Toward the Design of Chemical Libraries for Mass Screening Biased against Mutagenic Compounds

Oriol Llorens[†] and Juan J. Perez

Department of Chemical Engineering (UPC), The Barcelona School of Engineering, Av. Diagonal, 647, 08028 Barcelona, Spain

Hugo O. Villar*

Telik, Inc., Chemoinformatics Group, 750 Gateway Boulevard, South San Francisco, California 94080

Received October 30, 2000

The ability to develop a chemical into a drug depends on multiple factors. Beyond potency and selectivity, ADME/PK and the toxicological profile of the compound play a significant role in its evaluation as a candidate for development. Those factors are being brought into bear earlier in the discovery process and even into the design of libraries for screening. The purpose of our study is the comparative analysis of simple physical characteristics of compounds that have been reported to be mutagens and nonmutagenic ones. The analysis of differences can lead to the development of knowledge-based biases in the libraries designed for massive screening. For each of four Salmonella strains, TA-98, TA-100, TA-1535, and TA-1537, an analysis of the statistical significance of the deviance of the averages for a number of global properties was carried out. The properties studied included parameters, such as topological indices, and bit strings representing the presence or absence of certain chemical moieties. The results suggest that mutagens display a larger number of hydrogen bond acceptor centers for most strains. Moreover, the use of bit strings points to the importance of certain molecular fragments, such a nitro groups, for the outcome of a mutagenicity study. Development of multivariate models based on global molecular properties or bit strings point to a small advantage of the latter for the prediction of mutagenicity. The benefits of the bit strings are in accord with the use of fragment-based approaches for the prediction of carcinogenicity and mutagenicity in methods described in the literature.

Introduction

The use of high-throughput screening has represented a major shift in the lead discovery process. Whenever the paradigm is applicable, the use of robotics and large chemical libraries has resulted in the identification of multiple compounds that could potentially serve as leads for medicinal chemistry programs for many targets. From a chemical perspective, the selection of compounds used for screening has gone through several stages of maturation. Initially, compounds were taken from historical collections or peptide libraries or were built around a single chemical scaffold, all of which produced little structural variety.[1] The lack of variety spurred the development of techniques to assess the diversity of the libraries[2] used, mainly with the purpose of removing the significant structural biases present. Multiple approaches to characterize the diversity of chemical libraries were put forward in the literature.[3] Most of them use structural or physicochemical descriptors to evaluate similarity among the compounds, although other means have also been described.[4] The massive screening of diversified libraries has yielded uneven results that are quite dependent on the families of targets but could result in the identification of multiple hits for an assay.[5,6] In many cases, the hits found had to be subsequently discarded during the optimization stage because of their toxicological profile or poor bioavailability.[6,7] The result has been a perception of misplaced effort that led to the latest stage in library design, with the tailoring of the libraries to suit their intended use as pharmacological agents.[8] If compounds are ultimately to be used as drugs, certain biases are necessary, or at least helpful, to limit the range of properties of the compounds used for building up a library to those relevant to pharmaceuticals. Even within these constraints, maximal variability is still sought.[9]

The definition of the limits to be imposed on libraries has been part of active research.[10−12] For the most part, the limits have been based on the physicochemical properties of known drugs. Significant attention has been paid to setting rules for the prediction of bioavailability or the ability to formulate the compounds into drug products.[14−16] Toxicology is equally important in defining the viability of a chemical as a candidate for drug development. However, it is multifactorial and can only be approached in stages. One aspect of the toxicological package of a compound is its mutagenicity.

A positive result in a mutagenicity test causes, as a rule, the discontinuation of work on that family of compounds. Consequently, the exclusion during the library design process of likely mutagens could help avoid wasteful screening of compounds and is consistent

* To whom correspondence should be addressed. Permanent address: Triad Therapeutics, 5820 Nancy Ridge Rd., Suite 200, San Diego, CA 92121. E-mail: hvillar@triadt.com. Fax: (858) 455-0140.

[†] Work carried out while visiting Telik, Inc.

10.1021/jm0004594 CCC: $20.00
Published on Web 07/20/2001

with the current trend to bring ADME/PK and toxicological considerations early on in the discovery process.[1]

For several decades, the Ames *salmonella* mutagenicity assay has been widely applied in research laboratories and by regulatory agencies as a useful tool for the detection of chemical mutagens found in the environment, natural products, food, pesticides, and drug candidates.[17] The test represents one of the most widely used in vitro short-term screens for mutagenicity. Since the earlier application of the Ames test, different *Salmonella* strains have been generated to improve assay sensitivity and its range of applicability. These modifications have given the test enough versatility for the detection of different types of mutagenic compounds or compounds that become mutagenic upon metabolization by humans.[18]

For the development and maintenance of a reference data set, the use of standard tester strains has been recommended,[19] although neither a gold standard nor a battery of strains are considered as fully satisfactory. Nevertheless, mutagenicity data in strains TA-98, TA-100, TA-1535, and TA-1537 have been proposed in the literature as a starting screening routine.[19]

The classification of compounds as mutagenic or nonmutagenic results from combining the results of screening the different strains. Compounds are considered mutagenic if they give a positive result in any of the strains, with or without metabolic activation. In our analysis, the results for each strain will be independently scrutinized to capture different aspects of the mutagenic potential of a compound. The properties that each of the strains encapsulates are likely to be somewhat different, and some trends could be missed if all data sets were considered simultaneously.

Unfortunately, the molecular mechanisms of mutagenic processes are not well characterized and are still under study. The lack of a molecular understanding complicates the development of structure−property relationships. Nonetheless, a significant amount of literature has been devoted over the years to the prediction of the mutagenicity and carcinogenicty by computational means.[20] While the specifics of each approach vary greatly, methods for their prediction fall within two main camps: pattern recognition techniques or knowledge-based systems. The better known among the pattern recognition programs, CASE[21] and its successor MULTICASE,[22] TOPKAT,[23] and ADAPT,[24] predict toxicity based on the chemical structures. MULTICASE compares the distribution of molecular fragments found in a given molecule to those found in a database of fragments found in toxic substances. On the basis of the statistical weight of the compounds, it predicts toxicity or activity. Knowledge-based systems, such DEREK[25] and HAZARDEXPERT,[26] are based on the ability of the programs to identify molecular fragments or substructures that were previously known to give rise to various forms of toxicity. These programs constitute best efforts to accurately predict the activity of specific compounds and are extremely valuable later in the computer-assisted ligand design and optimization process.

The purpose of this article is not the accurate prediction of the mutagenic potential of individual compounds, but the analysis of the properties and structural features that can be found in mutagenic compounds and less frequently in nonmutagenic ones. The present work should be regarded as a first step toward the definition of molecular properties ranges, which could serve to introduce biases that enrich libraries in compounds with diminished genotoxic potential. Consequently, the types of properties that will be considered are the simple easily computed parameters that have been in use for library design, such as topological descriptors and bit strings. While global molecular properties provide a comprehensive view of a molecule, each bit in a bit string is associated to a certain molecular fragment. We will present the results of applying different statistical tools to investigate the biases present in libraries of compounds found to be active in the Ames test for each of the different *Salmonella* strains, with and without activation. Subsequently, we study the predictive power of each class of properties, as a means to evaluate the strategies that can be used during library design.

Mutagenecity Data

The CCRIS database[28] was mined to select groups of active and inactive compounds in the Ames test for each of the four test strains recommended for use: TA-98, TA-100, TA-1535, and TA-1537. Results were retrieved for both: nonmetabolically activated compounds and compounds that underwent metabolic activation using rat liver S9 mix test. The data assembled was restricted to these two types to improve consistency. The classification of the results into positive (active) or negative (inactive) for each compound is not always straightforward because of ambiguities in the data reported by different laboratories. To minimize the uncertainties, only molecules that had a consistent rating in over 80% of the studies reported in the database were taken into consideration.

CCRIS lacks structural information on the compounds; therefore, structures for the compounds were obtained by merging the data with the structures in the CMC and ACD (MDL Inc., San Leandro, CA) databases, using the CAS number as a linker which resulted in the database used for this study. This step further reduced the number of compounds available for study as not all entries could be successfully assigned a structure.

For each molecule, two classes of molecular descriptors were utilized. On one side, topological indices together with other global molecular properties were computed; while on the other, a set of bit strings that has been described in the literature as particularly useful for diversity and similarity classifications,[29] MDL's MOLSKEYS, was used.

First, a set of descriptors generated with the Molconn-X v 2.0 (Hall Associates, Quincy, MA) was computed for all the compounds in the database. These descriptors basically encode aspects of molecular size, shape, branching, and, to some extent, polarity. The property set monitored included formula weight (FW), number of rings (NRINGS), graph diameter (SDIAM), number of hydrogen bond donors (HBD) and acceptors (HBA), flexibility index (PHIA), Wiener number (W), Wiener *p* number (WP), total Wiener number (WT), Platt *f* number (PF), and Bonchev−Trinajstic indexes

Table 1. Characteristics of Each of the Eight Data Sets Used for the Study[a]

strain	metabolic activation	activity	no. of compds	MW	C	N	O
TA-98	activated	positive	335	212.37	11	2	2
		negative	1306	207.10	10	1	2
	not activated	positive	231	218.92	10	2	3
		negative	1708	209.13	10	1	2
TA-100	activated	positive	322	201.16	10	1	2
		negative	1170	209.70	10	1	2
	not activated	positive	232	199.36	8	2	3
		negative	1557	212.45	11	1	2
TA-1535	activated	positive	93	173.33	8	1	2
		negative	838	206.95	10	1	2
	not activated	positive	61	183.23	7	1	2
		negative	1082	208.59	10	1	2
TA-1537	activated	positive	82	226.19	13	1	2
		negative	794	210.84	10	1	2
	not activated	positive	59	231.59	12	2	3
		negative	908	213.06	10	1	2

[a] MW is the average molecular weight, while C, N, and O show the average count of those atoms in the data set.

(SIDC, SIDW). The octanol−water partition coefficient (LogP) was also computed based on the ACD software (ACD, Ontario Canada).

The second set of descriptors was MDL's MOLSKEYS.[32] Each of the 166 bits encodes the presence or absence of determined substructural patterns. Those MOLSKEYS are automatically created by the ISIS package (MDL, Inc., San Leandro, CA) at the time of the generation of a database. Each substructure bit was treated during the statistical analysis as a separate binary descriptor.

All statistical calculations were carried out using the algorithms embedded in the S-Plus 200 package (MathSci, Bothell, WA). Recursive partitioning and Wilcoxon rank−sum test were carried out as embodied in the S-PLUS "tree" and "wilcox.test" functions, respectively. Graphics were generated using S-PLUS capabilities.

Statistical Analysis of the Properties

Table 1 shows the general characteristics of each of the databases generated for this study and shows that they are comparable in their gross characteristics. The data sets are divided by *Salmonella* strain and whether the compounds have been metabolically activated, generating a total of eight groups to be analyzed independently. The strains TA-98 and TA-100 display a similar frequency of positive compounds, while the number is much smaller for the TA-1535 and TA-1537. In every other respect, the gross molecular characteristics of the compounds that compose each data set appear comparable.

The Wilcoxon rank sum test for two independent variables was carried out for each of the global properties computed to determine if statistically significant differences existed in the properties of positive and negative compounds. The Wilcoxon rank−sum test is widely used to analyze the differences between independent and not-paired data sets, taking into account their magnitude.[30] The test ranks all data assigning a value of 1 for the lowest value increasingly to the largest one, then detects how these rankings are distributed between the two groups in a way that if high or low ranks are found predominantly in one sample, this means the two populations are not identical.

A Z test statistic is calculated to establish whether the two distributions are significantly different, as follows:

$$Z = (R - \mu_R)/\sigma_R$$

where

$$\mu_R = ((n_1 n_2 (n_1 + n_2 + 1))/2)^{1/2}$$

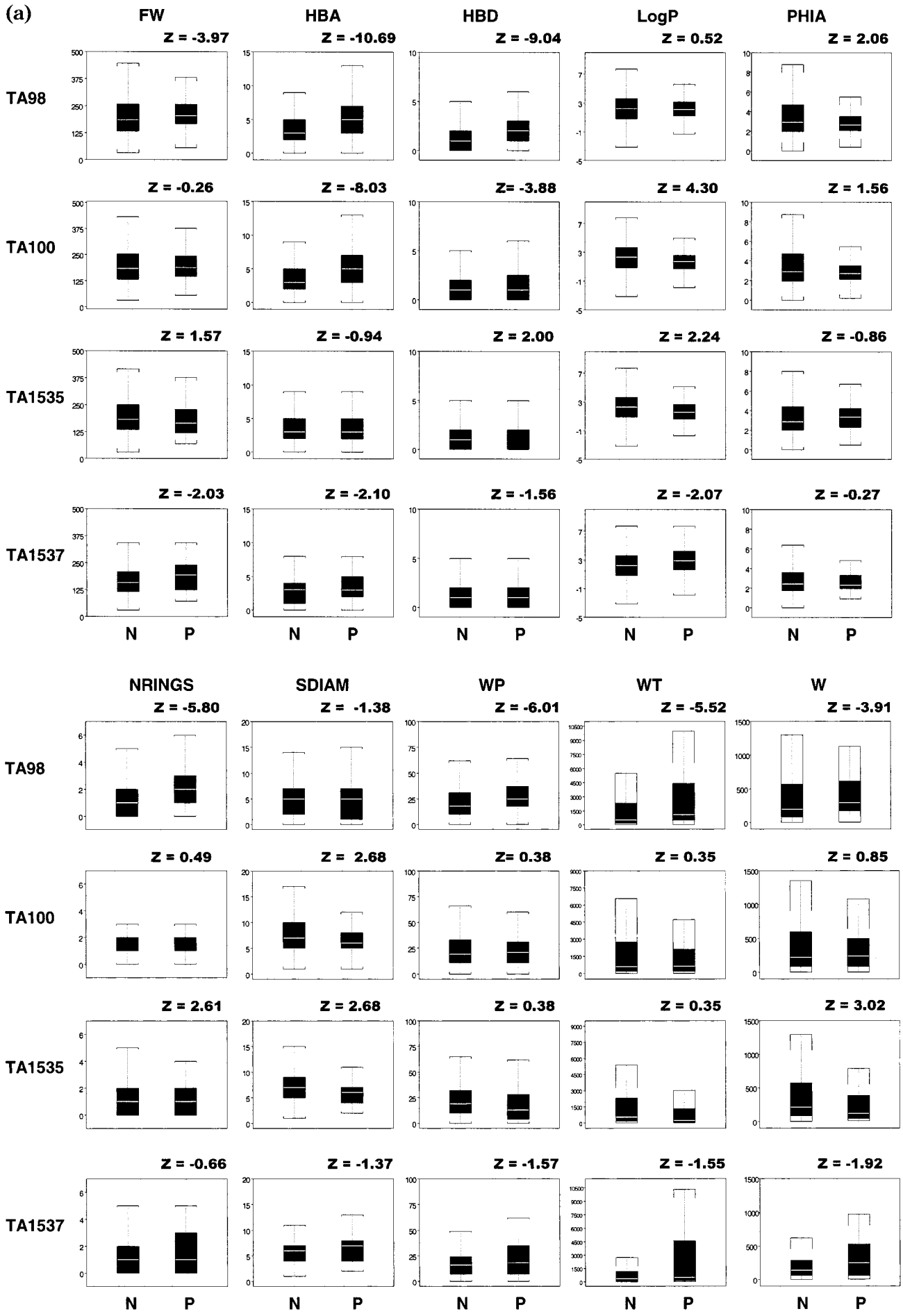
n_1 is the size of the reference sample; n_2 is the size of the sample for comparison; R is sum of ranks of the reference sample; σ_R is the standard deviation of R.

At 95% significance, the critical Z values for a two-tailed test are ± 1.96.[30] Values of Z outside the critical region lead to the rejection of the null hypothesis, and the two distributions are considered different. The Wilcoxon test has several advantages for data analysis over other conventional statistical tests. The most convenient aspects of the test are that the data does not need to follow a normal distribution, and the sizes of the populations do not need to be equal.

(a) Global Molecular Properties. Figure 1 shows a box plot of the global properties used for positive and negative compounds for each of the eight data sets. The box plot shows the median as a stripe, the upper and lower quartiles of the data distribution as the box, and the whiskers show the extent of the data beyond the quartiles. The box plot allows a rapid visualization of the data sets. For example, if the upper and lower quartiles of the box plot are at about the same distance from the median stripe, the data is distributed symmetrically around the stripe in the box. The results of the Wilcoxon analysis for the global properties are also part of Figure 1. The Z test data can be analyzed in three different ways taking into account (i) the statistical significance of the differences observed; (ii) the sign of Z, and (iii) the magnitude of Z. The two data sets are considered to be different if $|Z| > 1.96$. Positive Z values indicate that the negative compounds show higher values for that descriptor. Larger absolute values for Z point to more significant differences between the two distributions.

The data shows that clear, statistically significant differences exist in the properties of positive and negative compounds for each of the strains. It is also interesting that the properties of the compounds found to be positive in each strain are somewhat distinct. The ability to form hydrogen bonds is, in the case of the molecules that did not undergo activation, a clear discriminating factor. Compounds found to be positive with the TA-98, TA-100, and TA-1537 strains have a larger number of hydrogen bond acceptors, while the number of hydrogen bond donor centers is an important discriminant for TA-98 and TA-100.

The hydrophobicity of the compounds as characterized by their calculated LogP, has no bearing on the activity of the compounds for the TA-98 strain. This is a somewhat surprising result, because of the emphasis that has been given to the cLogP in the prediction of mutagenicity.[31] Nevertheless, the hydrophobic character of the molecule has influence for TA-100 and TA-1537. For the TA-100, the positive compounds are less lipophilic, while the opposite is true for the TA-1537 strain.


(a)
FW
HBA
HBD
LogP
PHIA
TA98
Z = -3.97
Z = -10.69
Z = -9.04
Z = 0.52
Z = 2.06
TA100
Z = -0.26
Z = -8.03
Z= -3.88
Z = 4.30
Z = 1.56
TA1535
Z = 1.57
Z = -0.94
Z = 2.00
Z = 2.24
Z = -0.86
TA1537
Z = -2.03
Z = -2.10
Z = -1.56
Z = -2.07
Z = -0.27
N
P
NRINGS
SDIAM
WP
WT
W
TA98
Z = -5.80
Z = -1.38
Z = -6.01
Z = -5.52
Z = -3.91
TA100
Z = 0.49
Z = 2.68
Z= 0.38
Z = 0.35
Z = 0.85
TA1535
Z = 2.61
Z = 2.68
Z = 0.38
Z = 0.35
Z = 3.02
TA1537
Z = -0.66
Z = -1.37
Z = -1.57
Z = -1.55
Z = -1.92
N
P

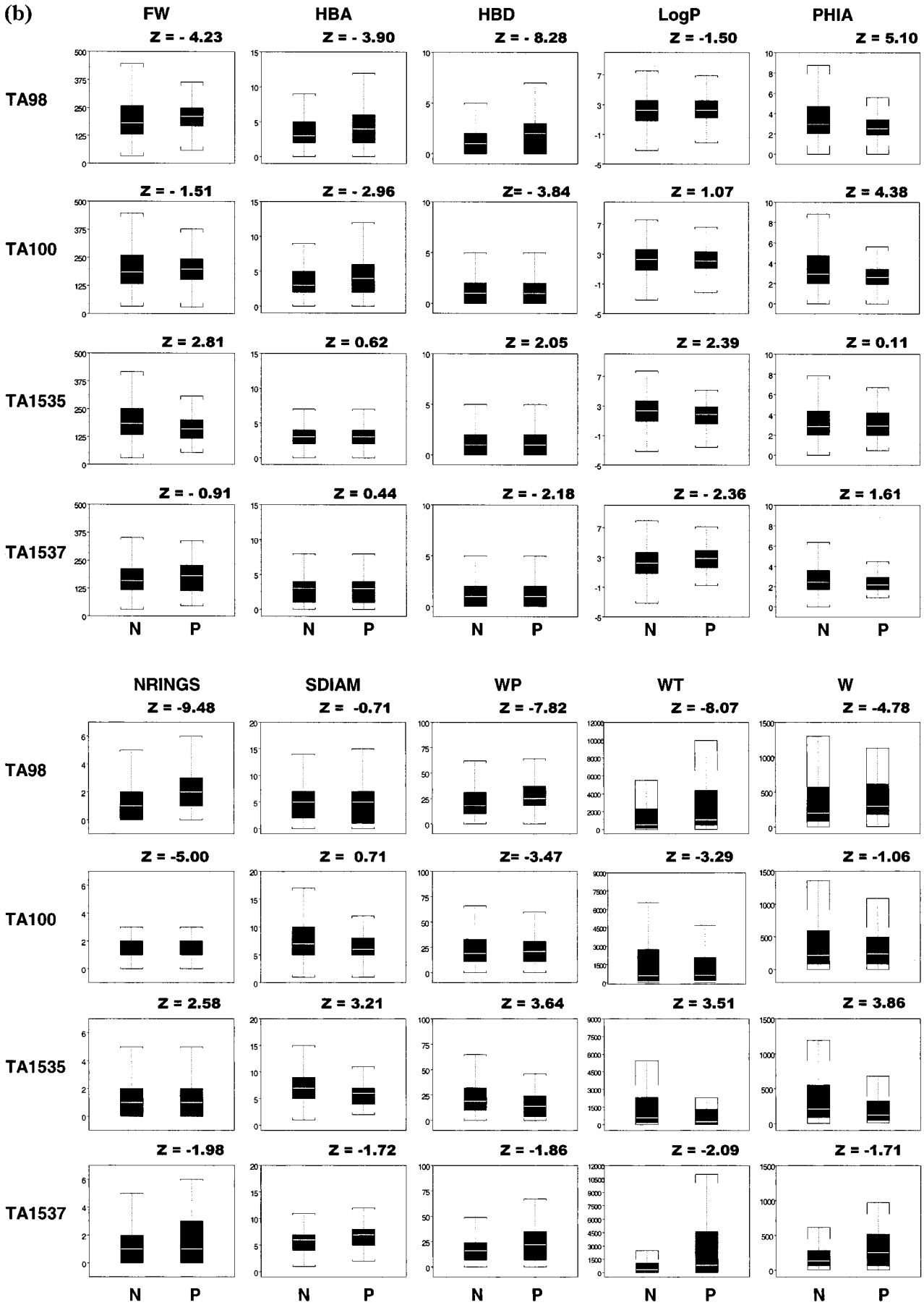
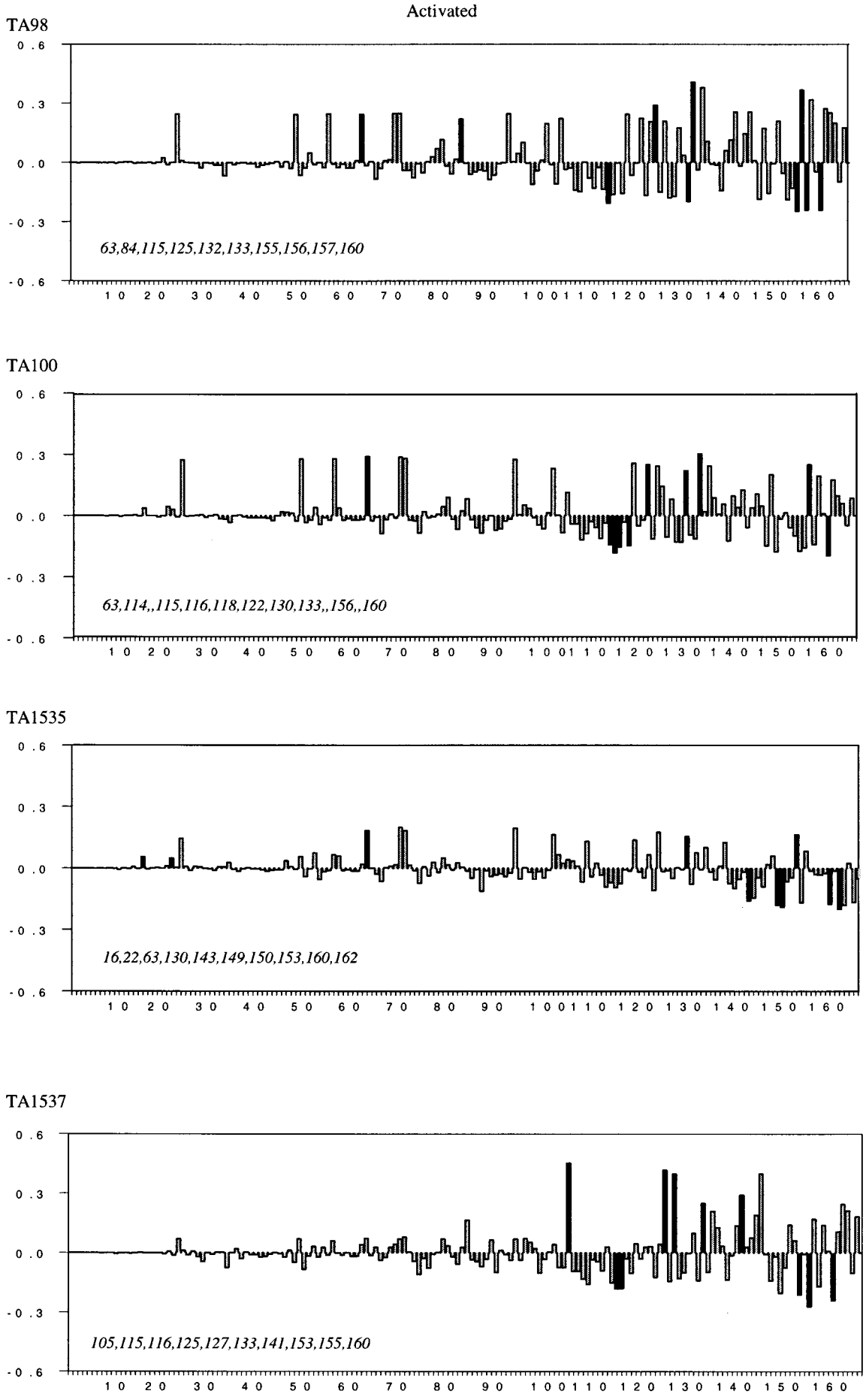


Figure 1. Box plots for each of the properties and the strains studied: (a) compounds did not undergo activation; (b) with the compounds undergoing metabolic activation. P stands for compounds giving a positive response in the assay, while N is for the negatives. The stripe in the box indicates the value of the mean.


Activated
TA98
0.6
0.3
0.0
-0.3
-0.6
63,84,115,125,132,133,155,156,157,160
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160
TA100
0.6
0.3
0.0
-0.3
-0.6
63,114,,115,116,118,122,130,133,,156,,160
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160
TA1535
0.6
0.3
0.0
-0.3
-0.6
16,22,63,130,143,149,150,153,160,162
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160
TA1537
0.6
0.3
0.0
-0.3
-0.6
105,115,116,125,127,133,141,153,155,160
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160

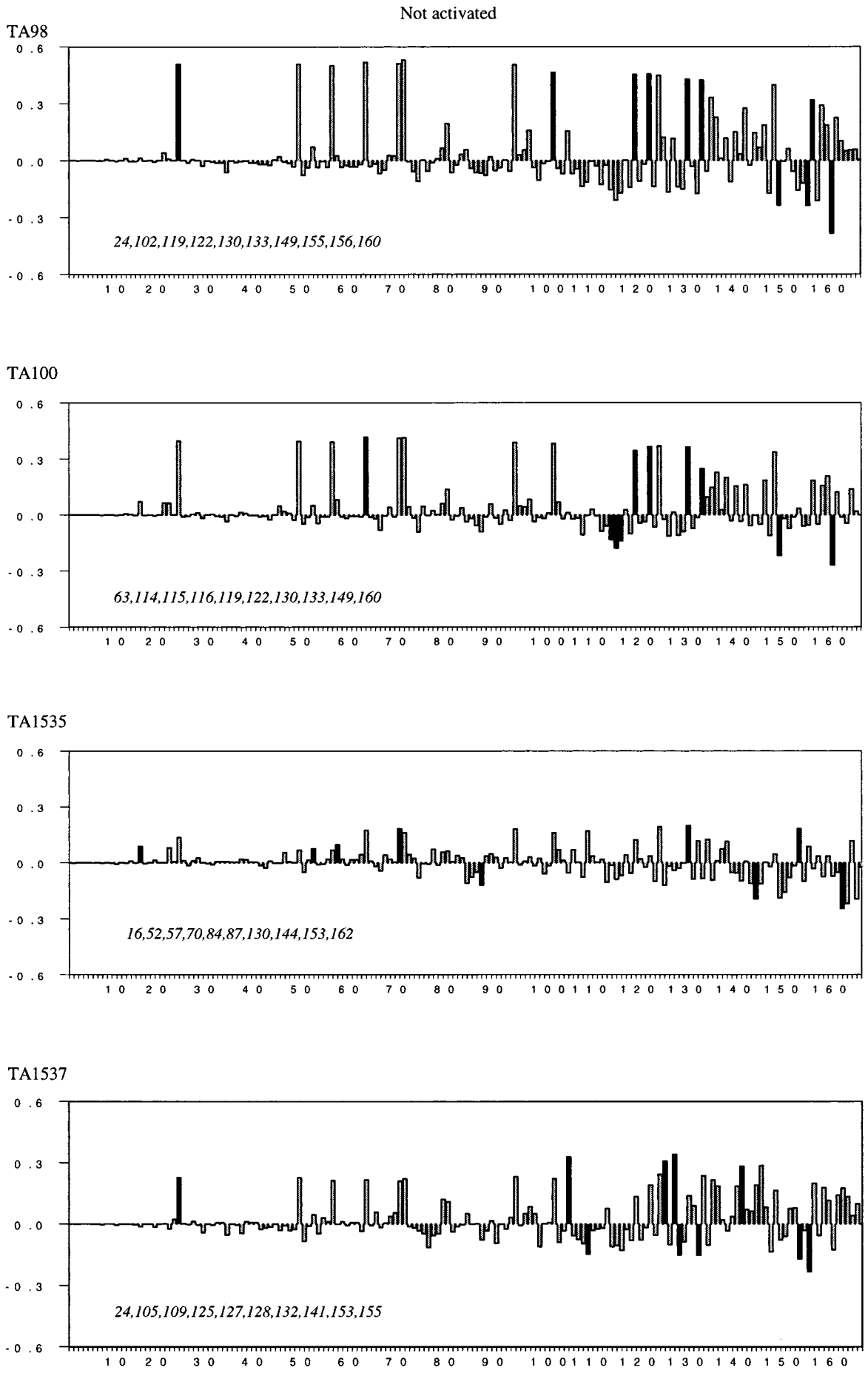


Figure 2. Frequency of the distribution of each ISIS/MOLSKEYS for each strain with or without activation.

Table 2. Identity of Certain Particular Bits in ISIS-MOLSKEYS[a]

bit no.	key	bit no.	key
16	QAA@1	22	three-membered ring
24	N−O	52	NN
57	O in heterocycle	63	N═O
70	QNQ	71	NO
84	NH_2	87	X!A$A
105	A$A($A)$A	109	A CH_2O
114	CH_3CH_2A	115	CH_3ACH_2A
116	CH_3AACH_2A	118	$ACH_2CH_2A > 1$
122	AN(A)A	125	aromatic rings > 1
127	A$A! > 1	128	ACH_2AAACH_2A
130	QQ > 1	132	$OACH_2A$
133	A$A!N	141	$CH_3 > 2$ and other rare features
143	A$A!O	144	A-not%-A%A-not%-A
149	$CH_3 > 1$	150	A!A$A!A
153	QCH_2A	155	A!CH2!A
156	NA(A)A	157	C−O
160	CH_3	162	aromatic

[a] Q stands for any atom other than C or H, A any atom other than H. $ represents a bond in a Ring, ! a bond in a chain, *@n* atom *n* attached to cycle.

These results may indicate that there is a window of hydrophobiticity that nonmutagenic compounds fall within.

When the compounds are activated using the rat liver S9 mix assay, the properties that distinguish between positive and negative compounds are not the same. The number of rings is a factor that shows a different distribution in positive activated compounds, the TA-98 and TA-100 strains being larger among the positive compounds. The number of hydrogen bond donors is quite significant as a discriminant for all strains but TA-1535, which appears to be dependent on more subtle factors that are better captured by topological parameters such as the Wiener index.

A comparison of the parameters that show divergence between positive and negative compounds makes it clear that the compounds that are detected as mutagenic by each strain have different physicochemical properties. The most significant is that the average number of hydrogen bond donors or acceptor groups appears to be larger for the positive compounds.

(b) Bit String Properties and Fragment Analysis. The bar plots in Figure 2 show the difference in the frequency with which each of the 166 bits described by the MOLSKEYS appears between positive and negative compounds. The sign of that bar plot determines whether the particular bit string is more abundant in the positive or in the negative set of compounds, respectively. Table 2 shows some of the MOLSKEYS of interest to this study and their associated moieties. While some bits appear with similar frequency in positive and negative compounds, others show a marked preference for either kind of compounds. The result in itself could be expected, because bit strings are associated to the presence of certain substructures or moieties in chemicals, and fragment-based methods are the basis of techniques described to predict the mutagenic character of compounds. The bar plots show that for the TA-98 and TA-100 strains keys 24, 63, 70, 71, and 122, related with NO_2 and NO groups, are overrepresented. All these keys are strongly paired with positive activity. Interestingly, some bits appear less frequently in the positive compounds. The implication would be that certain fragments might actually decrease the potential that a given compound would be positive in the Ames test. Keys 149, 157, and 160, encoding substructures related with aliphatic character and carboxyl groups, respectively, are most common in negative compounds.

Strain TA-1537 shows an overrepresentation of bits numbered 105, 125, and 127 and 141, encoding aromatic rings and heteroatoms in a cycle strongly related with activity. Conversely, for the same strain, the presence of alkyl chains (MOLSKEYS 153 and 155) is again associated with the negative compounds. Finally, TA-1535 presents MOLSKEYS 70 and 130, encoding the presence of heteroatoms related with positive activity. Surprisingly, aromaticity (MOLSKEY 162) is correlated with negative activity for this strain.

As it was the case for global properties, each strain is responding to compounds containing different molecular fragments. TA-98 and TA-100 have the closest similarity in the fragments that are overrepresented, but still the two strains are sufficiently different to be able to discriminate among them.

On the basis of the results obtained from the bits overrepresented in the MOLSKEYS, a subset of chemical functionalities was selected to determine whether they were at least in part responsible for the higher frequency of particular bit strings. Table 3 shows the results of searching for the presence of specific substructural moieties in positive compounds in assays carried out with and without activation for each of the bit strings. The table shows the number of positive and negative molecules containing each fragment for each strain and assay, as well as the binomial probability (*p* binomial) that the distribution could be due to chance. The lower the binomial probabilities, the more significant the difference in the frequency of appearance of that functional group or moiety compared to what could be expected if it was a random event. For TA-1535 and TA-1537, the limited number of positive compounds does not permit achieving statistical significance in some cases where a bias appears to be present.

Again, different strains discriminate the mutagenic potential of distinct chemical classes. TA-98 and TA-100 show some degree of correlation because, for both strains, $PhNO_2$ groups and N = X groups are disproportionately represented among the positive. TA-1535 identified epoxy and haloalkanes compounds as mutagenic, but interestingly all other strains show only a slight bias against these compounds. Certain functional groups appear to be less likely to be positive against any strain in assays with or without activation. For instance, compounds containing carboxylic, propyl, and even methyl groups are less frequent in positive compounds for any of the strains, compared to what could be expected considering their abundance in the overall library. Results in Table 3 are consistent with previous structural alerts about mutagenicity or carcinogenicity.[32] While there are global properties and fragments that can be found as biased in all assays, in a majority of cases, the results from each strain are different.

Mutagenicity Prediction

While our aim is not the prediction of the activity of individual mutagenic compounds, whether the properties discussed are indeed able to discriminate active from inactive compounds is important in their evalua-

Table 3. Molecular Fragments Found in Positive (P) and Negative (N) Compounds, Based on the Results for Compounds without and with Metabolic Activation

Without Metabolic Activation

	TA-98			TA-100			TA-1535			TA-1537		
moieties	P (231)	N (1708)	*p*	P (232)	N (1789)	*p*	P (61)	N (1082)	*p*	P (59)	N (908)	*p*
$PhNH_2$	**34**	136	<0.05	**17**	161	<0.05	4	**115**	<0.05	8	**76**	<0.05
$PhNO_2$	**118**	106	≪0.05	**89**	112	≪0.05	**8**	117	<0.05	**17**	81	<0.05
Ph-O-C	20	**176**	<0.05	13	**165**	≪0.05	**8**	129	<0.05	5	**88**	<0.05
Ph-S-C	**5**	21	<0.05	**4**	24	<0.05	1	20	>0.05	1	**15**	<0.05
$C\text{-}CH_2X$	8	**76**	<0.05	**28**	37	<0.05	**13**	28	<0.05	1	41	>0.05
$C\text{-}CH_2\text{-}O\text{-}CH_2\text{-}C$	1	36	>0.05	1	**30**	<0.05	1	17	0.05	0	13	>0.05
$HO\text{-}N\text{-}C_2$	0	1	>0.05	0	1	>0.05	0	0		0	0	
Ph Ph	**69**	317	<0.05	**42**	328	<0.05	9	**231**	<0.05	**30**	168	≪0.05
C-COOH	21	**339**	≪0.05	27	**285**	<0.05	6	**214**	<0.05	8	182	<0.05
N═C, N═O, N═N	**149**	263	≪0.05	**123**	258	≪0.05	**19**	224	<0.05	**21**	180	<0.05
$CH_2\text{-}CH_2\text{-}CH_2$	5	**244**	≪0.05	8	**229**	≪0.05	8	**162**	<0.05	3	**146**	<0.05
$C\text{-}CH_3$	47	**788**	≪0.05	46	**722**	≪0.05	25	**493**	<0.05	18	**433**	<0.05
epoxy	5	22	>0.05	**15**	15	<0.05	**7**	7	<0.05	0	10	>0.05
sulfonamide	0	**21**	<0.05	1	22	>0.05	0	13	>0.05	0	10	>0.05
methylsulfoxide	0	11	>0.05	1	9	>0.05	**2**	3	<0.05	1	2	>0.05
$-NH_2$	6	23	>0.05	5	28	<0.05	**0**	13	>0.05	**2**	22	<0.05
-NHC	0	5	>0.05	4	21	>0.05	0	3	>0.05	0	3	>0.05
C-CO-NCC	4	33	>0.05	3	26	~0.05	0	16	>0.05	0	15	>0.05

With Metabolic Activation

	TA-98			TA-100			TA-1535			TA-1537		
moieties	P (335)	N (1306)	*p*	P (322)	N (1170)	*p*	P (93)	N (838)	*p*	P (82)	N (794)	*p*
$PhNH_2$	**91**	70	≪0.05	**59**	96	<0.05	**11**	97	<0.05	**19**	63	<0.05
$PhNO_2$	**95**	92	≪0.05	**97**	84	≪0.05	**13**	73	<0.05	**11**	61	<0.05
Ph-O-C	29	**130**	<0.05	22	**113**	<0.05	8	**97**	<0.05	7	69	~0.05
Ph-S-C	**8**	11	<0.05	6	12	>0.05	1	15	>0.05	2	12	>0.05
$C\text{-}CH_2X$	8	53	~0.05	**26**	30	<0.05	**20**	18	≪0.05	3	**34**	<0.05
$C\text{-}CH_2\text{-}O\text{-}CH_2\text{-}C$	0	**34**	<0.05	1	**27**	<0.05	2	13	>0.05	0	13	~0.05
$HO\text{-}N\text{-}C_2$	1	2	>0.05	0	**16**	<0.05	0	0		0	0	
Ph Ph	**144**	215	<0.05	**96**	218	<0.05	19	**180**	<0.05	**49**	134	≪0.05
C-COOH	21	**290**	≪0.05	32	**233**	<0.05	8	**174**	<0.05	7	**166**	<0.05
N═C, N═O, N═N	**134**	164	≪0.05	**137**	184	≪0.05	**29**	155	<0.05	**20**	141	<0.05
$CH_2\text{-}CH_2\text{-}CH_2$	4	**198**	≪0.05	8	**185**	≪0.05	8	**121**	<0.05	4	**132**	<0.05
$C\text{-}CH_3$	80	**617**	<0.05	**87**	544	≪0.05	28	**383**	<0.05	23	**383**	<0.05
epoxy	5	17	>0.05	**13**	10	<0.05	**6**	3	<0.05	1	7	>0.05
sulfonamide	0	**19**	<0.05	1	**20**	<0.05	0	6	>0.05	0	8	>0.05
methylsulfoxide	0	10	>0.05	1	9	>0.05	0	3	>0.05	1	2	>0.05
$-NH_2$	3	22	>0.05	3	**19**	<0.05	0	**14**	<0.05	1	12	>0.05
-NHC	0	5	>0.05	0	5	>0.05	0	3	>0.05	0	3	>0.05
C-CO-NCC	3	23	>0.05	2	**17**	<0.05	0	10	>0.05	0	12	~0.05

[a] The binomial probability (*p*) for the frequency is shown. The total number of positive and negative compounds in each of the sets is shown in parentheses for each column. Bolded numbers highlight the moieties that are found more frequently than expected, based on the fraction of positive compounds and the abundance of that moiety in the library.

tion. The best way to achieve this goal is by developing predictive models that use the properties in question and analyzing their performance in the characterization of compounds not used for the development of the models.

Recursive partitioning is a technique used for the classification of the data starting from set of compounds. In our case, clusters of 'pure' positive and negative compounds are expected to be generated after few partitions with the aid of a decision tree to create a series of descriptor planes. These partitions filter the compounds into cohesive blocks containing the least interclass mixing possible. When a node or a branching point is defined by selecting a descriptor and a corresponding threshold at which to divide the data set, compounds for which the descriptor value is above the threshold are assigned to one branch, while compounds below the threshold are assigned to the other. The choice of descriptor and threshold is designed to 'best' purify the mixture. In statistical terms, the split is made so as to minimize the sum of the squared deviances in the left (L) and right (R) branches.

$$\text{minimize}\{\sum_{i\in R}(y_i - \mu_L)^2 + \sum_{i\in R}(y_i - \mu_D)^2\}$$

For the positive or negative case, y_i is assigned a value of 1 or 0, respectively, and the means μ_L and μ_R are simply the fraction of active compounds in each of the branches.

The partitioning procedure is repeated to form additional branches with a new splitting criterion at each node. A single descriptor may be used more than once in the construction of the tree, but a unique threshold is selected each time. The partition ends when no further reduction in the deviance is possible. If perfect separation is achieved, the resulting tree will terminate in leaves consisting of groups of purely active or inactive compounds.

Once a tree has been grown using a training set of compounds, it can be used to classify an external set of

compounds into active or inactive leaves. Since rarely all the leaves in a tree are pure, classification must be done in terms of probability, so a leaf is considered active when it contains more active compounds than inactives and vice versa. When a balanced training set is used, the composition needs only to be greater than 50% in either direction to assign the leaf class membership.

In any classification approach, an appropriate training set is critical to the accuracy of the predictions. This is more important, indeed, if the model is required for the prediction of large databases typically found in chemical libraries designed for combinatorial chemistry and for high-throughput screening or the ones generated for the present work. In our database, the active and inactive groups are unbalanced in a way that the population of negative molecules is much bigger than the population of positive compounds. Such unbalance in the training set can result in unbalanced predictions, giving high accuracy when the algorithm returns for every compound a negative attribute. To improve sensitivity, multiple numbers of balanced training sets were generated containing equal number of positives and negatives.[27]

The construction of trees using global descriptors computed using Molconn-X and MOLSKEYS bit strings (properties) was undertaken only with uncorrelated descriptors to avoid giving more weight to one given set of properties. The calculation of correlations between all Molconn-X descriptors revealed high correlation in all the strains between Bonchev-Trinajstic indexes and Wiener number and between Wiener *p* and Platt *f* numbers. Weiner and Platt *f* numbers were selected as representatives in the tree models performance.

The recursive partitioning models were derived for each strain, with and without activation, and based on the results of the training sets picked by random selection of an equal number of positive and negative compounds. The models derived were used to predict the activity of the remaining compounds. For balanced classifications, five different training sets were selected at random, containing an equal number of positive and negative compounds. For TA-1535 and TA-1537 strains where few active compounds were available, all the active compounds were included to generate the model, and thus, no predictions were performed for the active molecules.

All models were cross-validated to have reliable estimates of the number of nodes that could be considered statistically significant, as well as of the variables that provide an adequate partition of the data. Table 4 shows the percentage of compounds fitted by the cross-validated models and the fraction of compounds accurately predicted by the best model using the global molecular properties. The models selected are those that show a similar rate of misclassification for the fitted and the predicted data sets and are therefore considered more reliable. Figure 3 shows an example of the results of a recursive partition of the global properties for the TA-98, since it is for this strain where some degree of reliability in the models has been found.

When the compounds are not activated, the TA-98 and TA-100 strains show a rate misclassification in the prediction that ranges from 63% to 74%. The results can be considered satisfactory given the simplistic nature of the global properties considered. Also, for the activated compounds in the TA-98 strain, results are equally good. For TA-100, the best model found has a higher misclassification rate than any of the models for the TA-98 strain, using the global properties. Consequently, the predictions are not as satisfactory as with

Table 4. Percentages of the Number of Compounds Accurately Predicted by the Recursive Partition Method, Using Global and Bit String Properties

	activated				not activated			
	model		prediction		model		prediction	
strain	P	N	P	N	P	N	P	N
Global Descriptors								
TA-98	66	76	64	70	78	70	74	63
TA-100	61	73	57	57	79	67	67	64
TA-1535	48	85		74	46	87		78
TA-1537	52	82		79	73	61		62
Bit String Descriptors								
TA-98	80	70	73	68	64	95	68	91
TA-100	59	79	59	74	51	91	51	91
TA-1535	55	84		65	57	73		62
TA-1537	60	93		85	68	80		72

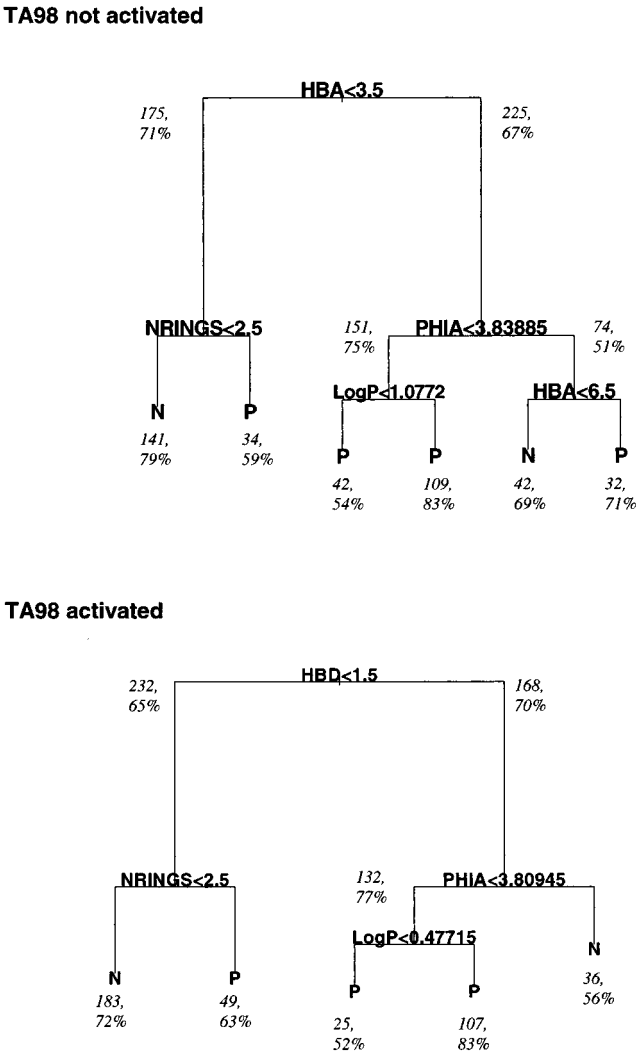


Figure 3. Recursive partition tree for the TA-98 strain, with and without undergoing metabolic activation. Each branch shows the number of compounds in the split and the fraction of compounds within the same class.

TA-98. For the TA-1535 strain, the global properties appear not to be able to generate a robust or reliable model for the results found for this strain. In the case of the TA-1537, the models appear to be reliable in the case of the not-activated compounds, but not so when the compounds are activated. The most likely reason for the inability of the method to generate reliable models is the small number of positive compounds available in the training sets.

The recursive partitioning trees reinforces some of the properties singled out by the Wilcoxon analysis. For instance, the number of hydrogen bond accepting centers is the first partitioning variable for the TA-98 and TA-100 strains when the compounds are not activated. Compounds with few hydrogen bond accepting groups and large number of rings are likely to be deemed mutagenic. The PHIA parameter that relates to the overall flexibility of the substance, whenever used by the models, is larger for nonmutagenic compounds. The property is present in several of the tree models and consistently shows lower values for positive compounds. Thus, more rigid chemicals have a tendency to be picked up as positive. The results are consistent with a smaller fraction of active compounds with akliphatic character. While these are not set rules with clear cutoffs, they are trends that should be considered in designing libraries.

In the derivation of the models using MOLSKEYS, attention was paid to the large number of variables that are part of the bit strings. To avoid overfitting the data, only 10 of the 166 bits present in the MOLSKEYS were selected for each database as variables to develop the models.[27] The bits used for the models are shown as black bars in Figure 2 and indicated at the foot of each bar plot. The most overrepresented or underrepresented bits were scrutinized, and then the degree of correlation between the bits was analyzed. Some bits over- or underrepresented in the set point to the same functional group or moiety, which results in correlation of the bits. Therefore, the 10 bits selected for each of the models were those that were significantly over or underrepresented and at the same time uncorrelated with each other. Table 2 provides a description of each of the bits used by the models.

The bit strings provide a complementary picture to that presented by the global descriptors and in line with the differences observed in the distribution of frequencies for individual bits. For instance, nitro groups set on several different bits in the MOLSKEYS. For the nonactivated TA-98 and TA-100, as well as TA-1537, the bits selected correspond to some of those that would be set by nitro groups. The importance of the bits associated to nitro groups as discriminants of mutagenicity is consistent with the large fraction of nitro-containing molecules that are mutagenic, as shown in Table 3. Moreover, their relevance to assess the mutagenic potential of a compound has been known for many years, even when those results were derived with smaller data sets.[32]

When the results of the models derived based on the MOLSKEYS or the 2D properties are compared, the misclassification rates are smaller for the former than the latter. The bit strings are able to generate predictions with 9% misclassification rates for the best models, where the global properties only reach 20% at best. TA-98 and TA-100 show plainly the trend when the predictions are considered. TA-1535 and TA-1537 show a similar trend in the cross-validated models; however, their validity is harder to assess, because of the lack of sufficient positive compounds for this analysis.

The performance of the bit strings is consistent with the use of fragment-based approaches by the pattern recognition techniques used for the prediction of mutagenicity and carcinogencity.

Conclusions

The purpose of this paper was the analysis of biases in the structure and molecular properties of the compounds that are found to be mutagens. The properties most commonly used for library design, bit strings and global properties, show their ability to discriminate among Ames positive and negative compounds for each strain selected. These properties can then be used for diversity analysis but also to provide constraints to further tailor libraries toward compounds more likely to be developed into drugs.

Previous efforts have focused on the overall mutagenic character of a compound, while in this study, an analysis of each strain was separately performed. Clear biases are found in the properties and structures of the positive compounds for each of the strains. Each strain shows unique variations in the properties of the compounds reported to be positive. The variations observed are consistent with the use of multiple assays to determine the overall mutagenicity of a compound.

Excellent and abundant literature exists on the accurate prediction of the mutagenic potential of a chemical. Our interest is in the analysis of properties and molecular fragments more likely to discriminate a positive compound so that they could be incorporated into the design of chemical libraries for screening. A few simple rules can be derived to decrease the chances that a compound may display genotoxicity. For instance, introduction of certain functional groups, such as aliphatic chains or carboxylic groups, could decrease the mutagenic potential of a compound. Also, replacement of other functional groups, such as NO_2, or decreasing the overall number of hydrogen bond accepting centers could also improve the odds of having a nonmutagen.

The use of molecular fragments appears to be more predictive than the use of global molecular properties. Consequently, methods based on the identification of structural elements such as LASSOO[33,34] or others described in the literature[35,36] to bias against certain chemical classes based on structure should be able to carry out efficient work in weeding out undesirable ligands. Even with the continuous addition of constraints in the process of compound selection, there should be no risk of having limited diversity in the libraries, as the number of potential chemicals is still enormous and not likely to be dented by these restraints.

Acknowledgment. O.J.L. acknowledges a grant in aid from the Generalitat de Catalunya to visit Telik Inc.

References

(1) Leach, A. R.; Hann, M. M. The in-silico world of Virtual Libraries. *Drug Discovery Today* **2000**, *5*, 326−336.

(2) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring diversity: experimental design of combinatorial libraries for drug discovery. *J. Med. Chem.* **1995**, *38*, 1431−436.

(3) Martin, Y. C.; Brown, R. D.; Bure, M. G. In *Combinatorial chemistry and molecular diversity in drug discovery*; Gordon, E. M., Kerwin, J. F., Jr., Eds.; Wiley-Liss: New York, 1998; pp 369−385.

(4) Dixon, S. L.; Villar, H. O. Bioactive diversity and screening library selection via Affinity Fingerprinting. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1192−1203.

(5) Lahana, R. How many hits from High Throughput Screening? *Drug Discovery Today* **1999**, *4*, 447−448.

(6) Newton, C. G. In *Molecular Diversity in Drug Design*; Dean, P. M., Lewis, R. A., Eds.; Kluwer Academic Publishers: Dordrecht, 1999.

(7) Watt, A. P.; Morrison, D.; Evans, D. C. Approaches to higher throughput pharmacokinetics in drug discovery. *Drug Discovery Today* **2000**, *5*, 17−24.

(8) Martin, E. J.; Critchlow, R. E. Beyond mere diversity: tailoting combinatorial libraries for drug discovery. *J. Comb. Chem.* **1999**, *1*, 32−45.

(9) Villar, H. O.; Koehler, R. T. Commernts on the design of chemical libraries for screening. *Mol. Diversity* **2000**, *5*, 13−24.

(10) Ajay; Walters, W. P.; Murcko, M. A. Can We Learn to Distinguish between "Drug-like" and "Nondrug-like" Molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.

(11) Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.

(12) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

(13) Kauvar, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engqvist Goldstein, A.; Bukar, R.; Bauer, R. E.; Dilley, H.; Rocke, D. M. Predicting Ligand Binding To Proteins By Affinity Fingerprinting. *Chem. Biol.* **1995**, *2*, 107−118.

(14) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeny, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Drug Delivery Rev.* **1997**, *23*, 3−25.

(15) Eddershaw, P. J.; Beresford, A. P.; Bayliss, M. K. ADME/PK as part of a rational approach to drug discovery. *Drug Discovery Today* **2000**, *5*, 409−414.

(16) Oprea, T. I. Property Distribution of Drug-Related Chemical Databases. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251−264.

(17) Josephy, P. D.; Gruz, P.; Nohmi, T. Recent Advances in the Construction of Bacterial Genotoxicity Assays. *Mutat. Res.* **1997**, *386*, 1−23.

(18) Basic Mutagenicity Tests: UKEMS Recommended Procedures. Kirkland, D., Gatehouse, D. G., Scott, D., Cole, J., Richold, M., Eds.; Cambridge University Press: Cambridge, 1990.

(19) Gatehouse, D.; Haworth, S.; Cebula, T.; Gocke, E.; Kier, L.; Matsushima, T.; Melcion, C.; Nohmi, T.; Ohta, T.; Venitt, S.; Zieger, E. Recommendations for the performance of Bacterial Mutation Assays. *Mutat. Res.* **1994**, *312*, 217−233.

(20) Lewis, D. F. V. *Computer Assisted Methods in the Evaluation of Chemical Toxicity in Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: 1992; Vol. 3, pp 173−222.

(21) Klopman, G. Artificial intelligence approach to Structure−Activity Studies, Computer Automated Structure Evaluation of Biological Activity of Organic Molecules. *J. Am. Chem. Soc.* **1984**, *106*, 7315−7320.

(22) Klopman, G., MULTICASE: A Hierarchical Computer Automated Structure Evaluation Program. *Quant. Struct.-Act. Relat.* **1992**, *11*, 176−182.

(23) Enslein, K.; Gombar, V. K.; Blake, B. J. Use of SAR in computer-assisted prediction of carcinogenicity and mutagenicity of chemicals by the TOPKAT program. *Mutat. Res.* **1994**, *305*, 47−61.

(24) Henry, D. R.; Jurs, P. C.; Denny, W. A. Structure-antitumor activity relationships of 9-anilinoacridines using pattern recognition. *J. Med. Chem.* **1982**, *25*, 899−908.

(25) Sanderson, D. H.; Earnshaw, C. G. Computer Prediction of Possible Toxic Action from Chemical Strcuture; the DEREK system. *Hum. Exp. Toxicol.* **1991**, *10*, 261−273.

(26) Benfenati, E.; Gini, G.; Computational predictive programs (expert systems) in toxicology. *Toxicology* **1997**, *119*, 213−225.

(27) Dixon, S. L.; Villar, H. O. Investigation of Classification Methods for the Prediction of Activity in Diverse Chemical Libraries. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 533−545.

(28) CCRIS database can be found through TOXNET, http://toxnet-.nlm.nih.gov.

(29) Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand−Receptor Binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.

(30) Triola, M. F. *Elementary Statistics*; Addison-Wesley Publishing Co.: Reading, MA, 1992.

(31) Debnath, A. K.; Debnath, G.; Shusterman, A. J.; Hansch, C. A QSAR Investigation of the Role of Hydrophobicity in Regulating Mutagenicity in the Ames Test: 1. Mutagenicity of Aromatic and Heteroaromatic Amines in Salmonella Typhimurium TA-98 and TA-100. *Environ. Mol. Mutagen.* **1992**, *19*, 37−52.

(32) Ashby, J.; Tennant, R. W. Definite Relationships among Chemical Structure, Carcenogenicity and Mutagenicity for 301 Chemicals Tested by the U.S. NTP. *Mutat. Res.* **1991**, *257*, 229−306.

(33) Koehler R. T.; Dixon, S. L.; Villar, H. O. LASSOO: A Generalized Directed Diversity Approach To The Design And Enrichment Of Chemical Libraries. *J. Med. Chem.* **1999**, *42*, 4695−4704.

(34) Koehler, R. T.; Villar, H. O. Design of Screening Libraries Biased for Pharmaceutical Discovery. *J. Comput. Chem.* **2000**, *21*, 1145−1152.

JM0004594