CLASSIFY

The InfoChem

Reaction Classification Program

V.2.5

InfoChem GmbH

Version 06/02

InfoChem GmbH
Dr. Troll-Str. 14
82194 Gröbenzell
Tel: + 89 58 30 02
Fax:+ 89 58 03 839

InfoChem GmbH
Landsberger Str. 408
81241 München
Tel: + 89 58 30 02
Fax:+ 89 58 03 839

1. Introduction

InfoChem's Reaction Classification Program CLASSIFY categorizes reactions according to the type of chemical transformation they represent. The resulting reaction "ClassCodes" may be used to cluster large reaction databases and to select subsets of reactions or to define groups of analogous reactions in large sets. In particular, the ClassCode may be used to link different reaction databases.

2. Functionality

CLASSIFY assigns to each processed reaction a set of numeric values (ClassCodes) characterizing the chemical transformation that occurs in the reaction.
CLASSIFY first determines atom mappings and reaction centers (sites) by its own built-in algorithm and uses this information to define the reaction transformation taking place.
Information on the chemical transformation is generated in three different levels of precision. CLASSIFY produces a pair of ClassCodes for each precision level.

2.1. Level 1: Broad

For the first level CLASSIFY uses information on the reaction center (site) atoms and bonds only, and encodes them into the first ClassCode.
A classification of reactions based on this level of ClassCodes results in a Broad selection, i.e. all reactions having only equal reaction center atoms are grouped together.

2.3. Level 2: Medium

The second level includes information on the reaction center (site) atoms and bonds and, in addition, on neighbouring atoms of each reaction center. These combined data are encoded into the second ClassCode.
These ClassCodes yield a Medium selection, by selecting all reactions with equal reaction center atoms and equal immediate neighbour atoms.

2.3. Level 3: Narrow

On this level CLASSIFY encodes information on the reaction center (site) atoms and bonds, their immediate neighbour atoms, and on the neighbours' neighbouring atoms. This procedure results in a third ClassCode.
Using these ClassCodes, the selection of reactions becomes Narrow, since all selected reactions are required to have the reaction center atoms and two spheres of neighbouring atoms in common.

3. Generalizations

- At the Medium and Narrow levels, information of neighbouring atoms is included, forming "atom clusters". Only the atom types of the second period of the periodic table of the elements (Be to F) are treated differently from the other members of the corresponding element group. I.e. all higher homologues within one element group are considered as one of eight "group atoms".
- Transition group elements are totally standardized, i.e. all transition group elements are treated as one "transition metal atom".
- All atoms of the first sphere around reaction center atoms are used to form the "atom clusters" of the Medium level.
- Building the "atom clusters" for the Narrow level by adding the neighbour atoms of the second sphere to the clusters, regular tetravalent carbon atoms (sp3-atoms) are ignored. All hetero atoms and carbon atoms with multiple bonds or in aromatic rings are considered to be different from regular atoms.
- The resulting "atom clusters" of the reactant and product site are used to generate the ClassCodes for the reaction type.
- Multiple occurrences of clusters are treated as only one occurrence. That means reactions showing several identical chemical transformations will be classified as equivalent to the reaction showing the chemical transformation only once.
- If a reaction contains two or more product molecules, one set of output lines is generated for each product molecule. The classification ClassCodes refer to the transformation leading to this particular product.

4. Technical Provision of CLASSIFY

CLASSIFY reads reactions from an MDL RD file (ISIS / Host) and writes reaction numbers and the determined ClassCodes to an output RD file.

The program CLASSIFY is available for the use as **batch** application on MS-Windows desktops, Sun Solaris, and Open-VMS Alpha AXP Hardware.

In addition, an **interactive** version is provided as a separate Windows-DLL for PC Windows systems and is integrated as an Add-In function into ISIS / Draw (MDL Information Systems). CLASSIFY is started with an additional button added to ISIS / Draw allowing the fast, interactive determination of the ClassCodes of individual reactions drawn in ISIS / Draw at the speed of a mouse click.

- Batch Version

Disk Space:	500 KByte
Input:	Filename / RDfile
Output:	Filename / RDfile including RXN-ClassCodes
Hardware:	PC-Windows / Intel; Sun; Alpha AXP
Software:	Windows NT / 2000; Solaris; Open-VMS

- Interactive Windows Version (DLL)

Hardware:	PC-Windows / Intel;
Software:	WindowsNT / 2000
Required:	ISIS / Draw from MDL

5. Commercial Applications of CLASSIFY

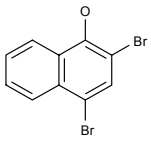
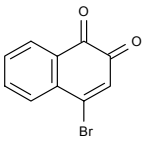
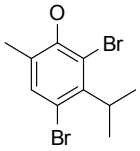
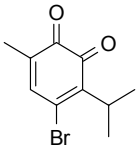
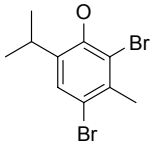
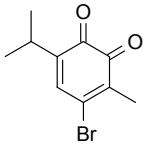
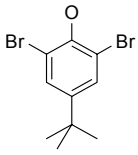
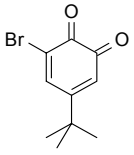
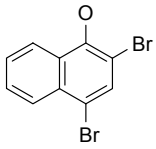
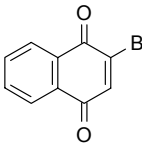
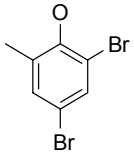
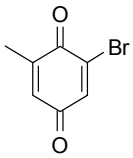
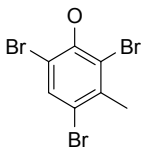
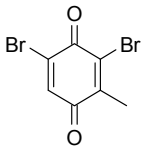
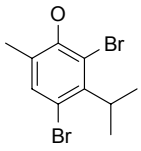
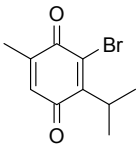
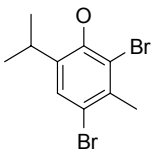
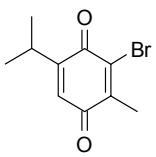
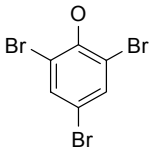
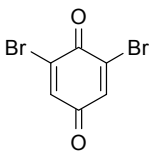
- CLASSIFY is used by MDL in the RXL-Browser to cluster the results in the hit lists of reaction queries.
- All reaction databases distributed by MDL are classified and include the ClassCodes.
- All reactions included in the Beilstein database will be classified to contain the ClassCodes.
- The ClassCodes are the only way to link different reaction databases. The product "*integrated* Major Reference Works" distributed by MDL is based on this concept and architecture.
- Several chemical and pharmaceutical companies are using CLASSIFY to process their internal proprietary reaction databases (e.g. BASF, Bayer, Merck, Boehringer, Novartis, a.o.m) to determine the ClassCodes.

6. Example

As an example a set of 10 reactions from the literature has been processed using CLASSIFY in order to detect groups of closely related, analogous reactions, and to rearrange the list according to similarity in chemical transformations.

The classification results on three levels (Broad, Medium, Narrow) are given for each reaction in the following diagrams. For simplicity reasons the numeric ClassCodes have been replaced by one-character symbols (A, B, C, ...) to increase the readability of the diagrams.

Obviously, the same symbols represent the same numeric values calculated. The example has been taken from the literature: 'A New Method for the Conversion of Halophenols and Halonaphthols to Quinones', Perumal P. T. and Bhatt M. V., Synthesis 1979, 205-206.

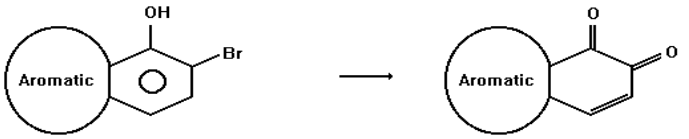
			Broad	Medium	Narrow
1.)			A	B	E
2.)			A	B	F
3.)			A	B	F
4.)			A	B	F
5.)			A	C	G
6.)			A	D	H
7.)			A	D	H
8.)			A	D	H
9.)			A	D	H
10.)			A	D	H

The symbols A ... H represent complex numeric classification ClassCodes. For the example given above the following results are obtained:

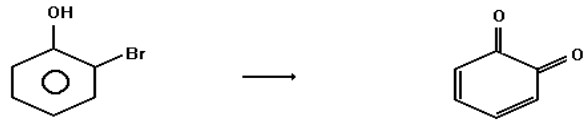
Broad:	1 Cluster (A)
Medium:	3 Cluster (B, C, D)
Narrow	4 Cluster (E, F, G, H)

Evaluating the classification ClassCodes for the level 'Narrow' four different reaction types are identified:

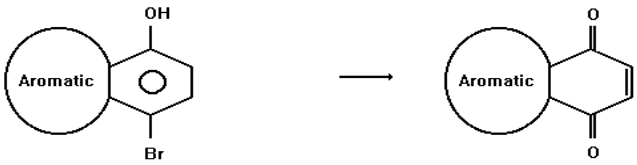
Type ABE



Type ABF



Type ACG



Type ADH