The program eliminates unmatchable leaders each time the free virtual processors are exhausted. If enough processors are reclaimed, then processing continues without increasing the number of virtual processors. Otherwise, the number of virtual processors is increased as described above.

This final optimization of the program used only 1M virtual processors, and run time was reduced by more than a factor of 2 on our data.

THE APPLICATION

This work was performed using 230 092 compounds from the National Cancer Institute Development Therapeutics Program Repository. The 116 706 resulting cluster leaders comprise a resource of diverse compounds for testing in a new screen that is now in operation. Such a large set would be used in conjunction with a program to test for activity and novelty in the screen. This latter program can be based on sufficient early testing in the screen as described in ref 5. If desired, other members of selected clusters can be retrieved.

The estimated CPU time for this job on an IBM 3090 mainframe was 20–34 h. The total run time (wall clock time) was 2 h, 35 min on the 16K Connection Machine. If a larger Connection Machine with more processors had been used, the time would have been reduced to about 1 h, 18 min (for a 32K machine) or 39 min (for a 64K machine).

The use of cluster leaders as input to a prioritizing program circumvents the bunches of similar compounds that would effectively prevent the collection of a diverse subset. Moreover, the set of cluster leaders can be used repeatedly for different biological tasks.

It is now almost feasible to do a comprehensive literature surveillance by clustering the Chemical Abstracts Service 10 million compound file. This job is estimated to take about 800 h on the large Connection Machine. Fortunately, like the application reported here, it would only need to be done once.

ACKNOWLEDGMENT

Connection Machine is a registered trademark of Thinking Machines Corporation. Sun is a registered trademark of Sun Microsystems, Inc. IBM is a registered trademark of International Business Machines Corporation. We thank Henry Dardy for providing access to the Connection Machines at the Naval Research Laboratory.

REFERENCES AND NOTES

(1) 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.

(2) Hillis, W. Daniel. *The Connection Machine.* MIT Press: Cambridge, MA, 1985.

(3) Belloch, G. E. Scans as Primitive Parallel Operations. *Proceedings of the International Conference on Parallel Processing*; 1987.

(4) Hillis, W. Daniel; Steele, Guy L., Jr. Data Parallel Algorithms. *Commun. ACM* **1986**, *29* (12), 1170–1183.

(5) Hodes, L. Computer Aided Selection for Large Scale Screening. In *Comprehensive Medicinal Chemistry, Vol. 1. General Principles*; Hansch, Sammes, Taylor, Eds.; Pergamon Press: New York, 1990; Chapter 3.3; p 279.

Clustering a Large Number of Compounds. 3. The Limits of Classification

LOUIS HODES* and ALFRED FELDMAN

National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892

Received December 18, 1990

Clustering is normally used to group items that are similar. In this application of obtaining a diverse sample from the 230 000 compounds in the National Cancer Institute Repository, we cluster to select compounds that are different from the rest, to optimize screening for new leads. With these constraints, our approach yielded many singleton clusters. We can interpret these results as evidence for a limit to classification, contrary to the customary view of chemistry as a study of classes of compounds.

INTRODUCTION

The first paper[1] announced the intention of clustering a set of over 230 000 diverse compounds. The objective was to extract a representative sample of compounds from the National Cancer Institute file for testing in the new NCI primary screen.[2] That objective was accomplished with the aid of a Connection Machine.[3]

This paper makes some general observations on the classification of chemical structures, based primarily on the results of the large clustering. Our results show that a diverse set of compounds will exhibit a dual nature, some clusters and some scattered.

We also showed that the systematic clustering method can provide examples for an attempt at classifying the file. In the reported case, less than half the file was successfully classified in this way; the remainder being probably too diverse to classify.

CLASSIFICATION AND CLUSTERING

Classification of compounds is a form of clustering. By performing reasonable clustering on sets of diverse compounds we have obtained evidence with respect to the validity of classification. One can imagine that, if compounds fall into natural classes, then this phenomenon would show up upon clustering a large set of compounds, even if they are diverse. Instead, we get a persistent occurrence of singletons in addition to the expected increase in large clusters.

These singletons are important for testing novel compounds in our screening program. In contrast, Willett[4] avoids singletons by assigning them each to its closest cluster. Lawson and Jurs[5] achieve a similar effect by their choice of clustering method.

This forcible treatment of singletons is an example of an implicit belief that compounds ought to belong to classes. Clustering is a means for classifying, and classifying has worked well in many areas of chemistry.

Perhaps the most fundamental achievement in chemistry was the classification of elements according to the periodic table. The study of organic chemistry under the traditional, or nonsystematic, nomenclature is heavily linked to classes. At the macro level, there is the classification of biological species. Back in chemistry, the classification of natural products derives from that of species.

However, chemists can synthesize almost any structure satisfying valence and geometry constraints. Therefore, it is not completely surprising that there are so many singletons in our results. Thus, there is an evident divergence to an unlimited number of "classes". This becomes obvious when one considers the possibility of synthesizing arbitrary compounds, restricted only by valence and geometry.

METHODS

Please see ref 1 for a detailed description of the variable sized molecular fragments used to represent compounds. Generally, each bond in the molecule is used as the center of a fragment, with the more common carbon bonds extending to larger fragments.

Each distinct fragment i in compound j gets a weight $w(i,j)$ which depends on its multiplicity in j, its size, and its incidence in the file. The weight of a compound is the sum of its distinct fragment weights, $\sum_i w(i,j)$. For two compounds, j and k, we use the weight of the fragments in common divided by the weight of the weightier compound as a similarity measure to determine matching. This measure has the formula $\sum_i \min[w(i,j),w(i,k)]/\max[\sum_i w(i,j),\sum_i w(i,k)]$. When the similarity measure exceeds a preset fraction or cutoff threshold the compounds j and k are a successful match.

Thus, the clustering method involved a similarity measure between two compounds and a threshold cutoff on the similarity measure in order to determine whether the two compounds matched. The cutoff was set at a stringent level so that the weight of the molecular fragments in common must exceed 65% of the weight of the weightier compound.

The reason for such a restriction on matching was that small differences in molecular structure can yield large differences in activity. It was felt that the sheer size of the set of compounds should greatly increase chances of matching at a fixed threshold cutoff. And moreover, the large sample resulting from the fine clustering can be prioritized by running through programs that predict activity and estimate novelty.[6]

Please see the earlier papers[1,3] for a more complete discussion of the methods, including use of the leader algorithm for clustering. The leader algorithm is perhaps the only way to deal efficiently with this amount of data. The leaders become the natural cluster representatives, and by sorting the file in the order of compound weight, the leaders are in some sense a simplest member of their clusters.

The first study will compare results on the 230 000 compounds with those on the initial sample of 5000 compounds and an intermediate set of 50 000 compounds. Although the degree of clustering increases substantially with the number of compounds, there are still large numbers of singletons at the 230 000 level.

We concede that our matching criterion of fragments in common is not the same criterion a chemist would use to detect matches. A chemist would consider certain structural parts to be more important than others, depending on past experience. We can say that the chemist would be doing intuitive or traditional clustering. This can be considered closer to classifying. Our algorithmic or mechanical clustering serves as an approximation.

Let us call an error of type M a match that most chemists would dispute and an error of type F a failure to match where most chemists would match. Our strict matching criterion with 65% threshold has practically eliminated errors of type M at the expense of errors of type F. For example, in the cancer file, the several hundred anthracyclines split into perhaps a dozen clusters. Some of them are singletons if the side chains are sufficiently novel.

Table I. Comparison of Overall Statistics

no. of compds	4979	49005	230092
overlap	21	1392	10151
apparent compds	5000	50397	240243
% overlap	0.4	2.8	4.4
clusters	3785	27453	116706
av compds/cluster	1.32	1.83	2.06
median cluster size	1	2	3

Table II. Incremental Clusters and Compounds for Three Runs

compds/cluster	5000 compds		50 000 compds		230 000 compds	
1	3 053	3 053	18 283	18 283	73 757	73 757
2	492	984	4 710	9 420	20 399	40 798
3–4	177	578	2 790	9 285	13 183	44 122
5–8	58	338	1 318	7 870	6 405	38 648
9–16	5	47	230	3 720	2 277	25 695
17–32			75	1 516	584	12 509
33–64			6	231	93	3 860
65–128			1	72	7	582
129–256					1	248
total	3 785	5 000	27 413	50 397	116 706	240 243

A study of a separate set of 12 900 diverse compounds collected to test against HIV showed that even when we lower the cutoff threshold down to 20% we still get errors of type F. But here there are many errors of type M, as evidenced by a large increase in overlapping clusters. In the following section we report results on this set of compounds at 20%, 30%, 40%, and 50% cutoff thresholds.

This set of compounds was also subjected to a more traditional classification process, so we can compare the two types of results.

STUDY 1: THE LARGE CANCER FILE

Table I presents some overall statistics from three runs using about 5000 compounds, 50 000 compounds, and the entire collection of 230 000 compounds, all performed under the same 65% threshold matching criterion.

A sign of good clustering discussed in ref 1 is a relatively small amount of overlap among the clusters. The fragment weight options were chosen so as to minimize overlap in the initial 5000-compound run. This overlap, measured as the number of multiple matching of compounds, was down to 0.4% of the total matching. Notice that merely increasing the number of compounds in Table I without changing the matching threshold greatly increases the rate of overlap. It jumps to 2.8% at 50 000 compounds and 4.4% at 230 000 compounds.

Table II presents a breakdown of cluster statistics by cluster size, i.e., the compounds per cluster. We list the number of new clusters and the number of compounds in them as we step up the cluster size by factors of 2. The first line starts with the singleton clusters; the second line gives pair clusters; then the next line includes clusters of size 3 and 4; etc. For example, there are no clusters larger than size 16 in the 5000-compound set, but there are 75 clusters from size 17 to size 32 in the 50 000-compound set and 584 such clusters in the 230 000-compound set. This is another example of the changing nature of the clustering results due solely to an increase in the number of compounds in a diverse set.

Table III presents the same data as Table II in an upside down cumulative manner. Thus, we begin with a total for all the clusters, and at each level we note the remaining number of clusters and compounds. The presentation in Table III provides an interesting graph in Figure 1.

Figure 1 is plotted on a log scale, and the space between the remaining clusters and the remaining compounds is filled. The subtraction, compounds minus clusters, gives the height of the filled graph. Since these numbers are scaled as logarithms, this difference (the height of the graph) is also the log

Table III. Remaining Clusters and Compounds for Three Runs

compds/ cluster	5000 compds		50 000 compds		230 000 compds	
1	3 785	5 000	27 453	50 397	116 706	240 243
2	732	1 947	9 170	32 114	42 949	166 486
3–4	240	963	4 460	22 694	22 550	125 688
5–8	63	385	1 670	13 409	9 367	81 566
9–16	5	47	352	5 539	2 962	42 918
17–32			82	1 819	685	17 223
33–64			7	303	101	4 690
65–128			1	72	8	830
129–256					1	248

Table IV. Overall Statistics on 12900 AIDS Compounds

% match	20	30	40	50
overlap	5898	2607	1276	694
apparent compds	18798	15507	14176	13594
% overlap	31.5	16.8	9.0	5.1
clusters	2658	4182	5600	6790
singletons	778	1906	3209	4463
compds/cluster	7.6	3.7	2.5	2.0

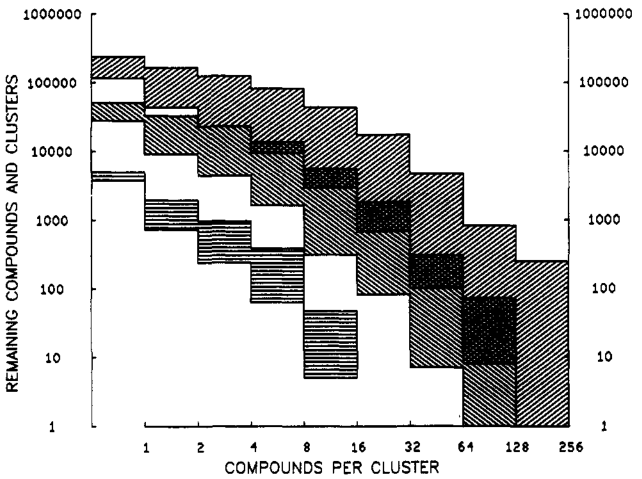

Figure 1. Log plot of the data in Table III for three runs on the cancer file. The height of the boxes yields the remaining compounds per cluster after each step.

of the ratio (compounds divided by clusters, or compounds per cluster).

Thus, the height of the filled boxes at any step is a measure of the log of the remaining compounds per cluster. For example, the initial heights show the overall compounds per cluster at the three levels. These are the same numbers, 1.32, 1.83, and 2.06, that occur in Table I. The height of the boxes between 1 and 2 gives the log of the compounds per cluster for all clusters of size greater than 1 and so forth.

STUDY 2: THE AIDS FILE

This study contrasts the clustering method with traditional work in classifying compounds. For this purpose we used 12 900 compounds tested in the anti-HIV screen.

In our earlier results, the usual chemical classes were split up to a large extent. Matching was influenced by fragment features other than the "central" ring or other features that normally characterize the compounds.

This splitting was largely caused by our deliberate use of a high-matching threshold. Perhaps by setting a lower threshold we can approximate the more traditional classification.

Thus, the 12 900 compounds were clustered at matching thresholds of 20%, 30%, 40%, and 50%. Table IV shows some overall results, with a breakdown of the cluster statistics shown in Table V and Figure 2.

Table V. Incremental Clusters for 12900 AIDS Compounds

compds/ cluster	% match			
	20	30	40	50
1	778	1906	3209	4463
2	476	795	1060	1120
3–4	446	675	686	693
5–8	385	424	392	355
9–16	316	254	188	125
17–32	163	89	49	27
33–64	72	32	12	6
65–128	14	7	4	1
128–256	8			
total	2658	4182	5600	6790

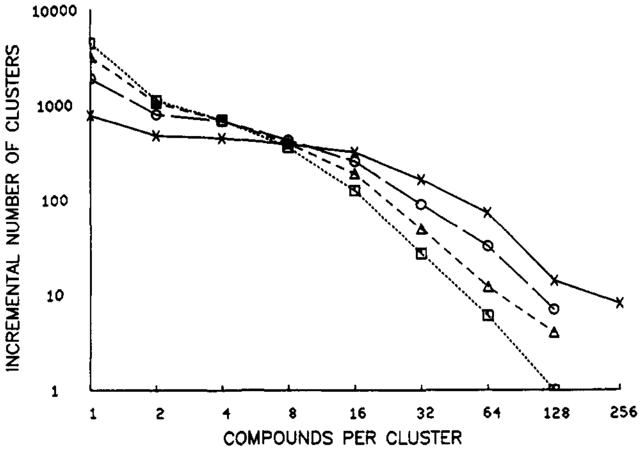

Figure 2. Log plot of the data in Table V for all four runs of the HIV compounds.

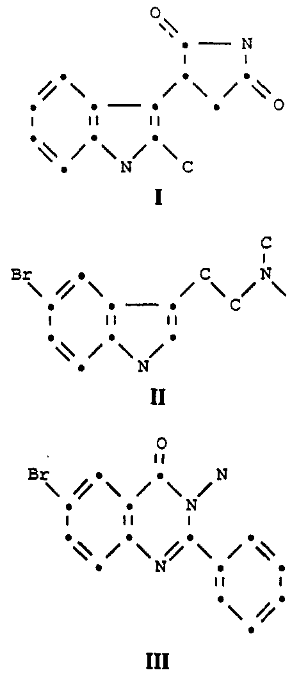

Figure 3. Example showing both types of mismatches at the 20% threshold cutoff level. Compounds I and II fall in different clusters, but compounds II and III do match.

At the 50% threshold there was a reasonably low 5% overlap, but a large number, 4463, of singletons. At the 20% threshold there was a much smaller number, 778, of singletons but a 31% overlap. There is no obvious satisfactory threshold where a good compromise between overlap and singletons would indicate a natural level for classification.

Indeed, with the matching threshold set at 20% there are still many examples of 'like' compounds split, errors of type F. For example, the two indoles, compounds I and II of Figure

3, are separated. Meanwhile, compound III, which does not belong to that class, was placed in the cluster with compound II, a more expected error of type M.

A traditional form of categorizing this set of compounds was performed as follows. The 517 cluster leaders of clusters bigger than size 9 were perused to identify chemically central features with large representation. These 25 or so features were then used for substructure search on the whole file of 12 900 compounds.

The 25 classes obtained in this way ranged in size from 100 to 400 compounds and together comprised about 40% of the 12 900 compounds. By a comparison to Table V, we see that these classes are broken up a lot. Even at the 20% threshold where the clusters are not so pure, it is clear that those sizes cannot be accommodated. So, the machine clustering can be used as a first step in a more traditional classification procedure. It may be that substantially more extensive categorization would be difficult due to diversity.

From this study we shed some light on the difference between mechanical and traditional clustering. Again we see it is difficult to classify a set of diverse compounds.

REFERENCES AND NOTES

(1) 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.

(2) Boyd, M. R. National Cancer Institute Drug Discovery and Development. In *Accomplishments in Oncology. Vol. 1, No 1. Cancer Therapy: Where Do We Go From Here?* Frei, E. J., Freireich, E. J., Eds.; J. B. Lippincott: Philadelphia, 1986; pp 68–76.

(3) Whaley, R.; Hodes, L. Clustering a Large Number of Compounds. 2. Using the Connection Machine. *J. Chem. Inf. Comput. Sci.* **1991**, *31* (preceding paper in this issue).

(4) Willett, P.; Winterman, V.; Bawden, D. Implementation of Nonhierarchic Cluster Analysis Methods in Chemical Information Systems: Selection of Compounds for Biological Testing and Clustering of Substructure Search Output. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 109–118.

(5) Lawson, R. G.; Jurs, P. C. Cluster Analysis of Acrylates to Guide Sampling for Toxicity Testing. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 137–144.

(6) Hodes, L. Computer Aided Selection for Large Scale Screening. In *Comprehensive Medicinal Chemistry. Vol. 1. General Principles*; Hansch, Sammes, Taylor, Eds.; Pergamon Press: New York, 1990; pp 279–284.

COMPUTER SOFTWARE REVIEWS

Winword

CLEMENS JOCHUM

Beilstein Institute, 40-42 Varrentrappstrasse, D-6000 Frankfurt, Germany

Received January 30, 1991

I have to confess that I have always been a Wordstar[1] addict. When I started working with personal computers in the late seventies under the CP/M operation system, Wordstar was the only decent word-processing program around. When I switched to IBM-compatible DOS-based PC's I was glad that Wordstar was also ported (although it worked slower than on my 4-MHz Z-80 machine!) and I did not have to learn a new word-processing program.

In the following years the competition for Wordstar became continuously stiffer. Within a short time other word-processing programs such as Wordperfect,[2] Microsoft Word (for DOS),[3] or Chemtext[4] had surpassed Wordstar in features and speed. They did not, however, offer so many more features that I considered it worthwhile switching. (Apparently Wordstar Corp. was convinced that they had only customers like myself, i.e., who are too lazy to learn a new word-processing program, so the company never did very much to improve Wordstar.)

During the last several years my colleagues and I tried out various other word-processing programs, but, as mentioned above, we could not convince ourselves to switch. Having a large local area network (LAN) with several servers, a switchover would mean that all other word-processing users at the Beilstein Institute would have to standardize on a new program, all Wordstar files would have to be converted, etc.

My initial experience with Word for Windows (Winword) version 1.0[5] was mixed. On one hand I liked all the features that make Winword not only a very good word-processing program, and it also contains many features usually found only in Desktop publishing (DTP) programs. On the other hand Windows 3.0 was not yet available and running it under Window-286 or Windows-386 (version 2.11) was cumbersome since one was constantly running out of memory. The whole situation changed with the release of Windows 3.0. This graphics user interface (GUI) and Winword (now version 1.1) are such a powerful combination (provided that you have the right hardware) that I finally said goodbye to Wordstar.

Let me begin with the negative points: You need fairly powerful hardware to run Winword efficiently. You should have at least a 12-MHz 80286-based computer (it is more fun to use Winword on a 20-MHz 80386 machine) and at least 2 MBytes of RAM. 4–8 MBytes of (extended—not expanded!) memory make all Windows-based programs much more efficient since several MBytes can be assigned to Windows disk-caching device driver Smartdrive. This speeds up Winword considerably. This review was written with Winword on a Compaq SLT-286 with 3.5 MBytes of RAM whereby 1 MByte was assigned to Smartdrive. The speed is absolutely sufficient for my two-finger-typing.

Another negative aspect of the program is that the very powerful Basic-like Macro language is not described in the manual. It is only described in the technical manual which has to be ordered separately. If you are an experienced Basic programmer you can figure out most of the language by simply using the sample macros. The macros also allow you to completely customize the user interface. You remove any command from any pull down menu or add other ones which you have written yourself with the Word-Basic macro language. This way you can customize one version for your secretary, one for scientific writing, one for programming, etc. However, I would not recommend changing the pull down