ences in the connection table; i.e., all of the entries in the ORIG connection table through the kth row refer to atoms for which correspondences have been found, and therefore have entries in the array F. The depth-first renumbering of the connection table referred to previously is applied exactly for this purpose: to produce a nonredundant connection table that contains no forward references. On this basis, selection of a candidate node corresponding to the $k + l$th atom proceeds as follows:

(1) For each atom m attached to atom $k + 1$ in ORIG do:
  (a) Find the corresponding atom, c = F[m], in COPY. Because there are no forward references, c is guaranteed to exist.
  (b) Form a set P_m containing all atoms p for which p is attached to c and p is in ATOMSET[k+1].
(2) The set of candidates corresponding to atom $k + 1$ is the intersection of all of the sets obtained in step b, above. Return one of these nodes as a candidate node, holding the rest in reserve for backtracking.

The argument behind these steps is straightforward: atom m is attached to $k + 1$; atom c corresponds to atom m; atom p is attached to atom c, and if it is contained in the set ATOMSET[k+1], then atom p is a possible correspondent to atom $k + 1$. Each connection m produces its own set P_m of possible correspondents p to k, hence step 2 is invoked, requiring that all of these sets P_m must jointly contain correspondents to atom $k + 1$.

The performance of the atom-by-atom search alone was timed for 100 iterations of the mapping of a given structure ORIG onto its COPY, using 22 structures spanning a size range of 10-120 atoms, on a microcomputer with an 80386D microprocessor (33 mHz clock). The structures were mostly ring assemblies, with a few large, branched alicyclic structures. Processing times per 100 iterations ranged from 2.09 s for the smallest structure to 63.61 s for the largest. A plot of log t vs log n was quite linear with some scatter. The approximate equation of the line was $t = n^{1.45/20}$. No great precision is claimed for these results, but they do give an order-of-magnitude estimate of performance.

REFERENCES AND NOTES

(1) Read, R. C.; Corneil, D. G. The Graph Isomorphism Disease. *J. Graph Theory* **1977**, *1*, 339. See also Carhart, R. E. Erroneous Claims Concerning the Perception of Topological Symmetry. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 108-110.
(2) Liu, X.; Balasubramanian, K.; Munk, M. E. Computational Techniques for Vertex Partitioning of Graphs. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 263-269.
(3) Rücker, G.; Rücker, C. Computer Perception of Constitutional (Topological) Symmetry: TOPSYM, a Fast Algorithm for Partitioning Atoms and Pairwise Relations among Atoms into Equivalence Classes. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 187-191.
(4) Rücker, G.; Rücker, C. Isocodal and Isospectral Points, Edges, and Pairs in Graphs and How To Cope with Them in Computerized Symmetry Recognition. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 422-427.
(5) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107.
(6) Zamora, A. An Algorithm for Finding the Smallest Set of Smallest Rings. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 40.
(7) Nijenhuis, A.; Wilf, H. S. *Combinatorial Algorithms*; Academic Press: New York, 1978.
(8) Sussenguth, E. H., Jr. A Graph-Theoretic Algorithm for Matching Chemical Structures. *J. Chem. Doc.* **1965**, *5*, 36.
(9) Figueras, J. Substructure Search by Set Reduction. *J. Chem. Doc.* **1972**, *12*, 237.
(10) Willett, P.; Wilson, T.; Reddaway, S. F. Atom-by-Atom Searching Using Massive Parallelism. Implementation of the Ullmann Subgraph Isomorphism Algorithm on the Distributed Array Processor. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 225-233.

Limits of Classification. 2. Comment on Lawson and Jurs

LOUIS HODES

National Cancer Institute, Bethesda, Maryland 20892

Received February 20, 1991

Lawson and Jurs[1,2] have two recent papers on clustering using data on a diverse set of 143 published acrylates. From our work[3] on clustering large diverse sets of compounds, we have found that a diverse set of compounds will generally have a dual nature—some clustered and some scattered compounds. Are the five published[2] clusters a good classification of that set of compounds? A further analysis throws more insight into the interrelations among the acrylates.

INTRODUCTION

Lawson and Jurs[1,2] have two recent papers on clustering using data on a diverse set of 143 published acrylates. Their first paper[1] explores clustering tendency with a modified Hopkins method. Their second paper[2] performs the clustering by the Kmeans and Isodata methods.

Their purpose in clustering was to guide the sampling of compounds for toxicity testing. One compound from each cluster can be chosen on the assumption that the other compounds in the same cluster should have similar toxicity.

From our work[3] on clustering large diverse sets of compounds, we have found that a diverse set of compounds will generally have a dual nature—some clustered and some scattered compounds, the distribution dependent on the criteria for clustering. It is our contention that algorithms which insist on a complete clustering of a diverse set will often force disparate compounds into unsuitable clusters.

There is a question as to how far this phenomenon, which we called a limit to classification, shows up in a relatively small set such as the 143 diverse acrylates. Are the five published[2] clusters a good classification of that set of compounds?

Here we fill in some gaps in the Lawson and Jurs work. A more precise analysis results in more flexibility in achieving their goal of sampling acrylates for toxicity testing.

Our relevant paper,[3] the third in a series on clustering a large number of compounds, was subtitled "the limits of classification". This paper can be considered number 4 in the series on clustering, but it is more fitting as the second paper dealing with the problem of classification. Also, this paper deals intimately with the data and methods of Lawson and Jurs[1,2] to the extent that it requires the reader to be somewhat

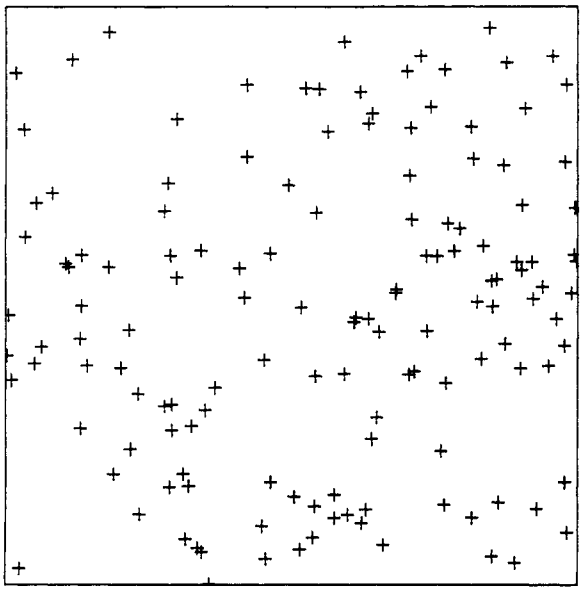

Figure 1. First two dimensions of 140 eight-dimensional points randomly generated to check out the Hopkins statistic. It did average 0.5.

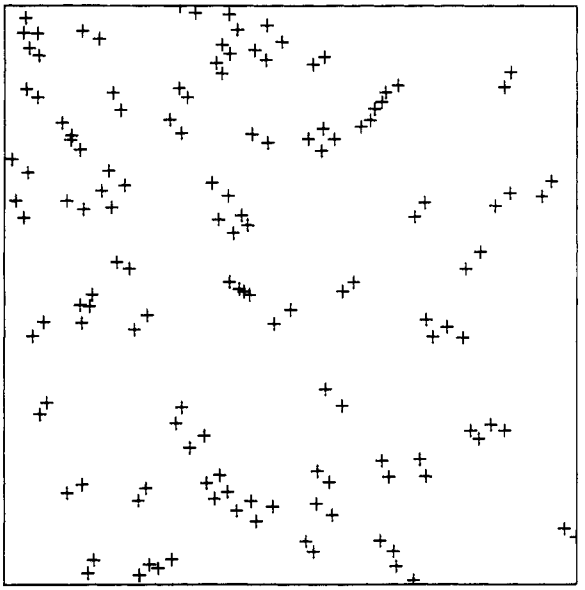

Figure 2. First 70 random points of Figure 1 paired with 70 close points. The Hopkins statistic indicates clustering at an average value of 0.9.

familiar with those papers. Therefore, it can be considered a commentaria on their work.

This work also touches on a philosophical issue—the nature of chemical structure data and the treatment of outliers. In mathematics, a datum that does not fit is called a counterexample and is sufficient to negate an entire proposition. Also, in physics, an anomalous effect can change a whole theory, for example, from Newtonian to Einsteinian. However, in laboratory experiments, outliers are usually considered noise and avoided. In sociology, for example, judging olympic and beauty contests, the highest and lowest scores are automatically eliminated. Everyone knows in a democracy that the majority usually wins and that minority rights need to be protected.

Thus, the treatment of outliers is generally related to the hardness or exactness of the science. Chemical structure data is usually precise. And, moreover novel compounds or outliers are often important. One should be careful in regard to the treatment of outliers when clustering chemical structures.

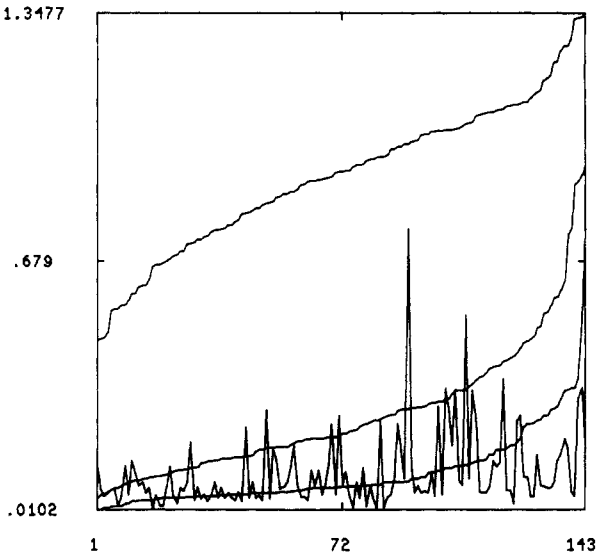

Figure 3. Plot of the distance to the closest neighbor for all the range-scaled acrylate compounds (jagged line). The lowest curve is the same data, sorted. The upper curve is a sorted sample of 143 random Hopkins distances to the nearest acrylate. Between these two is a sorted sample of 143 modified Hopkins distances to the nearest acrylate.

Table I. Comparison of Calculated and Reported Statistics

	calculated[b]		reported[c]	
descriptors[a]	mean	SD	mean	SD
1 SSS 2	1.4	0.69	0.6	0.8
2 CHIS 6	3.2	2.06	3.6	2.2
3 MOLC 7	0.65	0.62	0.7	0.6
4 KAPA 3	7.92	3.99	7.8	4.3
5 PATH 2	62	50	30	47
6 ALLP 1	334	357	320	370
7 TSCH 1	2.2	1.4	2.2	1.4
8 CLGP 0	2.85	1.88	3.0	2.0

[a] Descriptor symbols from ref 2. [b] Mean and standard deviation calculated from data table in ref 2 corrected by ref 6. [c] Mean and standard deviation reported in ref 1. Descriptors appeared in different order.

MOLECULAR FRAGMENTS VS UNIVERSAL INDICES

One striking difference of the Lawson and Jurs work from ours is their use of physicochemical constants and graphical indices rather than solely molecular fragments[4] as features. Their features are universal in the sense that they apply to every compound, whereas it is generally useful to consider molecular fragment features as applying only to the compounds in which they occur.

There is evidence[5] that the features they use are more relevant to general properties such as cohesion and transport rather than specific receptor structure–activity.

Universal features allow one to work in a space with Euclidean distances rather than merely similarity measures on pairs of data items. The choice of eight universal features gives Lawson and Jurs a reasonable dimensionality to display projected data, e.g., onto the first two principal components. It also allows one to easily simulate data to satisfy certain statistical hypotheses. Lawson and Jurs claim that the projected data do not show evidence of clustering, but they do make use of the Euclidean nature of the data space with many simulations and their use of the Hopkins method.

It would be a good exercise to compare the clustering using solely molecular fragment features. However, the kind of features Lawson and Jurs used may be more appropriate for

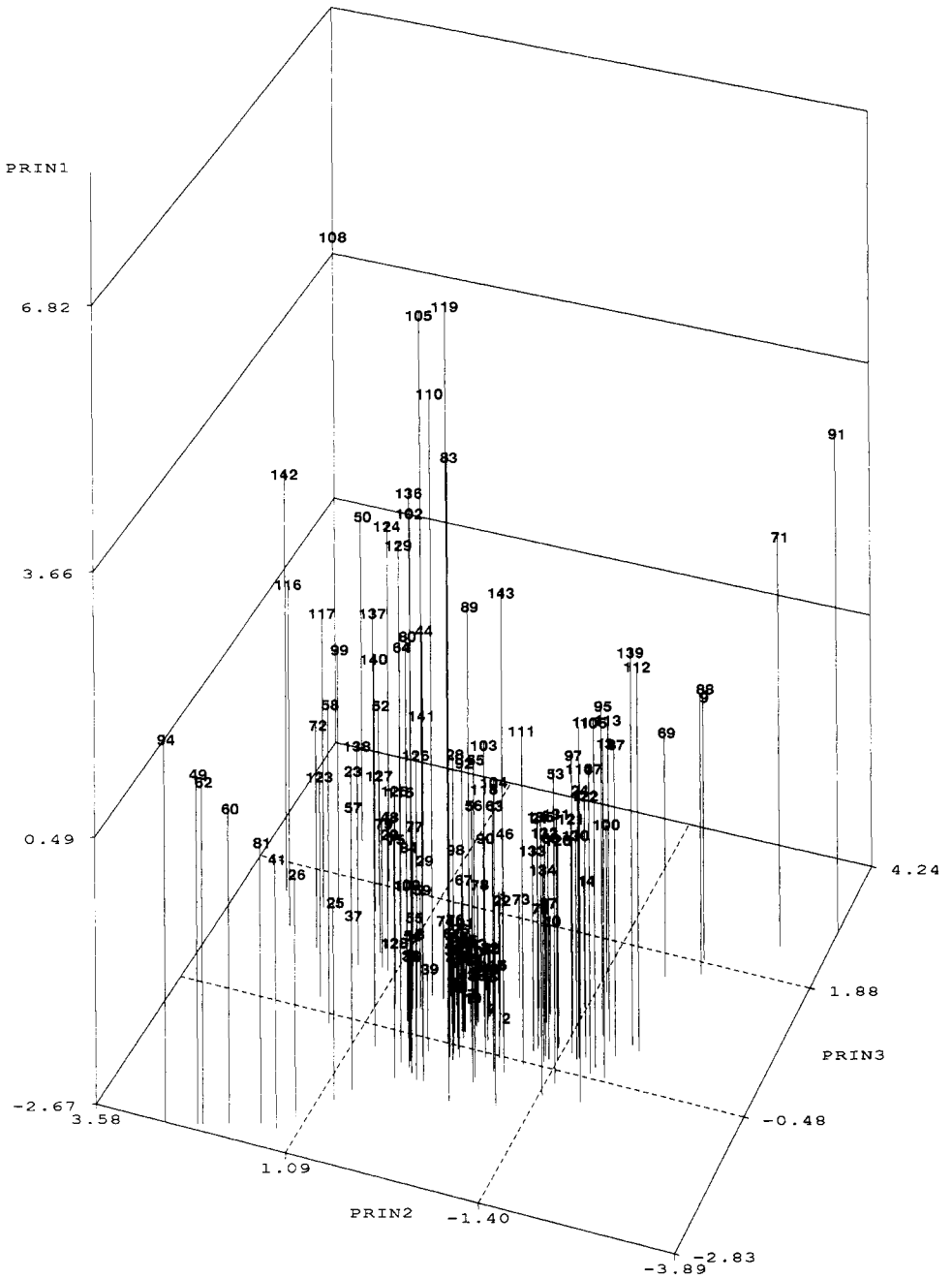

Figure 4. Projection of the autoscaled acrylate data onto the first three principal component axes. The compounds are numbered according to the table in ref 2.

this data set that is not quite so large or diverse as we have been treating.

THE TROUBLE WITH HOPKINS

Lawson and Jurs[1] use a modified Hopkins statistic as a measure of clustering tendency. The Hopkins statistic says that a data set is probably clusterable if the distance, U_i, of a random point in data space to the closest point in the data is, on the average, significantly greater than the distance, W_i, of an arbitrary data point to its nearest neighbor.

More precisely, the Hopkins statistic is the value H given by

$$H = \sum U_i / (\sum U_i + \sum W_i)$$

where i ranges over a sample of some given number of pairs of points and H itself is averaged over several iterations. See ref 1 for many examples.

Note that H will be about 0.5 for random data such as that shown in Figure 1 and will grow toward 1 for clustered data. Lawson and Jurs noticed that outliers in the data tended to artificially concentrate the main body of the data, producing spurious increases in H. They modified the method so that the "random" points were to be sampled from the data distribution independently in each dimension.

Lawson and Jurs show the results of Hopkins on their acrylate data in ref 1. The H values for the acrylate data averaged 0.82 originally and 0.65 under the modified Hopkins.

Compared to experiments with random and clustered artificial data, they get an unexpected wide range of H values for the acrylate data. One can see that such a wide range is due to outliers in the data.

A more basic problem with the Hopkins statistic arises from reliance on pairwise distances. For example, the value of 0.5 for the 140 random points of Figure 1 is easily increased to 0.9 for the 70 random pairs of points in Figure 2. Thus, the

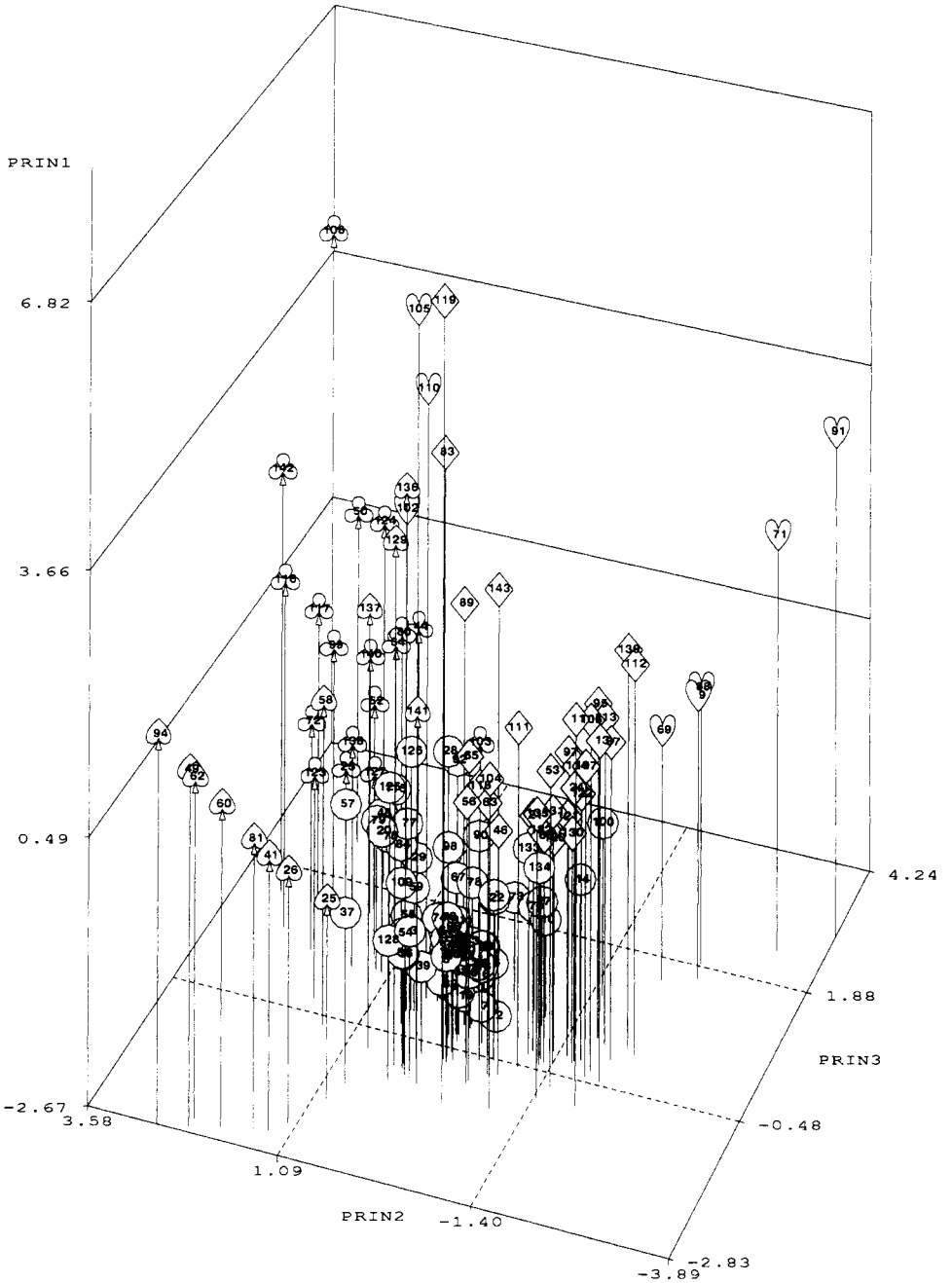

Figure 5. Clusters from ref 2 superimposed on Figure 4. Clusters: 1, heart; 2, spade; 3, balloon; 4, diamond; 5, club.

Hopkins statistic as well as the modified Hopkins statistic will indicate clustering if most data points have at least one close neighbor. This is usually true of chemical data such as the acrylate data.

Figure 3 shows the distance to the closest neighbor for all 143 compounds taken from the data in ref 2. For visual comparison we sort the distances, giving a histogram on the same plot. A sorted sample of 143 random Hopkins distances, i.e., distances of 143 random points to the closest data points are also shown, as well as a sorted modified Hopkins sample. The behavior of the Hopkins statistic is a measure of the disparity between the data curve and the relevant random sample curve.

THE ACRYLATE DATA

The numerical data for all eight variables for the 143 compounds were published in the May 1990 article.[2] In May 1991, Lawson and Jurs[6] published corrections for four of the data points. The following has been redone to incorporate these corrections.

The February article[1] discussed the data using the same eight variables, along with a table giving their means and standard deviations. There was substantial discrepancy when we computed the means and standard deviations of the May data. Both sets of values are shown in Table I. We believe the February values to be in error, e.g., the first variable, the multiplicity of the acrylate group, which varies from 1 to 5 cannot have a mean value of 0.6.

We next show a projection of the data onto the first three principal component axes. The three axes are somewhat superior to the first two axes, accounting for 76% of the variance as opposed to 61%. The data are scaled to mean zero and variance one. The numbering of the compounds in Figure 4 comes from the data table in ref 2.

Lawson and Jurs do not present a principal component plot, saying that the data do not show any obvious clustering along any two axes. However, Figure 4 provides an excellent in-

Table II. Acrylates Sorted by First Principal Component[a]

IDNO	CLNO	PRIN1	PRIN2	PRIN3	SSS 2	CHIS 6	MOLC 7	KAPA 3	PATH 2	ALLP 1	TSCH 1	CLOGP 0
2	3	−2.7	−0.44	0.208	1	0.727	0.048	3.000	21	21	0.742	0.750
16	3	−2.5	−0.10	0.542	1	1.095	0.048	5.000	25	36	1.256	−0.060
7	3	−2.5	−0.34	0.009	1	0.956	0.048	3.840	22	28	0.850	1.280
19	3	−2.3	−0.14	−0.005	1	1.123	0.048	5.000	25	36	0.808	1.260
6	3	−2.3	−0.27	0.534	1	1.790	0.166	3.556	34	60	1.151	−0.130
42	3	−2.3	0.041	0.217	1	1.364	0.048	5.878	28	45	1.083	0.560
17	3	−2.2	−0.15	−0.23	1	1.372	0.048	5.000	25	36	0.920	1.800
82	3	−2.1	−0.02	−0.27	1	1.403	0.048	5.878	28	45	0.889	1.790
33	3	−2.1	0.261	0.289	1	1.802	0.048	7.000	30	55	1.356	0.350
5	3	−2.0	0.179	0.013	1	1.569	0.048	7.000	30	55	1.191	1.090
85	3	−2.0	0.124	0.444	1	1.515	0.113	7.000	31	55	1.645	0.330
43	3	−2.0	−0.02	0.414	1	1.979	0.177	5.531	31	55	1.498	0.310
61	3	−2.0	0.354	0.466	1	1.708	0.048	7.901	33	66	1.598	0.010
15	3	−2.0	−0.24	−0.09	1	1.690	0.284	5.000	27	36	1.010	1.580
8	3	−1.9	−0.01	−0.46	1	1.725	0.048	5.878	28	45	0.995	2.330
38	3	−1.9	−0.29	−0.27	1	1.860	0.215	4.500	28	45	1.081	2.110
30	3	−1.8	−0.33	0.118	1	2.469	0.166	3.265	46	100	1.255	0.990
32	3	−1.8	0.135	0.163	1	2.175	0.364	7.000	31	55	1.107	0.650
12	3	−1.8	−0.31	0.097	1	0.500	0.227	7.000	31	55	1.875	2.040
101	3	−1.8	0.171	0.388	1	1.758	0.090	7.438	37	78	1.897	0.590
18	3	−1.8	−0.46	−0.28	1	1.938	0.116	3.265	51	106	0.988	2.060
51	3	−1.7	−0.12	−0.22	1	2.141	0.209	5.325	37	78	1.070	1.770
66	3	−1.7	−0.46	−0.35	1	2.547	0.166	2.651	45	85	1.068	2.220
74	3	−1.7	0.339	0.229	1	2.528	0.364	7.901	34	66	1.172	0.310
39	3	−1.6	0.161	−0.69	1	2.079	0.048	7.000	30	55	1.070	2.860
11	4	1.89	−1.9	−0.89	1	2.238	1.649	6.596	118	820	2.247	6.050
89	4	1.89	0.192	0.699	2	6.248	2.302	9.694	65	253	2.704	2.900
94	2	1.90	2.66	−2.8	1	7.382	0.048	21.960	75	325	2.194	7.000
143	4	1.98	−0.15	0.969	2	6.420	1.902	6.428	110	448	2.745	2.040
112	4	2.00	−2.3	0.154	1	1.405	2.135	7.090	132	990	2.247	3.660
139	4	2.10	−2.2	0.227	1	1.733	2.135	7.460	135	1035	2.247	3.290
71	1	2.43	−3.3	2.46	1	−0.228	1.729	5.600	114	741	9.286	3.890
142	5	2.58	2.98	1.38	3	7.200	1.028	18.840	0	666	4.660	2.370
137	2	2.59	0.803	−0.83	2	6.917	0.751	10.210	0	1260	3.057	5.690
102	1	3.03	0.883	0.485	2	6.380	0.813	12.720	144	804	4.361	3.400
91	1	3.55	−3.9	2.94	1	−0.475	2.061	6.201	132	990	10.930	4.359
129	2	3.59	0.393	−0.98	2	8.199	1.482	9.877	0	1381	3.168	6.490
108	5	3.91	3.58	4.24	5	7.710	1.655	16.000	111	703	4.661	−0.840
136	2	4.32	0.196	−1.1	2	9.111	1.753	9.671	0	1573	3.420	7.000
83	4	4.58	−0.13	−0.69	2	7.114	1.149	9.408	238	1217	2.857	5.870
110	1	4.71	0.490	0.224	2	7.063	0.726	12.500	255	1426	3.830	3.420
105	1	6.28	0.184	−0.79	2	8.339	1.149	13.270	282	1739	3.540	6.000
119	4	6.82	−0.42	−1.4	2	11.060	2.247	9.057	274	1589	2.247	7.000

[a] Compound number and cluster number from ref 2, the first 3 principal components and the eight variables. Lowest 25 compounds and highest 18 compounds.

dication of the distribution of the data. There is some clustering, but the clusters are not normally distributed. There are many scattered compounds in assorted groupings. Contrary to Lawson and Jurs, the picture is quite similar for the first two principal components but does not provide as much information. The data do show the type of distribution found in large diverse sets of compounds.

There is one obvious outlier in Figure 4, compound 108. Compounds 105 and 119 stand out at the top in PRIN1, the first principal component. There are two short sequences of compounds that stick out—69, 88, 9, 71, 91 at low PRIN2 and high PRIN3 and 26, 41, 81, 60, 82, 49, 94 at high PRIN2 and low PRIN3. This is a more complex picture than implied by the clustering of Lawson and Jurs.

THE SO-CALLED NATURAL CLUSTERING

Lawson and Jurs speak of a 'natural' clustering according to five different properties. These are high total charge, hydrophobicity, small compounds, halogenated compounds, and large or polymeric compounds. The problem is that some of these categories are not exclusive, e.g., halogen, which is not even an explicit variable. Moreover, the five clusters in ref 2 do not respect their natural boundaries very well. Lawson and Jurs speak of preponderance of their natural distinctions, but why does a compound which belongs in one cluster fall into another? Are there no compounds that are naturally separate?

Table III. Some Rough Relations between Ref 2 Cluster Number and the First Three Principal Components

cluster	PRIN1	PRIN2	PRIN3
1	high	low	high
2	high	high	low
3	low		
4	high	low	center
5	high	high	high

In Figure 5 we superimpose on Figure 4 the clustering of Lawson and Jurs from May 1990. The outlier, 108, is assigned to cluster 5, probably the closest cluster. The two protruding sequences form parts of clusters 1 and 2, respectively. However, the high PRIN1 compounds 105 and 119 are assigned to clusters 1 and 4, respectively, but are separate from other members of those clusters. There are several more cases of apparent anomalies. For example, cluster 1 has, in addition to the five-compound sequence labeled by hearts, three more compounds (hearts) that seem quite separate. All the other clusters have similar inconsistencies.

Are there more errors in the data or are the methods deficient? One way to look at the data is to sort by the three principal components in turn and examine the data at the low and high ends. For example, the compounds sorted by PRIN1

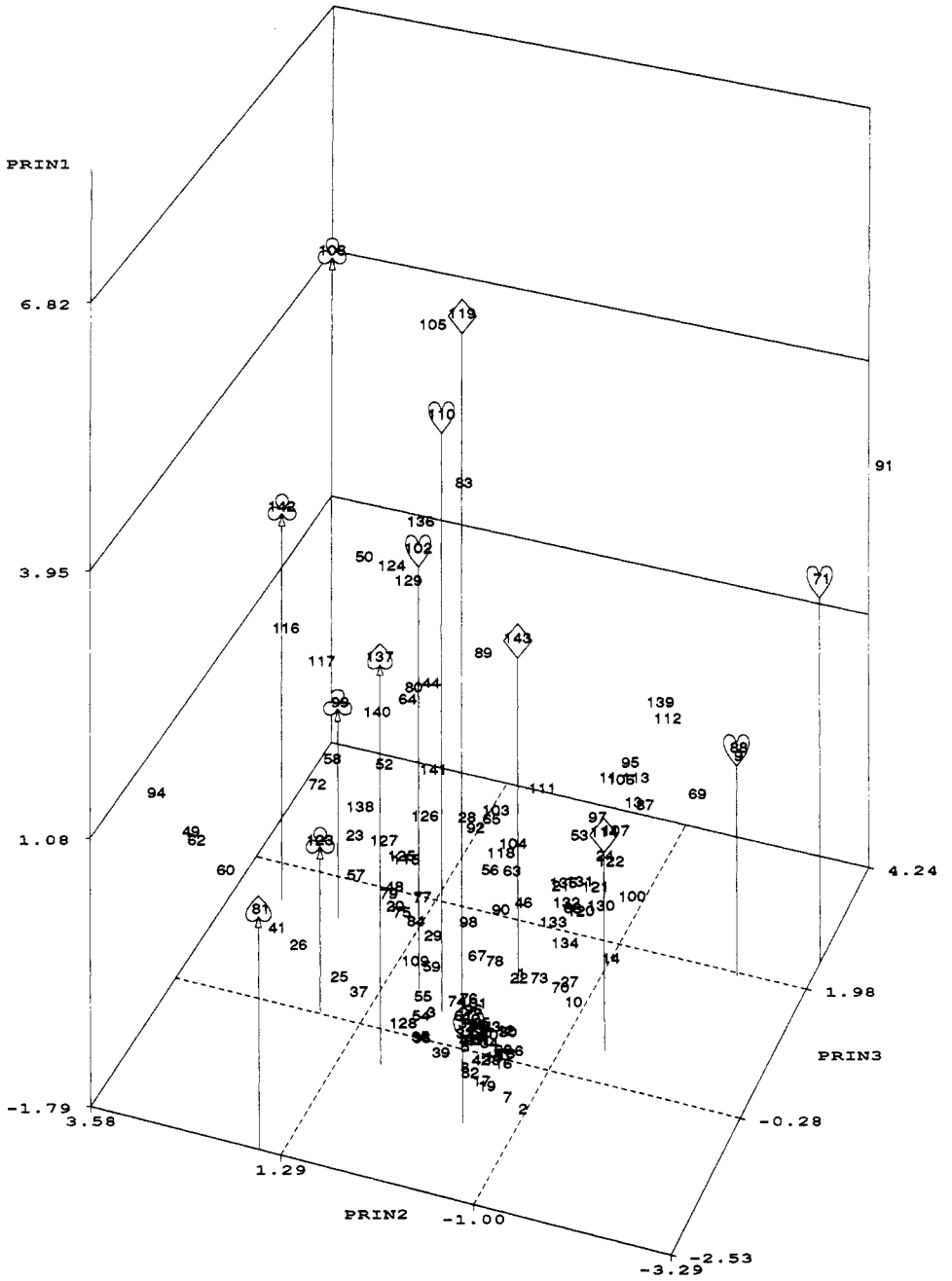

Figure 6. Samples from the 14 clusters shown in the principal component data space.

are shown in Table II. Some association between the five clusters and the three principal components is shown in Table III. However, we can examine the data more comprehensively by using hierarchical clustering.

HIERARCHICAL CLUSTERING

Lawson and Jurs say that hierarchical clustering on this amount of data would be too complex for simple interpretation. Actually, the clustering tendency of the data as shown by the principal component projection and the preponderance of close pairs of points reduces the complexity of the hierarchical clustering so that it can be reasonable. On the other hand, the methods used by Lawson and Jurs are intended for data that exhibit good clustering. The results on the mixed chemical data are very difficult to interpret, especially in more than two or three dimensions. Moreover, hierarchical clustering can provide flexibility in choosing levels of clusters to yield a desired number of representative compounds for testing.

The SAS package offers a bewildering assortment of 11 different hierarchical clustering methods. After some experimenting, we chose the centroid method as the most straightforward. At each step the two closest clusters are combined, where the distance is defined as the distance between the cluster centroids weighted by the number of cluster elements.

Table IV shows the output from centroid clustering all 143 compounds. If we take clusters at the level where the two major clusters join—a cluster of 100 is formed from clusters of 68 and 32—then we get the 14 clusters of Table V. This would roughly correspond to the level of clustering in Lawson and Jurs. The cluster number from ref 2 is included in the table. Notice that 5 of the 14 clusters are singletons.

It is instructive to choose a representative sample from each of the 14 clusters. We show these on the principal component graph in Figure 6. The 14 compounds provide a good covering of the data space.

One can cluster the 14 samples to see how the 14 clusters relate. Here we get the tree diagram of Figure 7. At this point we can reduce the number of clusters somewhat by

Table IV. Step-by-Step Formation of Clusters of the 143 Acrylates by the Hierarchical Centroid Method

no. of clusters	clusters joined		frequency of new cluster	normalized centroid distance	tie	no. of clusters	clusters joined		frequency of new cluster	normalized centroid distance	tie
142	24	122	2	0.039587		71	CL77	30	18	0.204942	
141	49	62	2	0.047056		70	116	117	2	0.206230	
140	17	82	2	0.057710		69	3	128	2	0.208082	
139	20	79	2	0.063875		68	23	72	2	0.214744	
138	33	85	2	0.066898		67	CL73	CL86	14	0.218289	
137	75	84	2	0.068320		66	129	136	2	0.220791	
136	97	114	2	0.073982		65	CL90	94	4	0.223476	
135	95	113	2	0.074573		64	125	126	2	0.225181	
134	34	45	2	0.074696		63	CL95	CL135	6	0.225258	
133	112	139	2	0.075758		62	76	78	2	0.228697	
132	CL138	61	3	0.076227		61	CL69	CL82	5	0.231883	
131	41	81	2	0.076655		60	CL67	CL71	32	0.249136	
130	7	19	2	0.077467		59	CL76	57	10	0.253003	
129	31	51	2	0.078132		58	CL75	CL136	8	0.253726	
128	10	70	2	0.078392		57	CL60	12	33	0.254590	
127	68	120	2	0.080238		56	22	CL62	3	0.262948	
126	CL139	48	3	0.082036		55	CL78	135	6	0.269244	
125	4	86	2	0.082206		54	CL57	73	34	0.271559	
124	47	93	2	0.084899		53	CL80	118	3	0.283916	
123	15	38	2	0.085629		52	CL83	100	5	0.284837	
122	32	74	2	0.086361		51	11	CL63	7	0.288652	
121	CL132	101	4	0.086909		50	CL54	CL87	38	0.303559	
120	8	CL140	3	0.088213		49	28	92	2	0.306011	
119	18	66	2	0.095271		48	CL51	CL133	9	0.309525	
118	CL134	40	3	0.096912		47	44	CL101	3	0.310490	
117	CL126	CL137	5	0.099412		46	CL50	CL56	41	0.319356	
116	132	133	2	0.101183		45	CL47	52	4	0.320822	
115	CL127	121	3	0.101230		44	CL46	CL84	44	0.323269	
114	36	39	2	0.102009		43	CL44	CL61	49	0.328758	
113	25	37	2	0.103293		42	CL74	104	3	0.348437	
112	5	42	2	0.104892		41	CL113	CL94	5	0.350196	
111	CL120	CL123	5	0.105102		40	CL68	CL81	4	0.363534	
110	21	CL115	4	0.106365		39	71	91	2	0.380725	
109	CL125	CL129	4	0.107898		38	CL48	CL58	17	0.381386	
108	130	131	2	0.110011		37	89	143	2	0.386930	
107	13	106	2	0.112372		36	CL49	CL64	4	0.404608	
106	CL119	CL118	5	0.113300		35	CL59	CL36	14	0.365584	
105	115	127	2	0.118206		34	CL42	111	4	0.420204	
104	2	CL130	3	0.120604		33	CL53	53	4	0.422524	
103	CL117	77	6	0.121243		32	CL45	50	5	0.428095	
102	CL116	134	3	0.122581		31	CL43	CL35	63	0.429455	
101	64	80	2	0.122964		30	137	141	2	0.442591	
100	CL109	CL111	9	0.126651		29	CL32	CL70	7	0.444088	
99	CL112	CL121	6	0.127336		28	CL72	88	3	0.444569	
98	CL99	43	7	0.115690		27	CL40	138	5	0.448462	
97	CL107	87	3	0.127566		26	CL31	CL52	68	0.450890	
96	55	59	2	0.129281		25	1	CL33	5	0.454465	
95	CL97	107	4	0.129425		24	CL25	CL34	9	0.449824	
94	26	CL131	3	0.133475		23	83	110	2	0.478986	
93	CL128	27	3	0.140359		22	CL23	105	3	0.448499	
92	CL106	96	6	0.141334		21	CL27	CL29	12	0.495881	
91	6	16	2	0.142174		20	CL24	CL38	26	0.506572	
90	CL141	60	3	0.148106		19	CL41	CL65	9	0.513785	
89	CL103	29	7	0.148645		18	CL21	124	13	0.530377	
88	54	CL96	3	0.149893		17	CL20	CL55	32	0.544913	
87	CL88	109	4	0.143413		16	CL66	CL30	4	0.566147	
86	CL98	CL122	9	0.156463		15	CL19	58	10	0.568821	
85	67	90	2	0.156585		14	CL18	103	14	0.646199	
84	CL85	98	3	0.148620		13	CL17	CL26	100	0.665777	
83	CL93	14	4	0.157488		12	CL13	CL14	114	0.746119	
82	35	CL114	3	0.157509		11	CL22	119	4	0.761177	
81	99	140	2	0.157700		10	CL15	123	11	0.787753	
80	46	56	2	0.158468		9	CL28	CL39	5	0.789850	
79	CL100	CL124	11	0.163437		8	CL37	102	3	0.793081	
78	CL108	CL102	5	0.164240		7	CL12	CL8	117	0.801030	
77	CL79	CL92	17	0.173332		6	CL7	CL10	128	0.843667	
76	CL89	CL105	9	0.179257		5	CL16	142	5	1.020612	
75	CL110	CL142	6	0.179489		4	CL6	CL5	133	1.095481	
74	63	65	2	0.183032		3	CL4	CL9	138	1.254093	
73	CL104	CL91	5	0.189957		2	CL3	CL11	142	1.587360	
72	9	69	2	0.191314		1	CL2	108	143	1.740513	

Table V. Clusters Produced by the Acumulation Tree of Table IV at the 14 Cluster Level, i.e., when the Number of Clusters Becomes 13 in Table IV

IDNO	LAJNO	CASNO	SSS 2	CHIS 6	MOLC 7	KAPA 3	PATH 2	ALLP 1	TSCH 1	CLOGP 0
					Cluster = 1					
1	3	95396	1	3.874	0.517	2.083	83	231	1.043	2.610
11	4	383073	1	2.238	1.649	6.596	118	820	2.247	6.050
13	4	423825	1	1.477	1.649	5.857	113	741	2.247	4.990
21	4	1492871	1	2.524	1.035	5.538	79	378	2.247	3.470
24	4	1893523	1	1.724	1.317	5.327	95	528	2.247	3.980
46	4	3741773	1	4.982	0.992	3.213	69	178	2.247	4.330
53	4	5888335	1	5.740	1.730	1.473	115	333	1.426	4.090
56	4	7347195	1	5.576	0.992	5.018	76	250	2.247	4.770
63	4	15419940	1	3.999	0.387	4.496	133	488	1.987	3.870
65	4	16432818	1	4.594	0.387	6.094	142	626	2.328	4.310
68	4	17329792	1	1.972	0.986	4.899	77	351	2.247	4.054
87	4	25268773	1	1.323	1.698	5.627	111	703	2.247	3.920
95	4	48077958	1	1.980	1.803	6.980	117	780	2.247	2.820
97	4	49859703	1	2.227	1.471	6.612	99	561	2.247	2.350
104	4	54449740	1	5.141	1.033	4.250	135	475	1.372	5.080
106	4	58920313	1	2.030	1.698	6.359	115	780	2.247	4.410
107	4	59071102	1	1.601	1.483	5.579	104	630	2.247	4.760
111	4	65983315	1	5.581	0.694	2.525	186	732	1.535	2.680
112	4	66008682	1	1.405	2.135	7.090	132	990	2.247	3.660
113	4	66008693	1	1.653	1.803	6.584	114	741	2.247	3.190
114	4	66008706	1	1.900	1.471	6.178	96	528	2.247	2.720
118	4	66671225	1	4.918	1.380	6.817	72	270	1.861	3.490
120	4	67584558	1	1.817	1.035	4.646	75	325	2.247	3.520
121	4	67584569	1	1.694	1.200	4.855	84	406	2.247	3.760
122	4	67584570	1	1.570	1.366	5.087	93	496	2.247	4.000
130	4	68084628	1	1.446	1.532	5.354	0	595	2.247	4.230
131	4	68227974	1	2.153	1.532	6.112	0	666	2.247	4.180
132	4	68227985	1	2.277	1.366	5.878	0	561	2.247	3.940
133	3	68227996	1	2.401	1.200	5.689	0	465	2.247	3.710
134	3	68298066	1	1.848	1.151	5.087	0	435	2.247	4.290
135	4	68298602	1	2.362	1.483	6.343	0	703	2.247	5.820
139	4	72276052	1	1.733	2.135	7.460	135	1035	2.247	3.290
					Cluster = 2					
25	2	2156969	1	3.846	0.048	11.930	45	120	1.444	5.510
26	2	2156970	1	4.554	0.048	13.940	51	153	1.594	6.570
37	3	2664553	1	3.493	0.048	11.000	42	105	1.369	4.980
41	2	3076048	1	4.907	0.048	15.000	54	171	1.669	7.000
49	2	4813574	1	6.675	0.048	19.950	69	276	2.044	7.000
58	2	13048345	2	4.572	0.960	15.260	60	210	2.056	5.030
60	2	13402023	1	5.968	0.048	17.950	63	231	1.894	7.000
62	2	13533181	1	6.309	0.048	19.950	69	276	1.954	7.000
81	2	21643425	1	5.261	0.048	15.940	57	190	1.744	7.000
94	2	48076386	1	7.382	0.048	21.960	75	325	2.194	7.000
					Cluster = 3					
2	3	96333	1	0.727	0.048	3.000	21	21	0.742	0.750
3	3	103117	1	3.251	0.252	7.101	38	91	1.293	4.320
4	3	106638	1	2.228	0.456	5.878	28	45	0.994	2.200
5	3	106741	1	1.569	0.048	7.000	30	55	1.191	1.090
6	3	106901	1	1.790	0.166	3.556	34	60	1.151	−0.130
7	3	140885	1	0.956	0.048	3.840	22	28	0.850	1.280
8	3	141322	1	1.725	0.048	5.878	28	45	0.995	2.330
10	3	356865	1	0.451	0.402	4.152	40	91	2.693	2.380
12	3	407476	1	0.500	0.227	7.000	31	55	1.875	2.040
14	3	424646	1	0.327	0.568	3.769	48	136	3.516	2.620
15	3	689123	1	1.690	0.284	5.000	27	36	1.010	1.580
16	3	818611	1	1.095	0.048	5.000	25	36	1.256	−0.060
17	3	925600	1	1.372	0.048	5.000	25	36	0.920	1.800
18	3	937417	1	1.938	0.116	3.265	51	106	0.988	2.060
19	3	999553	1	1.123	0.048	5.000	25	36	0.808	1.260
20	3	1070708	2	2.450	0.096	9.373	42	105	1.607	1.860
22	3	1663394	1	2.738	1.160	5.878	28	45	1.171	1.980
27	3	2160896	1	0.484	0.408	4.889	45	105	3.061	3.100
28	3	2223827	2	3.788	1.303	7.875	45	120	1.755	2.130
29	3	2274115	2	1.743	0.096	7.438	37	78	1.505	1.250
30	3	2399486	1	2.469	0.166	3.265	46	100	1.255	0.990
31	3	2426542	1	2.398	0.206	6.250	37	78	1.244	1.550
32	3	2439352	1	2.175	0.364	7.000	31	55	1.107	0.650
33	3	2478106	1	1.802	0.048	7.000	30	55	1.356	0.350
34	3	2495354	1	2.297	0.166	4.000	52	123	0.912	2.520
35	3	2499583	1	1.786	0.048	9.000	36	78	1.219	3.920
36	3	2499958	1	2.432	0.048	7.901	33	66	1.144	3.390
38	3	2998085	1	1.860	0.215	4.500	28	45	1.081	2.110
39	3	2998234	1	2.079	0.048	7.000	30	55	1.070	2.860

Table V (Continued)

IDNO	LAJNO	CASNO	SSS 2	CHIS 6	MOLC 7	KAPA 3	PATH 2	ALLP 1	TSCH 1	CLOGP 0
40	3	3066715	1	2.901	0.166	3.265	51	106	1.143	2.780
42	3	3121617	1	1.364	0.048	5.878	28	45	1.083	0.560
43	3	3326907	1	1.979	0.177	5.531	31	55	1.498	0.310
45	3	3530367	1	2.607	0.166	4.688	55	141	0.977	2.840
47	3	3953104	1	2.517	0.252	5.289	34	66	1.143	3.260
48	3	4074888	2	2.356	0.096	10.290	45	120	1.846	1.320
51	3	5390545	1	2.141	0.209	5.325	37	78	1.070	1.770
54	3	7251903	1	2.338	0.048	9.000	36	78	1.336	2.150
55	3	7328178	1	2.181	0.048	9.917	39	91	1.532	1.160
57	3	13048334	2	3.157	0.096	11.320	48	136	1.756	2.920
59	3	13282821	1	2.916	0.177	9.373	42	105	1.899	1.330
61	3	13533056	1	1.708	0.048	7.901	33	66	1.598	0.010
66	3	16868136	1	2.547	0.166	2.651	45	85	1.068	2.220
67	3	16969101	1	3.112	0.245	6.370	63	201	1.892	1.670
70	3	17527310	1	0.761	0.402	4.889	42	105	2.696	2.010
73	3	17977092	1	2.525	0.588	3.960	43	105	1.925	0.670
74	3	18526073	1	2.528	0.364	7.901	34	66	1.172	0.310
75	3	18621766	2	2.140	0.096	9.373	42	105	1.489	1.310
76	3	18933921	1	3.081	0.623	6.400	34	66	1.232	3.040
77	3	19485031	2	2.612	0.263	8.082	43	105	1.698	1.640
78	3	19660163	1	3.553	0.614	5.531	31	55	2.247	2.140
79	3	19721370	2	2.356	0.096	10.290	45	120	1.702	1.810
82	3	23916338	1	1.403	0.048	5.878	28	45	0.889	1.790
84	3	24493536	2	2.097	0.096	8.333	39	91	1.538	1.330
85	3	24615847	1	1.515	0.113	7.000	31	55	1.645	0.330
86	3	24910847	1	2.597	0.375	5.531	31	55	1.168	1.860
90	3	30697406	1	3.358	0.280	5.878	81	300	2.454	1.700
92	3	37275471	2	3.635	0.803	6.479	51	153	2.217	0.670
93	3	44914036	1	2.348	0.337	5.531	31	55	1.068	2.730
96	3	48145046	1	2.533	0.116	5.481	58	160	1.329	2.500
98	3	51727505	1	3.541	0.340	7.259	69	246	2.213	2.590
100	3	52591272	1	0.514	0.734	4.250	60	210	4.345	2.479
101	3	52607815	1	1.758	0.090	7.438	37	78	1.897	0.590
109	3	63225536	1	2.788	0.090	10.290	44	120	2.113	2.170
115	3	66028306	2	2.885	0.225	9.679	52	153	2.007	0.910
125	3	67905082	2	3.971	0.359	7.934	54	171	1.988	3.620
126	3	67905413	2	4.555	0.504	6.667	88	282	1.821	3.440
127	5	67905480	2	3.355	0.225	10.850	50	153	2.242	1.360
128	3	67952492	1	3.435	0.337	8.333	0	91	1.293	4.320
					Cluster = 4					
23	5	1680213	2	2.968	0.096	13.290	54	171	2.188	1.380
44	5	3524683	3	4.094	0.850	8.889	64	231	2.788	−0.340
50	5	4986894	4	4.747	0.900	11.340	75	325	3.036	1.180
52	5	5459381	3	2.931	0.262	9.600	54	171	2.315	1.540
64	5	15625895	3	4.289	0.851	8.889	64	231	2.465	2.180
72	5	17831719	2	3.580	0.096	16.200	63	231	2.529	1.450
80	5	19778859	3	4.246	1.071	9.562	61	210	2.394	1.650
99	5	52408421	2	4.833	0.354	16.260	71	300	3.416	−0.160
103	5	53417291	2	3.440	0.803	6.479	51	153	2.540	−1.850
116	5	66028328	3	5.306	0.321	15.580	81	378	3.072	3.060
117	5	66028340	3	4.590	0.321	13.810	75	325	2.923	2.000
124	5	67893009	3	4.949	0.356	11.560	139	645	3.515	0.610
138	5	71412356	2	4.585	0.596	12.250	0	231	2.654	2.310
140	5	72928428	2	4.657	0.521	14.080	68	276	3.469	0.040
					Cluster = 5					
9	1	307982	1	−0.167	1.232	4.396	84	406	6.811	3.555
69	1	17527296	1	0.266	1.066	4.543	78	351	5.992	2.950
88	1	27905459	1	0.019	1.398	5.035	9	528	7.639	3.420
					Cluster = 6					
129	2	67952505	2	8.199	1.482	9.877	0	1381	3.168	6.490
136	2	70146053	2	9.111	1.753	9.671	0	1573	3.420	7.000
137	2	70495395	2	6.917	0.751	10.210	0	1260	3.057	5.690
141	2	84732285	1	6.448	0.579	8.945	0	1001	3.781	5.400
					Cluster = 7					
71	1	17741605	1	−0.228	1.729	5.600	114	741	9.286	3.890
91	1	34395249	1	−0.475	2.061	6.201	132	990	10.930	4.359
					Cluster = 8					
89	4	30145518	2	6.248	2.302	9.694	65	253	2.704	2.900
143	4	87320056	2	6.420	1.902	6.428	110	448	2.745	2.040
					Cluster = 9					
83	4	24447787	2	7.114	1.149	9.408	238	1217	2.857	5.870
105	1	56361558	2	8.339	1.149	13.270	282	1739	3.540	6.000
110	1	64448686	2	7.063	0.726	12.500	255	1426	3.830	3.420

Table V (Continued)

IDNO	LAJNO	CASNO	SSS 2	CHIS 6	MOLC 7	KAPA 3	PATH 2	ALLP 1	TSCH 1	CLOGP 0
					Cluster = 10					
119	4	66710972	2	11.060	2.247	9.057	274	1589	2.247	7.000
					Cluster = 11					
123	5	67892993	1	5.531	0.166	15.750	0	445	2.962	1.280
					Cluster = 12					
102	1	52723963	2	6.380	0.813	12.720	144	804	4.361	3.400
					Cluster = 13					
142	5	85412540	3	7.200	1.028	18.840	0	666	4.660	2.370
					Cluster = 14					
108	5	60506812	5	7.710	1.655	16.000	111	703	4.661	−0.840

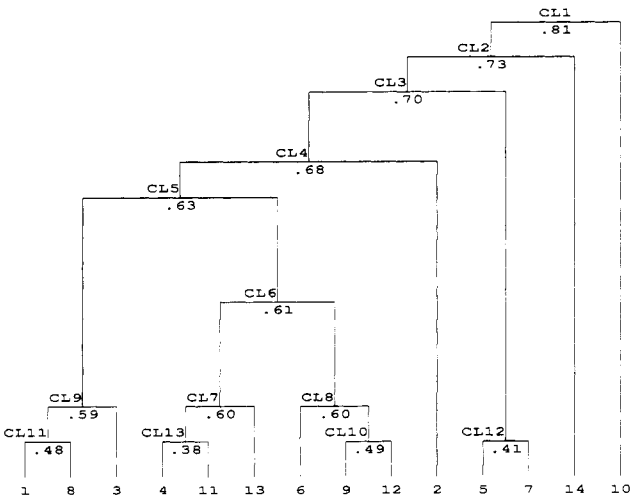

Figure 7. Clustering the 14 samples of Figure 6 depicts the relation among the 14 clusters.

choosing a cutoff at the 0.5 level. Here we get 10 clusters with just three singletons.

REMARKS AND CONCLUSIONS

The routine practice of assigning singleton clusters, i.e., outliers, to their closest cluster may be even worse than discarding them for the objective of sampling the data. The outlier can be inadvertently chosen as a representative sample of its cluster. The correct procedure would be to consider singletons separately, according to other nonstructural information relevant to the importance of the compound, e.g., its prevalence when testing for toxicity.

Hierarchical clustering allows a great deal of flexibility in selecting samples for toxicity testing. To increase the number of diverse samples one can simply lower the distance cutoff level. This will divide the clusters in a natural order.

We have shown shortcomings in the Lawson and Jurs work. These range from inexplicable errors to the choice of poor methods and neglect of appropriate methods. Their attempt to simplify a body of chemical data proved to be a lack of respect for its complexity.

ACKNOWLEDGMENT

SAS/STAT was used to do the principal components and the clustering. Karen Malley of ARC supplied programming assistance with SAS/GRAPH.

REFERENCES AND NOTES

(1) 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.
(2) Lawson, R. G.; Jurs, P. C. Clustering Analysis of Acrylates To Guide Sampling for Toxicity Testing. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 137–144.
(3) Hodes, L. Clustering a Large Number of Compounds. 3. The Limits of Classification. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 347–350.
(4) 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.
(5) See, e.g., Hodes, L. A Two-Component Approach To Predicting Antitumor Activity from Chemical Structures in Large-Scale Screening. *J. Med. Chem.* **1986**, *29*, 2207–2212.
(6) Errata. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 361.