PRODUCT DESCRIPTIONS

Chemical Fragment Generation and Clustering Software[§]

John M. Barnard*,[†] and Geoff M. Downs[‡]

Barnard Chemical Information Ltd., 46 Uppergate Road, Stannington, Sheffield, S6 6BX, UK

Received June 27, 1996[⊗]

Barnard Chemical Information Ltd.'s software products for generation of fragment-dictionary-based chemical structure fingerprints and for hierarchical and nonhierarchical clustering of large files of structures are described.

THE COMPANY

Barnard Chemical Information Ltd. (BCI) is a British company, which was formed in 1985 to provide highly-specialized services in the design and implementation of chemical structure-based information systems worldwide. In addition to consultancy, contract research, and development work, and computer programming contracts, the company is increasingly developing and marketing its own software products. These fall into two groups: the first concerned with structure "fingerprint" generation, and the second being a set of efficient implementations of a number of hierarchical and nonhierarchical clustering methods. All programs are available for a variety of hardware platforms, including Unix workstations and desktop PCs.

FINGERPRINT GENERATION SOFTWARE

BCI's fingerprint generation programs allow the creation of structural "fingerprints" suitable for use in diversity analysis work, which are bitstrings based on the presence or absence of 2-D structural features of a molecule, listed in a fragment dictionary. Programs are available to build and maintain such dictionaries, for use with different databases or applications. Six different families of fragment can be generated: **Augmented Atoms** (atom with its immediate neighbors and connecting bonds), **Atom/Bond Sequences** (atom/bond paths up to a user-specified maximum length, with stereospecific bond types to distinguish configurations around double bonds and at adjacent ring substitution positions), **Atom Pairs** ("topological distances" between pairs of atoms[1]), **Ring Composition Fragments** (atom/bond sequences around rings in the Extended Set of Smallest Rings (ESSR)[2]), **Ring Fusion Fragments** (sequence of ring connectivities around ESSR rings), and **Ring Ortho Fragments** (stereo configuration at nonplanar orthofusion junctions). Fragments are initially generated with fully-specified atoms and bonds and are then progressively generalized by using user-defined intermediate-level atom and bond types. (e.g., "any ring bond" or "any halogen"), and "any atom" and "any bond" types.

The **MAKEFRAG Program** generates all fragments in the required classes for a file of connection tables and their generalizations and outputs a list of all fragments found along with their incidence and occurrence frequencies. This output can be edited to form a list of fragments required for a dictionary, and the **MAKEDICT Program** used to sort it, assign screen numbers, and create a dictionary file. The **MAKEBITS Program** can then be used, with the fragment dictionary file, to generate fingerprints for large file of compounds. Various input connection table formats are available including MDL's SDfile,[3] and a direct interface to TDT files in the Daylight Chemical Information System[4] is scheduled. Output is to a proprietary BCI-format screen file, though direct writing of fingerprints to Daylight TDT files is also scheduled. A program called **PICKFRAG** is currently under development and will allow semiautomatic selection of fragments from a MAKEFRAG output file which are to be included in a dictionary, using user-specified criteria relating to minimum and maximum frequencies, and fragment co-occurrence.

CLUSTERING SOFTWARE

Figure 1 shows a basic classification of clustering methods with boxes round the methods for which implementations are available from BCI. The **Jarvis−Patrick Method** is widely used for chemical structure applications and is a nonhierarchical method based on common sets of nearest neighbors between compounds. The standard method[5] uses fixed-length lists of the nearest neighbors for each compound and is deliberately designed to be space distorting, clustering disparate objects together in sparsely-populated areas of space, whilst subdividing large clusters in densely-populated regions. This can cause problems when clustering large chemical structure databases, and BCI's implementation includes modifications in which a user-specified similarity threshold can be applied to produce variable-length nearest-neighbor lists.[6]

Two agglomerative hierarchical clustering methods are available: **Ward's Method** and the **Group-Average Method**. The standard (stored matrix) algorithm for these methods requires $O(N^3)$ time and $O(N^2)$ space, which makes them impractical for large datasets. However, BCI's implementation uses Murtagh's Reciprocal Nearest-Neighbor (RNN) algorithm,[7] which reduces the complexity to $O(N^2)$ time and $O(N)$ space. The programs can thus be used to cluster

---

[§] Product Review presented at the Fourth International Conference on Chemical Structures, Noordwijkerhout, The Netherlands, June 2−6, 1996.

[†] E-mail: barnard@bci1.demon.co.uk.

[‡] E-mail: downs@bci2.demon.co.uk.

[⊗] Abstract published in *Advance ACS Abstracts,* December 15, 1996.

S0095-2338(96)00090-X CCC: $14.00

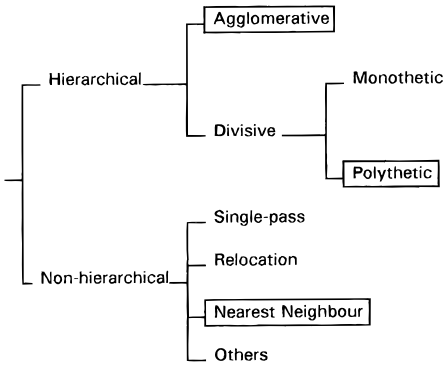

Figure 1. A classification of clustering methods.

datasets of up to 200 000 compounds (which require a few CPU days on powerful UNIX workstations).

The **Minimum Diameter Method** is a polythetic divisive hierarchical method and operates by iteratively dividing the largest-diameter cluster at each stage (starting with the entire dataset) into two subclusters, with the subdivision performed such that diameter of larger subcluster is a minimum. The method tends to give well-balanced hierarchy and can be useful if only a few clusters are needed. The algorithm of Guènoche *et al*.[8] has complexity O(N^2 log N) time and O(N^2) space, which makes the method slower than those implemented in BCI's other programs, and in the present version it is limited to 16 000 compounds, though an increase to 64 000 is planned.

Input to clustering programs is from a BCI format fingerprint file (a simple ASCII file listing screen (bit) numbers for fragments present in each structure) as generated by the MAKEBITS program or from Daylight[4] TDT files containing fingerprint data. Various options are available for output of clusters, including a list of the members of each cluster in the final partition and a list of the members of each cluster at all levels of the hierarchy. Cluster numbers for the final partition can also be added directly to Daylight TDT files as appropriately-formatted CL⟨⟩ data items.

A recent study[9] has suggested that hierarchical clustering methods (especially Ward's) can give better separation of active from inactive compounds in a dataset with known activity, when compared with nonhierarchical methods, and that the modified version of the Jarvis−Patrick can perform better than the standard version on diverse datasets.

REFERENCES AND NOTES

(1) Carhart, R. E.; Smith, D. H.; Venkataragharvan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.
(2) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F.; Computer storage and retrieval of generic chemical structures in patents. 9. An algorithm to find the Extended Set of Smallest Rings (ESSR) in structurally-explicit generics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 207−214.
(3) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. 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.
(4) Daylight Chemical Information Systems, Inc., 27401 Los Altos, Suite #370, Mission Viejo, CA 92691, USA. http://www.daylight.com.
(5) Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shared nearest neighbours. *IEEE Trans. Comput*. **1973**, *C-22*, 1025−1034.
(6) 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.
(7) Murtagh, F. Multidimensional clustering algorithms. *COMPSTAT Lectures*, Physica-Verlag: Vienna, 1985; Vol. 4.
(8) Guènoche, A.; Hansen, P.; Jaumard, B. Efficient algorithms for divisive hierarchical clustering with the diameter criterion, *J. Classification* **1991**, *8*, 5−30.
(9) 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.

CI960090K