ELSEVIER

Chemometrics and Intelligent Laboratory Systems 60 (2002) 5–11

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

Mixture deconvolution and analysis of Ames mutagenicity data

S. Stanley Young [a,*], Vijay K. Gombar [a], Michael R. Emptage [a], Neal F. Cariello [a], Christophe Lambert [b]

[a] *Department of Research Computing, Glaxo Wellcome Inc., 5 Moore Drive, Research Triangle Park, NC 27709, USA*
[b] *Golden Helix Inc., 716 South 20th Avenue, Suite 102, Bozeman, MT 59718, USA*

Abstract

Mixtures abound in chemistry: two or more compounds may be present in the same sample, the same biological effect may be produced by two different mechanisms, or two compounds might bind to a receptor in different orientations or even in different places. Sometimes, results are given in summary form. For example, a chemical may be declared a mutagen due to any of several assay results from an Ames test. Clearly, a single mathematical model is not going to hold for data sets where such multiplicity of phenomena are represented. We need molecular descriptors and statistical methods which enable us to deconvolute such mixtures. Our idea is to combine topological chemical descriptors—augmented atoms and through-bond distance measures—with a statistical technique, segmentation recursive partitioning, that is capable of dealing with mixtures. The benefit is the ability to develop structure–activity relationships for large, heterogeneous data sets. We successfully demonstrate the effectiveness of the above descriptors and the technique of recursive partitioning with Ames test results taken from public sources.

Keywords: Mutagenicity; Ames test; Recursive partitioning; Mixtures; Segmentation; QSAR

1. Introduction

Large structure–activity data sets are now becoming available for analysis. Unfortunately, these large data sets are often heterogeneous. Some of the compounds are active through one mechanism while others produce their effect by another. The Ames test for mutagenicity is prototypical in that typically four strains of bacteria are tested, with and without metabolic activation. A compound is declared positive if any of the eight results are positive. Therefore, not only can mutations in single strain/activation combination be induced by different mechanisms, but each strain is likely to operate somewhat differently. Clearly, multiple mechanisms are operable across the whole system of testing.

Unfortunately, simple linear-based methods of statistical analysis are likely to have trouble with multiple mechanism data sets. In this paper, we describe the use of recursive partitioning for the analysis of large chemistry data sets.

2. Data

The method of recursive partitioning requires relatively large data sets to be effective. We secured mutagenicity data (Ames test) from several public sources including the US EPA, NIH and the open

* Corresponding author.
E-mail addresses: ssy0487@glaxowellcome.com, genetree@bellsouth.net (S.S. Young), lambert@goldenhelix.com (C. Lambert).

0169-7439/02/$ - see front matter
PII: S0169-7439(01)00181-2

literature. After harmonizing the data from different sources, the data were cleaned by removing duplicate compounds, by excluding inorganic and organometallic compounds, by deleting studies on mixtures and by substituting structures of free acids/bases for their salts. The data set that we analyze here consists of a total of 2018 Ames test results from single, organic compounds. A copy of the data set (CAS number, 1/0—Ames positive/negative, and SMILES string) can be obtained by contacting the first author.

3. Methods

3.1. Descriptors

We build up a vector descriptor for each molecule from simple parts. The basic descriptor is an atom distance–atom descriptor. The two focus atoms name the element of the vector and the distance is the value of the element. Focus atoms are typed by their attached nonhydrogen atoms. For example, C(CC) denotes a focus atom, carbon, that is attached to exactly two other carbons. O(N) is an oxygen attached to a nitrogen. These augmented atoms, AAs (also termed atom-centered fragments), were first proposed by Adamson et al. [1]. As originally proposed, an AA was defined as a nonhydrogen atom, its nonhydrogen adjacent atoms and their bonds. AAs were used by CAS for chemical structure searching and have also been used for pattern recognition and QSAR studies (see, for example, Chu et al. [2]). We simplify the AA by omitting the nature of the connecting bonds. The distance is taken as the number of bonds in the shortest path between the focus atoms. We create a compound descriptor whose name is composed of two AAs. The value of the variable is missing (?) if one or both of the focus atoms do not appear in the structure, or the value of the variable is the number of bonds in the shortest path between the two focus atoms. See Fig. 1 for examples of augmented atoms.

Because there can be a second incidence of one of the focus atoms, two variables are created and the distance is either "high" or "low" depending upon the shortest path distance to the furthest focus atom or the nearest focus atom, respectively.

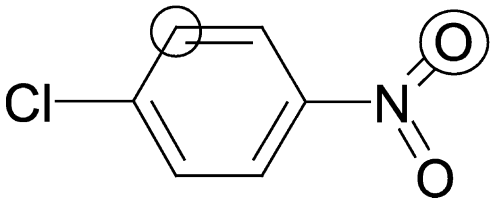

Fig. 1. Augmented atom pairs are defined by the focus atoms, in this case a carbon and an oxygen and the atoms attached to these atoms and the number of bonds separating the focus atoms. We use the notation C(CC)−4−O(N), where the first carbon is attached to two carbons, in this case, one on each side, and the oxygen, attached to only one nitrogen. There are four bonds between the focus atoms. In our representation, the nature of the bonds is not noted.

The motivation for AA pairs is that the interaction of a molecule with a biological system can be influenced by specific atom environments and the distance between those environments.

3.2. Segmentation recursive partitioning

Recursively splitting a data set into homogeneous subsets was first proposed by Morgan and Sonquist [3]. Methods for univariate recursive partitioning (RP) are described by Hawkins and Kass [4], Breiman et al. [5] and Quinlin [6]. Hawkins et al. [7] and Rusinko et al. [8] applied RP to QSAR problems. In RP, all potential predictor variables are examined and the single variable that will best split the entire data set into two or more daughter data sets is selected and the split made. In the case of a binary predictor, two-way splitting, those compounds with the feature go to the right daughter node and those without the feature go to the left. Where the descriptor is continuous, rather than feature present/absent, one or more "cut points" are determined, segmentation, and a two-way or multi-way split is effected. We use ChemTree™ [9], whose algorithms find the optimal multi-way split, such that the sum of squared deviations from the mean of each segment is minimized over all possible segmentations. An example of a segmentation is given in Table 1. There are a number of strategies one might use for selecting the number of segments. Yao [10] proposed selecting the number of change points by examining the residual sum of squares after fitting means to each segment. The ChemTree software first,

Table 1
This is an example of how segmentation might occur

Distance	Number of bonds between focus atoms											
	?	1	2	3	4	5	6	7	8	9	10	11
No. active	235	12	2	7	13	12	7	4	4	3	1	0
No. tested	612	44	7	20	19	25	17	10	13	10	2	2
Percent active	38	27	29	35	**68**	48	41	40	31	30	50	00

First, note that a molecule may not have the two focus atoms at issue. We note the absence of one or both focus atoms by (?). In this case, there are 612 molecules under consideration that do not have the focus pair of atoms. Of the 612 molecules, 235 are active, giving 38% active. There are 44 focus pairs where there is one bond between them; 12 are active, giving 27% active. There are seven molecules with two bonds between the focus atoms. Note the percent active. Those atoms having the focus atoms one, two or three bonds apart have approximately the same percent active molecules (27, 29 and 35). Note that percent active increases to 68% for molecules with four bonds between the focus atoms. The segmentation algorithm searches for the optimal grouping. The underlining gives a possible grouping. Note that the (?) group is statistically compared to each of the other groups and, if it is statistically reasonable, combined with one of the groups.

through an exhaustive search, finds the best k segments and then carries out $k-1$ tests comparing the response data for consecutive segment pairs. If the smallest test statistic is Bonferroni significant, then the optimal $k-1$ segment fit to the data is made, and the testing process repeated. The process stops when the smallest test is not statistically significant, indicating that the remaining segments are significantly different from one another.

4. Results

Recursive partitioning is capable of identifying multiple chemical classes of compounds from a data set, deconvoluting mixtures [7,8]. Fig. 2 gives a skeleton of the recursive partitioning tree. Figs. 2 and 3 and Table 2 are cross-referenced through the node number. Table 2 gives the number of compounds in each node (the number of compounds in daughter nodes adds up to the number of compounds in the parent node). The percent of positive compounds in a node is given. For example, in the initial node, Node 1, 50% of the compounds are positive. The splitting rules are given in Table 2. Consider the first split,

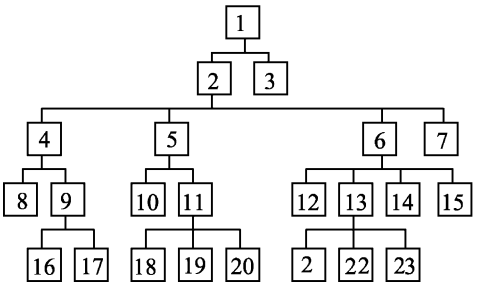

Fig. 2. Tree.

Node 1 splitting into Nodes 2 and 3. If a compound has a carbon connected to exactly two other carbons, C(CC), and that carbon is more than three bonds away from an oxygen connected to exactly one nitrogen, O(N), then these compounds go into Node 3 and there

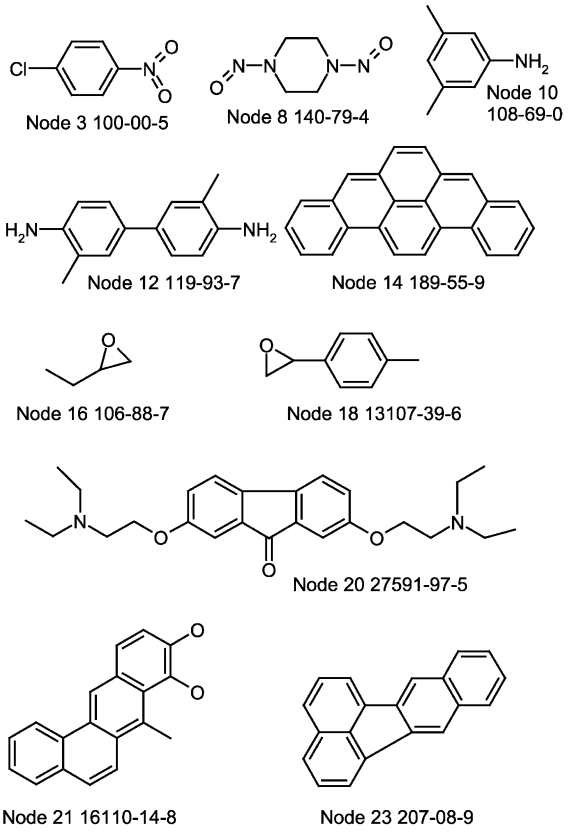

Fig. 3. Representative molecules from selected active nodes. Also given are the CAS numbers for the molecules.

Table 2
For each node, we give the number of compounds in the node and the percent that are Ames test positive

Node	*N*	Percent positive	Augmented atoms	Distance	Typical compound
1	2018	50			
2	1723	43	C(CC)−O(N)	$X \le 3$ or ?	
3	295	90	C(CC)−O(N)	$3 < X$	100-00-5
4	1258	36	C(CCC)−C(CCC)	$X \le 1$ or ?	
5	236	50	C(CCC)−C(CCC)	$1 < X \le 4$	
6	180	82	C(CCC)−C(CCC)	$5 < X \le 7$	
7	49	24	C(CCC)−C(CCC)	$7 < X$	
8	29	100	N(NO)−O(N)	any	140-79-4
9	1229	35	N(NO)−O(N)	?	
10	84	82	C(CC)−C(CCN)	$X \le 2$	108-69-0
11	152	33	C(CC)−C(CCN)	$2 < X$ or ?	
12	20	70	C(CC)−C(CCC)	$X \le 6$	119-93-7
13	90	97	C(CC)−C(CCC)	$6 < X \le 7$	
14	59	75	C(CC)−C(CCC)	$7 < X \le 9$	189-55-9
15	11	20	C(CC)−C(CCC)	$9 < X$ or ?	
16	64	86	C(CCO)−C(CO)	$X \le 1$	106-88-7
17	1165	32	C(CCO)−C(CO)	$1 < X$ or ?	
18	7	100	C(CCO)−C(CO)	$X \le 1$	13107-39-6
19	135	24	C(CCO)−C(CO)	$1 < X \le 6$ or ?	
20	10	100	C(CCO)−C(CO)	$6 < X$	27591-97-5
21	20	100	C(C)−O(C)	$X \le 9$	16100-14-8
22	3	0	C(C)−O(C)	$9 < X \le 11$	
23	67	100	C(C)−O(C)	$11 < X$ or ?	207-08-9

Also given are the augmented atoms and the number of bonds between them. Finally, selected compounds are noted that reside in the node.

is a 90% chance that the compound is a mutagen. The compounds where that distance is less than or equal to three or where one or both of the two AAs do not occur, the (?) group, go into Node 2. Also given in Table 2 are the CAS numbers of representative compounds from a number of the active terminal nodes. These compounds are displayed in Fig. 3. Fig. 1 gives an example of a compound that satisfies the split rule for the splitting of Node 1 into Nodes 2 and 3. There is a carbon attached to two carbons, C(CC), four bonds away from an oxygen attached to a nitrogen, O(N). The focus atoms are circled in the molecule, CAS number 100-00-5. Representative compounds from active terminal nodes are noted in Table 2 and are given in Fig. 3.

We present only four levels of the recursive partitioning tree here. In the full analysis, there are several additional splits. Representative compounds from additional positive terminal nodes are given in Fig. 4. Again, the focus atoms are noted with circles.

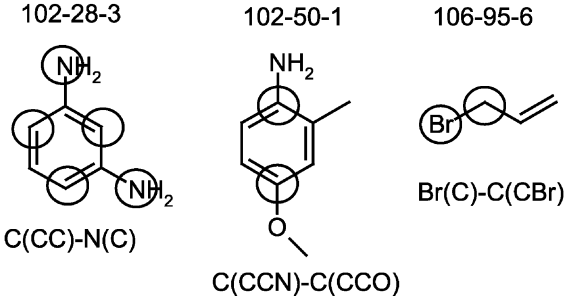

Fig. 4. Active compounds identified in additional terminal nodes.

5. Discussion

The key problem to be overcome in the analysis of heterogeneous data sets is that there are likely to be multiple, biological mechanisms. In the case of mutagenicity, four strains of bacteria are used and the compounds are either activated or not with a liver

enzyme preparation to mimic metabolic transformation of the starting compound. There are many steps in the replication of DNA. It is expected that a diverse set of compounds would include multiple modes of actions. Some would act through one mechanism and others by different mechanisms.

Most statistical methods assume that there is one underlying model of a single process. If there are several processes that lead to mutations, e.g. the compound fitting between two bases of DNA or the alteration of protein kinases, then the features important for one process are very unlikely to be important for another. Most statistical methods, e.g. linear regression, will average the effect for each feature over the multiple processes. Statistical predictions are likely to be compromised and could be entirely misleading.

Recursive partitioning is a simple statistical method that is capable of dealing with multiple mechanisms. A molecular feature is identified that is associated with active compounds and the data split based upon this feature. Compounds with the feature go to one daughter node; those without to another. If the feature is important for a specific mechanism, then compounds with that feature are separated out from the main body of the data. At this point, the analysis in this arm of the evolving tree is limited to just these compounds; other compounds in the data set have no affect on the subsequent analysis. In this manner, multiple mechanisms can be split off and identified.

Two augmented atoms when combined with a bond distance give a very flexible structural descriptors. The segmentation algorithm used for splitting a node adds to the search flexibility because acceptable distance ranges can be identified. It is interesting to note that when the bond distance is one, the two AAs then are confined to a small neighborhood and are capable of describing special functionality. We have noted the epoxide functionality. The AA Br(C)-1-C(CBr) defines a bromine attached to a CH_2. This Br is typically labile making the compound containing such a bromine reactive.

The tree given in Fig. 2 is only one of many statistically valid trees. It is termed the "greedy" tree in that at each point where a choice of splits is possible the split with the smallest adjusted *p*-value is made (also to conserve space, only the first four levels of splits are shown). Consider the first split, Node 1 going to Nodes 2 and 3. There are a large number of statistically valid split variables. Many of these are correlated. For example, eight of the top 10 variables involve oxygen connected to nitrogen. One is O(N)—O(N) with a bond distance of two, obviously NO_2. The combination of O(N) with C(CC) with a bond distance greater than three selects more compounds, 295 versus 1723, and has a smaller adjusted *p*-value. The O(N)—C(CC) rule is not structurally pure however in that it also includes nitrosoamines, NN=O, as well as nitro groups, NO_2. There are other splits that involve different structural features. It is likely that compounds with other features are acting by other mechanisms. The split of Node 2 into four nodes starts the splitting out of large planar compounds; Nodes 16 and 18 include compounds with a three member ring containing oxygen, C(CCO)—C(CO) with a bond distance of one. This is the epoxide functional group and it is typically quite reactive.

The compounds in Nodes 18 and 20, all 17 positive, typically contain three positive rules, C(CCC)-2, 3 or 4-C(CCC), C(CC)-2<X-C(CCN) and C(CCO)—C(CO). The build up of several rules can define a complex molecular substructure. We note that compounds containing the first two rules, Node 11, were 33% Ames test positive. Only when the third rule was added was the incidence of mutagenicity high. This points out that the SAR rules can be complex and indicates that simple structural alerts based upon only a few atoms could be misleading.

The segmentation algorithm used in ChemTree includes a feature call "predictive missing." Remember that many of the compounds examined for a particular pair of AAs will not contain the two AAs. One or both of the two AAs will be missing. The computer algorithm notes this fact. It is possible that if a compound does not have a particular feature that that information tells us something about the biological action of the compound. Examine Node 4 which is split into Nodes 8 and 9. Molecules without N(NO)—O(N) are less likely to be Ames test positive than those containing that AA pair. The segmentation algorithm forms a group of compounds where the AAs under consideration for a split are missing and then either leaves those molecules in their own group, as is the case for the splitting of Node 4, or groups those molecules with molecules with a similar level of biological action (see Node 2).

The analysis search algorithm examines each variable, pair of AAs, in turn and examines each possible segmentation with up to 10 cuts. This is very computation-intensive. A *p*-value is computed for each segmentation and the segmentation with the smallest *p*-value is saved. Next, for each variable, a *p*-value is computed. These *p*-values are adjusted for multiple testing. When many statistical tests are computed, it is important to adjust the statistical procedure so that the probability of a false positive result is controlled [10]. Table 3 gives the AA pairs, the number of compounds in the active node and the percent positive for the 10 possible splits with the smallest adjusted *p*-values for the splitting of Node 1. One AA is always O(N) or, if marked with an *, N(COO), which is a molecular synonym for NO_2. Usually, but not always, these notations are for a NO_2 group. The number of compounds covered, selected out, by these rules varies from 175 to 307. The percent positive varies from 85% to 91%. There are very many compounds in common to all these groups. The algorithm is selecting among correlated variables. Some of these rules in combination effectively select compounds that have the nitro group. It is interesting to note that rules including C(CC) give better coverage and higher positive rates. This suggests that the simple NO_2 structural alert can be improved upon. It also points out that there are many rules that are roughly equivalent. So is it the O of the NO_2 and the C of the C(CC) or the N of the NO_2 and the C that is attached to the NO_2? These methods do not distinguish and the data may allow such a distinction.

Table 3
First 10 valid splits possible at a Node 1 where the first AA is O(N) or *N(COO)

2nd AA	*N*	Percent positive
(1) C(CC)hi	295	90
(2) C(CC)lo	307	89
(3) O(N)lo	249	86
(4) O(C)	175	91
(5) N(COO)	237	86
(6) C(CC)lo *	212	87
(7) C(CC)hi *	203	88
(8) C(CCN)	228	85
(9) O(N)hi	249	86
(10) C(CCN) *	193	85

An important problem with the analysis of biological data is that assay results for individual compounds may be in error; in the case of Ames test data, experts examine the data and make an expert judgment on the "call." Tested compounds may not be pure and impurities may be responsible for the positive result. In the case of negatives in a positive node, there is the possibility that the compound is unstable and hence was not really tested. Recursive partitioning does not depend upon a single-assay value. Recursive partitioning is driven by averages of compounds with a specific feature and averages are much more stable than single assay values. The node average is the average of all the compounds that have the features that lead compounds to that node. Because the recursive partitioning process is driven by averages, the derived structure activity rules can have great statistical validity; *p*-values less than 10^{-54} were seen in this study.

A great deal of effort has been expended to make this code fast—essential for exploratory analysis. ChemTree made each split in few seconds for this data set, where there are 2018 compounds and 1714 AA variables. This speed has proven to be very useful. Obviously, time is money so completing an analysis quickly can help understand a complex data set. Just as important is that the speed can be used to explore alternative analyses. Medicinal chemists and biologists can interact with the data in real time increasing the likelihood that alternatives are considered and good decisions are made. The statistical methods are rigorous, e.g. *p*-values are adjusted for multiple testing, Miller [11], and help keep the exploratory analysis soundly based. For this data set the adjusted p-values ranged from 5.8×10^{-3} to 1.3×10^{-54} and were sufficiently small to be considered "real."

Augmented atoms could be criticized as too simple to be of use for structure activity determination. It is clear that binding into a protein is a three dimensional process; optical isomers often have very different effects. Knowledge of the binding conformation would seem to be essential for good SAR determination. Theoretical considerations would argue against the use of augmented atom descriptors. It is clear for this data set that AAs when combined with topological distance do capture sufficient structural information to reproduce or improve the SAR rules that have

appeared in the literature [12,13]. Our empirical results demonstrate that these simple descriptors, coupled with segmentation recursive partitioning, are effective in building simple and useful structure–activity models.

6. Conclusions

The combination of augmented atom descriptors and segmentation recursive partitioning found prediction rules for mutagenicity, deconvoluted a mixed data set.

References

[1] G.W. Adamson, M.F. Lynch, W.G. Town, J. Chem. Soc. C 22 (1971) 3702–3706.
[2] K.C. Chu, R.J. Feldman, M.B. Shapiro, G.F. Hazard Jr., R.J. Geran, J. Med. Chem. 18 (1975) 539–545.
[3] J.A. Morgan, J.N. Sonquist, J. Am. Stat. Assoc. 58 (1963) 415–434.
[4] D.M. Hawkins, G.V. Kass, in: D.M. Hawkins (Ed.), Topics in Applied Multivariate Analysis, Cambridge Univ. Press, Cambridge, UK, 1982, pp. 269–302.
[5] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth, New York, NY, 1984.
[6] J.R. Quinlin, C4.5 Programs for Machine Learning, Morgan Kaufmann Publisher, San Francisco, CA, 1992.
[7] D.M. Hawkins, S.S. Young, A. Rusinko III, Quant. Struct.-Act. Relat. 16 (1997) 1–7.
[8] A. Rusinko III, M.W. Farmen, C.G. Lambert, P.L. Brown, S.S. Young, J. Chem. Inf. Comp Sci. 39 (1999) 1017–1026.
[9] ChemTree™, The smart HTS solution. Golden Helix, 2000. www.goldenhelix.com.
[10] Y.-C. Yao, Estimating the number of change points via Schwartz's criterion, Stat. Probab. Lett. 6 (1988) 181–189.
[11] R.G. Miller, Simultaneous Statistical Inference, Springer-Verlag, New York, NY, 1981.
[12] J. Ashby, R.W. Tennant, Mutat. Res. 204 (1988) 17–115.
[13] J. Ashby, R.W. Tennant, Mutat. Res. 207 (1991) 229–306.