

# Structure Based Clustering of NCI's Anti-HIV Library

András Borosy - Ivax Drug Research Ltd.  
Ferenc Csizmadia - ChemAxon Ltd.  
András Volford - ChemAxon Ltd.

## Introduction

To support clustering, new software called *JKlustor* has been developed as an add-on module for ChemAxon's chemical database handling system, *JChem*. The application can generate 2D hashed fingerprints for molecules, but real number descriptors may also be used during the calculation. The clustering process applies a version of the *Jarvis-Patrick method*, which is based on variable-length neighbor lists. In the case of fingerprint input, the measure of similarity is the Tanimoto coefficient. Another clustering module, applying *Ward's minimum variance method* and using Murtagh's *reciprocal nearest neighbor (RNN) algorithm* as a heuristic, is also introduced.

## CLUSTERING METHODS

### (A) Ward Method

Hierarchical agglomerative clustering methods build a hierarchical grouping of the objects in a data set by agglomerations using a dissimilarity measure. A single partition, i.e., a cut across the hierarchy, can then be taken at any level to give the desired number of clusters. Ward's method [1,2] minimizes the variance of groups. The dissimilarity measure used in the Ward algorithm is the *Euclidean distance*. Usually, hierarchical clustering is very demanding in computational resources, with time and memory complexity of  $O(N^2)$  and  $O(N^2)$ , respectively, where  $N$  is the number of objects to be clustered. Applying *Murtagh's reciprocal nearest neighbor (RNN) algorithm* [3], the memory and time demand can be decreased to  $O(N^2)$  and  $O(N)$ , respectively.

### (B) Variable-length Jarvis-Patrick method

The algorithm used is a modified version of the original Jarvis-Patrick method [4]. The procedure is the following:

- For each structure, collect the nearest neighbors that have a dissimilarity less than a  $T$  threshold value.
- Two structures cluster together if
  - they are in each other's list of nearest neighbors
  - they have at least  $R_{min}$  of their nearest neighbors in common, where  $R_{min}$  is a ratio of the length of the shorter list

$T$  and  $R_{min}$  are adjustable parameters. The advantages of the variable-length clustering method are explained by Brown and Martin [6] and Barnard and Downs [7].

## FINGERPRINTS

The fingerprint of a molecule is bit string (a sequence of "0" and "1" digits) that contains information on the structure. Similar molecules should have "similar" fingerprints.

## Results and Discussion

### DATASET

Biological screening data and structures of 42,687 compounds were obtained from National Cancer Institute (NCI). The data are from a cell-based assay measuring protection from HIV-1 infection; and results are categorized as confirmed active (CA), confirmed moderately active (CM), or confirmed inactive (CI) in each group with 534, 1111, and 41042 molecules, respectively [5].

### SOFTWARE

*JKlustor* is a software for diversity calculations and clustering. Version 1.5.9 was used during the calculations.

General features:

- written in Java - it runs under practically every operating system.
- input/output: text files or database tables (practically all database engines having an SQL interface are supported).

At present, *JKlustor* includes the following command-line tools:

- GenerFP* generates 2D hashed fingerprints from molecule files (SDF, SMILES, etc.)
- Compr* compares two compound libraries using diversity and dissimilarity calculations.
- Jarp* and *Ward* perform Jarvis-Patrick and Ward type clustering, respectively, based on fingerprints and/or other data stored in a database table or a text file.
- MarvinView* displays structures and other data in structure data files (SDF files, SMILES files, etc.)

In the case database input, fingerprints are taken from structure tables (that may be generated by *JChem*).

See <http://www.chemaxon.com/products.html> for more details.

### FINGERPRINT PARAMETERS

*JKlustor* uses a proprietary method for generating fingerprints, which detects all patterns (linear sequences of atoms and bonds) up to a given size in the molecule and switches on bits that represent these patterns.

- Fingerprint length (number of bits used)
- Maximum pattern length
- Number of bits representing patterns

### TANIMOTO COEFFICIENT

Nearest neighbor searching in *Jarp* finds molecules that are similar to the query object. The Tanimoto (or Jaccard) coefficient is calculated by the following formula in the case of fingerprint input:

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

where  $N_A$  and  $N_B$  are the number of 1 bits in the fingerprint of A and B, respectively,  $N_{A\&B}$  is the number of common 1 digits in the two fingerprints. The dissimilarity measure used for the Jarvis-Patrick method was  $1 - T(A,B)$ .

### EVALUATING THE CLUSTERING METHODS

After *leaving out singletons* (clusters with only one member) from the data set, the following sets and parameters were determined:

The whole data set is denoted as  $D$ . The *proportion of actives* is calculated as

$$P_a = \text{number of actives in } D / \text{number of structures in } D$$

The *active cluster subset* ( $D_a$ ) contains compounds from all clusters that contain at least one active compound. As in [6], the *proportion of actives in the active cluster subset* is calculated by

$$P_{a,a} = \text{number of actives in } D_a / \text{number of structures in } D_a$$

The *enrichment ratio* expresses how richer is  $D_a$  in actives than  $D$ .

$$E = P_{a,a} / P_a$$

Since the number of actives in  $D_a$  equals to the number of actives in  $D$ ,

$$E = \text{number of structures in } D / \text{number of structures in } D_a$$

A greater value of  $E$  ensures a better separation. Selecting parameters that tighten the clusters usually increase  $E$ . In that case however, especially when the Jarvis-Patrick method is used, the number of singletons also grows, which is a negative effect.

### PARAMETERS VARIED

The following parameters were varied during the evaluation:

- Fingerprint length
- Max. pattern length used for generating the fingerprints
- Number of clusters (*Ward*)
- $T$ -dissimilarity threshold (*Jarp*)
- $R_{min}$ -minimum ratio of common neighbors (*Jarp*)



## CONCLUSIONS

- Ward provided better results than Jarp.
- Increasing the fingerprint length improved the separation in the case of Jarp. The reason is that longer fingerprints provide less chance for overlapping bit placements for different patterns.
- On the other hand, increasing the fingerprint length decreased the separation in the case of Ward in most cases. Longer fingerprints contain more 0 bits. Corresponding 0 bits have much less significance in structural similarity than corresponding 1 bits. Since Euclidean distance doesn't distinguish between the above cases, it is less efficient when the proportion of 0 values are high.
- Smaller clusters improved the results. (The average cluster size can be decreased by increasing  $R_{min}$  and decreasing  $T$  in the case of Jarp).
- If the average cluster size is small, clustering provides too many singletons, especially in the case of Jarp.
- Jarp is significantly faster than Ward. Calculations with Jarp took less than 2 hours, while Ward run for almost one day using a 850Mhz Celeron under Linux, using IBM's JDK 1.3.

These results are preliminary. We continue testing the effects of the various fingerprint parameters and plan to publish our results in a new paper.

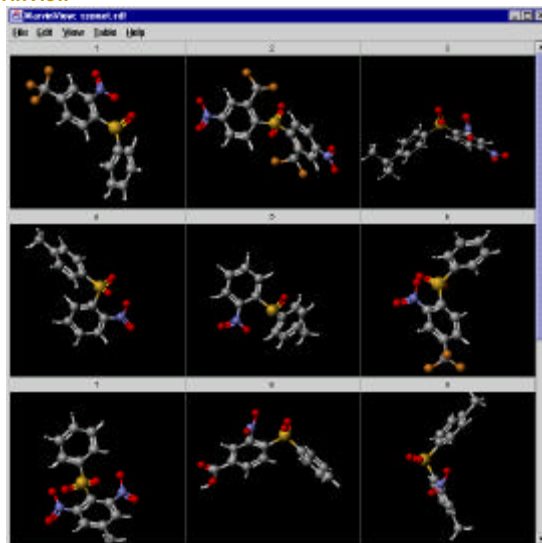
### CLUSTERING RESULTS

with Jarp	fingerprint length	max.pattern length	$T$	$R_{min}$	$E$
1	512	5	0.1	0.3	8.74
2	512	5	0.1	0.5	9.15
3	512	5	0.1	0.8	13.54
4	512	5	0.12	0.8	11.14
5	512	5	0.12	0.5	7.00
6	512	5	0.12	0.3	8.15
7	512	5	0.14	0.3	4.45
8	512	5	0.14	0.3	5.05
9	512	5	0.14	0.3	9.74
10	512	5	0.16	0.8	8.40
11	512	5	0.16	0.5	4.01
12	512	5	0.16	0.3	3.67
13	512	5	0.18	0.3	2.80
14	512	5	0.18	0.5	3.10
15	512	5	0.18	0.8	6.61
16	512	5	0.2	0.8	4.86
17	512	5	0.2	0.5	2.62
18	512	5	0.2	0.3	2.34
19	1024	6	0.1	0.3	12.47
20	1024	6	0.1	0.5	13.57
21	1024	6	0.1	0.8	15.25
22	1024	6	0.12	0.8	14.92
23	1024	6	0.12	0.5	10.72
24	1024	6	0.12	0.3	10.08

### CLUSTERING RESULTS

with Ward	fingerprint length	max.pattern length	cluster size	$E$
1	512	5	1000	2.61
2	512	5	5000	10.20
3	512	5	10000	18.19
4	512	5	15000	24.54
5	512	5	20000	28.33
6	1024	5	1000	2.71
7	1024	5	10000	17.21
8	2048	6	1000	2.57
9	2048	6	10000	17.69
10	2048	6	15000	24.12
11	2048	6	20000	28.93

Example for displaying a cluster using MarvinView



## References

- 1 Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. J. Am. Stat. Assoc. 1963, 58, 236-244.
- 2 Ward, J. H.; Hook, M. E. Application of an Hierarchical Grouping Procedure to a Problem of Grouping Profiles. Educ. Psychol. Meas. 1963, 23, 69-82.F.
- 3 Murtagh, A Survey of Recent Advances in Hierarchical Clustering Algorithms Computer Journal, 26, 354-359 (1983),
- 4 Jarvis, R. A.; Patrick, E. A. Clustering Using a Similarity Measure Based on Shared Nearest Neighbors IEEE Trans. Comput. 1973, C22, 1025-1034
- 5 National Cancer Institute, Bethesda, MD, USA, [http://dtp.nci.nih.gov/docs/aids/aids\\_data.html](http://dtp.nci.nih.gov/docs/aids/aids_data.html)
- 6 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
- 7 Barnard, J. M.; Downs, G. M. Chemical Fragment Generation and Clustering Software J. Chem. Inf. Comput. Sci. 1997, 37, 141-142.

