,
Pennsylvania 16802, and The Procter & Gamble Company, Miami Valley Laboratories, P.O. Box 538707,
Cincinnati, Ohio 45253-8707

Received May 27, 2003

A data set of 297 diverse organic compounds that cause varying degrees of chromosomal aberrations in Chinese hamster lung cells is examined. Responses of an assay are categorized as clastogenic (>10% aberrant cells) and nonclastogenic (<5% aberrant cells). Each of the compounds is represented by calculated structural descriptors that encode topological, geometric, electronic, and polar surface features. A genetic algorithm (GA) employing a *k*-nearest neighbor (*k*NN) fitness evaluator is used to iteratively search a reduced descriptor space to find small, information-rich subsets of descriptors that maximize the classification rates for clastogenic and nonclastogenic responses. To further improve modeling, a similarity measure using atom-pair descriptors is employed to create more homogeneous data subsets. Three different data sets are examined. Results for a set of 297 compounds using the GA-*k*NN method were 86.5% and 80.0% correct classification in the training set and prediction set, respectively. Results for a subset of 279 compounds in model 2 are 85.7% and 85.7% for the training and prediction sets, respectively. Results for a subset of 182 compounds in model 3 are 91.5% and 94.4% for the training and prediction sets, respectively. Creating smaller, more topologically similar data sets result in improved classification rates.

INTRODUCTION

Methods to assess the genotoxic potential of chemicals and drugs of interest to industry are costly and time-consuming. Early in the stage of drug or product development, there can be large numbers of chemicals that need to be screened for genotoxicity, which would be impractical to test using standard gentoxic assays. In addition, regulatory agencies often face the challenge of prioritizing thousands of chemicals for regulatory evaluation, oftentimes when there is very little data. Methods to accurately and quickly predict the genotoxic potential of chemicals and drugs are greatly needed. Based on this, our laboratory has been investigating computational methods to predict which chemicals may induce structural chromosomal aberrations in vitro since this is one of the key endpoints assessed in regulatory testing strategies worldwide. In a previous paper,[1] a model that accurately predicted the outcome of the in vitro chromosome aberration assay in Chinese hamster lung cells using a 24- or 48-h treatment regimen (in the absence of metabolic activation S9) based on data from 901 compounds obtained from *Compilation of Chromosomal Mutation Test Data*[2] was described. In this paper, we describe a predictive classification model that links molecular structure or organic compounds to their negative or positive assay responses in a short 3-h treatment regimen in the presence and absence of metabolic activation S9. Only those compounds that provided a positive assay response without S9 were considered as positives (clastogenic) for this study. Compounds that resulted in a negative assay response (with or without S9) were considered negative (nonclastogenic) for this study.

The data used in this study were from the National Toxicology Program[3] and the *"Data Book of Chromosomal Aberrations* In Vitro*"*.[4] The purpose of this study was to build predictive classification models to link the molecular structure of organic compounds to their negative or positive assay responses. Responses were from a standard assay and deemed positive or negative relevant to the number of chromosomal aberrations that took place after exposure to an organic compound.

Compounds were represented by structural descriptors to encode several molecular properties such as topology, geometry, electronic environment, and polar surface features. Classification models were developed using a genetic algorithm[5] (GA) employing a *k*-nearest neighbor (*k*NN) fitness evaluator on training set compounds, and the predictive ability of the models was examined using external prediction sets. The methodology described in this study has been successfully applied to classifications of other biological data.[6−10]

EXPERIMENTAL SECTION

Data Set. Assay responses were compiled from two sources[3,4] and are the result of exposure of Chinese hamster lung (CHL) or Chinese hamster ovary (CHO) fibroblast cells to organic compounds. Cells and test compounds were incubated for 3 h both with and without S9 rat liver homogenate, at which time cells were removed by centrifugation and incubated for 21 h. Cell division was halted at metaphase by the addition of Colcemid, slides were prepared as described previously,[4] and the number of chromosome

* Corresponding author phone: (814)865-3739; fax: (814)865-3314; e-mail: pcj@psu.edu.
[†] The Pennsylvania State University.
[‡] The Procter & Gamble Company.

10.1021/ci034104f CCC: $25.00
Published on Web 08/27/2003

aberrations (gaps, breaks, and exchanges) was tallied for a total percentage of aberrations. A complete discussion of assay methodology is found in "*Data Book of Chromosomal Aberrations* In Vitro".[4] Assay responses were categorized as clastogenic or nonclastogenic. Clastogenic responses are defined as induction of 10% or more aberrant cells, whereas nonclastogenic responses are those where fewer than 5% aberrant cells were induced. A few compounds that caused between 5 and 10% aberrations (equivocal response) were disregarded in favor of a two-class problem.

This study used a data set of 297 diverse organic compounds having a molecular weight range of 42−959 amu (mean = 218 amu). A list of compounds by their CAS numbers is shown in Table 1. Of the 297 compounds, 68 were clastogenic (23%) and 229 were nonclastogenic (77%). The clastogenic compounds were from the assays using no S9 rat liver homogenate, while nonclastogenic compounds were those causing <5% aberrations with and without S9. Structural makeup of the compounds was quite varied, with 201 compounds containing oxygen, 148 containing nitrogen, 33 containing sulfur, 14 containing phosphorus, 83 containing halogens, and 219 containing rings. Compound types ranged from small industrial solvents to pesticides and small drug molecules. No congeneric series were present. In a few instances where a test compound was in the salt form, the cation was removed and replaced with a hydrogen atom to create a neutral molecule for modeling purposes.

Structure Entry and Optimization. All structures were sketched on a Pentium-III PC using HyperChem (Hypercube, Inc. Waterloo, ON, Canada) to store two-dimensional information such as atom types and connectivity. Two-dimensional structure information for each compound was passed to the semiempirical molecular orbital package MOPAC,[11] where the PM3 Hamiltonian[12] was employed to create low-energy three-dimensional geometries, and the AM1 Hamiltonian[13] was used in a 1SCF calculation to calculate charge information. Previous work has shown the preference for choosing these methods for their respective advantages.[14]

Compounds were initially placed into a training set (TSET) or prediction set (PSET). TSET compounds were used for objective feature selection and to guide the GA-*k*NN classification routine. PSET compounds (approximately 10% of the data sets) were classified by a model after training in order to test the model's generalizability in predicting a classification response for unknown compounds. Compounds were placed pseudorandomly into one of these two subsets with the stipulation that the global ratio of clastogenic to nonclastogenic responses was maintained within the subsets.

Descriptor Generation. A total of 289 descriptors were calculated for each compound using the ADAPT (Automated Descriptor Analysis and Pattern recognition Toolkit) software package. Of those, 185 were topological, 24 were geometric, 10 were electronic, and 70 were polar surface feature descriptors. Topological descriptors encoded geometry-independent information from structure connection tables, including atom and bond counts, substructure counts, and branching information. Geometric descriptors gave information concerning the three-dimensional aspects of a molecule such as surface area, volume, and shape. Electronic descriptors represented features such as highest occupied and lowest unoccupied molecular orbitals and dipole moments. Polar surface feature descriptors combined the previous descriptor types for more site-specific structural details, such as partially charged surface areas[15] and hydrogen bonding characteristics.[16] Specific details on descriptors appearing in models are described below.

Objective Feature Selection. Several of the 289 calculated structural descriptors contained little or no information or were highly correlated with one or more descriptors. To remove such descriptors, objective feature selection was performed using TSET compounds. Identical tests removed descriptors containing redundant or zero information across 80% of the descriptor range. The remaining descriptors were correlated with all other descriptors. If two descriptors had a pairwise correlation value >0.90, one of those two descriptors was randomly removed. The identical test and correlation cutoff values were adjusted, depending on the number of compounds and descriptors, to ensure that the final reduced pool of descriptors contained a number of descriptors no greater than 60% of the number of compounds in the TSET. This guideline has been shown to reduce the possibility of chance correlations in model development.[17] Once a reduced pool of descriptors was created, subjective feature selection using an evolutionary search algorithm began.

***k*-Nearest Neighbor.** *k*NN is a fast and algorithmically simple supervised learning method, which assigns a class to a compound based upon its Euclidean through-space distance to its *k*-nearest neighbors in *n*-dimensional descriptor space. This method was utilized as a fitness evaluation tool in a genetic algorithm to iteratively guide the selection of small descriptor subsets. Starting with three descriptors, the GA-*k*NN used the TSET compounds to determine the classification rate of a selected model. This process was repeated through 1000 iterations, at which point the three-descriptor model that resulted in the highest classification rate was recorded. The above processes were repeated for sequentially increasing number of descriptors, until the addition of a descriptor yielded no marked improvement in the TSET classification rate. This final model was then applied to the PSET compounds in order to assess its predictive ability.

Because of the high proportion of nonclastogenic to clastogenic compounds, a GA-*k*NN containing a weighting factor was also employed. The procedures outlined above were followed, but the weighted GA-*k*NN applied a penalty to models having a high misclassification rate for clastogenic compounds. This approach resulted in models with slightly lower global classification rates but with higher clastogenic classification rates compared to the nonweighted GA-*k*NN.

Similarity Considerations. The selection of the original 297 compounds used in this study was guided by the need to choose compounds that fell within the assay conditions under consideration (e.g., use of S9 or not, time of assay) and were also within the constraints of our approach (e.g., fewer than 46 heavy atoms, neutral compounds, etc.). As a result, the 297 usable compounds were quite diverse in their representation of different chemical features and presented a particularly difficult modeling environment. The first data set made use of all 297 compounds, while subsequent modeling focused upon creating more homogeneous data subsets from the 297 available compounds to improve classification rates. To accomplish this, atom-pair descriptors were calculated for each compound.

Table 1. CAS Numbers and Calculated Class Memberships of the 297 Compounds Used in This Study

no.[a]	CAS registry no.	calc class model 1[b]	calc class model 2[b]	calc class model 3[b]	no.[a]	CAS registry no.	calc class model 1[b]	calc class model 2[b]	calc class model 3[b]
1	100-40-3	−	−	x−	76	1634-78-2	−	−	x−
2	100-41-4	−	−	−	77	1746-01-6	−	p−	−
3	100-42-5	−	−	−	78	1777-84-0	−	−	−
4	10097-16-2	−	−	−	79	1825-21-4	−	−	−
5	101-61-1	−	−	−	80	1836-75-5	p−	−	−
6	101-67-7	−	p+*	x−	81	1912-24-9	−	x−	x−
7	101-96-2	p−	−	x−	82	19666-30-9	+*	−	x−
8	1024-57-3	−	−	−	83	19700-21-1	−	−	x−
9	102-71-6	−	−	−	84	2164-17-2	−	−	x−
10	103-33-3	−	p-*	−	85	2216-51-5	−	−	x−
11	105-55-5	−	−	p−	86	22350-70-5	−	p−	p−
12	105-60-2	+*	−	x−	87	226-36-8	−	−	−
13	105-87-3	−	−	x−	88	230-27-3	−	−	−
14	106-46-7	p−	−	−	89	2425-85-6	+*	−	−
15	107-21-1	−	+*	+*	90	2432-99-7	−	−	−
16	108-30-5	−	−	x−	91	25013-15-4	−	−	−
17	108-42-9	−	−	−	92	25104-18-1	−	−	x−
18	108-67-8	−	−	−	93	25620-78-4	−	−	+*
19	108-78-1	−	−	x+*	94	25956-17-6	−	−	−
20	108-88-3	−	−	−	95	15056-34-5	p−	x−	x−
21	108-90-7	−	−	−	96	271-89-6	−	−	−
22	109-69-3	p−	+*	x+*	97	5859-11-0	−	−	−
23	109-99-9	−	−	p−	98	2832-40-8	−	p−	−
24	110342-24-0	−	−	−	99	298-00-0	p−	−	−
25	110-82-7	−	p−	−	100	299-42-3	−	−	−
26	110-86-1	−	−	−	101	30516-87-1	−	−	x−
27	111-42-2	−	−	−	102	319-84-6	−	−	−
28	111-76-2	−	−	x+*	103	33229-34-4	−	−	x−
29	115-29-7	−	−	−	104	333-41-5	−	−	x−
30	115-32-2	−	−	−	105	34522-32-2	−	−	p−
31	1156-19-0	p−	−	−	106	13613-55-3	−	−	−
32	115-77-5	−	x−	x−	107	3648-21-3	−	−	−
33	115-96-8	−	−	x−	108	389-08-2	−	−	x−
34	116-06-3	−	x−	x−	109	396-01-0	+*	−	x+*
35	1163-19-5	−	x−	x−	110	420-04-2	−	x−	x−
36	117-81-7	−	−	−	111	422-05-9	−	x−	x+*
37	118-52-5	+*	−	x−	112	434-07-1	−	p−	−
38	118-74-1	−	−	−	113	434-13-9	−	−	−
39	119-53-9	p−	p−	−	114	464-49-3	−	−	x−
40	119-84-6	−	−	−	115	50-29-3	−	−	−
41	120-32-1	−	−	−	116	50-44-2	p−	−	x+*
42	120-40-1	−	−	p−	117	50-55-5	−	−	x−
43	120-83-2	−	p+*	−	118	50-81-7	−	−	x−
44	121-14-2	−	−	−	119	510-15-6	−	−	p−
45	1212-29-9	−	−	x−	120	513-37-1	p−	−	x+*
46	122-34-9	−	+*	x−	121	15958-19-7	−	−	−
47	122-39-4	−	−	−	122	518-47-8	−	−	−
48	123-91-1	−	−	−	123	520-45-6	−	−	x−
49	124-48-1	−	−	p−	124	535-87-5	−	p−	−
50	125-33-7	−	−	x−	125	536-33-4	−	−	x−
51	126-73-8	−	−	x−	126	536-90-3	+*	+*	+*
52	72214-01-8	−	−	x−	127	5392-40-5	−	−	x−
53	127-18-4	p+*	−	−	128	540-73-8	+*	x+*	x−
54	127-69-5	−	−	x−	129	542-18-7	−	−	−
55	128-37-0	−	−	x−	130	542-78-9	−	−	x+*
56	128-66-5	−	p−	+*	131	54724-00-4	−	−	−
57	129-15-7	−	−	−	132	56-23-5	−	−	−
58	131-11-3	−	−	−	133	56-38-2	−	−	−
59	132-32-1	+*	−	−	134	57-41-0	−	−	−
60	132-64-9	−	−	−	135	57-66-9	−	p−	x−
61	1330-78-5	p−	−	x−	136	57-74-9	p−	−	−
62	13366-73-9	−	−	−	137	58-25-3	−	−	−
63	13537-32-1	−	x−	x+*	138	58-55-9	−	−	x−
64	135-88-6	−	−	−	139	58-89- 9	−	−	−
65	136-77-6	−	−	x−	140	58-93-5	−	−	x−
66	95-86-3	+*	+*	−	141	59-42-7	−	p−	p−
67	139-13-9	−	−	x−	142	59820-43-8	p−	−	x−
68	140-11-4	p−	−	p−	143	598-55-0	+*	−	x+*
69	142-46-1	−	x−	x−	144	5989-27-5	−	−	−
70	147-47-7	−	p−	−	145	599-79-1	−	−	−
71	147-84-2	−	−	x+*	146	604-75-1	−	−	−
72	1510-16-3	−	−	−	147	60-57-1	−	−	−
73	156-59-2	−	−	−	148	60-87-7	−	−	−
74	1582-09-8	−	−	x−	149	6153-64-6	−	−	x−
75	1596-84-5	−	−	x−	150	630-20-6	−	−	−

Table 1. (Continued)

no.[a]	CAS registry no.	calc class model 1[b]	calc class model 2[b]	calc class model 3[b]	no.[a]	CAS registry no.	calc class model 1[b]	calc class model 2[b]	calc class model 3[b]
151	6358-85-6	−	p−	x−	225	96-45-7	p−	−	−
152	637-07-0	+*	−	x−	226	96-69-5	−	−	x−
153	89-65-6	−	−	−	227	968-81-0	−	−	−
154	6471-49-4	p−	−	−	228	98-95-3	−	−	−
155	64-77-7	−	−	−	229	99-56-9	+*	−	−
156	671-16-9	−	−	x−	230	100-01-6	+	−*	p+
157	67-64-1	+*	−	x+*	231	100-22-1	+	+	+
158	67-72-1	p−	−	−	232	101-70-2	+	−*	+
159	69-65-8	−	−	−	233	101-77-9	p+	p-*	+
160	70699-77-3	−	−	x−	234	101-80-4	+	−*	+
161	71-43-2	−	−	p−	235	101-90-6	−*	−*	x-*
162	71-55-6	−	−	−	236	103-90-2	−*	−*	+
163	7177-48-2	−	p−	−	237	104-94-9	+	+	+
164	7235-40-7	−	−	x−	238	106-87-6	+	−*	x-*
165	72437-42-4	−	−	−	239	106-88-7	+	+	+
166	72-54-8	−	−	−	240	106-89-8	p+	+	+
167	72-55-9	−	−	−	241	106-92-3	+	+	x-*
168	72-56-0	−	−	−	242	107-07-3	+	p-*	p+
169	72-92-9	−	+*	x+*	243	107-13-1	+	x-*	x+
170	7320-37-8	−	−	−	244	108-91-8	−*	−*	−*
171	73-22-3	−	−	x−	245	110-00-9	+	−*	x-*
172	74-31-7	−	−	−	246	111-30-8	+	+	x-*
173	7450-62-6	−	−	−	247	1116-54-7	−*	p-*	−*
174	74-96-4	−	x−	x+*	248	113-45-1	p+	+	x-*
175	75-09-2	−	x−	x+*	249	121-69-7	+	+	+
176	7512-17-6	p−	−	x−	250	121-88-0	+	+	+
177	75-25-2	−	−	−	251	26148-68-5	−*	−*	+
178	75-27-4	−	−	−	252	127-00-4	+	+	+
179	75-34-3	−	p−	p−	253	131-17-9	−*	−*	−*
180	75-52-5	−	x−	x+*	254	154-23-4	−*	−*	x-*
181	75-65-0	−	x−	x+*	255	156-43-4	+	+	+
182	756-79-6	+*	−	+*	256	15972-60-8	+	−*	x-*
183	76-01-7	−	−	−	257	1897-45-6	+	−*	+
184	76-44-8	p−	−	−	258	2052-01-9	+	+	+
185	76-57-3	+*	p−	x−	259	2244-16-8	+	−*	−*
186	77-73-6	−	+*	x−	260	2698-41-1	+	p-*	x-*
187	78-11-5	−	x−	x−	261	302-17-0	−*	−*	x-*
188	78-42-2	−	−	x−	262	42397-64-8	+	+	+
189	78-59-1	+*	−	x−	263	50-37-3	p+	+	x-*
190	78-79-5	−	−	x−	264	51-79-6	−*	−*	x+
191	79-01-6	p−	−	+*	265	518-82-1	+	−*	p+
192	79-11-8	+*	+*	x+*	266	53-21-4	+	+	x-*
193	79-34-5	−	−	p−	267	53-96-3	−*	−*	−*
194	80-05-7	−	−	−	268	542-56-3	+	+	x+
195	80-07-9	−	−	−	269	556-52-5	+	−*	+
196	81-15-2	−	−	x−	270	57-06-7	+	x-*	x-*
197	834-28-6	−	−	x+*	271	5711-40-0	−*	+	x-*
198	83-79-4	−	+*	x−	272	57-14-7	−*	x+	x-*
199	842-07-9	−	−	−	273	58-90-2	p-*	+	−*
200	84-66-2	−	p−	−	274	598-72-1	−*	+	+
201	84-74-2	−	−	−	275	609-20-1	+	+	−*
202	85-01-8	p−	−	−	276	62450-07-1	+	p-*	+
203	85-44-9	−	−	−	277	62-53-3	−*	−*	+
204	85-68-7	−	−	−	278	62-73-7	+	p+	x-*
205	86-30-6	−	−	p−	279	67-20-9	−*	−*	x-*
206	86-50-0	−	−	x−	280	67977-01-9	−*	−*	p-*
207	87-08-1	−	−	−	281	75321-20-9	p+	+	+
208	87-29-6	−	−	−	282	75-56-9	+	+	x+*
209	87-61-6	−	p−	−	283	77-47-4	−*	+	+
210	88-06-2	p−	+*	−	284	78-87-5	+	+	+
211	88-72-2	−	−	−	285	79-06-1	+	−*	x+*
212	88-96-0	−	−	−	286	8003-22-3	−*	+	−*
213	89-25-8	−	−	−	287	80-62-6	+	−*	x+*
214	9012-76-4	−	−	−	288	828-00-2	+	p-*	x-*
215	90-30-2	−	−	−	289	868-85-9	−*	−*	+
216	90-43-7	−	−	−	290	87-68-3	+	+	+
217	91-20-3	−	−	−	291	87-86-5	−*	+	+
218	93-15-2	−	−	x−	292	924-42-5	+	−*	x-*
219	94-20-2	−	−	p−	293	77-10-1	−*	+	x-*
220	94-74-6	−	−	−	294	96-12-8	−*	−*	x+*
221	94-75-7	−	p−	−	295	98-01-1	p+	−*	x-*
222	95-50-1	−	−	−	296	99-57-0	+	+	+
223	95-79-4	−	−	−	297	99-98-9	+	+	+
224	961-11-5	+*	−	x−					

[a] Compounds 1−229 are nonclastogenic; compounds 230−297 are clastogenic. [b] Calculated class memberships are denoted by a "+" for clastogenic and "−" for nonclastogenic. Any result containing a "p" denoted a prediction set compound. Any results followed by a "*" denotes a misclassification. Model 1 results are from a 7-descriptor weighted GA-*k*NN model using 297 compounds; model 2 results are from a 6-descriptor nonweighted GA-*k*NN model using 279 compounds; model 3 results are from an 8-descriptor nonweighted GA-*k*NN model using 182 compounds. For models 2 and 3, any results containing a "x" denote an exclusion set compound.

Table 2. Seven Descriptors for Model 1 of 297 Compounds Using a Weighted GA-*k*NN

	range		mean (SD)	
descriptor[a]	clastogenic	nonclastogenic	clastogenic	nonclastogenic
NAB	0−24	0−29	4.57 (5.66)	5.54 (6.00)
KAPA-3	0.00−6.00	0.0−16.00	2.84 (1.31)	3.80 (2.98)
MDEC-24	0.00−23.10	0.00−73.50	1.23 (3.40)	2.47 (7.19)
ENEG	3.13−6.01	3.19−6.68	4.74 (0.72)	4.88 (0.72)
SHDW-3	14.5−58.1	10.8−139	28.0 (10.1)	36.2 (16.7)
SAAA-3	0.00−0.51	0.00−0.79	1.24 (3.40)	0.17 (0.16)
NITR-4	0.00−0.30	0.00−0.50	0.04 (0.06)	0.02 (0.06)

[a] Explanation: NAB, the count of aromatic bonds; KAPA-3, the κ-index of three-bond fragments;[20] MDEC-24, molecular distance-edge between secondary and quaternary carbons;[21] ENEG, electronegativity as 0.5*(HOMO+LUMO); SHDW-3, molecular shadow area projected on YZ plane;[22] SAAA-3, relative acceptor atom surface area ($\Sigma SA_{acc}/SA_{tot}$);[16] NITR-4, relative surface area of nitrogen ($\Sigma SA_{nitrogen}/SA_{tot}$).[10]

Table 3. Confusion Matrix for Model 1 of 297 Compounds Using a Seven-Descriptor Weighted GA-*k*NN

	TSET = 86.5%		PSET = 80.0%	
actual class	clast[a]	nonclast[b]	clast[a]	nonclast[b]
clast	41	20	4	3
nonclast	16	190	3	20

[a] Clastogenic classification: TSET = 67.2%, PSET = 57.1%. [b] Nonclastogenic classification: TSET = 92.2%, PSET = 87.0%.

Carhart et al.[18] define atom-pairs as topological descriptors for use in structure−activity relationships. Each atom in a molecule is paired with every other atom, and the atom-pair descriptor is denoted by each atom type and the shortest bond distance between them. Once atom-pairs were calculated, the information was used to create a similarity index[18,19] value between any two compounds. For this application, the similarity index was defined as

$$SI = 2\Sigma AP_{A+B}/\Sigma AP_A + \Sigma AP_B \qquad (1)$$

where ΣAP_{A+B} is the atom-pairs shared by compounds A and B, and ΣAPA and ΣAPB are all atom-pairs in compounds A and B, respectively. In comparing any two compounds, the SI value ranges from $0 < SI < 1$, where $SI = 0$ represents two compounds sharing no atom−pairs, and $SI = 1$ represents a compound compared with itself or its stereoisomer. An $n \times n$ matrix of SI values was created ($n = 297$), from which all SI values for each compound could be assessed. If a compound had no SI values greater or equal to the chosen cutoff value (excluding itself), it was removed to an exclusion set. By choosing a cutoff limit of SI values, the overall similarity of a data subset could be manipulated. It will be shown later that by using this approach in subset generation, more homogeneous data sets were produced that resulted in increased classification ability. The models discussed below were trained and validated using SI cutoff values of 0.00 (set of 297 compounds), 0.25 (set of 279 compounds), and 0.50 (set of 182 compounds) and are labeled model 1, model 2, and model 3, respectively.

RESULTS AND DISCUSSION

Model 1. The first model contained all 297 compounds, with 267 compounds in the TSET and 30 compounds in the PSET. After objective feature selection, a reduced pool of 82 descriptors was searched by subjective feature selection. Models ranging from three to ten descriptors were trained and validated using both the weighted and nonweighted GA-*k*NN. The best model contained seven descriptors (Table 2) using a 2:1 weighted GA-*k*NN and produced classification rates of 86.5% and 80.0% for the TSET and PSET, respectively (Table 3). Clastogenic compound classification rates were 67.2% in the TSET and 57.1% in the PSET, while nonclastogenic compound classification rates were 92.2% and 87.0% in the TSET and PSET, respectively. The discrepancy between clastogenic and nonclastogenic classification rates for this and all subsequent models was heavily influenced by the large proportion of nonclastogenic compounds in the data sets and the effect that proportion had on the *k*NN algorithm, both with and without a weighting scheme.

NAB is a count of aromatic bonds in each compound. Nonclastogenic compounds had a slightly larger range of values and on average had a greater number of aromatic bonds than clastogenic compounds. KAPA-3 is the κ_3 index,[20] which represents the degree of branching for three-bond fragments. In general, lower values represent short linear fragments, and higher values denote longer linear fragments or more compact, highly branched fragments. Clastogenic compounds had a range of KAPA-3 values from 0.0 to 6.0, while nonclastogenic compounds had a range of 0.0−16.0 and a higher average value. On average, both NAB and KAPA-3 values suggest that nonclastogenic compounds are larger and contain more aromatic character. This agrees with the average molecular weights of each class: clastogenic mean amu = 172 and nonclastogenic mean amu = 231. MDEC-24 is a molecular distance-edge descriptor[21] that relates interactions between secondary and quaternary carbons. The average MDEC-24 value for nonclastogenic compounds was twice the average clastogenic value. ENEG is the electronegativity value calculated as 0.5*(HOMO+LUMO). The ranges and averages for ENEG class values do not differ as dramatically, though nonclastogenic compounds are slightly higher on average. SHDW-3 is the molecular shadow area[22] projected on to a Cartesian YZ plane, which describes the size of each molecule in that orientation. SHDW-3 values are almost 30% larger on average for nonclastogenic compounds. SAAA-3 is a hydrogen-bonding descriptor defining the relative surface area of acceptor atoms[16] on each molecule. This is calculated as all acceptor atom surface area over total molecular surface area ($\Sigma SA_{acc}/SA_{total}$). Clastogenic compounds have much higher values on average than nonclastogenic compounds, suggesting that clastogenic compounds have more acceptor atom surface area per total surface area than nonclastogenic compounds. NITR-4 is a nitrogen-specific charged partial surface area (CPSA)[10,15] descriptor representing the relative

Table 4. Six Descriptors for Model 2 of 279 Compounds Using a Nonweighted GA-*k*NN

	range		average (SD)	
descriptor[a]	clastogenic	nonclastogenic	clastogenic	nonclastogenic
MDEC-11	0.00−1.84	0.0−16.9	0.13 (0.32)	0.426 (1.51)
WTPT-5	0.00−11.7	0.00−19.8	2.45 (2.74)	3.09 (4.04)
N6CH	0−17	0−47	1.8 (2.8)	2.5 (5.1)
ENEG	3.13−6.01	3.19−6.68	4.74 (0.716)	4.88 (0.716)
WPSA-2	7.63−7.79 × 10^2	0.00−4.14 × 10^3	154 (155)	428 (624)
CHAA-2	−0.009−0.000	−0.014−0.002	−0.002 (0.001)	−0.002 (0.002)

[a] Explanation: MDEC-11, molecular distance-edge descriptor between primary carbons;[21] WTPT-5, weighted path value from nitrogens;[23] N6CH, Chi index count of sixth-order chains;[24] ENEG, electronegativity as 0.5*(HOMO+LUMO); WPSA-2, charge-weighted partial positive surface area $(\Sigma SA^+/\Sigma Q_{tot})*(SA_{tot})/1000$;[15] CHAA-2, relative acceptor atom charges $(\Sigma Q_{acc}/SA_{tot})$.[16]

Table 5. Confusion Matrix for Model 2 of 279 Compounds Using a Six-Descriptor Nonweighted GA-*k*NN

	TSET = 85.7%		PSET = 85.7%		XSET = 83.3%	
class	clast[a]	nonclast[b]	clast[a]	nonclast[b]	clast[a]	nonclast[b]
clast	33	25	4	3	1	2
nonclast	11	182	1	20	1	14

[a] Clastogenic classification: TSET = 56.9%, PSET = 57.1%, XSET = 33.3%. [b] Nonclastogenic classification: TSET = 94.3%, PSET = 95.2%, XSET = 93.3%.

surface area of nitrogens on each molecule. NITR-4 is calculated as SAAA-3 but considers only nitrogen atoms, and as above clastogenic compounds have higher values on average.

Model 2. The second model employed 279 compounds with SI values >0.25, with 251 compounds in the TSET and 28 in the PSET. The 18 compounds that did not meet the cutoff value (i.e., were dissimilar) were placed into an exclusion set (XSET) for later prediction. Objective feature selection created a reduced pool of 86 descriptors, and models ranging in size from three to 10 descriptors were trained and validated using both a weighted and nonweighted GA-*k*NN routine. The best model was created by a nonweighted GA-*k*NN and contained six descriptors (Table 4), and classification rates for the TSET and PSET were 85.7% (Table 5). Clastogenic classification rates were 56.8% and 57.1% for TSET and PSET compounds, respectively. Nonclastogenic classification rates were 94.3% and 95.2% for the TSET and PSET compounds, respectively. The 18 compounds in the XSET were then predicted using this model for a classification rate of 83.3%, with 33.3% and 93.3% classification rates for clastogenic and nonclastogenic compounds, respectively. The 18 compounds of the XSET were those not meeting the similarity cutoff value of 0.25; therefore, the high classification rate of nonclastogenic compounds is due in large part to the high proportion of nonclastogenic compounds in the training model.

MDEC-11 is a molecular distance-edge descriptor[21] of the interactions between primary carbons. WTPT-5 is the sum of weighted paths[23] starting from nitrogen atoms, which encodes the degree of branching in nitrogen-containing compounds. N6CH is a Chi index[24] of sixth-order chain counts in a molecule, such as six-membered rings, five-membered rings with one attached heavy atom, etc. For MDEC-11, WTPT-5, and N6CH, nonclastogenic compounds have higher average values than the clastogenic compounds, suggesting that nonclastogenic compounds are larger and more branched than clastogenic compounds. ENEG is the electronegativity, described above. WPSA-2 is a CPSA[15] descriptor of charge-weighted positive surface area, calculated as $(\Sigma SA^+/\Sigma Q_{tot})*(SA_{tot})/1000$. SA^+ is positive surface area, Q_{tot} is total molecular charge, and SA_{tot} is total molecular surface area. Nonclastogenic compounds have a much higher average value (180% greater) than clastogenic compounds. CHAA-3 is a hydrogen bonding descriptor[16] of the relative charge on acceptor atoms $(\Sigma Q_{acc}/SA_{tot})$, where Q_{acc} is the acceptor atom charge, and SA_{tot} is total molecular surface area. Almost no difference is noted between clastogenic and nonclastogenic compounds, but removing this descriptor from consideration results in poorer classification results.

Model 3. The third model contained 182 compounds having SI values >0.50, with 164 and 17 compounds in the TSET and PSET, respectively. A total of 115 compounds not meeting this cutoff were removed to the XSET for later prediction. A reduced descriptor pool of 89 members was searched by a weighted and nonweighted GA-*k*NN routine to find three to 10 descriptor models. The best model using a nonweighted routine contained eight descriptors (Table 6), with classification rates of 91.5% and 94.4% in the TSET and PSET, respectively (Table 7). Clastogenic classification rates for TSET and PSET compounds were 77.8% and 75.0%, respectively. Nonclastogenic classification rates were 95.3% and 100% for TSET and PSET, respectively. The classes of the 115 compounds in the XSET were the predicted using this model, with an overall classification rate of 65.2%. Clastogenic and nonclastogenic classification rates were 25.0% and 62.1%, respectively. The lower rates for XSET compounds for model 3 were expected, as these compounds were removed as being dissimilar to those compounds used in model training. Again, as with model 2 XSET results, the higher proportion of nonclastogenic compound represented by the training model resulted in a better nonclastogenic compound classification rate of the excluded compounds.

SSS-6 is a simple substructure count of ether moieties (R−O−R). EDIF-1 is the difference in maximum and minimum atomic electrotopological-state[25] values. WTPT-2 is the average molecular ID[23] number. 3SP2 is a count of all sp^2-hybridized carbons attached to three other carbon atoms, which relates a degree of branching and size of the molecule. Nonclastogenic compounds have higher average values for EDIF-1, WTPT-2, and 3SP2, revealing the trend for larger, more branched compounds to be inactive as above. SHDW-4 is a normalized projected molecular shadow area[22] on the XY plane. Almost no difference was noted between average values of each class. HOMO is the energy of the highest

Table 6. Eight Descriptors for Model 3 of 182 Compounds Using a Nonweighted GA-*k*NN

	range		average (SD)	
descriptor[a]	clastogenic	nonclastogenic	clastogenic	nonclastogenic
SSS-6	0.0−4.0	0.0−7.0	0.59 (0.89)	0.56 (1.1)
EDIF-1	0.836−16.6	0.00−19.5	9.45 (4.02)	11.0 (4.36)
WTPT-2	1.67−2.17	1.60−2.15	1.90 (0.116)	1.93 (0.108)
3SP2	0.0−7.0	0.00−10.0	0.84 (1.6)	0.94 (1.4)
SHDW-4	0.437−0.693	0.350−0.810	0.537 (0.60)	0.532 (0.068)
HOMO	−11.7 − −7.36	−12.4 − −7.64	−9.62 (1.12)	−9.71 (0.995)
RNCG	0.108−0.556	0.04−1.00	0.257 (0.103)	0.205 (0.135)
NITR-4	−32.3−8.17	−202−33.2	−2.94 (7.61)	−7.64 (25.8)

[a] Explanation: SSS-6, count of R−O−R substructures; EDIF-1, difference in maximum and minimum electrotopological-state values;[25] WTPT-2, average molecular ID (molecular ID/no. atoms);[23] 3SP2, count of sp^2-hybridized carbons bonded to three other heteroatoms. SHDW-4, normalized molecular shadow area projected on the XY plane;[22] HOMO, energy of the highest occupied molecular orbital; RNCG, relative negative molecular charge, $Q^{-}_{\max}/Q^{-}_{\mathrm{tot}}$;[15] NITR-4, relative surface area of nitrogen atoms ($\Sigma SA_{\mathrm{nitrogen}}/SA_{\mathrm{tot}}$).[10]

Table 7. Confusion Matrix for Model 3 of 182 Compounds Using an Eight-Descriptor Nonweighted GA-*k*NN

	TSET = 91.5%		PSET = 94.4%		XSET = 65.2%	
class	clast[a]	nonclast[b]	clast[a]	nonclast[b]	clast[a]	nonclast[b]
clast	28	8	3	1	7	21
nonclast	6	122	0	14	33	54

[a] Clastogenic classification: TSET = 77.8%, PSET = 75.0%, XSET = 25.0%. [b]Nonclastogenic classification: TSET = 95.3%, PSET = 100.0%, XSET = 62.1%.

Table 8. 15-Model Averages for 297-, 279-, and 182-Member Data Sets

set	15 model av class rate (%)	SD	set	15 model av class rate (%)	SD
297 TSET	86.6	1.7	279 PSET	80.3	3.0
297 PSET	76.6	4.5	182 TSET	90.7	2.5
279 TSET	86.7	1.5	182 PSET	81.9	6.3

Table 9. Majority Rule Classification Rates of Three Predictions for Each Data Subset

actual class	clast	nonclast	class rate (%)
	Calculated Class − 297 Compds		
clast	28	40	41.2
nonclast	22	207	90.4
		overall =	79.1
	Calculated Class − 279 Compds		
clast	29	36	44.6
nonclast	10	204	95.3
		overall =	83.5
	Calculated Class − 182 Compds		
clast	18	22	45.0
nonclast	6	136	95.8
		overall =	84.6

occupied molecular orbital from MOPAC[11] output. Nonclastogenic compounds had slightly more negative HOMO energies than clastogenic compounds. RNCG is a CPSA[15] descriptor of the relative negative molecular charge, or $Q^{-}_{\max}/Q^{-}_{\mathrm{tot}}$, where $Q^{-}_{\max}$ is the most negative charge on the molecule and Q^{-}_{tot} is the total negative charge. Clastogenic compounds seem to have a bit more relative negative charge than nonclastogenic compounds. NITR-4 is a nitrogen-specific CPSA descriptor, discussed above, and again clastogenic compounds have higher values for this descriptor than for nonclastogenic compounds.

Set Membership. The methodology described above made use of fixed TSET/PSET pairs for model building and validation. For each of the three TSET/PSET pairs (297, 279, and 182 members), compounds were placed pseudorandomly into one of the two subsets while maintaining the global clastogenic to nonclastogenic response ratio. These subsets remained fixed for the remainder of the model training and validation, as it was important to maintain subset continuity in order to compare models of differing sizes.

An additional experiment was performed to determine the effect of varying TSET/PSET membership on overall classification rates. Using a leave-20%-out approach, five TSET/PSET pairs were created for each of the three data sets. Compounds were placed randomly into a TSET or PSET but in such a manner that each compound in a data set was present four times in a TSET and once in a PSET. This procedure was repeated for a total of three times, so that each compound was in a PSET thrice. This procedure is based loosely on work by Breiman[26] and has been successfully implemented in previous studies from this group.[27,28] Using the modeling approach above, 15 GA-*k*NN models were constructed for each data set (297, 279, and 182 compounds, respectively) and varied in size from five to nine descriptors.

Table 8 shows the TSET and PSET results for a 15-model average for each of the three data sets. The overall results are not exactly comparable with models 1−3 above because of the different compound distributions needed to accommodate the leave-20%-out approach; however, in general classification rates improved with smaller, more similar data sets. TSET classification rates improved from 86.6% to 90.7%, while PSET classification rates improved from 76.6% to 81.9%. Table 9 shows the results of all compounds when they were placed into a PSET, and are the majority rules result of three predictions. Classification rates improved from 79.1% to 84.6% by increasing the similarity of subsets. The clastogenic classification rates are much lower, and the nonclastogenic classification rates are much higher than models 1−3 above, but the individual class rates and the overall classification rates again show improvement using more similar data sets. The smaller representation of clastogenic compounds in the data sets does affect the preference of the kNN models to choose nonclastogenic class membership.

Results from this experiment also showed that regardless of set memberships, certain descriptors or classes of descriptors were chosen for the reduced pool of descriptors using objective feature selection, and these descriptors appeared in many models using the GA-*k*NN approach. Table 10 lists

Table 10. Common Descriptors from the Multiple TSET Experiment

descriptor	% ofmodels	% of descriptors
ENER	51	7.2
ENVR-6	29	4.1
HOMO	27	3.8
WTPT-2	27	3.8
RNCG	24	3.5
NITR-5	22	3.1
SHDW-4	22	3.1
SAAA-2	20	2.8

some of the common descriptors chosen for the multiple TSET experiment. ENEG, the electronegativity, appeared in 51% of all 45 models, representing 7.2% of the 318 descriptors chosen in those 45 models. HOMO appeared in 27% of the models. Overall, these two electronic descriptors or their counterparts (LUMO and HARD) appeared in 71% of the 45 models. Other important descriptors were ENVR-6, a path-1 count of ether moieties; WTPT-2, the average molecular ID number;[23] RNCG, the molecular relative negative charge;[15] NITR-5, the relative atomic-weighted surface area of nitrogens;[10] SHDW-4, a standardized shadow area on the XY plane;[22] and SAAA-2, the average acceptor atom surface area.[16] These descriptors, or descriptors of the same class, also appeared in models 1−3.

Randomization Experiments. Scrambling experiments were performed on each data set above to ensure that classification results were due to a structure−activity relationship and not chance effects. The dependent variables were randomly assigned to a compound, and models of the same size were trained and validated as above. This procedure was repeated 10 times, and the classification rates of PSET compounds were averaged. For simplification, randomization experiment models were labeled as models 1a, 2a, and 3a to correspond with the data sets used for models 1, 2, and 3. The PSET classification rates were 65.3 ± 13.3% for model 1a, 70.4 ± 9.1% for model 2a, and 62.2 ± 12.2% for model 3a. All three PSET classification rates fell within one standard deviation unit of true random, which supported a relationship between structure and activity in models 1−3, rather than chance effects.

CONCLUSIONS

The exact mechanism for each compound to cause an aberration is not known; therefore, the interpretation of the model descriptors is limited. However, it seems clear that there are features consistently chosen by models 1−3 and those in the multiple TSET experiment that relate structure to activity. A class of electronic descriptors appeared in models 1−3 and 71% of the 45 multiple TSET experiment models. CPSA descriptors appeared in models 1−3 and 78% of the multiple TSET experiment models. Hydrogen bonding descriptors appeared in models 1 and 2 and in 58% of the multiple TSET experiment models. In all, it seems that relationships exist between the electronics of a compound (encoded via ENEG and HOMO, e.g.) and that compound's ability to cause an aberration. This relationship is further enhanced by the appearance of hybrid descriptors, which are calculated using partial atomic charges, to relate charge distribution over specific surface areas of each compound (positive or negative surface area descriptors) or the atoms involved in hydrogen bonding (oxygen, nitrogen, etc.). By combining these descriptors with other factors encoding size and shape, successful classification models were created. Descriptor class averages also suggest that clastogenic compounds are smaller or less branched and have a greater proportion of their surface area dedicated to acceptor atoms or partial negative charges.

Similarity measures used to remove dissimilar compounds from data sets resulted in improved classification results while choosing similar descriptors or classes of descriptors in all models, regardless of data set size. By increasing the data set similarity, the overall classification rate as well as the clastogenic and nonclastogenic rates were improved. The skewness of the clastogenic to nonclastogenic compound ratio presented a difficult modeling environment and resulted in a nonclastogenic class preference by the kNN algorithm in both the single TSET and multiple TSET experiments. However, the models presented here can be used for future screening applications of organic compounds to approximate their clastogenicity, provided that those compounds are similar to the ones used for model training. Atom-pair descriptors could also be used to determine the similarity of a set of query compounds toward the compounds used in model development, increasing the confidence that a particular model will calculate proper clastogenic responses for a set of unknown compounds.

ACKNOWLEDGMENT

We wish to thank Procter & Gamble for support of this study and to convey special thanks to Donna Morrall and Marilyn Aardema for their feedback and review of this manuscript.

REFERENCES AND NOTES

(1) Serra, J. R.; Thompson, E. D.; Jurs, P. C. Development of Binary Classification of Structural Chromosome Aberrations for a Diverse Set of Organic Compounds from Molecular Structure. *Chem. Res. Toxicol.* **2003**, *16*, 153−163.

(2) Ojiama, T.; Hayashi, S.; Matsuoka, A. *Compilation of Chromosomal Mutation Test Data*; Life Science Information Center: Japan, 1998.

(3) NIEHS. National Toxicology Program. http://ntp-server.niehs.nih.gov (accessed 2001).

(4) *Data Book of Chromosomal Aberration Test In Vitro*; Sofuni, T., et al., Eds.; Elsevier Science Publishing Co.: New York, 1988.

(5) Luke, B. T. Evolutionary Programming Applied to the Development of Quantitative Structure−Activity Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1279−1287.

(6) Bakken, G. A.; Jurs, P. C. Classification of Multidrug- Resistance Reversal Agents Using Structure-Based Descriptors and Linear Discriminant Analysis. *J. Med. Chem.* **2000**, *43*, 4534−4541.

(7) Mattioni, B. E.; Jurs, P. C. Development of Quantitative Structure−Activity Relationship and Classification Models for a Set of Carbonic Anhydrase Inhibitors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 94−102.

(8) Kauffman, G. W.; Jurs, P. C. QSAR and k-Nearest Neighbor Classification Analysis of Selective Cyclooxygenase-2 Inhibitors Using Topologically-Based Numerical Descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1553−1560.

(9) McElroy, N. R., et al. QSAR and Classification of Murine and Human Soluble Epoxide Hydrolase Inhibition by Urea-like Compounds. *J. Med. Chem.* **2002**, *46*, 1066−1080.

(10) Kauffman, G. W. The Development of Predictive Models for Physical and Biological Properties from Molecular Structure and the Analysis of Data from a Conducting Polymer Chemiresistive Sensor Array. Doctoral Thesis. The Pennsylvania State University, University Park, PA, 2002.

(11) Stewart, J. P. P. *MOPAC 6.0, Quantum Chemistry Program Exchange*; Program 455; Indiana University: Bloomington, IN, 1990.

(12) Stewart, J. J. P. MOPAC: A Semiempirical Molecular Orbital Program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1−105.

(13) Dewar, M. J. S., et al. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(14) Aleman, C.; Luque, F. J.; Orozco, M. Suitability of the PM3-Derived Molecular Electrostatic Potentials. *J. Comput. Chem.* **1993**, *14*, 799−808.

(15) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure−Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323−2329.

(16) *Handbook of Molecular Descriptors*; Todeschini, R.; Consonni, V., Eds.; Wiley-VCH: Weinheim, Germany, 2000.

(17) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative-Structure Property Relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.

(18) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atoms Pairs as Molecular Features in Structure−Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.

(19) Jaeger, E. P. Computer-Aided Investigations of Chemical Structure-Biological Activity Relationships. Ph.D., The Pennsylvania State University, University Park, PA, 1989.

(20) Kier, L. B. Shape Indexes of Orders One to Three from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1986**, *5*, 1−7.

(21) Liu, S.; Cao, C.; Li, Z. 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.* **1998**, *38*, 387−394.

(22) Stouch, T. R.; Jurs, P. C. A Simple Method for the Representation, Quantification, and Comparison of the Volumes and Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 4−12.

(23) Randic, M. On Molecular Identification Numbers. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164−175.

(24) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.

(25) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*; Academic Press: San Diego, CA, 1999.

(26) Brieman, L. Bagging Predictions. *Machine Learning* **1996**, *24*, 123−140.

(27) Mattioni, B. E., et al. Predicting the Genotoxicity of Secondary and Aromatic Amines Using Data Subsetting to Generate a Model Ensemble. *J. Chem. Inf. Comput. Sci.* **2003**, in press.

(28) Mattioni, B. E.; Jurs, P. C. Prediction of Dihydrofolate Reductase Inhibition and Selectivity Using Computational Neural Networks and Linear Discriminant Analysis. *J. Mol. Graphics Model.* **2003**, *21*, 391−419.

CI034104F