# Rational Screening Set Design and Compound Selection: Cascaded Clustering

Paul R. Menard,*,† Richard A. Lewis,‡ and Jonathan S. Mason†

Computer-Assisted Drug Design, New Lead Generation, Rhône-Poulenc Rorer, Collegeville, Pennsylvania 19426 and Dagenham, RM10 7XS, UK

Received January 12, 1998

The use of cascaded clustering is reported. This technique was developed to permit the application of Jarvis-Patrick clustering based on structural fingerprints to large chemical databases, while keeping the maximum cluster size and the number of singletons produced at reasonable levels. The basis for the algorithm, its implementation, and validation are described. In the first part of the paper, the approach is used to create a representative subset of compounds for biological testing from the corporate compound repository. A variation of the method is then used for the comparison of relatively large databases. Finally, compound selection using cascaded clustering is shown to be complementary to the Diverse Property-Derived approach, which is based on partitioning by six molecular descriptors.

## INTRODUCTION

With the advent of high-throughput screening (HTS), many more compounds can be subjected to bioassay than was possible with traditional screening methods. Researchers must therefore identify more compounds for screening in order to satisfy this increased capacity. One of the best sources for screening candidates is the corporate compound repository. A second source is external compound collections: compounds different from those in the corporate repository can be identified and acquired. More recently, combinatorial chemistry techniques have allowed the production of large numbers of compounds both quickly and relatively inexpensively. One approach to screening is simply to schedule all available compounds for bioassay in random order or by some other arbitrary selection process. This may result in large batches of structurally similar but inactive compounds being screened sequentially, with little resultant return in terms of biological activity.

A major goal of pharmaceutical research is to identify new biological leads as quickly and efficiently as possible. Furthermore, not all bioassays are appropriate for HTS methodology. In some cases, the assay cost may become prohibitive when dealing with large numbers of compounds. By identifying a small, structurally diverse or representative set of compounds for priority screening from a larger compound collection and by augmenting this set regularly with new dissimilar compounds, the chances of finding a lead quickly should be increased, and such a set should be relevant for both high-throughput and traditional screening methods. If such a method also defines relationships between similar compounds, this would facilitate the process of selecting compounds for lead follow-up.[1] The cascaded clustering technique has been developed in order to address these needs. This approach is based on clustering using 2D structural fingerprints and incorporates novel methodology for dealing partially with clustering problems inherent to the Jarvis-Patrick nonhierarchical method; such problems are too wide a range of cluster sizes, heterogeneous clustering, and excessive numbers of compounds which fail to cluster at all. The method is equally applicable to clustering based on other similarity metrics.

**Background.** There is no universally agreed-upon definition of chemical diversity, and there are several approaches for identifying chemically diverse or representative collections of compounds. We therefore start by defining our terms: "diverse" will be defined as a compound collection spanning as wide a range of values as possible relative to some descriptor derived from a compound's structure. "Representative" will mean a compound subset reflecting the distribution of values for some descriptor shown by the parent collection from which it is derived. Descriptors which can be used for diversity profiling can be derived from one-dimensional chemical properties (1D, e.g., molecular weight), two-dimensional (2D, e.g., topological descriptors such as structural fingerprints or substructural keys, or physicochemical descriptors such as clogP), or three-dimensional (3D, e.g., shape indices or pharmacophoric keys). For a fuller review, see refs 2−9.

Furthermore, either partitioning or clustering may be used to select compounds relative to a given descriptor. In partitioning, the descriptor must be capable of being defined by an absolute value such as a real number (e.g., clogP). The possible range of values for the descriptor is then typically broken down into subdivisions of equal size, and compounds are selected so as to fill each subdivision. Advantages are easy control of granularity, the ability to identify property space not represented by any structures in the set being considered, and straightforward comparison of compound populations. Partitioning is particularly useful in designing diverse collections. In clustering, compounds are grouped together based on their similarities, normally a distance-based calculation using some descriptor. Since results are relative to the compound collection being analyzed and not against absolute values as in partitioning, it is difficult to determine what property space is unrepresented or to

† Collegeville, Pa.
‡ Dagenham, UK.

S0095-2338(98)00003-1 CCC: $15.00 
Published on Web 04/04/1998

1) For each molecule in the database

compute the molecular descriptor

2) For each molecule **i** in the database

for each molecule **j**

compute the similarity between **i** and **j**

maintain a dynamic list of the **K** nearest neighbors of **i**

print out the list of nearest neighbors

3) Cluster the nearest neighbors list using the rule:

**A** and **B** are in the same cluster if they have **J** out of their **K** nearest neighbors in common

**Figure 1.** Daylight fingerprint/clustering methodology

compare populations. However, the relative nature of the analysis is well suited to defining representative subsets.[10]

In an earlier paper in this series,[11] the Diverse Property-Derived (DPD) approach was presented. Briefly, the DPD approach involves partitioning compounds over a property space derived from six pairwise noncorrelated molecular and physicochemical descriptors. The descriptors used were hydrogen bond acceptor count, hydrogen bond donor count, flexibility index, clogP, aromatic density, and electrotopological index. These are intuitively satisfying to the pharmaceutical researcher, since they relate directly to factors important for ligand−receptor binding (electrostatics and shape) as well as to absorption/cell wall transport. This method was initially used to create a small diverse set (1000 compounds) for biological testing, and is currently being applied in lead follow-up studies and external compound acquisitions.

In traditional pharmaceutical research, relatively large numbers of structurally related compounds may be made during the exploration of structure−activity relationships. On the other hand, externally purchased (or available) compounds may be quite different from those produced by in-house research and may be the only representatives of their chemical families in the repository. However, compounds having widely different structures may have similar physicochemical properties, and using these properties alone to characterize and select compounds may fail to sample important chemical families or interesting structures. It was therefore decided to develop a topologically based method which would complement the DPD approach. This method would serve two critical needs: (1) the selection of structurally representative sets from compound collections and (2) the comparison of compound collections, with the capability of identifying compounds from a query collection which are structurally dissimilar to those in a reference collection.

**Daylight Fingerprints and Similarity.** The approaches to be discussed are based on the Daylight Clustering Package[12] which provides a full capability for clustering based on 2D structural descriptors. The strategy is outlined in Figure 1. Briefly, the Daylight fingerprint module generates a binary descriptor (fingerprint) representing the 2D structural characteristics of a molecule. All bond paths within the molecule containing typically between two and eight atoms are identified and converted to binary representations using a pseudorandom hashing algorithm, with each representation being added to the overall molecular fingerprint. The resulting fingerprint is then folded by successively dividing it in half and ORing the fragments together until a sufficient information density is achieved (in this study, at least 30% of the bits set). A potential advantage that this method has over fingerprints that are based on substructural keys is that the fingerprint is based solely on the topology of the molecule, not on a predefined set of functional group keys that may not be entirely appropriate for the compound collection being analyzed. A possible disadvantage is that the hashing algorithm may cause different functionalities to set the same bits. Furthermore, fingerprint folding, although advantageous for increasing the efficiency of disk and CPU utilization, can cause a slight loss of information as the fingerprint halves are merged. In our experience, neither of these two factors appeared to cause any substantial misclustering.

A nearest neighbors list is then generated based on the similarities of each molecule to all other molecules in the dataset. The Tanimoto metric is used to calculate similarity

$$T = N_{A\&B}/(N_A + N_B - N_{A\&B})$$

where $N_A$ and $N_B$ are the number of bits set in fingerprints A and B, and $N_{A\&B}$ is the number of bits common to A and B.

**Clustering Methods.** Two methods were considered for clustering: a hierarchical method based on Ward's algorithm[13] and the nonhierarchical Jarvis-Patrick method[14] provided by Daylight. In Ward's method, relationships are maintained not only between compounds in clusters but also between clusters themselves, including singletons (structures which cluster alone). This simplifies the task of finding related structures for bioassay once an active compound is found. It is easy to vary the number of clusters produced by modulation of the clustering parameters. The clusters produced are generally intuitively sensible to the chemists, the ratio of singletons to clusters is low, and maximum cluster sizes, even under forcing conditions, are reasonable. The drawback is rather demanding in terms of disk space and CPU requirements. The dissimilarity matrix can require up to $O(N^2)$ disk space for storage, and the standard algorithm requires $O(N^3)$ time for the clustering process.[15] Some workers have achieved improved performance by use of Murtagh's Reciprocal Nearest-Neighbor algorithm, which requires only $O(N)$ disk space and $O(N^2)$ time.[16] With the implementation we had in our hands at the time, Ward's method was best suited to the analysis of 10 000 structures or less, due to the time constraints imposed by our $O(N^2)$ implementation; the technique was therefore not used in this study. In fairness, it should be added that newer implementations can cluster up to 200K structures in a reasonable time.[17]

The Jarvis-Patrick algorithm clusters molecules using conditions which state that two structures will cluster together if (a) they are in each other's list of **K** nearest neighbors and (b) they have some integer value **J** of their **K** nearest neighbors in common. Both **J** and **K** are user-defined parameters and can be varied to provide optimum results based on the chemical database being studied and on the clustering goals. In the Daylight nearest neighbors routine, if some integer value **K** of nearest neighbors is requested for each structure then that number will be found, regardless

of how low a value of $T$ is required to satisfy this criterion. If a subsequent clustering method is based solely on nearest neighbors data, this may cause structures which are in reality quite dissimilar to cluster together. The Jarvis-Patrick algorithm has several computational and practical advantages for clustering chemical structures. It is computationally efficient from both a CPU ($O(N^2)$ time) and a disk utilization perspective and therefore is suitable for very large databases ranging upward of a million structures. Our experiments on an SGI IndigoII R4000 running Irix5.3 gave an approximate formula of $W$/CPU s = 1700*($N$/10 000)∧2, where $W$ is the run time, and $N$ is the number of objects. For $N$ = 100 000, $W \approx$ 48 CPU h. A comparison by Willett of nonhierarchical clustering methods for the analysis of chemical datasets based on substructural features showed the relative superiority of this approach.[18]

Unfortunately, the method also has drawbacks. Jarvis-Patrick is nonhierarchical, and no relationship between clusters is obtained. If **J** ≈ **K**, then the condition for clustering is very stringent, and a significant fraction of the data remains unclustered (singletons). If **J** ≪ **K**, the condition for clustering is less stringent, and a few, very large, clusters may form. A similar phenomenon occurs in hierarchical clustering as the fusion distance is increased; in this case, methods for determining the correct distance have been proposed.[19] Unfortunately they are not applicable to the Jarvis-Patrick method. The Daylight Jarvis-Patrick implementation (v441) did not support a minimum similarity threshold in the nearest neighbors calculation; therefore, some heterogeneous clusters may result as described above. We speculate that the use of a sensible cutoff may improve the quality of clusters but that the issues posed by the presence of singletons will still exist.

**Alternative Methods for Set Design and Compound Selection.** The selection of a method for set design or compound selection depends on a number of factors, such as the number of structures being analyzed, the size of the subset desired, available biological results and their possible correlation with a given descriptor, etc. Several studies have been published that look at the issues surrounding selection of representative/diverse compound subsets and comparison of compound collections; only a few will be mentioned here. Willett studied three nonhierarchical clustering methods based on 2D substructural fragments[18] and found the Jarvis-Patrick approach the most effective for grouping structures with similar physical, chemical, or biological properties.

Daylight fingerprints in conjunction with Jarvis-Patrick clustering, and readily calculated physicochemical descriptors such as partition coefficient and molar refractivity, were used by Shemetulskis to analyze and compare datasets.[20] Both approaches were found to be practical for use on moderately sized datasets (approximately 100 000 structures). Pearlman has developed software[21] to perform subset selection from large compound collections, using novel BCUT descriptors and more traditional structural fingerprints. An advantage to the BCUT approach is its applicability to very large datasets and its potential appeal to medicinal chemists (BCUT values are based on electronegativity, polarizability, hydrogen bond donor and acceptor ability, and other factors related to ligand-protein binding). Willett[22] has developed a rapid method for comparing datasets based on descriptors calculated from pairwise intermolecular similarities within each dataset, including but not limited to the number of heavy atoms, rotatable bonds or rings, or structural fingerprints. Each descriptor characterizes the dataset as a whole, not on a molecule-by-molecule basis, but is computationally efficient and also applicable to very large datasets. An extension of this method has been used to analyze the diversity of compound subsets relative to a parent database and thereby select a set of molecules representing much of the parent dataset's diversity. More recently, Clark[23] has described a stochastic algorithm (OptiSim) that uses dissimilarity to select molecules. There is an adjustable parameter that controls the subsample size and hence the balance between representativeness and diversity. However, this does require a comparison metric that is meaningful when measuring distance between quite dissimilar objects. Our anecdotal experience is that a measure of (1-T) is not suitable when folded Daylight fingerprints are used, as the distance can be affected by chance bit matches that are artifacts of the folding process.

## RESULTS AND DISCUSSION

**Objectives.** The standard compound collections we wished to analyze ranged from a few hundred structures to several hundred thousand. The initial goal of our work was to develop a clustering approach based on the Jarvis-Patrick algorithm which would meet our needs, namely to produce homogeneous, limited-size clusters with a small (<5%) percentage of singletons. Development would first center on producing structurally representative sets. Following validation, the method would be extended to the comparison of compound collections and the identification of complementary structures. Finally, it would be shown that the sets identified by this method and the DPD approach based on physicochemical properties were indeed orthogonal.

**Representative Set Generation.** The specific purpose of the first study was to develop a method for producing a small (approximately 10 000-compound) set of structures for biological screening that would represent the major structural types present in the company compound repository (CCR). The resulting compounds would also presumably be relatively structurally diverse and thus provide an excellent starting set for blind biological testing. If other relevant information were available, for example from known ligands, we would advocate a more directed approach toward compound selection.

A subset of 164 381 structures (CCR1) suitable for bioassay was selected from the overall CCR, which was maintained as a MACCS−II 2.2 database.[24] The structures were written out in SDF file format[25] and were converted to SMILES codes[26,27]using either **GEMINI** or **mol2smi**, both from Daylight. Salts were neutralized, functional groups that could be represented in more than one manner (e.g., nitro) were standardized, and duplicate structures were removed using a series of in-house-developed FORTRAN programs. Daylight 2D fingerprints were generated using the Daylight clustering module program **fingerprint** and nearest neighbors were calculated using **nearneighbors**, using an upper limit of 16 neighbors. The code version was 4.41. Clustering behavior was then studied using **jpscan** along with a variety of values for Jarvis-Patrick parameters **J** and **K**. Results are shown in Table 1.

**Table 1.** Variation in Clustering Results with Jarvis-Patrick **J**, **K** Parameters

| J/K | max. cluster size | no. of clusters | no. of singletons | % singletons |
|---|---|---|---|---|
| 2/14 | 111227 | 4751 | 6409 | 3.9 |
| 3/12 | 86326 | 6960 | 10013 | 6.1 |
| 5/10 | 361 | 16092 | 21367 | 13.0 |
| 5/12 | 30749 | 10664 | 15629 | 9.5 |
| 6/12 | 524 | 13813 | 19859 | 12.1 |
| 7/14 | 594 | 12236 | 18778 | 11.4 |
| 7/15 | 8020 | 10267 | 16323 | 10.0 |
| 7/16 | 42290 | 8618 | 14387 | 8.7 |
| 8/16 | 1001 | 10945 | 17855 | 10.9 |
| 9/16 | 329 | 13729 | 22430 | 13.6 |

The variation in results with minor changes in parameters is astounding. Simply requiring seven nearest neighbors in common from the top 16 instead of the top 15 increased the maximum cluster size by over 34 000. Requiring nine nearest neighbors in common rather than eight from the top 16 increased the number of singletons by 4500. Our goals were to discover a combination of parameters that would produce a maximum cluster size of approximately 1000 (we justify this figure later), with 10 000 clusters and not more than 5000 singletons. (Although the ultimate goal of the project was a 10 000-compound set, the compounds were not filtered for stock beforehand. A larger initial set was thus desired to account for compounds, particularly singletons, that would be out-of-stock and not have any other suitable replacements.) None of the parameter pairs gave an acceptable combination of clusters, singletons, and maximum cluster size. The number of singletons in the closest acceptable combinations, ranging from 10.9 to 13.6%, was far too high for inclusion into a small or moderately sized screening set, and in all cases exceeded our overall desired set size of 10 000. We return to this issue later.

Various techniques can be used to reduce the amount of singletons, but all suffer from some drawbacks. Less stringent clustering conditions (smaller values of the ratio **J**/**K**) will cause singletons to merge into clusters but will also produce large heterogeneous clusters. From Table 1, **J**/**K** values of 2/14 gave only 6409 singletons but produced only 4751 clusters as well, with the largest cluster containing more than 50% of the structures being analyzed. Reclustering singletons in a single separate run is effective, but the clusters produced can be heterogeneous and difficult to sample. Singletons can be merged into the cluster populated by the majority of their nearest neighbors, but they may still be markedly different from these compounds and cause sampling problems again. To further concentrate the singletons into natural clusters without forcing the remaining structures to consolidate into larger, heterogeneous clusters, the cascaded clustering technique was developed. The approach is as follows:

1. For a given data set, select values for **J** and **K** such that the Jarvis-Patrick clustering gives an acceptable value for maximum cluster size (or, optionally, for some other criterion such as total number of clusters or average cluster size);

2. Remove the singletons but retain all other clustering results;

3. Rerun the nearest-neighbors calculation on the singletons and recluster using mild Jarvis-Patrick conditions;

4. Repeat steps 2−3 until an acceptable number of singletons is obtained.

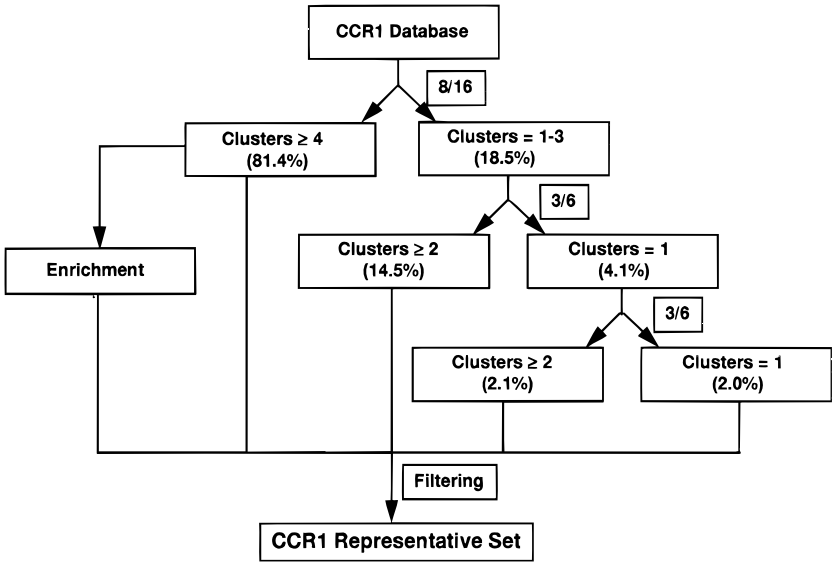
If a dataset contains a large number of very small clusters (e.g., less than five members), these can be concentrated as well by treating the members as singletons in steps 2−3. Mildness in the Jarvis-Patrick conditions should be distinguished from stringency. Stringency is achieved by increasing **J** relative to **K**; mildness is achieved by reducing **J** and **K**. The singleton set, by definition, is more disperse than the original set, so each object will have fewer meaningful neighbors; using lower values of **J**/**K** (we used 3/6 in many cases) enables the clusters to form without resorting to less stringent conditions. The purpose of using mild clustering conditions in step 3 is to ensure that the structures are not forced into unnatural clusters. The use of mild conditions tends to produce many small, rather than a few large, clusters; reducing stringency would have the opposite effect. Small clusters can be well represented by their centroids or other cluster members and should give a fair representation of the starting set, especially if the starting structures are not especially similar.

Scheme 1 shows the cascaded clustering technique as applied to CCR1. **J** and **K** values of 8 and 16 were chosen for the primary clustering step; both maximum (1000) and average (13.3) cluster sizes were reasonable, and review of a random group of larger clusters showed them to be chemically sensible in general. We postulated that a value of 1000 was a reasonable upper limit on the size of the largest analogue series in our corporate registry. We confirmed this through visual inspection of the cluster, by sorting our database by the project code of each compound, and by an analysis of structural families.[28] The percentage of compounds in clusters of size three or less was 18.5%; of these, 10.9% were singletons. Secondary clustering was done over these clusters/singletons using **J**/**K** values of 3/6; tertiary clustering was done over the resulting singletons at 3/6. The final percentage of singletons was 2.0% (3230) of total structures, more than a 5-fold reduction compared to primary clustering alone. The secondary clustering gave a maximum cluster size of 153 and average size of 5.5, and the tertiary a maximum of 34 and average of 3.1. These decreases are expected, since with each additional clustering similar compounds should be removed through their membership in clusters, and the compounds remaining should become increasingly dissimilar.

The screening set itself was constructed by taking the centroids of all retained clusters (i.e., those clusters not removed in any iteration of step 2) and singletons produced in the tertiary clustering and filtering the structures for availability, any other undesirable functionality,[11] and internal similarity. The last step was necessary because certain large structures such as steroids tended to form several small clusters based on small changes in functionality, even though the overall structures were obviously quite similar and the family needed to be represented only once. Enrichment candidates were also selected from some of the larger clusters using Chem-X software,[29] with clustering over a combination of 2D structural and 3D distance keys. Final set size was approximately 9000 compounds.

**Singletons.** The recurring issue of this work is how singletons and small clusters are handled. At the danger of repetition, the subject will be discussed again. Singletons

**Scheme 1.** Cascaded Clustering as Applied to the CCR1 Dataset




produced by nonhierarchical methods can pose a problem in biological screening, where one frequently wishes to test related structures when a bioactive compound is found. Furthermore, the percentage of singletons may be high (over 15% of total structures), and they may overwhelm what was meant to be a structurally representative set. Attempts to reduce the amount of singletons by altering **J** and **K** generally produce many large, heterogeneous clusters which are difficult to represent properly, even when multiple selections are taken from the larger clusters.

As with all classification methods, there is a degree of subjectivity about what constitutes an acceptable size of cluster, or fraction of singletons. Visual inspection indicated that, to a chemist's eye, some of the compounds were part of a larger series and did not deserve to be isolated. It was this problem that we sought to address. It is hard to give any guidelines about the natural percentage of singletons (as this will probably depend on the dataset used), although a threshold value of similarity may be used to determine the compounds that have no real neighbors from a single pass through the nearest neighbors list.

Although the discussion so far has focused on the number of singletons, we were also concerned as to the physical meaning of the smaller clusters (< five members) that perhaps ought to be part of larger clusters. However, changing the ratio of **J**/**K** did not give sensible results. We therefore also included the smaller clusters in the initial levels of our cascade. The results from secondary clustering (largest cluster size increased from 3 to 153, cluster checked by visual inspection) reinforced our prejudice that the results from primary clustering were unsatisfactory.

**Sampling of Large Clusters.** At this stage we returned to the larger clusters. The set produced from Scheme 1 is a 1/17 sample of the original CCR1. It is not unreasonable to sample the larger clusters at a similar fraction, rather than at 0.1% (1 in 1000). To overcome problems in homogeneity within the clusters under the original metric, we chose to use the CDL 2D and 3D keys to split the clusters up sensibly.

When selecting compounds for weighing, we discovered that stock constraints gave an attrition rate of 20%. Our strategy was to move progressively out from the centroids until acceptable in-stock compounds could be found. The first step was to remove toxic or reactive compounds[11] before inspection. Each cluster centroid, singleton or in-stock replacement was examined for biological acceptability by a medicinal chemist.

**Table 2.** Clustering of the CCR2 Database

| clustering step | unretained structures | retained structures | retained clusters | retained singletons |
|---|---|---|---|---|
| primary | 41369 | 181280 | 5681[a] | |
| secondary | 9368 | 32001 | 5454[b] | |
| tertiary | | 9368 | 1195[b] | 4781 |

[a] Clusters ≥ size 4 (average size 31.9). [b] Clusters ≥ size 2 (average size 5.9).

**Testing of the Clustering Conditions.** To compare the behavior of a second dataset when clustered under the same conditions, another large database (CCR2) was selected, containing 222 649 structures acceptable for bioassay. The structures were treated exactly as were the CCR1 structures: primary clustering at **J**/**K** values of 8/16 with clusters of size greater than or equal to 4 being retained, secondary at 3/6 with clusters greater than size 1 being retained, and tertiary clustering at 3/6. (If the goal of the study had been to generate a representative set, then conditions appropriate to and specific for this database would have been used.) Results are shown in Table 2.

In terms of percentages, the results are quite similar to those obtained with clustering of the CCR1 database. Primary clustering of the CCR2 structures gave 18.6% of compounds in clusters of size 3 or less (12.2% singletons), compared to 18.5% of CCR1 structures (10.9% singletons). One notable difference was the size of the largest cluster; for CCR2 it was 9158, compared to 1000 for CCR1. A minor adjustment in the **J**/**K** parameters, to either 8/15 or 9/16, would have reduced this value to less than 600. Cascaded clustering was equally effective here at reducing the number of singletons. Over the three steps the percentage of singletons dropped from 12.2 to 2.1%, compared to a drop from 10.9 to 2.0% for the CCR1 set.

**Representative Set Generation: Validation.** At this point two questions arose regarding the suitability of

**Table 3.** Comparison of CCR1 Dataset Clustering Results (**J**/**K** 8/16) to Derivative Datasets

| database | structures | % singletons | av cluster size | max. cluster size |
|---|---|---|---|---|
| CCR1 | 164381 | 10.9 | 13.4 | 1001 |
| CCR1 minus singletons | 146526 | 0.6 | 14.1 | 2449 |
| CCR1 representative set | 12434 | 49.9 | 9.0 | 284 |
| CCR1 random set | 12434 | 21.0 | 13.3 | 433 |

cascaded clustering for the definition of structurally representative sets. First, are the singletons truly interesting compounds worthy of special consideration, and second, will the reclustering behavior of a representative set reflect its expected structural diversity?

Singletons can arise from two sources: they might represent structurally different compounds, or they might simply be statistical artifacts of the clustering process. Certainly they would be of much less interest for inclusion in representative sets, particularly in sets proposed for biological screening, if the latter effect were to predominate. To delineate these factors roughly, the CCR1 dataset was clustered once using **J**/**K** values of 8/16. All singletons were removed, the nearest-neighbors calculation was redone on the remaining structures, and these were reclustered using **J**/**K** values of 8/16. If singletons are predominantly statistical artifacts, then the percentage of singletons produced by the reclustering of the reduced set should approach that given by the clustering of the entire dataset.

Results are shown in Table 3. The percentage of singletons produced by the reclustering is greatly reduced compared to the percentage produced by clustering the entire dataset, indicating that singletons from this clustering approach, to a large part, are indeed structurally different compounds.

Representative sets created from cascaded clustering should contain a variety of structural types and therefore be relatively diverse. The clustering behavior of these sets relative to the original datasets should reflect this; increases in the percentage of singletons and decreases in maximum and average cluster size might be expected. To test this, a set of 12 434 structures was created by taking all cluster centroids resulting from the cascaded clustering of the CCR1 dataset as described above, as well as all singletons resulting from the tertiary clustering step. This set was then clustered using **J**/**K** values of 8/16, and the results compared to the original dataset clustered under identical conditions. Results are in Table 3. The percentage of singletons is greatly increased, from 10.9% to nearly 50%. Both average and maximum cluster sizes are reduced as expected.

It could be argued that these changes result from the reduced size of the data set, which is only 7.6% of the size of the complete CCR1 data set. In a larger set, there is a greater likelihood that structurally related compounds will be present and will cluster together, reducing the number of singletons and increasing cluster sizes. To test this, a random set of 12 434 unique structures was selected from the complete CCR1 data set using a program based on the FORTRAN intrinsic function RAN (a random number generator). These were clustered under conditions identical to those of the representative set (see Table 3). The percentage of singletons is increased relative to the entire data set but is still less than half that seen with the representative set. The average cluster size is nearly identical to that of the entire set, whereas maximum cluster size is greatly reduced; in both cases the figures are substantially higher than those for the representative set, as would be expected for a less diverse set.

**Alternative Strategies for Singleton Processing.** There are several strategies for dealing with singletons. It is important to remember that singletons may really be singletons, in which case further processing would give spurious results. Visual inspection indicated that this was not the situation with the databases we examined. We have already discussed the effect of changes in the parameters **J** and **K** and concluded that this was not a satisfactory method, leading us to develop cascaded clustering. Singletons may also be rescued to the cluster containing the majority of its nearest neighbors. This can have an effect similar to single-linkage clustering and may merge (and lose) real clusters that exist outside of the context of the large clusters. This is what we have found using cascaded clustering, that real secondary clusters can be formed in the absence of the primary clusters; rescuing would lose this information. A similar argument can be made against lumping singletons into artificial orphan clusters, made up by aggregating any pairs of objects that are in each other's nearest neighbor list. The advantage of the cascaded cluster strategy is that it recognizes that all small clusters, not just singletons, could be artifacts of the values of **J** and **K** chosen to fix the size of the largest cluster and so ought to be reclustered within a separate context.

**Database Comparisons.** Daylight fingerprint-based clustering might be equally useful for comparing predefined compound collections such as screening sets to other collections (e.g., commercially available compounds, proposed combinatorial libraries) to identify structural types not already represented. This could be done by simply combining the reference and test compound collections, performing a single clustering, and identifying those test compounds that do not cluster with any reference compounds. An excessive number of singletons may pose a problem here as well: it may not be cost-effective to acquire or synthesize all test singletons.

A variant of the cascaded clustering method has been developed to compare effectively two sets of molecular structures, as follows:

1. Combine the reference and test datasets and perform the nearest-neighbors calculation;
2. Select values for **J** and **K** such that the Jarvis-Patrick clustering gives an acceptable value for maximum cluster size, total number of clusters, or average cluster size;
3. Identify and retain those clusters consisting solely of test compounds, reference compounds, and mixed clusters;
4. Rerun the nearest-neighbor calculation on the remaining structures (singletons and, optionally, small-cluster compounds) and recluster using mild Jarvis-Patrick conditions (**J**/**K** = 3/6);
5. Repeat steps 3−4 until an acceptable number of singletons is obtained.

Using this procedure, the CCR1 data set was compared against two external collections: the Available Chemicals Directory (ACD)[24] and the CCR2 database.

**Table 4.** Cascaded Clustering Results for Mixed CCR1/ACD Database

| clustering | structures | clusters (size ≥2) | singletons[a] (%) | av cluster size | max. cluster size |
|---|---|---|---|---|---|
| primary | 271K | 17655 | 28529 (10.5%) | 13.7 | 1602 |
| secondary | 48K | 6871 | 10534 (3.9%) | 5.5 | 266 |
| tertiary | 10K | 1320 | 5128 (1.9%) | 4.1 | 78 |

[a] As a percentage of total starting structures.

**Table 5.** Mixed CCR1/ACD Database Clustering: Classification of Clusters

| clustering (size retained) | % CCR1-only | % ACD-only | % mixed |
|---|---|---|---|
| primary (≥4) | 35.2 | 11.1 | 53.7 |
| secondary (≥2) | 44.7 | 16.9 | 38.4 |
| tertiary (≥2) | 39.9 | 16.2 | 43.9 |
| tertiary (1) | 63.0 | 37.0 | |

For the ACD, approximately 106K structures, filtered so as to include those possible for bioassay, were identified and converted to SMILES codes as described earlier. These were combined with the CCR1 SMILES codes to produce a single database of 271K structures. The cascaded clustering technique was then applied as described above. Primary clustering was done using **J**/**K** values of 8/16. Clusters of size four or greater were separated and classified as CCR1-only, ACD-only, or mixed (CCR1 and ACD). The remaining structures were subjected to secondary clustering at 3/6. The singletons resulting from this step were removed and reclustered at 3/6 (tertiary clustering). Results are shown in Table 4.

Results from the combined database are generally comparable to those seen when the CCR1 database was clustered alone under similar conditions. With each successive clustering stage, both the average and maximum cluster sizes drop as expected and the number of singletons is halved. The final percentage of singletons (1.9%) is quite similar to that of the CCR1 database (2.0%). Of more interest is the population breakdown of the clusters, shown in Table 5. Results are tabulated as the percentage of clusters retained in each step (i.e., those clusters not subjected to another level of clustering).

The original database contained only a moderate excess of CCR1 structures (61% CCR1 and 39% ACD). However, the percentage of CCR1-only clusters produced was generally 2−3 times that of ACD-only clusters at each stage and was comparable to that of the mixed clusters. This probably reflects the fact that most CCR1 compounds were synthesized for particular biological targets and to probe specific structure−activity relationships. As such, most CCR1 compounds can be grouped into logical structural families possibly not represented by any ACD compounds, and this is reflected by their clustering behavior. The ACD structures represent a more random mix; thus the generally low percentage of ACD-only clusters and greater participation in mixed clusters. The total number of ACD-only clusters and singletons produced was 4254. Examination revealed the structures to be sufficiently different from CCR1 compounds so as to be potential targets for acquisition.

A similar study was done by combining the CCR1 and CCR2 structures to give a new database containing approximately 387 000 structures (42.5% CCR1, 57.5% CCR2) and clustering under the same conditions. The percentage of singletons and average cluster sizes were nearly identical to those seen for the mixed CCR1/ACD study (Tables 4 and 6). The percentage of mixed clusters was also similar at each stage (Table 7). However, the ratio of CCR1-only to CCR2-only clusters was much closer to the ratio of these structures in the database, possibly reflecting the fundamentally different structural families present in these databases.

**Table 6.** Cascaded Clustering Results for Mixed CCR1/CCR2 Database

| clustering | structures | clusters (size ≥2) | singletons[a] (%) | av cluster size | max. cluster size |
|---|---|---|---|---|---|
| primary | 387K | 24395 | 44419 (11.5%) | 14.0 | 2224 |
| secondary | 73K | 10170 | 15824 (4.1%) | 5.6 | 437 |
| tertiary | 16K | 2016 | 7617 (2.0%) | 4.0 | 83 |

[a] As a percentage of total starting structures.

**Table 7.** Mixed CCR1/CCR2 Database Clustering: Classification of Clusters

| clustering (size retained) | % CCR1-only | % CCR2-only | % mixed |
|---|---|---|---|
| primary (≥4) | 21.7 | 24.6 | 53.7 |
| secondary (≥2) | 24.2 | 37.3 | 38.5 |
| tertiary (≥2) | 17.6 | 37.8 | 44.6 |
| tertiary (1) | 36.7 | 63.3 | |

It was next decided to study how the clustering behavior of test database structures versus a reference database would vary based on the size of the test database. This could be important if, for example, one wanted to evaluate compounds in a few 96-well plates taken from a larger commercially available collection. Specifically, would structures from a large test database that clustered in mixed clusters when combined with a reference database still show this behavior if the size of the test database were reduced? In other words, would the percentage of test singletons produced vary greatly with the size of the test database? If great variations were found, then it might be advantageous to develop specific size-based clustering conditions for studying mixed databases.

The mixed CCR1/CCR2 database was clustered once using **J**/**K** values of 8/16. All mixed clusters of size 2 or greater were identified and the CCR2 structures (144 257) were extracted. A new set of 12 021 CCR2 structures was defined by selecting every 12th structure from this list. This represents a reduction of over 18-fold when compared to the original CCR2 database. These structures were combined with the original CCR1 set and again clustered once at 8/16. Analysis showed that 79.6% of the CCR2 structures still remained in mixed clusters, which is very good consistency with the previous results. Only 16.8% appeared as new singletons, whereas 3.6% were found in new CCR2-only clusters.

**Comparison: Clustering vs. DPD.** One of the reasons for implementing clustering based on 2D structural descriptors was to have available a method for analysis that would complement the DPD approach, which is based on molecular and physicochemical descriptors. Studies were thus undertaken to determine whether the two methods are indeed complementary. A full description of DPD methodology can be found in ref 11. Only enough detail to permit comparison of the two approaches will be provided here.

**Table 8.** Clustering Comparison: DPD vs Random Set

| database | av cluster size | max. cluster size | clusters (size ≥2) | singletons (%) |
|---|---|---|---|---|
| DPD | 9.2 | 192 | 97 | 373 (29.3%) |
| random | 8.2 | 271 | 97 | 479 (37.4%) |

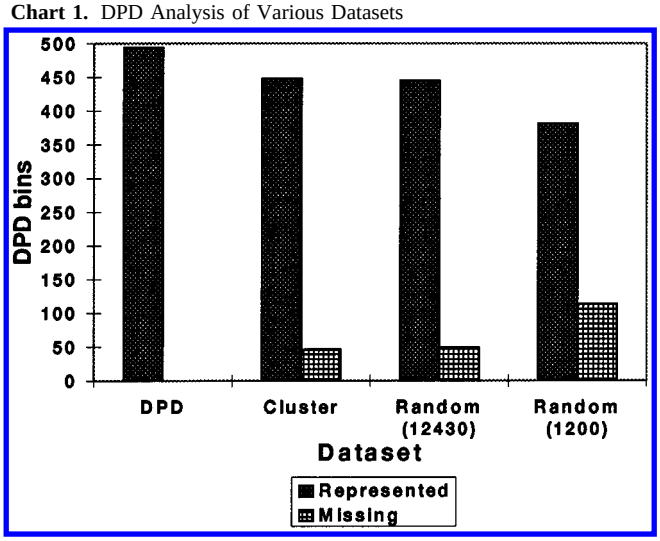
The DPD method involves calculating values for six pairwise noncorrelated molecular and physicochemical descriptors. The possible range of values for each descriptor is subdivided into predefined bins. Each bin (for each descriptor) is given a single-digit code number, beginning at 1 and being incremented by 1 for each succeeding bin. A molecule's values are calculated and translated into bin numbers, and these numbers are concatenated to give a six-digit DPD code. The following are the descriptors used to calculate DPD codes as well as the number of partitions each descriptor was subdivided into for this study: number of H-bond donors, 2; number of H-bond acceptors, 2; electrotopological index, 3; flexibility index, 3; clogP, 4; aromatic density, 4. The total number of DPD codes thus defined is 576, although not all are necessarily accessible. (For example, it is unlikely that a compound with a large number of hydrogen-bond donor and acceptor groups would also have a high clogP value.)

DPD codes were calculated for a filtered subset of molecules from the CCR1 database. The original goal of this exercise was to create a biological screening set of approximately 1000 compounds based on the DPD approach. Therefore, approximately three compounds were selected to represent each of the 576 theoretical DPD codes where possible. The three compounds were selected from all potential candidates based on stock amounts and so as to minimize 2D structural similarity. The DPD Screening Set defined in this manner contained 1281 compounds, which corresponded to 494 different DPD codes.

Two studies were done to compare the DPD set to fingerprint/cluster-based sets. The clustering was performed using the Jarvis-Patrick method and the Tanimoto coefficient of the Daylight fingerprints of the structures as the similarity metric. In the first, 1280 random structures were selected from the CCR1 database. This set and the DPD set were clustered once using **J**/**K** values of 8/16; results are in Table 8. Clustering behavior is similar; average cluster size and number of clusters produced are almost identical. The percentage of singletons is actually higher in the random set. The opposite would be expected if the DPD approach were effective in differentiating structures based on molecular connectivity, as does the fingerprint-based method.

Second, structures from the CCR1 representative set produced by cascaded clustering were classified according to their DPD codes. Random sets of 12 430 CCR1 structures (the approximate size of the representative set before filtering) and 1200 structures (the size of the DPD set) were also created and classified by DPD code. Results are shown in Chart 1. Of the 494 bins which could be filled, the representative set filled only 90%, which is almost exactly that filled by the random set of the same size. This shows that structural fingerprint-based clustering, although intuitively appealing to medicinal chemists, is no more effective than random sampling in identifying compounds having a range of physicochemical properties.

**Chart 1.** DPD Analysis of Various Datasets




## CONCLUSION

The cascaded clustering technique provides a means for using the computationally efficient Jarvis-Patrick clustering algorithm for producing representative sets based on 2D structural characteristics, while minimizing concerns created by excess singleton production. This technique is equally effective for comparing compound collections and has been readily applied to databases of over 350 000 structures. The results are complementary to the DPD approach, which is based on partitioning over molecular and physicochemical properties. Both methods are being used to evaluate the novelty of proposed combinatorial libraries relative to existing compound collections and to select diverse commercially available compounds for acquisition, together with newer 3D methods based on an analysis for all pharmacophores (three and four points).[2,4,30]

The major advantages of the cascaded clustering approach are thus its ease of use, efficiency, applicability to very large structural datasets, and ability to deal with singletons, making it particularly useful for the selection of representative sets and the analysis and comparison of large databases.

## ACKNOWLEDGMENT

We would like to thank the referees and our colleagues at Rhône-Poulenc Rorer, I. Morize, C. Luttmann, and R. Labaudiniere, for their valuable comments and suggestions regarding this work.

## REFERENCES AND NOTES

(1) Taylor, R. Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 59−67.
(2) Pickett, S. D.; Mason, J. S.; McLay, I. M. Diversity Profiling and Design Using 3D Pharmacophores: Pharmacophore-Derived Queries (PDQ). *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214−1223.
(3) Warr, W. A. Combinatorial Chemistry and Molecular Diversity. An Overview. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 134−140.
(4) Ashton, M. J.; Jaye, M. C.; Mason, J. S. New Perspectives in Lead Generation II: Evaluating Molecular Diversity. *Drug Discovery Today* **1996**, *1*, 71−78.
(5) Mason, J. S.; McLay, I. M.; Lewis, R. A. Applications of Computer-Aided Drug Design Techniques to Lead Generation. In *New Perspectives in Drug Design*; Dean, P. M., Jolles, G., Newton, C. G., Eds.; Academic Press: London, 1995; pp 225−253.
(6) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring Diversity: Experimental Design of

Combinatorial Libraries for Drug Discovery. *J. Med. Chem.* **1995**, *38*, 1431−1436.
(7) Briem, H.; Kuntz, I. D. Molecular Similarity Based on DOCK−Generated Fingerprints. *J. Med. Chem.* **1996**, *39*, 3401−3408.
(8) 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.
(9) *Molecular Similarity in Drug Design*; Dean, P. M., Ed.; Blackie Academic and Professional: London, 1995.
(10) Willett, P. Using Computational Tools to Analyze Molecular Diversity. In *Combinatorial Chemistry; A Short Course*; DeWitt, S. H., Czarnik, A. W., Eds; American Chemical Society: Washington, 1997.
(11) Lewis, R. A.; Mason, J. S.; McLay, I. M. Similarity Measures for Rational Set Selection and Analysis of Combinatorial Libraries: The Diverse Property-Derived (DPD) Approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 599−614.
(12) Daylight Chemical Information Systems, Inc., 27401 Los Altos, Suite 370, Mission Viejo, CA 92691 USA.
(13) Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Statistical Assoc.* **1963**, *58*, 236−244.
(14) Jarvis, R. A.; Patrick, E. A. A Clustering Using a Similarity Measure Based on Shared Nearest Neighbors. *IEEE Trans. Comput.* **1973**, *C*-22, 1025−1034.
(15) 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.
(16) Barnard, J. M.; Downs, G. M. Chemical Fragment Generation and Clustering Software. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 141−142.
(17) Barnard, J. M., BCI, personal communication, 1997.
(18) 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.
(19) Mojena, R. Hierarchical Grouping Methods and Stopping Rules: An Evaluation. *Computer J.* **1977**, *20*, 359−363.
(20) Shemetulskis, N. E.; Dunbar, J. B. Jr.; Dunbar, B. W.; Moreland, D. W.; Humblet, C. Enhancing the Diversity of a Corporate Database Using Chemical Database Clustering and Analysis. *J. Comput.-Aid. Mol. Design* **1995**, *9*, 407−416.
(21) Pearlman, R. S.; Stewart, E. L.; Smith, K. M.; Balducci, R. Novel Software Tools for Combinatorial Chemistry and Chemical Diversity. Paper given at the 1997 Charleston Conference *Advancing New Lead Discovery*, Isle of Palms, SC (March 1997).
(22) Turner, D. B.; Tyrrell, S. M.; Willett, P. Rapid Quantification of Molecular Diversity for Selective Database Acquisition. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 18−22.
(23) Clark R. D. OptiSim: An Extended Dissimilarity Selection Method for Finding Diverse Representative Subsets. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181−1188.
(24) MDL Information Systems Inc., 14600 Catalina Street, San Leandro, CA 94577, USA.
(25) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244−255.
(26) Weininger, D. SMILES 1. Introduction and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.
(27) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.
(28) Morize, I., Rhône-Poulenc Rorer, personal communication, 1995.
(29) Chemical Design Limited, Roundway House, Cromwell Park, Chipping Norton, Oxfordshire OX7 5SR, UK.
(30) Mason J. S.; Pickett, S. D. Partition-based selection. *Perspectives Drug Discovery Design* **1997**, *7*, 1−29.

CI980003J