Balancing Representativeness Against Diversity using Optimizable *K*-Dissimilarity and Hierarchical Clustering

Robert D. Clark* and William J. Langton

Tripos, Inc., 1699 South Hanley Road, St. Louis, Missouri 63144

Received June 6, 1998

When assessing the pharmacological potential of large libraries of compounds, it is often useful to start by determining the biochemical activities of some subset thereof. This is so whether the compounds in question have in fact already been synthesized or exist solely as virtual libraries. A suitable subset for this task must be structurally diverse, so as to minimize redundant testing, but must also be representative, so that valuable subgroups do not get overlooked. These two needs are intrinsically in conflict, with gains in one necessarily coming at the expense of the other. Results obtained using optimizable *K*-dissimilarity selection and clustering are described and compared with those obtained using more traditional agglomerative hierarchical clustering techniques.

When setting out to identify new lead chemistries active against some biochemical target, it is generally considered desirable to include a wide range of structural classes in the initial screening. This impulse together with the steady growth in the size of corporate compound databases, the advent of combinatorial chemistry, and the increasing commercial availability of screening libraries has prompted the creation of various tools for automated maximization of structural diversity in screening libraries. Once these tools began to be widely deployed, however, it became apparent that it is rather easy to get too much of a good thing where diversity is concerned.

For one thing, most diversity measures currently in use are biased in favor of large, complex molecules[1] which are less likely to make good primary leads. Then, too, though novel lead compounds certainly offer more opportunity for drug development from a patentability perspective, it is also the case that many are unusual for good reason—they may be hard to make, toxic, unstable, or too promiscuous in their biochemical effects. Substructural filters play a critical role in library management by excluding many such obviously unproductive chemistries, but removing them all without also eliminating offbeat but potentially valuable leads is not practical.

Simply assaying everything is generally not an economically viable approach. The alternative of assaying a random subset of everything is appealing, in part because it will often produce more hits than will a maximally diverse subset; this is particularly so when corporate databases are used to carry out retrospective analyses. Many of the hits obtained in this way will, of course, be redundant to a greater or lesser degree because they will be analogs synthesized as follow-up to some central lead from an earlier program (legacies of the "methyl ethyl butyl futile" syndrome). Where new targets are closely related to historical ones—e.g., looking for inhibitors of new 5HT subtypes or profiling a database being considered for acquisition—random sampling may be a logical and productive approach. It is not likely, however, to turn up structurally novel leads and is hard to justify when making selections from virtual libraries.

What is really needed for discovery screening is a way to identify library subsets which are diverse but still representative of the universe of potential drugs. In general, the compounds comprising such an array should be distinctive yet not bizarre, so that, taken together, they efficiently embody the potentiality of the entire library. Assaying such a subset would be an efficient way to obtain information about relationships between structures and activities for the entire library.[2] That is, one would like to be able to select compound subsets from realized or virtual libraries which would serve a purpose analogous to that of the U.S. Senate, which is intended to be an assemblage of distinguished and distinctive people who can together represent the interests of the nation as a whole.

The relevant dichotomy can be characterized as a balance between representativeness and diversity. In this context, a *representative* subset is one in which the distribution of members mirrors in some sense the distribution found in the population as a whole. Members of a *diverse* subset are readily distinguishable from one another—i.e., they are relatively dissimilar to each other. Problems arise when members of the subset bear so little similarity to other compounds in the population that assaying them provides only idiosyncratic information which cannot be extrapolated reliably to many other compounds.

When medicinal chemists are asked to create representative diverse subsets from chemical libraries, they typically proceed by sorting and clustering the constituent compounds based on substructures and some sense of bioisosteric groups. Indeed, hierarchical clustering based on molecular fields[3] or on 2D structural fingerprints[4] have both been shown to give intuitively satisfactory results; moreover, some clustering methods seem to give quantitatively better results than do others.[5,6] One thrust of this report is to explore how the hierarchical clustering method used affects the trade-off between representativeness and diversity.

* Corresponding author: e-mail: bclark@tripos.com. phone: 314-647-1099. fax: 314-647-9241.

10.1021/ci980107u CCC: $15.00
Published on Web 10/02/1998

Optimizable *K*-dissimilarity selection (OptiSim,[7] for short) is an alternative selection method in which each new selection is made from a subsample of *K* compounds drawn at random from the population. As was shown in the initial report,[8] varying *K* provides a direct and natural way to shift the balance between representativeness and diversity: low values of *K* produce more representative selections, whereas larger subsample sizes give more diverse selection sets. Here we go on to demonstrate that OptiSim selection sets can be used to cluster datasets in a useful and meaningful way, and to compare the distribution profiles obtained from such *OptiSim clustering* to those seen for hierarchical clustering. Interactions between subsample size *K* and the size *M* of the selection set are also discussed.

METHODOLOGY

Construction of the Test Set. A structured test set was constructed from three combinatorial sublibraries which were generated using Legion[9] as described in detail previously.[8] All members from a 6000-compound 2,3,4,6-substituted pyridine combinatorial were included in the full library, whereas the 500 1,3- and 1,3,4-substituted cyclohexanes and 100 2,4- and 2,4,5-substituted pyrimidines included were drawn at random from homologous 2244-compound combinatorials. Each class in the resulting 6600-member mixed library covers a similar (hyper)volume of structural space despite a wide range of representation—from 91% for pyridines to 7.6% for cyclohexanes and 1.5% for pyrimidines. The individual combinatorials are relatively well-separated in fingerprint space. Hence, by construction, pyridines are the most redundant class and so are most representative of the library as a whole but are the least diverse *as a group*. Pyrimidines, in contrast, are the least redundant and so are diverse but poorly representative *as a group*. Cyclohexanes exhibit intermediate properties.

For hierarchical studies, a 1000-member subset was drawn at random from the full mixed library. This subset was comprised of 892 pyridines (89%), 92 cyclohexanes (9%), and 16 pyrimidines (1.6%).[8]

Optimizable K-Dissimilarity Selection.[8] The OptiSim selection set is "seeded" with a compound selected at random from the dataset of interest. A subsample of *K* compounds is then drawn at random from the dataset for each subsequent selection step. That compound in each subsample which is most different from those which have already been selected is itself added to the selection set. Only compounds which exceed some preset minimum dissimilarity *R* with respect to those already selected are considered at each step, and the process is continued until a preset number of compounds *M* have been selected, or until no valid unselected candidates remain in the dataset. The redundancy threshold *R* generally corresponds to the neighborhood radius[10] for the metric upon which the selection is based.

Optimizable K-dissimilarity selection was carried out using ChemEnlighten[9] 1.0a or in a prerelease 2.0 version. In all commercial implementations of the method, subsamples are drawn from the population without replacement until each compound has been considered once for inclusion in the selection set. The default *R* value of 0.85 was used as the maximum acceptable Tanimoto similarity coefficient with respect to UNITY[9] 2D fingerprints. At this stringency, 5−10% of the compounds in each subsample from the mixed library typically are rejected as being too redundant when 500 compounds are being selected.

For selection series across subset and subsample sizes, a different random number seed was used for each value of *K*, but that same random number seed was used for all subset sizes.

K-Dissimilarity Clustering. As noted in the original publication,[8] it is straightforward to partition a population by creating clusters centered on the compounds in an OptiSim selection set. In particular, each compound in the population can be assigned to the cluster containing the selected compound to which it is most similar. For the work described here, datatables bearing OptiSim selection subsets were exported from ChemEnlighten to Molecular Spreadsheets in SYBYL;[9] simple macro scripts were then run to create the desired OptiSim cluster columns.

Note that OptiSim clustering is inherently fast, since it scales with $M \times N$, where *M* is the number of compounds in the selection set and *N* is the number of compounds in the entire population.

RESULTS

Categorical Distribution across Clusters. Different clustering methods can be compared qualitatively by clustering the selection sets obtained with each and examining how compounds from different structural classes are distributed across the corresponding dendrograms. Sets of 30 compounds were obtained from the 1000-member subset using the single-linkage, group average, and complete-linkage methods[11] in SYBYL Selector.[9] Each set of 30 compounds was then itself clustered to give the dendrograms shown in Figure 1; for the sake of consistency, this secondary clustering was done using complete linkage in each case. The class of compound making up each cluster is indicated at the base of the dendrogram, where each descending line corresponds to one selection and, hence, to one cluster.

Single linkage is clearly the most biased toward unusual structures (pyrimidines and cyclohexanes) and so gives the most diverse but least representative selection. Indeed, many of the single linkage clusters are pyrimidine singletons, and the 89% of the dataset comprised of pyridines are lumped together into just a few clusters (six out of 30 clusters = 20%). Complete linkage clustering, on the other hand, is skewed toward a more reasonable representation of the more common class of compounds (pyridines) and is correspondingly less diverse. Note, however, that the complete linkage selection is still significantly enriched in cyclohexanes (8/30 = 27%) and in pyrimidines (3/30 = 10%) versus their representation in the population as a whole (9 and 1.6%, respectively). Group average clustering provides an intermediate balance in discrimination.

Figure 2 shows analogous dendrograms obtained by applying hierarchical clustering to OptiSim selections made at $K = 1$, 3, 5, and 10. At the smallest subsample size (which corresponds to minimum dissimilarity selection[8]), the selection is almost purely representative. One pyrimidine and two cyclohexanes were selected, in quite reasonable agreement with their contributions to the population. The selection sets for $K = 3$ and $K = 5$ have the same relative distribution across the three compound classes, but a careful examination of the branching patterns in the respective dendrograms

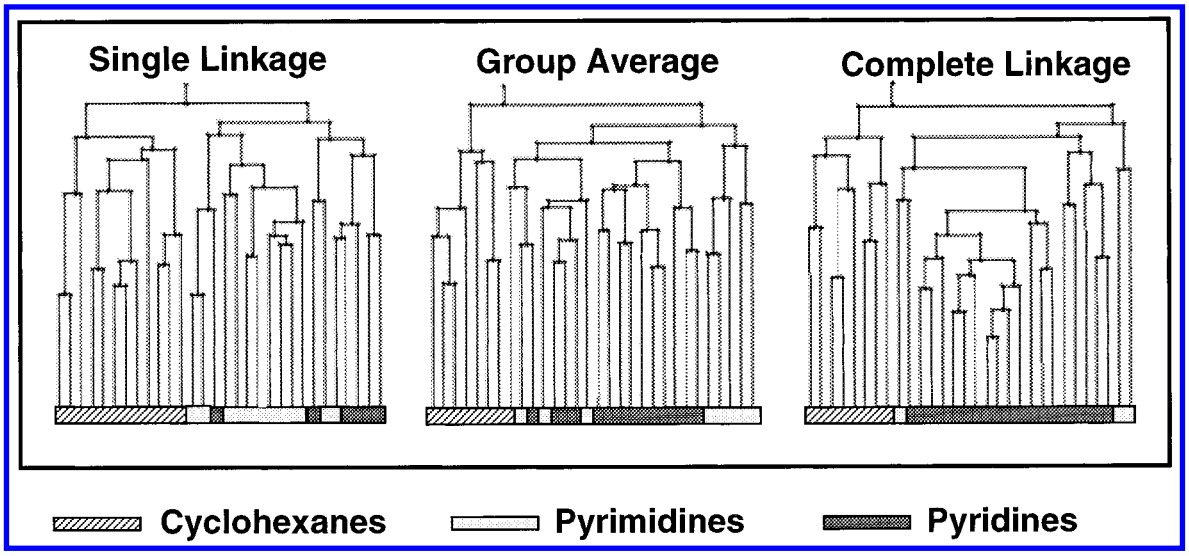

Figure 1. Dendrograms for different hierarchical cluster-based selection methods. The 1000-member subset was divided into 30 clusters using the single linkage, group average, or complete linkage method, and a representative was chosen at random from each cluster. The selection sets obtained were then themselves clustered using complete linkage to give the dendrograms shown. Stippled bars at the base of the dendrogram indicate clusters from which a pyrimidine was selected; hashing indicates cyclohexanes; and dark bars indicate pyridines.

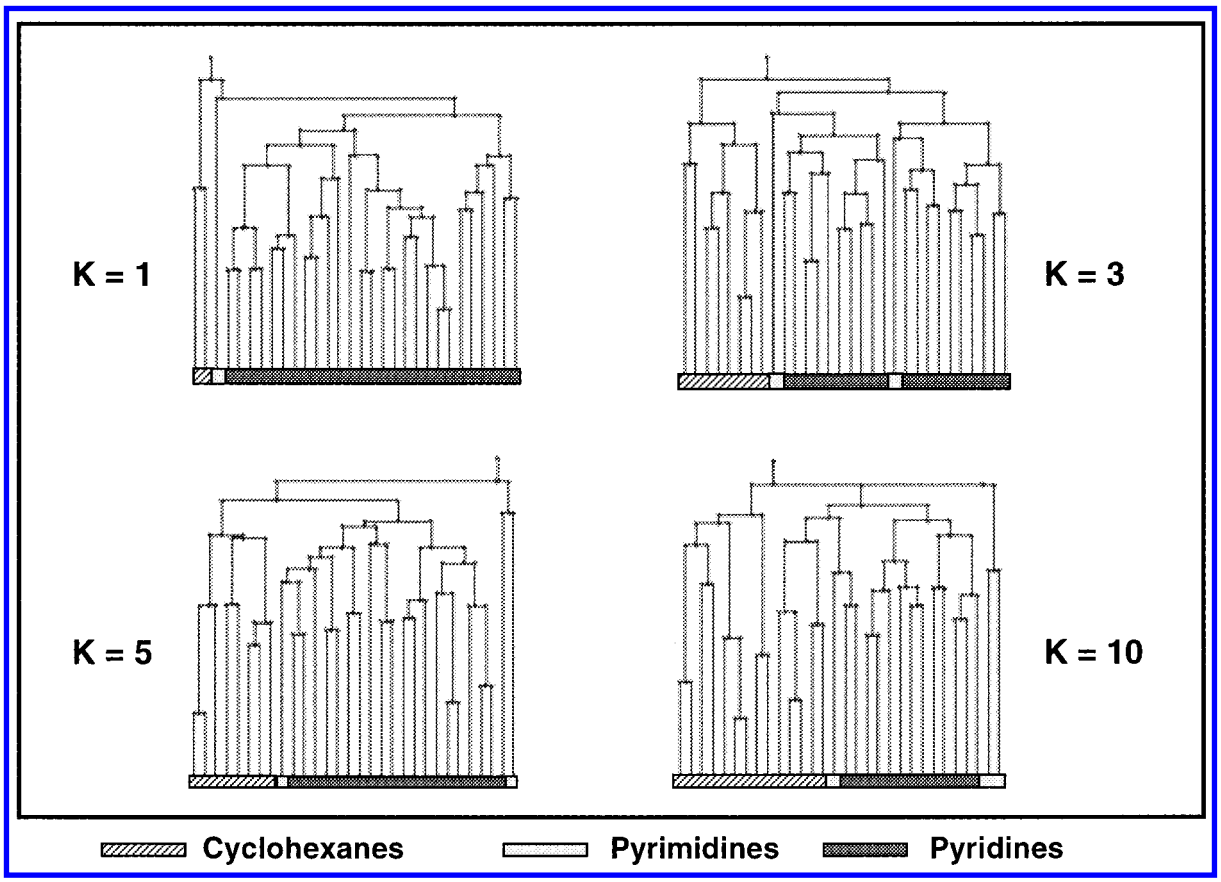

Figure 2. Dendrograms for OptiSim clustering. OptiSim was used to select 30 compounds from the 1000 member subset using $K = 1$, 3, 5, or 10. Each selection set was clustered using the complete linkage method to give the dendrograms shown. The class of compound selected for each cluster is as indicated for Figure 1.

indicates that the latter is most similar to that obtained for selections made using complete linkage (Figure 1). The $K = 10$ OptiSim selection set is clearly more skewed toward cyclohexanes, however, and hence is more diverse but less representative than is complete linkage. These results are in good qualitative agreement with those obtained using parametric measures of representativeness and diversity.[8]

Distribution Profiles. Dendrograms are good tools for seeing how compounds within a selection set relate to one another and thereby provide a good sense of a selection's diversity. They are not as effective at conveying the representativeness of a selection set, however. For that, it is important to know how many other compounds each selection represents.

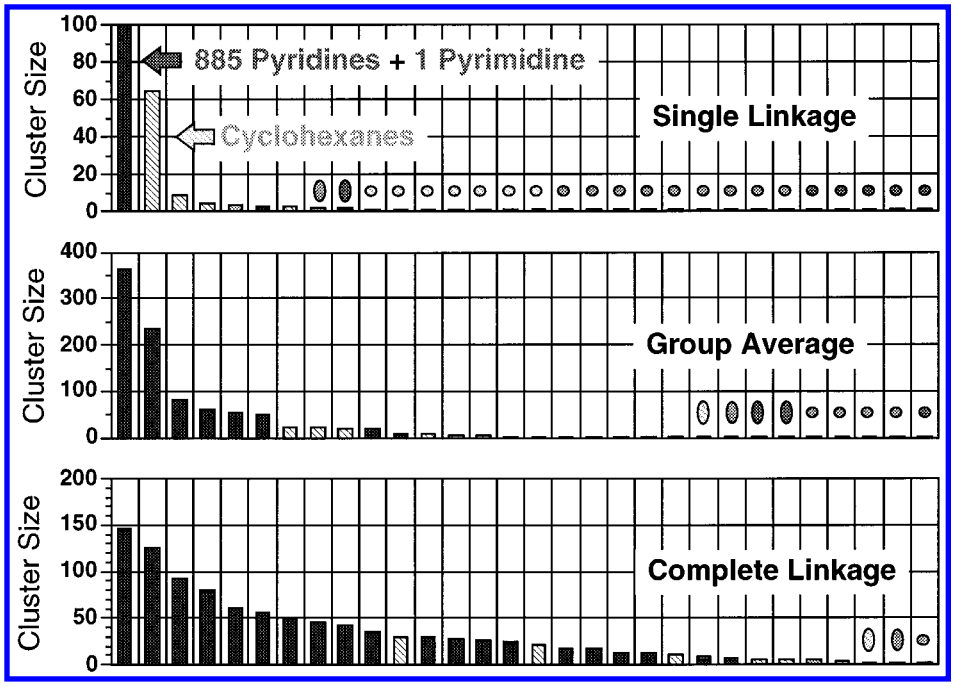

Figure 3. Distribution of cluster sizes for different hierarchical clustering methods. The number of members in each cluster from the analyses described in Figure 1 are shown, sorted from largest to smallest for each method. Bars are coded as for Figure 1. Circles indicate singletons and ellipses indicate doubleton clusters, respectively.

It is possible to expand dendrograms by directly incorporating cluster size information, but such plots are hard to compare meaningfully. It is easier to step back and simply look at the distribution of cluster sizes, bearing in mind that "size" in this context refers to membership count, not to how much structural space is covered by each cluster. Distribution profiles for single linkage, group average, and complete linkage clustering are shown in Figure 3. Again, pyridines are indicated by dark bars, whereas cyclohexanes are indicated by hashed bars and pyrimidines are indicated by stippled bars. Cluster sizes are shown in descending order, and, because the coding of the small bars is not decipherable, circles have been added to indicate singleton clusters; ellipses have been added to indicate doubletons. Note the differences in scale among the different plots.

For single linkage (top), most of the pyridines (more than 88% of the entire dataset) constitute a single large cluster, whereas the second largest cluster is comprised of only 65 compounds. In addition, more than two-thirds of the clusters are singletons. Such a distribution profile indicates a very uninformative selection set, because assaying compounds from a singleton cluster will provide little or no information about any other compound in the population. In addition, very large clusters are unlikely to be adequately represented by any single compound. The single linkage clustering is clearly too diverse and not representative enough.

The clusters obtained using group average clustering show a more useful level of resolution among the pyridines but still discriminate too thoroughly among the pyrimidines. Among the methods illustrated in Figure 3, the most intuitively appropriate balance between representativeness (in this case, discriminating between pyridines) and diversity (discriminating between cyclohexanes) is struck by complete linkage clustering (bottom).

Clusters can be obtained from OptiSim selection sets by assigning each compound in the population to that OptiSim selection which is most similar to it. In so doing, one obtains a partition composed of Voronoi cells[12]—i.e., soap bubbles in hyperspace. Distribution profiles obtained from such OptiSim clusterings are shown in Figure 4 for $K = 1$, 2, 3, 5, 10, and 15. For minimum dissimilarity ($K = 1$), the clustering is representative, in that the pyridines are well distributed across clusters. The uncommon compounds—here, cyclohexanes and pyrimidines—fall into just one or two clusters each, as seen in the corresponding dendrogram (Figure 1). Even a subsample size as small as 2 increases the resolution among the cyclohexanes appreciably. As the subsample size grows further, discrimination among the more distinctive compound classes increases, so that at $K = 5$ the distribution is quite similar to that seen for complete linkage clustering in Figure 4. Beyond $K = 5$, the selection sets become more biased toward outliers. They are still, however, qualitatively superior to group average clustering, in that there are fewer very small clusters.

There is one respect in which the OptiSim clustering for $K = 5$ does differ appreciably from that for complete linkage. Note that the cyclohexane clusters are considerably more even in size in the latter case. This reflects the fact that the OptiSim clusters can and will adjust their diameters (maximum pairwise dissimilarity within the cluster) dynamically to compensate for "local" variations in scale and/or spatial distribution. This flexibility is a potential advantage over hierarchical methods in general and complete linkage in particular. It arises because the OptiSim selection method is inherently progressive: the first selection by definition represents the whole dataset, and subsequent selections fill in details of areas not yet adequately represented. Indeed, such progressivity is one of the desirable hallmarks of divisive clustering methods,[6] which are not themselves generally well suited for application to large datasets.

The large cluster comprised of 343 pyridines and three pyrimidines which was obtained for $K = 10$ merits additional

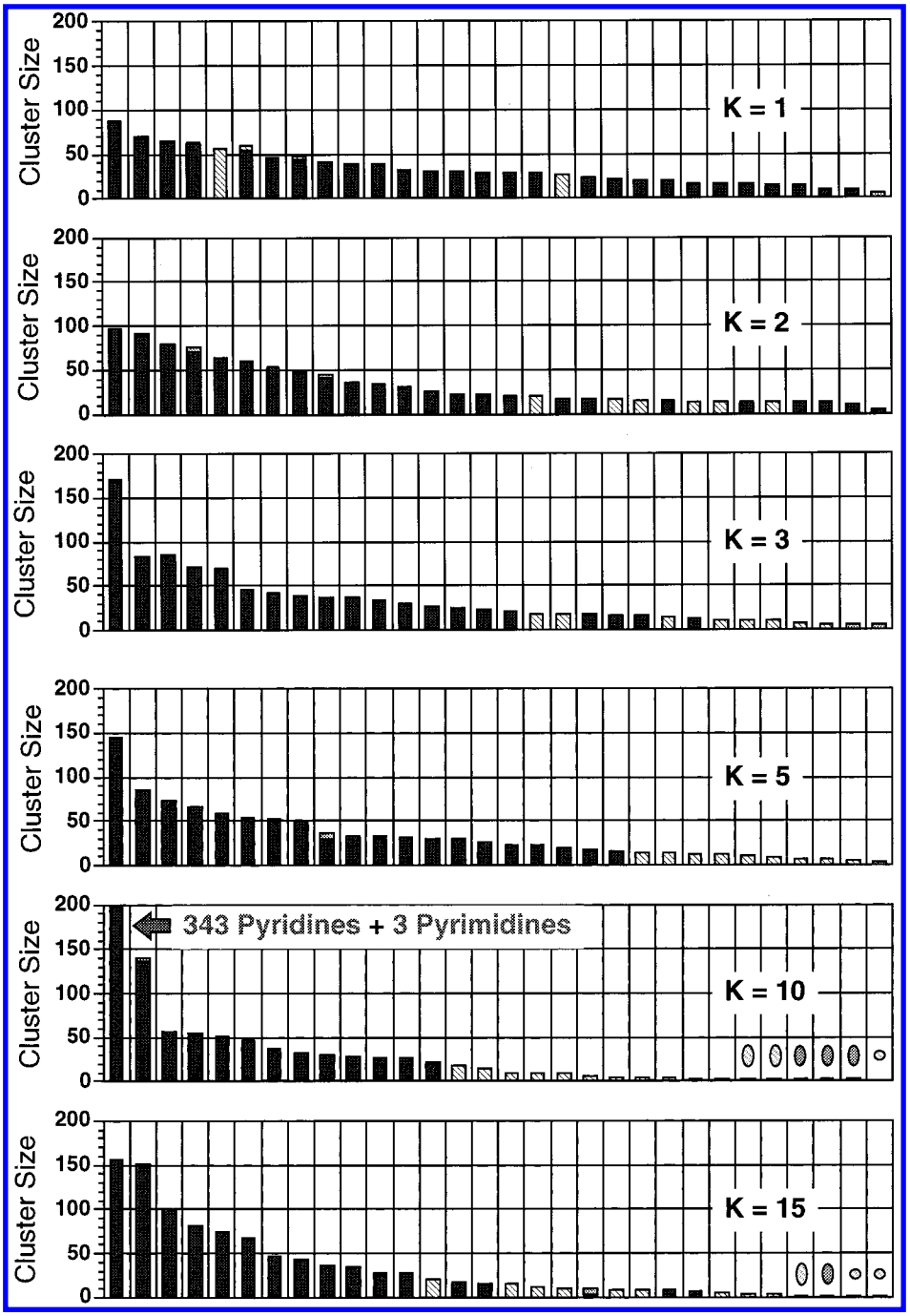

Figure 4. Distribution of cluster sizes for OptiSim clustering as a function of subsample size *K*. OptiSim selections were made from the 1000-member subset using $K = 1$, 2, 3, 5, 10, or 15. For each OptiSim selection set, every compound in the subset was allocated to the cluster centered on the selected compound to which it was most similar. The graphs show the distribution of sizes among those clusters. Bars are coded as for Figure 1. Circles indicate singletons and ellipses indicate doubleton clusters, respectively.

comment. By chance, one of the early OptiSim selections happened to lie near the "center" of the pyridine combinatorial and ended up appropriating 1/2 to 2/3 of the compounds from each of the four largest groups in the other OptiSim clusterings. The stochastic component is an integral part of the OptiSim selection method, so such nuisances cannot be prevented altogether. In cases where they cannot be tolerated, their effects can, however, be minimized in three ways:

(1) Run multiple OptiSim clusterings, each from a separate random number seed, and use that one which produces the smoothest distribution profile (e.g., the lowest χ^2 statistic with respect to a uniform distribution across clusters). The selection obtained in this way will have a "pure" OptiSim distribution but runs the risk of serious hidden biases.

(2) One can make two separate OptiSim selections, each at half the desired ultimate resolution (in this case, at $M = 15$), which is equivalent to sampling each cluster twice. The population can then be partitioned using the union of the two selection sets. Although such "double dipping" selection sets have some very attractive properties in other contexts, the properties of OptiSim clusters derived from them are likely to differ significantly from those of the corresponding hierarchical clusterings.

(3) Any cluster which exceeds a given membership (here, 200 would be appropriate) or diameter (maximum pairwise dissimilarity between members of the cluster) can itself be

submitted to OptiSim clustering. This option (recursive OptiSim clustering) may be the alternative of choice when working with heterogeneously distributed populations for which the extent of structural variation is misleadingly different in different regions of the structural space; an analogous approach has recently been proposed for improving the distributions obtained from Jarvis−Patrick clustering.[13]

Interaction between Subset Size *M* and Subsample Size *K*. The results described above were obtained for a relatively small population ($N = 1000$) and for a fixed subset size ($M = 30$). Figure 5 shows the results obtained when OptiSim was used to select from 5 to 50 compounds from the full mixed library of 6600 compounds. Figure 5A includes data for M = 5, 10, 20, and 50; representativeness ρ is plotted against diversity δ[8]

$$\delta = (1/M)\sum_{i=1}^{M}\min(d(\mathbf{s}_i,\mathbf{s}_j):\ 1 \le j \le M,\ i \ne j) \qquad (1)$$

$$\rho = (1/n)\sum_{i=1}^{n}\min(d(\mathbf{u}_i,\mathbf{s}_j):\ 1 \le j \le M) \qquad (2)$$

where $\mathbf{s}_i$ is one of the M members of the selection set, $\mathbf{u}_i$ is one of n compounds drawn from the $N - M$ compounds in the population which were not selected, and d is the dissimilarity metric (here, the Soergel distance[14] with respect to UNITY 2D fingerprints). Representativeness ρ is the average distance from OptiSim cluster members to their centers, so a small value is desirable. Diversity δ is the average nearest-neighbor distance between centers, so a large value indicates a high average information content. Therefore, the lower right corner is the best place to be in such a plot, because in this region compounds are broadly spread among well-spaced clusters.

Data points shown for each subset size correspond to average values for triplicate selections made with $K = 1$, 2,3, 5, 10, 15, 25, 35, 50, 65, 80, or 100 (chosen to correspond roughly to the natural squares up through 100), with the leftmost data points in all cases representing the lowest values of K. The curves obtained are approximately quadratic in form ($R^2 = 0.770-0.904$).

Diminishing returns begin to set in quickly with this dataset. The subsample size K needed to achieve a particular δ sharply increases as M increases, the maximum δ attainable falls off, and increases in ρ (loss of representativeness) set in at lower values of K. This effect arises because δ is an intensive variable which is a measure of the average *incremental* diversity,[15] i.e., the extent to which the diversity of the selection set is increased by addition of *each individual compound.* In fact, a substantial part of this incremental diversity is captured once a representative has been selected from each of the three major structural classes. For this dataset, the pyrimidines provide the limiting case by being present at a level of 1.5%. On average, then, one would anticipate that capturing incremental diversity will begin to impact representativeness near the point where $M \times K = 66$. Conversely, the reciprocal of $M \times K$ can be seen as specifying the nominal resolution at which the population is being examined. One would expect to begin to detect pyridines when $M \times K$ becomes substantially greater than

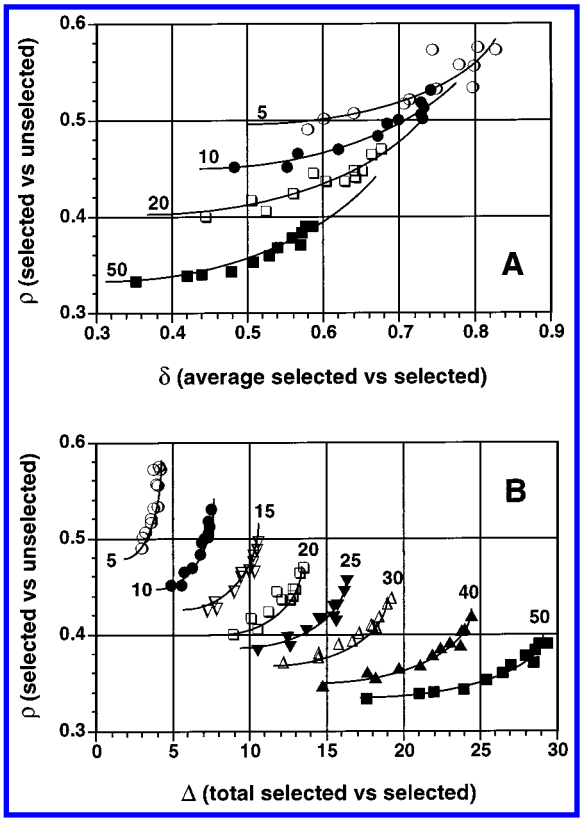

Figure 5. Representativeness versus diversity. OptiSim was used to select M compounds from the full mixed library of 6600 compounds with K set to 1, 2, 3, 5, 10, 15, 25, 35, 50, 65, 80, or 100. Data shown are averages of triplicate runs, where a different random number seed was used for each replicate at each value of K. In each case, the first point from the left corresponds to $K = 1$, with diversity generally increasing thereafter with increasing K. Statistics were calculated as described in the text. The curves shown approximate quadratic fits to the data obtained using nonlinear regression analysis. (A) Representativeness ρ versus incremental diversity δ: (○−○) $M = 5$; (●−●) $M = 10$; (□−□) $M = 20$; and (■−■) $M = 50$. (B) Representativeness ρ versus total diversity Δ: (○−○) $M = 5$; (●−●) $M = 10$; (▽−▽) $M = 15$; (□−□) $M = 20$; (▼−▼) $M = 25$; (△−△) $M = 30$; (▲−▲) $M = 40$; and (■−■) $M = 50$.

66, where the resolution is 1.5%. For $M = 20$ the effect begins to make itself felt above $K = 3$, which serves nicely to rationalize the optimality of $K \approx 5$ for this library at this selection set size.

Perhaps a more appropriate perspective on the situation is obtained by considering the relationship between ρ and the extensive variable Δ, which is a measure of the *total* diversity.

$$\Delta = \sum_{i=1}^{M}\min(d(\mathbf{s}_i,\mathbf{s}_j):\ 1 \le j \le M,\ i \ne j) \qquad (3)$$

The data from Figure 5A is replotted in this way in Figure 5B. This method of presentation makes it clear that the decrease in incremental diversity obtained as the selection set grows larger is more than offset by the increase in total diversity.

Hence, one needs to take the clumpiness of the population into account when creating selection sets and try to avoid unduly skewing the selection set toward outliers, particularly if the size of-the desired selection set is relatively large with respect to the size of the target population (>1%). Problems

should rarely arise with virtual libraries but may occur often with corporate and commercial libraries. The best way to avoid difficulties is to draw multiple representatives from each OptiSim cluster, either by clustering the entire library on a more appropriately sized selection set or (equivalently) by merging several OptiSim selections sets obtained using different random number seeds.

DISCUSSION

Choice of Dissimilarity Measure. Here, as in the initial report,[8] the "best" exemplar is chosen from each OptiSim subsample to maximize the nearest neighbor distance with respect to those compounds already selected (the MaxiMin selection criterion). The average and total nearest neighbor distances are also used as a diversity measure for the selection sets obtained. It has been pointed out[15] that diversity measures based on minimum spanning trees[16] may behave more appropriately than simple nearest neighbor functions when applied to highly redundant subsets. This may be important for methods which involve an analytical or evolutionary maximization of diversity but is not relevant to visualization (e.g., Figure 5) and has relatively modest impact on OptiSim results.

In fact, how the best compound is chosen from each OptiSim subsample is mostly a matter of convenience. The results obtained are qualitatively robust, so for cases where MaxiMin is not appropriate other, more CPU-intensive evaluation criteria can be used instead. Maximizing the average of all pairwise distances with respect to compounds already selected, for example, gives qualitatively similar results to using MaxiMin. The selection sets obtained are, however, less diverse in terms of nearest neighbor distances, just as was previously noted for the special case of maximum dissimilarity selection.[17] In addition, the effect of increasing selection set size M on incremental diversity is considerably larger (data not shown).

Implications for Other Clustering Methods. Though only three hierarchical clustering methods are considered directly here, Lance and Williams have shown that all agglomerative hierarchical clustering can in fact be formulated as special case applications of a generalized agglomerative update equation[18]

$$d(i\cup j,k) = \alpha_i d(i,k) + \alpha_j d(j,k) + \beta\, \mathrm{d}(i,j) + \gamma|\mathrm{d}(i,k) - \mathrm{d}(j,k)| \quad (1)$$

where $d(i,j)$ is the distance (dissimilarity) between sets i and j; $i\cup j$ denotes the set formed by the union of i and j; k is a third set; and α_i, β, and γ are weighting functions. Setting $\alpha_i = \alpha_j = 1/2$, $\beta = 0$, and $\gamma = -1/2$ gives single linkage, for which the dissimilarity between sets is the *minimum* pairwise dissimilarity between any element in one set and those in the other. This leads to a few large, distended sets and isolates the most distinctive compounds as singletons, thereby producing diverse but unrepresentative selection sets (Figures 1 and 3). Jarvis−Patrick, though not an agglomerative hierarchical method, is even more prone to creating skewed distribution profiles,[13] and performs correspondingly poorly in validation studies.[6]

Setting $\alpha_i = \alpha_j = 1/2$, $\beta = 0$, and $\gamma = 1/2$ gives complete linkage. Here, the dissimilarity between sets is the *maximum* pairwise dissimilarity between any compound in one set and those in the other. This is generally the preferred method for substructural clustering, since it tends to produce clusters of uniform diameter—i.e., clusters in which no two compounds are more dissimilar than some characteristic value shared by all of the clusters. A beneficial side effect is that compounds tend to be more evenly distributed among clusters than for other methods (Figure 3). In particular, the number of singletons obtained is minimized when complete linkage is used, as is the appearance of very large clusters.

Group average clustering is obtained for an intermediate case in which $\beta = 0$ and α_i, α_j, and γ are positive numbers determined by the relative sizes of clusters i, j, and k. The update equation for Ward's method is very similar except that β takes on a value between -1 and 0 which is determined by the relative sizes of clusters i, j, and k.[19] This strongly suggests that the results obtained using Ward's method in a Euclidean fingerprint space can be carried over into Tanimoto space (where it cannot be applied directly) by using OptiSim clustering with a suitable subsample size. In fact, Ward's method is an analytical approach to maximizing the cluster separation while minimizing cluster size, since it minimizes the increase in total variance within clusters at each agglomerative step.[19] This is closely akin to balancing representativeness against diversity using OptiSim.

A variation of the special case where OptiSim clustering is based on minimal dissimilarity selection ($K = 1$) has recently been published.[20]

CONCLUSIONS

Compound selection sets based on agglomerative hierarchical clustering have many good distributional properties, particularly when the complete linkage method is used. Unfortunately, applying this technique can be problematic for large datasets, particularly in non-Euclidean spaces where centroids are not well-defined. Such is the case for substructural 2D fingerprints, where the Tanimoto coefficient[14] is the similarity measure of choice.[21] Even in well-behaved spaces and for efficient methods, agglomerative hierarchical clustering scales[21] with N^2. Using OptiSim selections as cluster centers provides a fast and convenient alternative for producing clusters with equally good distributional properties from very large datasets, when the OptiSim parameters are chosen so as to mimic the results from hierarchical clustering. Indeed, the ability to fine-tune the OptiSim parameters means that in many cases OptiSim clusters can be produced which are better behaved than those obtained using any of the more traditional clustering techniques.

ACKNOWLEDGMENT

Roman Dorfman and Girish Bakhru of Tripos, Inc., helped turn OptiSim from an idea into a commercial product. Richard Cramer, David Patterson, and John Begemann, also of Tripos, Inc., have provided support and helpful suggestions along the way, as has Prof. Peter Willett of the University of Sheffield.

REFERENCES AND NOTES

(1) Martin, E. J.; Critchlow, R. E.; Spellmeyer, D. C.; Rosenberg, S.; Spear, K. L.; Blaney, J. M. Diverse Approaches to Combinatorial Library

Design. In *Pharmacochemistry Library*; Timmerman, H., Ed.; Elsevier Publishers: Amsterdam, 1998; Vol. 29, pp 133−146.
(2) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(3) Clark, R. D.; Ferguson, A. M.; Cramer, R. D. Bioisosterism and Molecular Diversity. In *3D QSAR in Drug Design*; Kubinyi, H., Martin, Y. C., Folkers, G., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1998; Vol. 2, pp 211−224.
(4) Cramer, R. D.; Clark, R. D.; Patterson, D. E.; Ferguson, A. M. Bioisosterism as a Molecular Diversity Descriptor: Steric Fields of Single Topomeric Conformers. *J. Med. Chem.* **1996**, *39*, 3060−3069.
(5) Brown, R. D.; Martin, Y. C. Use of Structure-Activity Data to Compare Structure-Based Clustering Methods and Descriptors for use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.
(6) Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.
(7) Patent pending. OptiSim is a registered trademark of Tripos, Inc., 1699 S. Hanley Rd., St. Louis, MO 63144.
(8) Clark, R. D. OptiSim: An Extended Dissimilarity Selection Method for Finding Diverse Representative Subsets. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181−1188.
(9) Legion, SYBYL, UNITY, ChemEnlighten, and Selector are available from Tripos, Inc., 1699 S. Hanley Road, St. Louis, MO 63144.
(10) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of Molecular Diversity Descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.
(11) Murtagh, F. A Survey of Recent Advances in Hierarchical Clustering Algorithms. *Comput. J.* **1983**, *26*, 354−359.
(12) Okabe, A.; Boots, B.; Sugihara, K. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*; Wiley: New York, 1992.
(13) Menard, P. R.; Lewis, R. A.; Mason, J. S. Rational Screening Set Design and Compound Selection: Cascaded Clustering. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 497−505.
(14) Gower, J. C. Measures of Similarity, Dissimilarity, and Distance. In *Encyclopedia of Statistical Sciences*; Kotz, S., Johnson, N. L., Eds.; John Wiley & Sons: New York, 1985; Vol. 5, pp 397−405.
(15) Waldman, M.; Li, H.; Hassan, M. Novel Metrics for the Optimization of Molecular Diversity of Combinatorial Libraries. Manuscript in preparation.
(16) Mount, J.; Ruppert, J.; Welch, W.; Jain, A. IcePick: A Flexible Surface Based System for Molecular Diversity. *J. Med. Chem.* In press.
(17) Holliday, J. D.; Willett, P. Definitions of Dissimilarity for Dissimilarity-Based Compound Selection. *J. Biomol. Screening* **1996**, *1*, 145−151.
(18) Lance, G. N.; Williams, W. T. A. A General Theory of Classificatory Sorting Strategies. I. Hierarchical Systems. *Comput. J.* **1967**, *9*, 373−380.
(19) Kaufman, L.; Rousseeuw, P. J. In *Finding Group in Data: An Introduction to Cluster Analysis*; Wiley-Interscience; New York, 1990; pp 230−243.
(20) Reynolds, C. H.; Druker, R.; Pfahler, L. B. Lead Discovery Using Stochastic Cluster Analysis (SCA): A New Method for Clustering Structurally Similar Compounds. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 305−312.
(21) Barnard, J. M.; Downs, G. M. Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 644−649.

CI980107U