DAYLIGHT

Chemical Information Systems, Inc.

Scaffold-Directed Clustering Use Cases

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 possibly activity. Here we analyzed several data sets using only structural information (non-supervised clustering).

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, since otherwise there is no basis but trial and error for selecting the clustering parameters. Thus, the results for these two methods are 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 Set[1]

(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.

CLUSTERING RESULTS

Briem Drug Mechanism Data Set[2]

(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 sdclusters, at least with JP, if the knowledge gained from sdclusters is used to select the parameters for the other clustering methods. The significant point is that sdclusters 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, just the relationship between the scaffold and the cluster members that you wish to achieve.

CLUSTERING RESULTS

Quinolones[3]

(158 unique compounds with MIC values; 83 actives; 75 inactives; 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 new 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, or that have been shown to be, active against a specific protein target, oxidosqualene-lanosterol-cyclase.[4] 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. This library is a product of Otava (Kyiv, Ukraine). We selected a second set comprised of all of the analgesics in the WDI database.[5] 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, sdclusters produced the following results.

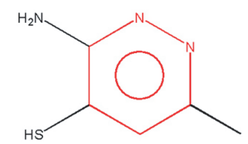
	Analgesics	Inhibitors
Number of Compounds	1823	1277
Clusters	185	51
Singletons	88	2
NAverage Cluster Size	9.4	25.0
Number of Compounds	127	272

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 traditional clustering methods where parameters are selected based on the number of clusters or the number of singletons.

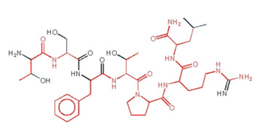
Examples of the scaffolds highlighted on the smallest molecule of a cluster are shown below.

ANALGESICS

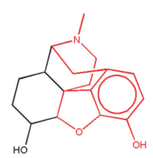
cluster with smallest scaffold
6 members
minimum coverage 0.3125



cluster with largest scaffold
5 members
minimum coverage 0.7581

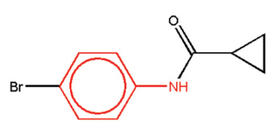


cluster with most compounds
127 members
minimum coverage 0.3061

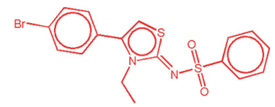


INHIBITORS

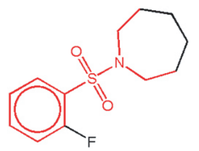
cluster with smallest scaffold
5 members
minimum coverage 0.9259



cluster with largest scaffold
5 members
minimum coverage 0.7581



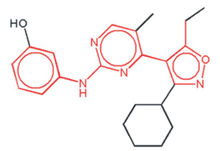
cluster with most compounds
272 members
minimum coverage 0.3636



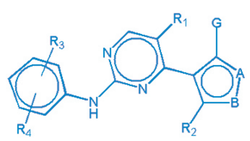
CASE C:
PATENT ANALYIS

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. The structures from these specific claims were clustered to provide the scaffolds shown below. These scaffolds 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.

Scaffold from the only Cluster



Markush Claim



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 that whose claims are less well supported.

References:

1) Weinstein et al., Science 258: 447, 1992; van Osdol et al., J. Nat. Cancer Inst. 86: 1853, 1994; http://dtp.nic.nih.gov/docs/cancer/searches.
2) Briem and Lessel, Perspectives in Drug Discovery Design 20: 231, 2000.
3) Gozalbes, et al Antimicrobial Agents and Chemotherapy, 2000 Prediction of Quinolone Activity against Mycobacterium avium: Molecular Topology and Virtural Computational Screening.
4) http://otavachemicals.com.
5) World Drug Index, Thompsom Scientific; http://scientific.thomson.com/products/wdi/.
6) US Patent 6689778 B2 2004, Inhbitiors of SRC and LCK Protein Kinases, Bemis et al. Assignee: Vertex Pharmaceuticals Inc; curation by GVK Biosciences, Hyderbad, India as a part of the Protein Kinase database.

Prepared by

Pat Bacha, Ph. D, Software Support Engineer, Daylight CIS, Inc. - pat@daylight.com

Terence Brunck, Ph. D, VP - Research & Development, Daylight CIS, Inc. - terry@daylight.com

Peter Nielsen, Director - Marketing & Business Development, Daylight CIS, Inc. - peter@daylight.com

DAYLIGHT

Chemical Information Systems, Inc.

www.daylight.com