

Clustering Methods and their Uses

Pat Bacha, PhD, Terry Brunck, PhD, and Jack Delany PhD

Daylight Chemical Information Systems, Inc.
Santa Fe, NM

October 2007

Copyright 2007 Daylight Chemical Information Systems, Inc.

Table of Contents

Introduction	3
Daylight Clustering Methods	5
Daylight Fingerprints	5
Example Use Cases	
Jarvis-Patrick	7
Sphere Exclusion	8
Scaffold-directed	9
SAR/Mechanism	9
Preferred Scaffolds	12
Patent Analysis	14
Two-stage Clustering	15

Introduction

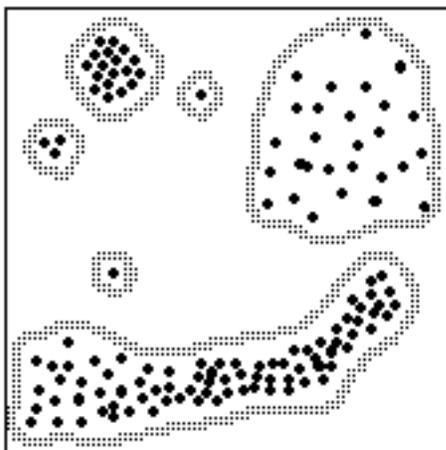
With the increase in the amount of chemical information available, methods that can organize chemical structures and their associated data are essential. A wide variety of clustering methods have been applied in the past with varied results. No one method appears appropriate for all uses so it is important to have a suite of methods that cover the use cases for structural clustering.

There are two important decisions to be made for each potential use of structural clustering: what are the appropriate structural descriptors for the intended use and what is the appropriate method.

Many types of descriptors have been used, including binary fingerprints, structural keys, and derived keys. While fingerprints will be described more fully below, a salient feature of fingerprints is that they cover all structural types by definition. In contrast, structural keys are a predefined set of substructures and descriptors that represent the presence or absence of the substructures in a particular structure. The disadvantage of predefined substructures is that they may not cover all structures equally or at all. Derived keys are keys that are generated from a particular set of structures. They are therefore guaranteed to optimally cover the data set in question. Their disadvantage is that each data set will have a different set of derived keys and direct comparison of key-based properties is impossible, whereas fingerprints and structural keys are portable. For the studies reported below, binary fingerprints and derived keys were selected.

Currently popular clustering methods include Wards,¹ Wards-Kelly,² Jarvis-Patrick,³ k-means,⁴ k-modes,⁵ and sphere exclusion.⁶ More recently, a method based on derived keys has been reported.⁷

Wards clustering is a hierarchical agglomerative method. For N structures, there will be $N(N-1)/2$ nodes in the hierarchy. The main problem is determining the level in the hierarchy where the clusters are optimally defined. The Wards-Kelly method provides such stopping criteria in terms of similarity so that a set of clusters can be identified. Both methods tend to form “globular” clusters.



Cartoon of Cluster Types: Top left is a dense globular cluster represented in a reduced-dimensionality chemical space. Top right is a more diffuse globular cluster. Bottom is a dense, “stringy” cluster

The Jarvis-Patrick method was found to be a good method for structural clustering because it can identify frequently occurring “stringy” clusters. It is not hierarchical and directly provides a set of clusters based on a similarity metric and a user specification of the proportion of nearest neighbors that must be in common between two structures in a cluster.

K-modes clustering, a derivative of k-means, is a rapid, non-hierarchical method wherein the number of clusters is specified and the clusters are seeded with an initial structure. Additional structures are added to clusters based on similarity and then structures may be relocated to other clusters iteratively, again based on similarity. The final clusters depend on the ordering of structures as they are placed in clusters. Clusters will tend to be globular and it must be known *a priori* how many clusters there are (or how many are desired).

Sphere exclusion clustering begins by selecting an initial structure, including in the first cluster all structures that meet a defined similarity threshold, and repeating this process until all structures are in clusters. The structure selection process can be random or directed by some preprocessing of the structures. As with Jarvis-Patrick clustering, clusters are defined by similarity and their number is not predetermined.

Scaffold-directed clustering is one of a relatively few methods that defines clusters in terms of common substructures (scaffolds), not similarity. It is an agglomerative method in which the stopping criterion is minimum coverage of the common substructure over the members of a cluster. Initially, each structure is placed in its own cluster. Clusters are merged if the resulting common substructure meets the minimum coverage requirement. Clustering stops when there are no two clusters suitable for merging. At each merge iteration, the two clusters that result in the largest scaffold coverage are selected for merging. As currently implemented by Daylight, the method is not deterministic in the event of ties in coverage. Scaffold-directed clustering uses derived keys as they are the type of key most easily convert to scaffolds.

References (Introduction):

- 1) Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Statist. Assoc.* **1963**, *58*, 236-244.
- 2) Kelley L.A., Gardner S.P., Sutcliffe M.J. An Automated Approach for Clustering an Ensemble of NMR-Derived Protein Structures into Conformationally-Related Subfamilies. *Protein Eng.* **1996**, *9*, 1063-1065.
- 3) Jarvis R. A., Patrick E. A. Clustering Using a Similarity Measure Based on Shared Near Neighbors. *IEEE Transact. Comput.* **1973** *C22*, 1025-1034.
- 4) Huang Z. Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values, *Data Min. Knowl. Discov.* **1998**, *2*, 283-304.

- 5) Chaturvedi A., Green P. E., Carroll J. D. K-modes Clustering. *J. Class.* **2001**, *18*, 35–56.
- 6) Wooton R., Cranfield R., Sheppy G. C., Goodford P. J. Physicochemical Activity Relationships in Practice. 2. Rational Selection of Benzenoid Substituents. *J. Med. Chem.* **1975**, *18*, 607-613.
- 7) Nicolaou C. Identification of Lead Compounds in Pharmaceutical Data Using Data Mining Techniques. In *Advances in Informatics, Lecture Notes in Computer Science*. Springer Berlin: Heidelberg, 2003; 2563, pp 141-156.

Daylight Clustering Methods

The Daylight clustering package consists of four descriptor/method combinations that were selected to cover nearly all use cases.

Fingerprint/Jarvis-Patrick
Fingerprint/k-modes
Fingerprint/sphere-exclusion
Derived-key/scaffold-directed

The package includes all software necessary to generate fingerprint and/or keys, to cluster structures, and to analyze clusters. Because Daylight fingerprints are portable and have several uses, they will be described in the next section. Daylight's derived keys are data set-dependent and are not used outside of scaffold-directed clustering and scaffold determination.

Daylight Fingerprints

The default structural characterization used in the Clustering Package is based on a binary structural fingerprint, derived as follows.

- Generate a number for each path in a structure. Starting with each atom, traverse all paths, branches, and ring closures up to a certain depth (typically 8). For each substructure, derive a hash-like number from unique, relatively prime, order-dependent contributions of each atom and bond type. Critical properties of this number are that it is reproducible (each substructure produces a single number) and its value and graph are not correlated (a linear congenital generator is used to ensure this).
- Map each resulting number into a large range (typically 2K-64K) to produce a redundant (typically 4 or 5 bits per pattern), large-scale, binary representation of the substructural elements. The resultant "fingerprint" contains a large amount of information at a low density.

Because the number of possible paths is so huge, it is not possible to assign a particular bit to each pattern.

- Iteratively "fold" the fingerprint by OR-ing the two halves of the fingerprint until the bit density reaches a minimum required value or until the fingerprint reaches a minimum allowable length. The resulting fingerprint now has a high information density with a minimal (and controllable) information loss.

Fingerprints thus generated approximate a complete characterization of substructural content. The quality of the approximation is determined by the information content and a user-defined density.

Fingerprints have several advantages over structural keys:

- Since fingerprints have no predefined set of patterns, one fingerprinting system serves all databases and all types of queries. Furthermore, the information content is not biased by the generation method—fingerprints work equally well on reagents, drugs, dyes, and insecticides without any "tweaking."
- More effective use is made of the bitmap. Structural keys are usually very "sparse" (mostly zeros) because a typical molecule has very few of the patterns that the structural key's bits represent. Although a mathematical analysis of fingerprint density is beyond the scope of this introduction, fingerprints can be relatively "dense" (20-40% ones) without losing specificity. The result is that a fingerprint can be much smaller than a structural key with the same discriminating power.
- The patterns that go into a fingerprint are highly overlapped—except for "lone atoms," each pattern shares portions of itself with at least one other pattern. The result is that the more complex a molecule gets, the more accurately its fingerprint characterizes it.

The main disadvantage of fingerprints is that the bits have lost their connection to specific structural features, making it difficult to impossible to reassemble a structure from its fingerprint.

Example Use Cases

The following examples show the situations in which the various Daylight clustering methods can be applied to advantage. These examples also provide some information about performance.

Jarvis-Patrick (JP) Clustering (Large Library Analysis)

JP clustering is a fast method to organize large data sets based on fingerprint similarity. There are many reasons to perform such clustering, including database profiling, database comparisons, and database organization. Profiling a database can give an indication of database diversity and can show the distribution of structures across clusters. These profiles can then be used to compare two databases in terms of diversity. A database of compounds for screening can be organized via clustering, which permits the immediate selection of structures similar to hits by selecting compounds from clusters populated with one or more hits.

To demonstrate the use of JP clustering, we have clustered four large data sets as examples. The databases selected range from ~200K to 800K compounds. For each data set we converted input SD or RDfiles to canonical SMILES using Daylight's Convert Package and removed duplicate structures. We generated fingerprints, nearest neighbor lists, and JP clusters using default parameters in each case. The results are shown in the table below.

	NCI ²	ACD ³	MCD ⁴	PubChem ⁵
Unique compounds	215,017	550,222	581,805	778,159
Timing ¹	1.5 h	7.9 h	12.5 h	19.2 h
Clusters	16,467	42,996	37,777	60,954
Singletons	38,855	58,961	42,623	106,363
Average cluster size	10.5	11.4	14.3	11.0
Largest cluster	217	223	558	711
Percent Singletons	18.1	10.7	7.3	13.7

In terms of performance, JP is an $O(N^2)$ algorithm with the largest data set requiring ~19 hours. The number of singletons, or un-clustered structures, in these cases is in line with the expectation of 10-20% of total structures. The fact that singletons are not forced into clusters is one of the advantages of JP clustering.

By way of database comparison, MCD appears to be the least diverse, having both the largest average cluster size and the lowest percentage of singletons. Using the same measure, NCI is the most diverse.

References (Jarvis-Patrick Clustering):

- 1) Timing was done on a machine with 3 Ghz processor and 2 GB RAM under Red Hat ES3
- 2) NCI - http://dtp.nci.nih.gov/docs/3d_database/Structural_information/structural_data.html (downloaded August 2003).
- 3) ACD - http://www.mdl.com/products/experiment/available_chem_dir/index.jsp (v2007.2).
- 4) MCD - www.gvkbio.com (October 2006).
- 5) PubChem - <http://pubchem.ncbi.nlm.nih.gov/> (downloaded August 2005).

Sphere Exclusion (Rapid Diverse Compound Selection)

A common need within lead discovery is to rapidly select a set of dissimilar compounds from a library. Fortunately, clustering can be used to select compounds that are similar by selecting from a single cluster, or it can be used to select compounds from all clusters to get a broad sampling. Sphere exclusion clustering is a very rapid method to organize compounds based on similarity for use in compound selection.

In this example, we consider the situation in which the need is to select about 1000 diverse structures for screening from a vendor library. We use the sphere exclusion clustering method, as it is a rapid method that forms clusters iteratively by selecting a compound as a cluster center, adding similar compounds to this cluster, and excluding dissimilar compounds. Once this cluster is completed, an excluded compound, not in any previous cluster, is selected and the process is repeated.

The libraries selected were from Maybridge¹ and Chembridge.² The similarity threshold parameter for Daylight's spherex clustering was adjusted to 0.35 for Maybridge and 0.4 for Chembridge, which gives approximately 1000 clusters for each library. The table below shows the clustering results. The speed of the method is clear from the clustering times.

	Maybridge	Chembridge
Number of Compounds	57,293	75,358
Clustering Time	26 s	62 s
Clusters	1049	1043
Singletons	146	345
Average cluster size	54.6	53.5
Largest cluster	1075	2491

References (Sphere Exclusion):

1) Maybridge -

http://www.maybridge.com/portal/alias__Rainbow/lang__en/tabID__146/DesktopDefault.aspx (supplied January 2002).

2) Chembridge - <http://www.chembridge.com/chembridge/compound.html> (supplied May 1999).

SDCluster (The Value of Scaffolds)

Several use cases are presented below to show the types of analyses that are appropriate for the Daylight's sdcluster method. In general, the method is appropriate when the number of structures to be clustered is moderate (up to about 20,000 structures) and when there is a need for scaffolds for structure-activity relationship (SAR) studies, mechanism studies, identification of preferred scaffolds, screening data analysis, or patent analysis. In cases where scaffolds are desired for a large set of structures, it is possible to use a two-step process as described in the next section.

Case A: Clustering for SAR or Mechanism Studies

When performing SAR or mechanism of action studies, it is important to use clustering methods that tend to produce clusters that are pure with regard to the mode of action and possible activity. Here we analyzed several data sets using only structural information (non-supervised clustering). For each of these data sets, we eliminated duplicate structures and entries that had no associated structures. For comparison purposes, we have included two traditional methods, Jarvis-Patrick and k-modes.

We have shown with each example the single coverage parameter for sdcluster. In each case, we adjusted the parameters for Jarvis-Patrick and k-modes to give approximately the same number of clusters and coverage of compounds in clusters as scaffold-directed clustering gave. This is a tremendous hint for selecting the parameters, because otherwise there is no basis but trial and error for selecting the clustering parameters. It would be rare to know ahead of time how many clusters are optimal given the desire to have scaffolds to represent a homogeneous set. The results for these two methods are thus partially optimized with the knowledge of the approximate number of structural clusters present in the various data sets.

Clustering Results:

NCI Anti-cancer Mechanism Data Set¹

121 unique compounds in 5 mechanistic classes

min_coverage = 0.3 (default)

	Scaffold-directed	Jarvis-Patrick	k-modes
Clusters	30	25	23
Singletons	28	20	17
Average cluster size	3.1	4.0	4.5
Largest cluster	13	10	13
Common substructure:			
Clusters with ≥10 atoms in common	23 (77%)	19 (76%)	10 (43%)
Clusters with <10 atoms in common	7 (23%)	5 (20%)	13 (57%)
Clusters with no atoms in common	0	1 (4%)	0
Purity:			
Clusters with 1 mechanism of action	23 (77%)	14 (56%)	9 (39%)
Clusters with 2 mechanisms of action	7 (23%)	9 (36%)	11 (48%)
Clusters with 3 mechanisms of action	0	1 (4%)	2 (9%)
Clusters with 4 mechanisms of action	0	1 (4%)	1 (4%)

This example shows that the clusters generated by sdcluster have a higher purity of mechanism than those generated by either JP or k-modes, even when the latter use the results of sdcluster to set parameters.

Briem Drug Mechanism Data Set²

377 unique compounds in 5 mechanistic classes

min_coverage = 0.4

	Scaffold-directed	Jarvis-Patrick	k-modes
Clusters	80	74	74
Singletons	75	89	66
Average cluster size	3.8	3.9	4.2
Largest cluster	12	12	25
Common substructure:			
Clusters with ≥15 atoms in common	70 (88%)	67 (91%)	64 (86%)
Clusters with <15 atoms in common	10 (22%)	7 (9%)	10 (14%)
Purity:			
Clusters with 1 mechanism of action	71 (89%)	66 (89%)	51 (69%)
Clusters with 2 mechanisms of action	9 (11%)	8 (11%)	23 (31%)

This example shows that it may be possible to achieve the purity of mechanism shown by sdcluster, at least with JP, if the knowledge gained from sdcluster is used to select the parameters for the other clustering methods. The significant point is that sdcluster uses a single intuitive parameter, the coverage of the scaffold as represented as a SMARTS, to generate its clusters. For any particular data set, you do not need to know how many clusters there are ahead of time; you only need to know the relationship between the scaffold and the cluster members that you wish to achieve.

Quinolones³

158 unique compounds with MIC values; 83 actives; 75 inactive structures; active means MIC<32uM

min_coverage = 0.4

	Scaffold-directed	Jarvis-Patrick	k-modes
Clusters	16	16	14
Singletons	4	20	6
Average cluster size	9.6	8.6	10.9
Largest cluster	50	26	43
Common substructure:			
Clusters with ≥20 atoms in common	14 (88%)	12 (60%)	8 (57%)
Clusters with <20 atoms in common	2 (12%)	4 (40%)	6 (43%)
Purity:			
Clusters with 100% actives	2 (13%)	1 (6%)	1 (7%)
Clusters with 67% to 99% actives	2 (13%)	4 (25%)	1 (7%)
Clusters with 34% to 66% actives	4 (25%)	4 (25%)	4 (29%)
Clusters with 1% to 33% actives	2 (13%)	2 (13%)	1 (7%)
Clusters with 0% actives	6 (38%)	5 (31%)	7 (50%)

In this example, we tried to push the performance of unsupervised clustering to see if it would be possible to generate clusters that were pure with respect to activity. The quinolone compounds in this data set are closely related. Still, 51% of the quinolone class compounds were found in clusters that were purely active or inactive compounds using sdcluster with the coverage parameter increased from the default. By comparison, only 37% of the JP clusters were pure, and again, this result is based on knowing the results of sdcluster. For unexplained reasons, use of k-modes on this data set (with parameters set based on sdcluster results) gave a very high proportion of clusters that were inactive.

Case B: Preferred Scaffolds

The sdcluster method can be used to identify preferred scaffolds. For this example, we selected a chemical library that was assembled to provide compounds for screening that would be likely to be, or that have been shown to be, active against a specific protein target, oxidosqualene-lanosterol-cyclase.⁴ This library is a product of Otava (Kyiv, Ukraine). We selected a second set composed of all of the analgesics in the WDI database.⁵ The scaffolds generated by sdcluster methodology represent, in an abstract way, the general families of compounds known to be active (or likely to be active) against this family of enzymes. The scaffolds as SMARTS queries can be applied against another library, corporate or vendor, to identify compounds that may also be active.

Clustering Results:

Using the default minimum scaffold coverage, sdcluster produced the following results.

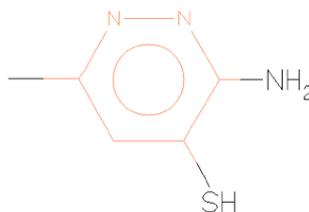
	Analgesics	Inhibitors
Number of Compounds	1823	1277
Clusters	185	51
Singletons	88	2
Average cluster size	9.4	25.0
Largest cluster	127	272

The data reduction from compounds to scaffolds varied from 10:1 to 25:1. The amount of data reduction is a feature of the compound sets and could not be easily predicted or estimated ahead of time. This shows the difficulty encountered when using traditional clustering methods where parameters are selected based on the number of clusters or the number of singletons.

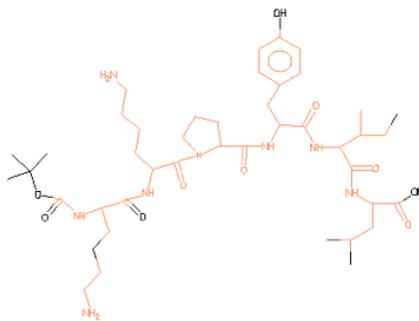
Examples of the scaffolds highlighted on the smallest molecule of a cluster are shown below.

Analgesics:

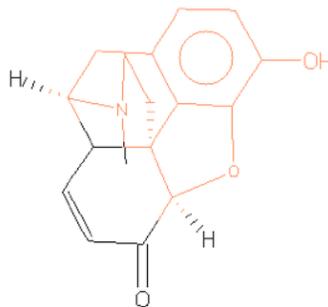
cluster with smallest scaffold
6 members
minimum coverage 0.3125



cluster with largest scaffold
5 members
minimum coverage 0.7581

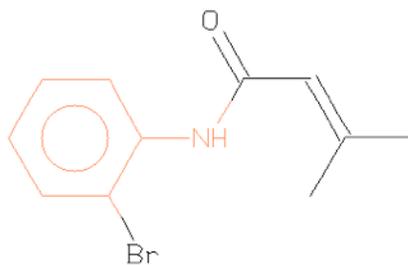


cluster with most compounds
127 members
minimum coverage 0.3061

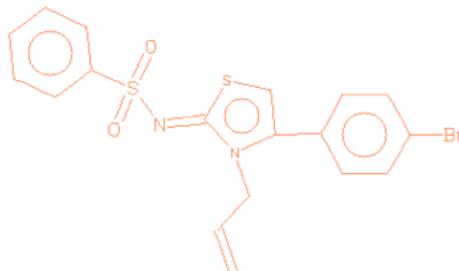


Inhibitors:

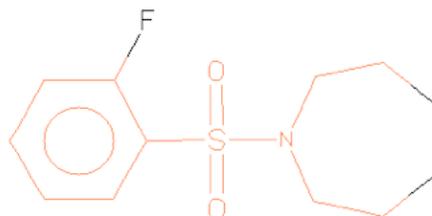
cluster with smallest scaffold
5 members
minimum coverage 0.9259



cluster with largest scaffold
5 members
minimum coverage 0.7581

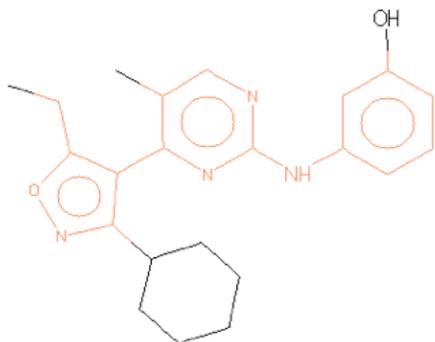


cluster with most compounds
272 members
minimum coverage 0.3636

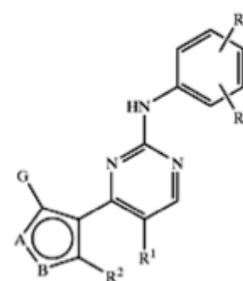


Case C: Patent Analysis

Chemical patents often contain claims of varying levels of generality along with claims for specific chemical structures. The sdcluster method can be used in two ways: to help generate Markush claims from a set of structures, and to analyze Markush claims in a patent. As an example of the latter, we selected a patent whose specific claims were curated by GVKBio.⁶ The structures from these specific claims were clustered to provide the scaffold shown below.



Scaffold from the only cluster



Markush Claim 1

This scaffold can then be compared to the Markush claims of the patent to determine if the general claims are well supported by the actual compounds that were reduced to practice. In this case, it appears that there was poor representation of some parts of Markush claim 1. For example, A and B in the Markush claim were only O and N, respectively, in the actual compounds. Also, G of the Markush claim always contains a methylene group adjacent to the ring.

This type of analysis can be used to construct patents that are more defensible or to challenge patents whose claims are less well supported.

References (SDCluster):

- 1) Weinstein J. N., et al., *Science* **1992**, 258, 447; van Osdol W. W., et al., *J. Nat. Cancer Inst.* **1994**, 86: 1853;
http://dtp.nci.nih.gov/docs/cancer/searches/standard_mechanism.html
(downloaded August 2003); mechanism classification taken from the publication as presented on the NCI website; combined DNA antimetabolite and RNA/DNA antimetabolite mechanism classes.
- 2) Briem H. and Lessel U. F., *Perspect Drug Discov Design* **2000**, 20, 231; structures taken from publication.
- 3) Gozalbes R., et al. Prediction of Quinolone Activity Against Mycobacterium avium: Molecular Topology and Virtual Computational Screening. *Antimicrobial Agents and Chemotherapy*, **2000**, 44, 2771; structures and

classification taken from publication; Compounds with MIC ($\mu\text{g/mL}$) < 32 were considered active; Compounds with MIC ($\mu\text{g/mL}$) ≥ 32 were considered inactive.

4) <http://otavachemicals.com>; supplied March 2007; Only used those structures from oxidosqualene-lanosterol-cyclase4 set.

5) World Drug Index, Thompson Scientific;
<http://scientific.thomson.com/products/wdi/> (v2007.2);

Only used those structures with activity listed as ANALGESIC.

6) US Patent 6689778 B2, 2004, Inhibitors of SRC and LCK Protein Kinases, Bemis et al. Assignee: Vertex Pharmaceuticals Inc; curation by GVK Biosciences, Hyderabad, India as a part of the Protein Kinase database; supplied August 2006.

Two-Step Clustering (Scaffolds from large datasets)

Sequential Clustering: k-modes \rightarrow sdcluster

When it is desirable to generate clusters with scaffolds for larger data sets, a two-step process can be used. The first step is preferably a rapid method that is used to crudely partition the data set. In this example, we used k-modes so that we could easily estimate the average size of the clusters after the first step. The second step clusters each partition from the first step and generates scaffolds for each cluster. This process could be used for large volumes of high-throughput screening (HTS) data where the clusters and scaffolds could be used to generate structure-activity relationships (SAR) or rules. The data sets below are of similar size to hit lists.

We have chosen five data sets of increasing size to demonstrate this approach. The first three sets approximate the size and nature of screening data sets. Each set was converted from SD or RDfile format to canonical SMILES using Daylight's Convert Package and only unique structures were used. The latter two sets were selected to show the feasibility of clustering large data sets. We chose the number of partitions in the first step with the goal of having about 1000 compounds per partition. Although the average cluster size was near the desired value, the largest clusters were from 2 to 20 times larger. The results of each step are shown in the table below.

	Nat_NCI ²	HIV ³	WDI ⁴	ACD ⁵	MCD ⁶
Unique compounds	34,943	42,001	73,578	556,222	581,805
k-modes:					
Timing ¹	0.3 min	0.67 min	1.9 min	30 min	69 min
Clusters	35	45	75	522	574
Singletons	0	0	0	0	1
Average cluster size	998.4	933.4	981.0	1,054.1	1,015.4
Largest cluster	2,494	5,384	6,258	12,545	20,019
Scaffold-directed clustering: (min_path = 0, default)					
Timing ¹	0.2 h	0.9 h	6.8 h	22.2 h	41.5 h
Clusters	1756	3921	5,702	12,451	20,572
Singletons	549	1868	6,891	12,639	25,711
Average cluster size	19.5	8.7	13.6	33.8	27.4
Largest cluster	563	334	474	3,456	1,691
Scaffold-directed clustering (min_path = 4):					
Timing	0.2 h	0.4 h	3.6 h	20.6 h	38.7 h
Clusters	3,456	7,848	5,675	12,266	20,624
Singletons	914	3,821	7,744	18,297	26,498
Average cluster size	19.0	8.7	13.1	31.5	27.2
Largest cluster	563	334	474	3,456	1,691

Note that the k-modes clustering shows the same qualitative results as did the Jarvis-Patrick method with regard to database diversity. MCD, the least diverse set, has the largest maximum cluster size.

It is difficult to predict the time required for scaffold-directed clustering since there is a strong dependence on structure and hence scaffold complexity. Clustering times can be shortened by eliminating the smaller scaffold fragments through the use of a minimum path length parameter.

It is possible that the same scaffolds might be learned from different clusters. It is also possible that a single compound might be represented by more than one scaffold. As a final step, duplicate scaffolds can be removed and all of the compounds can be matched against each scaffold using the coverage criteria. This will provide the most complete way to combine the clusters into sets having unique scaffolds. As a further step, the singletons can be combined and re-clustered using sdcluster to ensure that no scaffolds were missed by having

separated related molecules across the initial coarse clusters. All of these steps could be automated to make the processing of large data sets routine.

If the goal of clustering was HTS analysis, inactive compounds can be matched against the scaffolds learned from the hit list (using the same coverage criteria) so that the clusters will then contain both active and inactive structures. This makes the evaluation of leads from HTS assays much more thorough. Statistical analysis of the distributions of activities of a cluster can be compared to the activity distribution of the entire data set to provide a sound basis for lead selection. The selected clusters and their associated scaffolds can be used to generate R-Tables and SAR.

References (Two-Step Clustering):

- 1) Timing done on a machine with 2.8 GHz RHES 3.0 4.096 GB RAM.
- 2) Nat_NCI, Natural products - <http://dtp.nci.nih.gov/branches/npb/repository.html> (downloaded June 2001).
- 3) HIV, NCI compounds tested for anti-HIV activity; http://dtp.nci.nih.gov/docs/aids/aids_data.html (October 1999 release).
- 4) WDI – World Drug Index; <http://scientific.thomson.com/products/wdi/> (v2007.2).
- 5) ACD - http://www.mdl.com/products/experiment/available_chem_dir/index.jsp (v2007.2).
- 6) MCD - <http://www.gvkbio.com> (supplied October 2006).