Cluster Analysis of Acrylates To Guide Sampling for Toxicity Testing

RICHARD G. LAWSON and PETER C. JURS*

152 Davey Laboratory, Chemistry Department, Pennsylvania State University, University Park, Pennsylvania 16802

Received October 10, 1989

A set of 143 acrylates drawn from the TSCA inventory have been investigated for structurally defined clusters of compounds to simplify sampling for future toxicity screening. Each acrylate was represented by eight descriptors calculated from the molecular structure. Several standard clustering methods have been used to find five natural clusters of compounds. These five clusters are largely populated by compounds with similar chemical attributes with separate clusters formed for compounds with high absolute partial atomic charges, hydrophobic compounds, small compounds, halogenated compounds, and large or oligomeric compounds.

INTRODUCTION

Acrylates are an important class of polymers used in a wide variety of consumer products. Unfortunately, there is concern about the safety of a few of these materials as monomers.

basic acrylate structure

Some reports in the literature cite potential toxicity problems, and others raise questions about carcinogenic potential in animals.[1-4] Because of these questions the U.S. EPA is required to treat acrylates as potential carcinogens when new compounds are being considered for introduction into the marketplace.

There are more than 200 acrylates currently on the TSCA inventory. Since there are so many compounds in the acrylate family, it would be impractical to test every one for biological activity. There must be some way, however, through which researchers can assess the risks of the acrylates that exist now and the risks of new compounds. If the mechanism of toxicity or carcinogenicity for acrylates were understood thoroughly enough to allow accurate a priori predictions of hazardous compounds, this knowledge could be used directly. Research in this direction is proceeding, but it is not presently possible to even estimate the biological activity of untested compounds. Considering the structural diversity in the set of acrylates that we considered in this study alone, it will likely be some time before this goal is reached.

In the meantime, since it is important to know now which compounds are dangerous, or are at least likely to be dangerous, another strategy must be taken. If we cannot identify the actual relationships between the physical and chemical properties of acrylates and their biological activity, we can seek empirical relationships that bridge this gap. This is the foundation of structure–activity relationship (SAR) studies.[7-9] Few acrylates have known biological activity, however, so there is not yet enough information to formulate even an empirical equation. To pursue this structure–activity relationship strategy, the compounds under consideration would have to first be divided into groups of related physical and chemical properties, representative members could then be selected from these subsets for biological testing, and finally the connection between the physicochemical properties of the groups and their activity could be sought. The goal of this paper is thus to provide a means of sampling from the set of all acrylates in such a way that toxicity data can be gathered in an objective and efficient way.

There are many obvious ways to divide acrylates into groups: halogenated vs nonhalogenated, sulfonated vs nonsulfonated, oligomers vs simple monomers, by molecular weight, etc. These possibilities are all reasonable, but unless they are considered in a systematic way, some of the groups may be inconsistent or may overlap with other groups. Another consideration is that to derive groupings that are most likely to be related to biological activity, the discriminating features must be chosen as objectively as possible, from a pool of features that are intended to encode the reactivity of the compounds.

The general approach of using cluster analysis as an objective tool to divide compounds into meaningful groups of related compounds so that representative elements could be sampled for screening has been described by Hodes and Willett.[5,6] Willett applied cluster analysis to approximately 8500 compounds in a commercial inventory to provide an automated means of choosing trial compounds for activity screening. The compounds were separated on the basis of atom and bond-centered fragments. Hodes' work focused on a 4980-member subset of the 232 000 compounds in an National Cancer Institute's inventory. His objective was to derive natural groups of compounds from which test compounds could be selected for cell culture screening, again on the basis of molecular fragments. The reports of these projects focused mainly on the mathematical and computer aspects of the projects, but provided evidence for the validity and value of such strategies. This paper employs the same basic strategy, but will focus more on the chemical aspects of the clustering, especially the use of chemically meaningful descriptions of the compounds.

Our search for an objective clustering of the acrylates was done with the ADAPT software system. ADAPT has been described previously in the literature.[7,10] Briefly, the system is composed of more than 90 independent software modules that allow the user to enter, model, and store structures in computer disk files and then to calculate their geometric, electronic, physicochemical, and topological features in addition to simple fragments. The compounds can then be studied with multivariate statistical or pattern recognition methods. In the present study, clustering programs were employed. These programs partition the compounds into related groups based on the calculated molecular structure descriptors.

THE DATA SET

The 143 acrylates selected for this study were found via a CAS ONLINE search done with a profile designed to retrieve acrylates of commercial importance but to eliminate salts, metal-containing species, adducts, etc., from the list since these pose problems in structure entry and in calculations. A set of over 200 acrylates was originally obtained, but some compounds had to be excluded because they had too many atoms or atoms unsupported in the ADAPT software; 143 acrylates

0095-2338/90/1630-0137$02.50/0

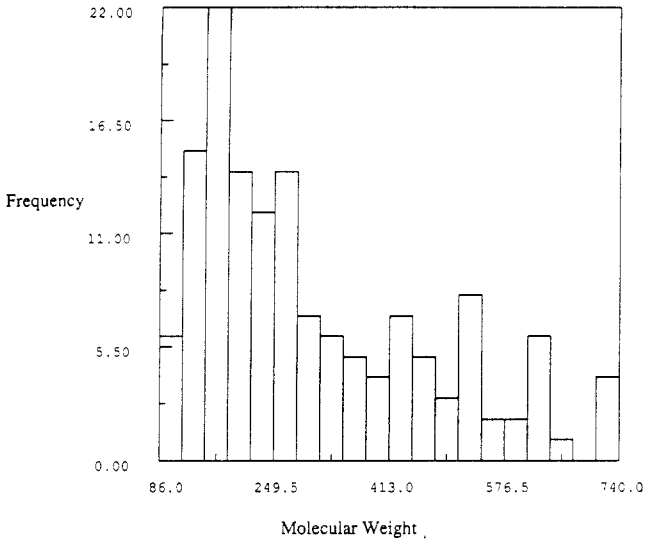

Figure 1. Histogram of the frequency of molecular weights for all of the acrylates.

remained after these restrictions. Substantial structural diversity is obvious in the data set with 40 halogenated compounds, 24 containing sulfur, and 37 containing nitrogen. The nominal molecular weights of the 143 compounds range from 86 to 740 amu, as shown in Figure 1. Many of the molecules have long alkyl side chains, which leads to great conformational mobility.

DESCRIPTOR GENERATION

Any structure–activity relationship is based on terms that describe the properties of the molecules accounting for activity, direct or indirectly. For example, to describe the solubility of a class of compounds empirically, one might use the polarity, density, and size of the compounds. Since it is not as clear which acrylate properties are related to toxicology, it is desirable to have a large number of features that could be related to reactivity, distribution in living systems, etc. For the 143 acrylates of interest, not only is their toxicity unknown, but there is no comprehensive body of data for any property. Therefore, it was necessary to calculate or estimate properties to describe the electronic, geometric, and physicochemical features of the compounds. Electronic features can be used to explain reactivity; the more highly charged a compound is, the more reactive it would be in general. The differences in charges on a particular atom, or atoms located within a substructure, for a set of compounds could possibly explain differences in their reactivity. In a similar way, the differences in the size or shape of molecules as measured by geometric descriptors could account for differences in their access to an active site or in their rate of transport through living systems.

The high degree of conformational flexibility in these compounds prohibited the use of geometric descriptors, however. To test the sensitivity of geometric descriptors to conformational mobility, we calculated the moments of inertia and other parameters dependent on conformation for several energetically reasonable conformations of several structurally dissimilar molecules. The range of descriptor values found was too great for these features to be considered for further use.

Electronic descriptors have been useful in past SAR studies.[12,13] However, the common ones—Hückel molecular orbital calculations and extended Hückel molecular orbital calculations—could not be used since the compounds were for the most part largely saturated, which ruled out simple HMO treatment, and they could not be modeled confidently, which ruled out extended HMO treatment. More sophisticated electronic descriptors could have been applied if that were

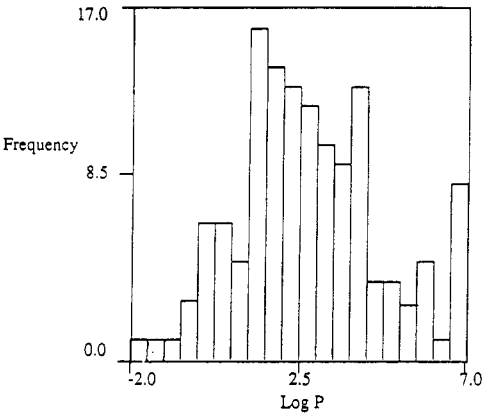

Figure 2. Histogram of the frequency of calculated log P values for the acrylates with a maximum value of 7.0.

necessary, but with 143 compounds, simple descriptors requiring little CPU time for calculation were favored. The results obtained below indicate that the simple descriptors were adequate. Some simple semiempirical calculations of σ charge density developed by Del Re are independent of geometry, however, so these were calculated.[16,17] These are based largely on dipole moments and polarizability.

Measured or estimated physicochemical descriptors such as log P (the logarithm of the partition coefficient between n-octanol and water) or molecular polarizability have been found to be related to biological activity in the past.[16,17] The few acrylate compounds that have partition coefficients reported in the literature have a wide range of values, especially considering the logarithmic scale.[16-19] For consistency and simplicity we calculated log P values using the CLOGP 3.5 program.[20] These values are displayed in Figure 2. Some of the acrylates strained the limits of the program's parameters, however, resulting in unrealistically high values (as high as 18). Because such values are unrealistic, and to prevent these values from having too much influence on later model development, we truncated the log P values for this study at 7.0.

Topological descriptors are based only on the identities of the atoms and their connections to each other and are independent of geometry. Simple counts of atoms or bonds can be used, or more complex graph theoretical molecular connectivity indices can be calculated.[21] The molecular connectivity indices provide estimates of the degree of branching, or mathematical measures of the size and shape of molecules that are especially useful for compounds that cannot be modeled confidently.

More than 60 descriptors of these types were calculated. Before clustering was begun, the variables were evaluated to determine which ones could help discriminate between different groups of compounds. For example, if the minimum σ charge for all compounds were identical, that feature would not help separate the acrylates. On the basis of these considerations, descriptors with a majority of zero values, or identical nonzero values, would be discarded. With the acrylate data set approximately 20 compound variables were eliminated on the basis of these considerations. The number of descriptors required to adequately describe the variance of the data was estimated by using principal components analysis.[22] Although nearly 40 variables were considered, only 5 variables were required to account for 95% of the variance, and 8 variables accounted for 98% of the variance. Examination of the eigenvalues to determine the variables accounting for the majority of the variance was difficult with so many variables. An alternative method to choose highly orthogonal variables is to use vector space analysis. This involves selecting a variable at random as a basis vector and adding the variable from the remaining set that is most orthogonal to the basis vector. These

	1	2	3	4	5	6	7	8
1	1.00							
2	0.47	1.00						
3	0.09	0.13	1.00					
4	0.47	0.58	-0.19	1.00				
5	0.16	0.34	0.43	0.06	1.00			
6	0.19	0.48	0.68	0.15	0.60	1.00		
7	0.22	0.01	0.47	0.14	0.31	0.47	1.00	
8	-0.24	0.40	0.31	0.23	0.27	0.52	0.14	1.00

Figure 3. Pairwise correlation matrix of the eight variables selected for clustering on the basis of variance.

Table I. Set of Eight Structural Descriptors for the Acrylates

1	SSS 2	count of SS-2 (acrylate)
2	CHIS 6	$^2\chi^v$: path 2 valence-corrected molecular connectivity
3	MOLC 7	$^3\chi_C$: cluster 3 molecular connectivity index
4	KAPA 3	$^3\kappa$: topological index measuring the importance of midchain branching
5	PATH 2	path environment for SS-2 (acrylate)
6	ALLP 1	number of molecular paths
7	TSCH 1	sum of absolute values of all σ charges
8	CLGP 0	calculated log P

two vectors form a plane. The remaining variables are added one at a time on the basis of the plane angle that they form in the space of increasing dimensionality. Repeating this procedure with a number of different basis vectors leads to a stable set of orthogonal variables which is generally the same as the set that would have been determined through principal components analysis. Highly orthogonal, i.e., minimally correlated, descriptors are desirable because they contain information that is not encoded by other descriptors. Since standard methods for feature selection in cluster analysis have not been established, selecting variables on the basis of maximum variance, or maximum information content, seemed to be a reasonable approach.

The set of eight structural descriptors found to support clustering on the basis of these criteria is shown in Table I. The number of molecular paths (ALLP 1) is a topological descriptor based on the hydrogen-suppressed graph. The path environment descriptor PATH 2 describes the steric surroundings of the acrylate substructure as imbedded within the molecules. Only paths of lengths up to 10 were included in the computation. TSCH 1 is a whole-molecule descriptor codifying degree of departure from electronic neutrality. It is the total absolute value of the σ charges in the molecule. MOLC 7 is a molecular connectivity index based on atoms with three bonds attached to them (cluster 3). CLGP 0 is the calculated value of log P from the Pomona Medicinal Chemistry project software.[20] SSS 2 is the count of the number of acrylate groups present in each molecule. CHIS 6 is the $^2\chi^v$ path 2 valence-corrected molecular connectivity index. KAPA 3 is $^3\kappa$, a measure of the shape of the molecular structure derived from the connection table

$$^3\kappa = 4(^3P_{max})(^3P_{min})/(^3P_i)$$

where $^3P_{max}$ is the maximum number of paths of length 3 for a molecule of the same size, $^3P_{min}$ is the minimum number of paths of length 3 for a molecule of the same size, and ^3P_i is the actual number of paths of length 3 in the molecule.[23] This index is reported to measure the presence of midchain branching in a molecular structure. Thus, each acrylate compound is represented by a point in this eight-dimensional space where each axis is associated with one of the structural descriptors. The problem of clustering of the acrylates can be stated as seeking natural clusters of points in this eight-dimensional feature space.

Table II. Natural Number of Clusters

clusters	iterations	mean SI	max SI	min SI
4	11	0.80	1.00	0.69
5	11	0.85	1.00	0.77
6	11	0.85	0.99	0.66
7	11	0.76	0.97	0.61

Some of the variables chosen are topological, and some are not continuous. The count of acrylate substructures is a discrete variable. Although discrete variables are not as commonly used in Euclidean space as continuous variables, their use can be just as valid. In the case of the count of acrylate substructures, which can only take on the values 1, 2, 3, 4, or 5, Euclidean distances are still meaningful. A compound with a value of 4 has twice as many acrylate substructures as a compound with a value of 2. The difference between these compounds is the same as the difference between a compound with a value of 5 and a compound with a value of 3. The more abstract topological indices are just as valid in Euclidean space for the same reasons and have been used in many other studies.[24-27]

CLUSTERING TENDENCY

Once a set of descriptors has been chosen, it is important to determine that the data are not distributed randomly throughout the space. With randomly distributed data any clustering results will certainly be artifacts. Hopkins' statistic has been used to test whether a spacial distribution shows aggregation or randomness.[28,29] The Hopkins' statistic is defined as

$$H = \frac{\sum U_j}{\sum U_j + \sum W_j}$$

where the U_j are distances from randomly selected locations in the sample space to their nearest neighbor points and W_j are distances from randomly selected data points to their nearest neighbors. The number of random locations and the number of data points selected are the same and should be 5–10% of the number of points in the data set.[29] This statistic compares the nearest-neighbor distance distribution of randomly selected locations to that for the randomly selected patterns. Values of H close to 1.0 indicate clustering tendency, and values of H close to 0.5 indicate random placement of the points in space. The probability of a given value of H arising due to chance can be computed as well. Thus, the Hopkins' statistic is easily calculated and is appealing in its simplicity and directness.

The Hopkins' statistic was computed 10 times, with 14 sampling points each time, for the acrylate data set. Mean values of 0.76–0.86 were found, depending on the type of descriptor scaling used (raw descriptor values as opposed as autoscaled desriptors).[30] These values indicate very strongly that the data set is clustered and justify proceeding with cluster analysis. The probability of a random data set giving H values in this range is only approximately 5%. While this computation provides very strong evidence that the data set is clustered, it says nothing about the populations, memberships, or numbers of clusters. These were determined by partitional clustering methods.

CLUSTERING STUDIES

Many clustering programs have been described in the literature, so it can be difficult to decide which program is best for a specific problem. A clustering method was selected for this study by using some simple and practical guidelines. First, with the relatively large number of compounds, and eight features associated with each of them, hierarchical methods, such as single-linkage or Ward's method, would be more

Table III. Acrylate Clustering Results

CAS no.	name	CAS no.	name
	Cluster 1, 8 Compounds		
000307982	1,1-dihydroperfluorooctyl acrylate	034395249	2-(perfluorododecyl)ethyl acrylate
017527296	2-(perfluorohexyl)ethyl acrylate	052723963	$C_{21}H_{26}N_2O_8$ diacrylate
017741605	2-(perfluorodecyl)ethyl acrylate	056361558	bisphenol A diethylene glycol diacrylate
027905459	1,1,2,2-tetrahydroperfluorodecyl acrylate	064448686	$C_{25}H_{28}O_8$ diacrylate
	Cluster 2, 13 Compounds		
002156969	decyl acrylate	048076386	eicosyl acrylate
002156970	lauryl acrylate	067952505	(1-methylethylidene)bis[4,1-phenyleneoxy(1-methyl-2,1-ethanediyl)] diacrylate
003076048	tridecyl acrylate		
004813574	octadecyl acrylate (stearyl acrylate)	070146053	(1-methylethylidene)bis[2-methyl-4,1-phenylene)oxy(1-methyl-2,1-ethanediyl)] diacrylate
013048345	decamethylene glycol, diacrylate		
013402023	hexadecyl acrylate	070495395	methylenebis[2,1-phenyleneoxy(1-methyl-2,1-ethanediyl)] diacrylate
013533181	oleyl acrylate		
021643425	tetradecyl acrylate	084732285	$C_{20}H_{25}N_3O_6$
	Cluster 3, 71 Compounds		
000095396	cyclol acrylate	004074888	diethylene glycol diacrylate
000096333	methyl acrylate	005390545	2-nitrobutyl acrylate
000103117	2-ethylhexyl acrylate	007251903	2-butoxyethyl acrylate
000106638	isobutyl acrylate	007328178	diethylene glycol ethyl ether acrylate
000106741	2-ethoxyethyl acrylate	013048334	hexamethylene glycol diacrylate
000106901	glycidyl acrylate	013282821	3-butoxy-2-hydroxypropyl acrylate
000140885	ethyl acrylate	013533056	diethylene glycol, monoacrylate
000141322	*n*-butyl acrylate	016868136	cyclopentyl acrylate
000356865	1,1-dihydroperfluoropropyl acrylate	016969101	3-phenoxy-2-hydroxypropyl acrylate
000407476	2,2,2-trifluoroethyl acrylate	017527310	perfluorobutyl acrylate
000424646	1,1-dihydroperfluorobutyl acrylate	017977092	2,2-dinitropropyl acrylate
000689123	isopropyl acrylate	018526073	3-(dimethylamino)propyl acrylate
000818611	2-hydroxyethyl acrylate	018621766	2-butene-1,4-diol, diacrylate
000925600	*n*-propyl acrylate	018933921	1,3-dimethylbutyl acrylate
000937417	phenyl acrylate	019485031	1,3-butylene glycoldiacrylate
000999553	allyl acrylate	019660163	2,3-dibromopropyl acrylate
001070708	1,4-butylene diacrylate	019721370	bis(ethylene acrylate) monosulfide
001663394	*tert*-butyl acrylate	023916338	2-butenyl acrylate
002160896	hexafluoroisopropyl acrylate	024493536	1,3-proanediol diacrylate
002223827	neopentylglycol diacrylate	024615847	2-carboxyethyl acrylate
002274115	ethylene glycol diacrylate	024910847	2,3-dichloropropyl acrylate
002399486	tetrahydrofurfuryl acrylate	030697406	monoacryloyloxyethyl phthalate
002426542	(diethylamino)ethyl acrylate	037275471	trimethylolpropane diacrylate
002439352	(dimethylamino)ethyl acrylate	044914036	2-methylbutyl acrylate
002478106	4-hydroxybutyl acrylate	048145046	phenoxyethyl acrylate
002495354	benzyl acrylate	051727505	2-hydroxypropyl acrylate
002499583	heptyl acrylate	052591272	nonafluorohexyl acrylate
002499958	hexyl acrylate	052607815	methylcarbamoyloxyethyl acrylate
002664553	nonyl acrylate	063225536	2-acryloyloxyethyl butylcarbamate
002998085	*sec*-butyl acrylate	066028306	*N*,*N*-bis(2-acryloyloxyethyl)formamide
002998234	amyl acrylate	067905082	$C_{14}H_{22}O_4$
003066715	cyclohexyl acrylate	067905413	$C_{14}H_{20}O_4$
003121617	2-methoxyethyl acrylate	067952492	2-methylheptyl acrylate
003326907	2-hydroxy-3-chloropropyl acrylate	068227996	$C_{13}H_{14}F_{11}NO_4S$
003530367	phenylethyl acrylate	068298066	$C_{12}H_{12}F_{11}NO_4S$
003953104	2-ethylbutyl acrylate		
	Cluster 4, 33 Compounds		
000383073	$C_{17}H_{16}F_{17}NO_4S$ monoacrylate	059071102	$C_{14}H_{12}F_{15}NO_4S$ acrylate
000423825	$C_{15}H_{12}F_{17}NO_4S$ monoacrylate	065983315	$C_{15}H_{20}O_3$ acrylate
001492871	$C_{12}H_{14}F_9NO_4S$ monoacrylate	066008682	$C_{17}H_{12}F_{21}NO_4S$ acrylate
001893523	$C_{13}H_{12}F_{13}NO_4S$ monoacrylate	066008693	$C_{15}H_{12}F_{17}NO_4S$ acrylate
003741773	2,4,6-tribromophenyl acrylate	066008706	$C_{13}H_{12}F_{13}NO_4S$ acrylate
005888335	isobornyl acrylate	066671225	neopentyl glycol acrylate benzoate
007347195	2-(2,4,6-tribromophenoxy)ethyl acrylate	066710972	$C_{25}H_{24}Br_4O_6$ acrylate
015419940	4-benzoyl-3-hydroxyphenyl acrylate	067584558	$C_{10}H_{10}F_9NO_4S$ acrylate
016432818	2-(4-benzoyl-3-hydroxyphenoxy) acrylate	067584569	$C_{11}H_{10}F_{11}NO_4S$ acrylate
017329792	$C_{11}H_{12}F_9NO_4S$	067584570	$C_{12}H_{10}F_{13}NO_4S$ acrylate
024447787	2,2-bis(4-acryloxyethoxyphenol)propane diacrylate	068084628	$C_{13}H_{10}F_{15}NO_4S$ acrylate
025268773	*N*-methyl perfluoroactane sulfanamidoethyl acrylate	068227974	$C_{15}H_{14}F_{15}NO_4S$ acrylate
030145518	acryloxypivalyl acryloxypivalate	068227985	$C_{14}H_{14}F_{13}NO_4S$ acrylate
048077958	2-[(perfluorodecyl)sulfonyl)]methylamino ethyl acrylate	068298602	$C_{16}H_{16}F_{15}NO_4S$ acrylate
049859703	$C_{14}H_{14}F_{13}NO_4S$ acrylate	072276052	$C_{18}H_{14}F_{21}NO_4S$ acrylate
054449740	*p*-α,α-dimethylbenzylphenyl acrylate	087320056	$C_{17}H_{26}O_6$ acrylate
058920313	$C_{16}H_{14}F_{17}NO_4S$ acrylate		
	Cluster 5, 18 Compounds		
001680213	triethylene glycol diacrylate	060506812	dipentaerythritol pentaacrylate
003524683	pentaerythritol triacrylate	066028328	$C_{19}H_{27}NO_7$
004986894	pentaerythritol tetraacrylate	066028340	$C_{17}H_{23}NO_7$
005459381	glycerol triacrylate	067892993	$C_{18}H_{32}O_8$
015625895	trimethylolpropane triacrylate	067893009	$C_{18}H_{21}N_3O_6$
017831719	tetraethylene glycol diacrylate	0679054580	2-hydroxy-1,6-hexanediyl diacrylate
019778859	trimethylolethane triacrylate	071412356	$C_{15}H_{24}O_6$
052408421	tetramethylenebis(oxy(2-hydroxytrimethylene) diacrylate	072928428	$C_{15}H_{24}O_8$
053417291	pentaerythritol diacrylate	085412540	$C_{24}H_{32}O_{12}$

computationally demanding than nonhierarchical methods, such as Kmeans or Isodata. In addition, the resultant dendrograms, which can be difficult to interpret or even misleading with much smaller data sets, would be too complex for simple interpretation. Finally, the objective in this study was to find a natural grouping of compounds, not the hierarchy of groupings that hierarchical methods provide. Two commonly used and well-established nonhierarchical partitioning methods, Kmeans and Isodata, were thus used in this study. Since hierarchical methods were not used, they will not be discussed further.

To divide the compounds into reproducible clusters, the first task is to determine the number of clusters that the data naturally comprise. The first step in this procedure involves the clustering program Kmeans.[31] It is initialized with a set number of data points as starting cluster centers, and then it assigns the compounds to those clusters on the basis of their proximity to the respective centers. The cluster centers are updated to the centroid of these newly formed clusters, and the points are reassigned, again on the basis of proximity. This procedure is repeated until there is no change in the cluster populations. The number of clusters is input at the beginning of the program and cannot change.

The simplicity of this Kmeans routine can be exploited by splitting the compounds into a fixed number of clusters repeatedly, with different initial cluster centers from one trial to the next. This procedure can then be repeated for a different number of clusters, so that the range of the expected number of clusters is covered. In general, the number of clusters that is closest to the natural number of clusters will have more similar results from one trial to the next. For example, if the routine divides the compounds into more clusters than they would fall into naturally, it would be forcing similar compounds into different groups. If this is repeated with different starting points, the compounds that are artificially separated will likely be different from those in the previous iteration simply because artificial distinctions are being forced. If this is repeated many times for this number of clusters, the similarity of one clustering trial to the next will be relatively low. A similar argument holds for dividing the compounds into fewer clusters than the natural number. Earlier work with these acrylate compounds suggested that the natural number of clusters would be near 4, so 11 clustering trials were performed for 4, 5, 6, and 7 clusters. Work by Jain and Moreau with artificial data sets allows an estimate of the number of experiments with random starting points that are required to have a specified probability of obtaining the true partitioning.[32] Eleven experiments were found to be necessary to have 99.5% probability of correctly partitioning the data into four clusters, and 95% probability of correctly partitioning the data into five clusters.

The results of comparing the 11 trials pairwise among themselves are shown in Table II. The similarity index presented in this table is a weighted average of a cluster by cluster comparison of the two partitions as shown here: $SI = \sum w_i((\text{cluster}(i)_1 \cap \text{cluster}(i)_2)/(\text{cluster}(i)_1 \cup \text{cluster}(i)_2)$. This index can range from near 0 for completely dissimilar clusters to near 1 for identical clusters. The form of this index relating the intersection of two corresponding clusters to their union was suggested by the form of the similarity index used in atom pairs work.[33] The results in Table II suggest, and statistical tests confirm, that the differences are significant, that five cluster partitions are more similar to each other than either four or seven, and six cluster partitions are also more similar to each other than four or seven. Although these experiments were not able to definitely provide us with the natural number of clusters, we now had evidence that it was either five or six clusters.

To better determine the proper number of clusters, and to refine the results, the clustering program Isodata was used.[34] This is a more sophisticated clustering routine that employs adjustable parameters to increase or decrease the number of clusters within a trial on the basis of the cluster populations, the distance between them, the spread within them, and the expected final number of clusters. Although the user does enter the number of clusters that is expected, the final number of clusters is determined by the chosen parameters and the characteristics of the data set. In this way, with the target number of clusters, reasonable values for the distance between clusters, and the separation within them all determined by the preliminary Kmeans experiments, Isodata was expected to determine the appropriate number of clusters, and the most stable populations. In general, using more than one clustering method on a data set will provide more confidence in the results. Although the algorithms of Kmeans and Isodata are both based on minimizing square error, and thus are not completely independent, using both programs provides more confidence than using either one alone. The use of a clustering program based on a completely different method would have provided even more confidence in the results. Partitional clustering programs other than Kmeans and Isodata were not available at the time of the study, however.

The first experiments with Isodata used six as the target number of clusters. The first 11 iterations were less consistent than anticipated, with a similarity coefficient of just 0.83, and several partitions had five final clusters instead of six. Some of the parameters were adjusted slightly in hopes of obtaining more consistent results with six clusters, but five final clusters were obtained 9 times in 11 trials. This was strong evidence that the natural number of clusters was actually five. In experiments with five as the target number of clusters the results were much more consistent: four clusters were obtained once and six clusters once, but 9 times of 11 there were five final clusters, with an average similarity of 0.87. The average similarity of 0.87 was determined to be statistically significantly different from the mean of 0.83 obtained earlier on the basis of the Mann–Whitney Wilcoxon rank-sum test.[35] This was not only strong evidence that five was the proper number of clusters, but also that these partitions were stable. Of the 11 partitions the one that had the highest average similarity to the other 10 was chosen as the representative clustering. The compounds comprising each of the five clusters are listed in Table III. Thus, the use of the Kmeans and Isodata clustering routines provided us with the memberships of the five clusters.

These clusters of acrylate compounds can be characterized in general terms based on the distribution of some of their descriptor values. Cluster one is comprised of only eight highly charged compounds. The minimum total σ charge in this group is larger than that for all except for 6 of the remaining 135 compounds. The average total σ charge within the group is 6.55, which is more than 3 standard deviations higher than the mean total σ charge value for the entire population (2.24 $\pm$ 1.4). Five of the compounds are monomeric and perfluorinated, and the remaining three are dimers. All are high molecular weight compounds.

Cluster 2 is made up of 13 hydrophobic compounds. All the log P values within this group are greater than 5.0, with seven compounds having the maximum value of 7.0. The average log P value for the entire population is 3.0. There are nine monomers and four dimers in this group.

The large cluster of 71 compounds comprising cluster 3 contains the simple acrylates; 57 are monomeric, 13 dimeric, and 1 trimeric. The few nonmonomers in this group have lower path and molecular connectivity values than average. All of the compounds have rather low total σ charge values; 65 of

Table IV. Descriptor Values for All Acrylates

no.	CAS no.	1	2	3	4	5	6	7	8
1	95396	1	3.874	0.517	2.083	83	231	1.043	2.61
2	96333	1	0.727	0.048	3.000	21	21	0.742	0.75
3	103117	1	3.251	0.252	7.101	38	91	1.293	4.32
4	106638	1	2.228	0.456	5.878	28	45	0.994	2.20
5	106741	1	1.569	0.048	7.000	30	55	1.191	1.09
6	106901	1	1.790	0.166	3.556	34	60	1.151	−0.13
7	140885	1	0.956	0.048	3.840	22	28	0.850	1.28
8	141322	1	1.725	0.048	5.878	28	45	0.995	2.33
9	307982	1	1.232	4.396	84.000	406	6	3.555	0.00
10	356865	1	0.451	0.402	4.152	40	91	2.693	2.38
11	383073	1	2.238	1.649	6.596	118	820	2.247	6.05
12	407476	1	0.500	0.227	7.000	31	55	1.875	2.04
13	423825	1	1.477	1.649	5.857	113	741	2.247	4.99
14	424646	1	0.327	0.568	3.769	48	136	3.516	2.62
15	689123	1	1.690	0.284	5.000	27	36	1.010	1.58
16	818611	1	1.095	0.048	5.000	25	36	1.256	−0.06
17	925600	1	1.372	0.048	5.000	25	36	0.920	1.80
18	937417	1	1.938	0.116	3.265	51	106	0.988	2.06
19	999553	1	1.123	0.048	5.000	25	36	0.808	1.26
20	1070708	2	2.450	0.096	9.373	42	105	1.607	1.86
21	1492871	1	2.524	1.035	5.538	79	378	2.247	3.47
22	1663394	1	2.738	1.160	5.878	28	45	1.171	1.98
23	1680213	2	2.968	0.096	13.290	54	171	2.188	1.38
24	1893523	1	1.724	1.317	5.327	95	528	2.247	3.98
25	2156969	1	3.846	0.048	11.930	45	120	1.444	5.51
26	2156970	1	4.554	0.048	13.940	51	153	1.594	6.57
27	2160896	1	0.484	0.408	4.889	45	105	3.061	3.10
28	2223827	2	3.788	1.303	7.875	45	120	1.755	2.13
29	2274115	2	1.743	0.096	7.438	37	78	1.505	1.25
30	2399486	1	2.469	0.166	3.265	46	100	1.255	0.99
31	2426542	1	2.398	0.206	6.250	37	78	1.244	1.55
32	2439352	1	2.175	0.364	7.000	31	55	1.107	0.65
33	2478106	1	1.802	0.048	7.000	30	55	1.356	0.35
34	2495354	1	2.297	0.166	4.000	52	123	0.912	2.52
35	2499583	1	2.786	0.048	9.000	36	78	1.219	3.92
36	2499958	1	2.432	0.048	7.901	33	66	1.144	3.39
37	2664553	1	3.493	0.048	11.000	42	105	1.369	4.98
38	2998085	1	1.860	0.215	4.500	28	45	1.081	2.11
39	2998234	1	2.079	0.048	7.000	30	55	1.070	2.86
40	3066715	1	2.901	0.166	3.265	51	106	1.143	2.78
41	3076048	1	4.907	0.048	15.000	54	171	1.669	7.00
42	3121617	1	1.364	0.048	5.878	28	45	1.083	0.56
43	3326907	1	1.979	0.177	5.531	31	55	1.498	0.31
44	3524683	3	4.094	0.85	8.889	64	231	2.788	−0.34
45	3530367	1	2.607	0.166	4.688	55	141	0.977	2.84
46	3741773	1	4.982	0.992	3.273	69	178	2.247	4.33
47	3953104	1	2.517	0.252	5.289	34	66	1.143	3.26
48	4074888	2	2.356	0.096	10.290	45	120	1.846	1.32
49	4813574	1	6.675	0.048	19.950	69	276	2.044	7.00
50	4986894	4	4.747	0.900	11.340	75	325	3.036	1.18
51	5390545	1	2.141	0.209	5.325	37	78	1.070	1.77
52	5459381	3	2.931	0.262	9.600	54	171	2.315	1.54
53	5888335	1	5.740	1.730	1.473	115	333	1.426	4.09
54	7251903	1	2.338	0.048	9.000	36	78	1.336	2.15
57	7328178	1	2.181	0.048	9.917	39	91	1.532	1.16
56	7347195	1	5.576	0.992	5.018	76	250	2.247	4.77
57	13048334	2	3.157	0.096	11.320	48	136	1.756	2.92
58	13048345	2	4.572	0.096	15.260	60	210	2.056	5.03
59	13282821	1	2.916	0.177	9.373	42	105	1.899	1.33
60	13402023	1	5.968	0.048	17.950	63	231	1.894	7.00
61	13533056	1	1.708	0.048	7.901	33	66	1.598	0.01
62	13533181	1	6.309	0.048	19.950	69	276	1.954	7.00
63	15419940	1	3.999	0.387	4.496	133	488	1.987	3.87
64	15625895	3	4.289	0.851	8.889	64	231	2.465	2.18
65	16432818	1	4.594	0.387	6.094	142	626	2.328	4.31
66	16868136	1	2.547	0.166	2.651	45	85	1.068	2.22
67	16969101	1	3.112	0.245	6.370	63	201	1.892	1.67
68	17329792	1	0.986	4.899	77.000	351	2	4.054	0.00
69	17527296	1	0.266	1.066	4.543	78	351	5.992	2.95
70	17527310	1	0.761	0.402	4.889	42	105	2.696	2.01
71	17741605	1	−0.228	1.729	5.600	114	741	9.286	3.89
72	17831719	2	3.580	0.096	16.200	63	231	2.529	1.45
73	17977092	1	2.525	0.588	3.960	43	105	1.925	0.67
74	18526073	1	2.528	0.364	7.901	34	66	1.172	0.31
75	18621766	2	2.140	0.096	9.373	42	105	1.489	1.31
76	18933921	1	3.081	0.623	6.400	34	66	1.232	3.04
77	19485031	2	2.612	0.263	8.082	43	105	1.698	1.64

Table IV (Continued)

no.	CAS no.	1	2	3	4	5	6	7	8
78	19660163	1	3.553	0.614	5.531	31	55	2.247	2.14
79	19721370	2	2.356	0.096	10.290	45	120	1.702	1.81
80	19778859	3	4.246	1.071	9.562	61	210	2.394	1.65
81	21643425	1	5.261	0.048	15.940	57	190	1.744	7.00
82	23916338	1	1.403	0.048	5.878	28	45	0.889	1.79
83	24447787	2	7.114	1.149	9.408	238	1217	2.857	5.87
84	24493536	2	2.097	0.096	8.333	39	91	1.538	1.33
85	24615847	1	1.515	0.113	7.000	31	55	1.645	0.33
86	24910847	1	2.597	0.375	5.531	31	55	1.168	1.86
87	25268773	1	1.323	1.698	5.627	111	703	2.247	3.92
88	27905459	1	0.019	1.398	5.035	9	528	7.639	3.42
89	30145518	2	6.248	2.302	9.694	65	253	2.704	2.90
90	30697406	1	3.358	0.280	5.878	81	300	2.454	1.70
91	34395249	1	2.061	6.201	132.00	990	10	4.359	0.00
92	37275471	2	3.635	0.803	6.479	51	153	2.217	0.67
93	44914036	1	2.348	0.337	5.531	31	55	1.068	2.73
94	48076386	1	7.382	0.048	21.960	75	325	2.194	7.00
95	48077958	1	1.980	1.803	6.980	117	780	2.247	2.82
96	48145046	1	2.533	0.116	5.481	58	160	1.329	2.50
97	49859703	1	2.227	1.471	6.612	99	561	2.247	2.35
98	51727505	1	3.541	0.340	7.259	69	246	2.213	2.59
99	52408421	2	4.833	0.354	16.260	71	300	3.416	−0.16
100	52591272	1	0.734	4.250	60.000	210	4	2.479	0.00
101	52607815	1	1.758	0.090	7.438	37	78	1.897	0.59
102	52723963	2	6.380	0.813	12.720	144	804	4.361	3.40
103	53417291	2	3.440	0.803	6.479	51	153	2.540	−1.85
104	54449740	1	5.141	1.033	4.250	135	475	1.372	5.08
105	56361558	2	8.339	1.149	13.270	282	1739	3.540	6.00
106	58920313	1	2.030	1.698	6.359	115	780	2.247	4.41
107	59071102	1	1.601	1.483	5.579	104	630	2.247	4.76
108	60506812	5	7.710	1.655	16.000	111	703	4.661	−0.84
109	63225536	1	2.788	0.090	10.290	44	120	2.113	2.17
110	64448686	2	7.063	0.726	12.500	255	1426	3.830	3.42
111	65983315	1	5.581	0.694	2.525	186	732	1.535	2.68
112	66008682	1	1.405	2.135	7.090	132	990	2.247	3.66
113	66008693	1	1.653	1.803	6.584	114	741	2.247	3.19
114	66008706	1	1.900	1.471	6.178	96	528	2.247	2.72
115	66028306	2	2.885	0.225	9.679	52	153	2.007	0.91
116	66028328	3	5.306	0.321	15.580	81	378	3.072	3.06
117	66028340	3	4.59	0.321	13.810	75	325	2.923	2.00
118	66671225	1	4.918	1.380	6.817	72	270	1.861	3.49
119	66710972	2	11.060	2.247	9.057	274	1589	2.247	7.00
120	67584558	1	1.817	1.035	4.646	75	325	2.247	3.52
121	67584569	1	1.694	1.200	4.855	84	406	2.247	3.76
122	67584570	1	1.570	1.366	5.087	93	496	2.247	4.00
123	67892993	1	5.531	0.166	15.750	0	445	2.962	1.28
124	67893009	3	4.949	0.356	11.560	139	645	3.515	0.61
125	67905082	2	3.971	0.359	7.934	54	171	1.988	3.62
126	67905413	2	4.555	0.504	6.667	88	282	1.821	3.44
127	67905480	2	3.355	0.225	10.850	50	153	2.242	1.36
128	67952492	1	3.435	0.337	8.333	0	91	1.293	4.32
129	67952505	2	8.199	1.482	9.877	0	1381	3.168	6.49
130	68084628	1	1.446	1.532	5.354	0	595	2.247	4.23
131	68227974	1	2.153	1.532	6.112	0	666	2.247	4.18
132	68227985	1	2.277	1.366	5.878	0	561	2.247	3.94
133	68227996	1	2.401	1.200	5.689	0	465	2.247	3.71
134	68298066	1	1.848	1.151	5.087	0	435	2.247	4.29
135	68298602	1	2.362	1.483	6.343	0	703	2.247	5.82
136	70146053	2	9.111	1.753	9.671	0	1573	3.420	7.00
137	70495395	2	6.917	0.751	10.210	0	1260	3.057	5.69
138	71412356	2	4.585	0.596	12.250	0	231	2.654	2.31
139	72276052	1	1.733	2.135	7.460	135	1035	2.247	3.29
140	72928428	2	4.657	0.521	14.080	68	276	3.469	0.04
141	84732285	1	6.448	0.579	8.945	0	1001	3.781	5.40
142	85412540	3	7.200	1.028	18.840	0	666	4.660	2.37
143	87320056	2	6.420	1.902	6.428	110	448	2.745	2.04

the 71 values are below the average for the entire population, and 10 of these values are below 1.0.

Cluster 4 contains 33 compounds, including 24 halogenated compounds (21 fluorinated and 3 brominated). The log P values tend to be slightly higher than average.

Cluster 5 contains 18 compounds that are mostly oligomeric: 1 monomer, 7 dimers, 8 trimers, 1 tetramer, and 1 pentamer. The relatively wide range of log P values includes the lowest log P in the entire data set. Four of the total six hydrophilic compounds are in this group.

CONCLUSIONS

In conclusion, five stable clusters of acrylates were found by using objective clustering methods based on calculated physical and chemical properties. The partitions that were

derived from the features we chose were not obvious divisions to make before we began this study, but at the same time they could make sense in terms of toxicology. At this stage we cannot make conclusions about a relationship between these groups and biological activity, but the objective sampling scheme itself has succeeded in its initial goals. Once samples selected on the basis of these clusters have passed through bioassays, it may be possible to present a SAR study to predict the toxicity of other compounds.

ACKNOWLEDGMENT

The financial support of the Chemical Manufacturers Association made this work possible. Log *P* values were calculated at Rohm and Haas by Dr. Clay Frederick and Mary Bright.

REFERENCES AND NOTES

(1) Hashimoto, K.; Aldridge, W. N. Biochemical Studies on Acrylamide, A Neurotoxic Agent. *Biochem. Pharmacol.* **1970**, *19*, 2591–2604.
(2) Pozzani, U. L.; Weil, C. S.; Carpenter C. P. Subacute Vapor Toxicity and Range-Finding Data for Ethyl Acrylate. *J. Ind. Hyg. Toxicol.* **1949**, *31*, 311–316.
(3) Treon, J. F.; Sigmon, H.; Kitzmiller, K. U. The Toxicity of Methyl and Ethyl Acrylate. *J. Ind. Hyg. Toxicol.* **1949**, *31*, 317–326.
(4) Borzelleca, J. F., et al. Studies on the Chronic Oral Toxicity of Monomeric Ethyl Acrylate and Methylmethacrylate. *Toxicol. Appl. Pharmacol.* **1964**, *6*, 29.
(5) Hodes, L. Clustering a Large Number of Compounds. 1. Establishing the Method on an Initial Sample. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 66–71.
(6) Willett, P.; Winterman, v.; Bawden, D. Implentation of Nonhierarchic Cluster Analysis Methods in Chemical Information Systems: Selection of Compounds for Biological Testing and Clustering of Substructures Search Output. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 109–118.
(7) Stuper, A. J.; Bruger, W. E.; Jurs, P. C. *Computer Assisted Studies of Chemical Structure and Biological Function*; Wiley-Interscience: New York, 1979.
(8) Jurs, P. C. Pattern Recognition Used to Investigate Multivariate Data in Analytical Chemistry. *Science* **1986**, *232*, 1219–1224.
(9) Varmuza, K. *Pattern Recognition in Chemistry*: Springer-Verlag: Berlin, 1980.
(10) Jurs, P. C. Computer Assisted Studies of Structure-Activity Relations Using Pattern Recognition. *Drug Inf. J.* **1983**, *17*, 219–229.
(11) Jurs, P. C.; Stouch, T. R.; et al. Computer-Assisted Studies of Molecular Structure–Biological Activity Relationships. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 296–308.
(12) Rose, S. L.; Jurs, P. C. Computer-Assisted Studies of Structure–Activity Relationships of N-Nitroso Compounds Using Pattern Recognition. *J. Med. Chem.* **1982**, *25*, 769–776.
(13) Stouch, T. R.; Jurs, P. C. Computer-Assisted Studies of Molecular Structure and Genotoxic Activity by Pattern Recognition Techniques. *EHP, Environ. Health Perspect.* **1985**, *61*, 329–343.
(14) Del Re, G. A Simple MO-LCAO Method for the Calculation of Charge Distributions in Saturated Organic Molecules. *J. Chem. Soc.* **1958**, 4031–4040.
(15) Del Re, G.; Pullman, B.; Yonezawa, T. Electronic Structure of the Alpha-Amino Acids of Proteins. *Biochim. Biophys. Acta* **1963**, 153–182.
(16) Autian, J. Structure–Toxicity Relationships of Acrylic Monomers. *EHP, Environ. Health Perspect.* **1975**, *11*, 141–152.
(17) Lawrence, W. H.; Bass, G. E.; et al. Use of Mathematical Models in the Study of Structure–Toxicity Relationships of Dental Compounds: I. Esters of Acrylic and Methacrylic Acids. *J. Dent. Res.* **1972**, *51*, 526–535.
(18) Tanii, H.; Hashimoto, K. Structure–Toxicity Relationships of Acrylates and Methacrylates. *Toxicol. Lett.* **1982**, *11*, 125–129.
(19) Fujisawa, S.; Masuhara, E. Determination of Partition Coefficients of Acrylates, Methacrylates, and Vinyl Monomers Using High Performance Liquid Chromatography (HPLC) *J. Biomed. Mater. Res.* **1981**, *15*, 787–793.
(20) Pomonoa College Med-Chem package ClogP 3.5.
(21) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure–Activity Analysis*; Wiley: New York, 1986.
(22) Watanabe, S. Karhunen-Loeve Expansion and Factor Analysis—Theoretical Remarks and Applications. *Proceedings of the 4th Conference on Information Theory*; Publishing House of the Czechoslovak Academy of Sciences: Prague, 1965.
(23) Kier, L. B. A Shape Index from Molecular Graphs. *Quant. Struct.–Act. Relat. Pharmacol., Chem. Biol.* **1985**, *4*, 109.
(24) Randic, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.
(25) Hansen, P. J.; Jurs, P. C. Prediction of Olefin Boiling Points from Molecular Structure. *Anal. Chem.* **1987**, *59*, 2322–2327.
(26) Rohrbaugh, R. H.; Jurs, P. C. Prediction of Retention Indexes for Diverse Drug Compounds. *Anal. Chem.* **1988**, *60*, 2249–2253.
(27) Randic, M.; Jurs, P. C. On a Fragment Approach to Structure–Activity correlations. *Quant. Struct.–Act. Relat.* **1989**, *8*, 39–48.
(28) Hopkins, B. A. New Method of Determining the type of Distribution of Plant Individuals. *Ann. Bot.* **1954**, *18*, 213–226.
(29) Jain, A. K.; Dubes, R. C. *Algorithms for Clustering Data*; Prentice Hall: Englewood Cliffs, NJ, 1988.
(30) Lawson, R. G.; Jurs, P. C. A New Index for Clustering Tendency and Its Application to Chemical Problems. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 36–41.
(31) MacQueen, J. Some Methods for Classification and Analysis of Multivariate Data. *Proceedings of the 5th Berkeley Symposium on Probability and Statistics*; University of California Press: Berkeley, 1967.
(32) Jain, A. K.; Moreau, J. V. Bootstrap Techniques in Cluster Analysis. *Pattern Recognit.* **1987**, *20*, 547–568.
(33) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure–Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.
(34) Ball, G. H.; Hall, D. J.; Isodata, an Iterative Method of Multivariate Analysis and Pattern Classification. *Proceedings of the AFIPS Fall Joint Computer Conference*; Spartan Books: Washington, DC, 1965; Vol. 2, pp 329–330.
(35) Ott, L. *An Introduction to Statistical Methods and Data Analysis*; Duxbury Press: Boston, 1984.