NAME

RDKitClusterMolecules.py - Cluster molecules using 2D fingerprints

SYNOPSIS

RDKitClusterMolecules.py [--butinaSimilarityCutoff <number>] [--butinaReordering <yes or no>] [--clusteringMethod <Butina, Centroid, CLink...>] [--fingerprints <MACCS166Keys, Morgan, PathLength...>] [--fingerprintsType <IntVect, BitVect, or Auto>] [--infileParams <Name,Value,...>] [--numClusters <number>] [--outfileMode <SingleFile or MultipleFiles>] [--outfileParams <Name,Value,...>] [--overwrite] [--paramsFingerprints <Name,Value,...>] [--similarityMetric <Dice, Tanimoto...>] [-w <dir>] -i <infile> -o <outfile>

RDKitClusterMolecules.py -h | --help | -e | --examples

DESCRIPTION

Cluster molecules based on a variety of 2D fingerprints using Butina [Ref 136] or any other available hierarchical clustering methodology and write them to output file(s).

The supported input file formats are: Mol (.mol), SD (.sdf, .sd), SMILES (.smi, .txt, .csv, .tsv)

The supported output file formats are: SD (.sdf, .sd), SMILES (.smi), CSV/TSV (.csv, .tsv, .txt)

OPTIONS

-b, --butinaSimilarityCutoff <number> [default: 0.55]

Similarity cutoff to use during Butina clustering. The molecule pairs with similarity value greater than specified value or distance less than '1 - specified value' are considered neighbors. This value is only used during 'Butina' value of '-c, --clusteringMethod' option and determines the number of clusters during the clustering of molecules. It is ignored for all other clustering methods.

--butinaReordering <yes or no> [default: no]

Update number of neighbors for unassigned molecules after creating a new cluster in order to insure that the molecule with the largest number of unassigned neighbors is selected as the next cluster center.

-c, --clusteringMethod <Butina, Centroid, CLink...> [default: Butina]

Clustering method to use for clustering molecules. Supported values: Butina, Centroid, CLink, Gower, McQuitty, SLink, UPGMA, Ward. Butina is an unsupervised database clustering method to automatically cluster small and large data sets. All other clustering methods correspond to hierarchical clustering and require a priori specification of number of clusters to be generated.

-f, --fingerprints <MACCS166Keys, Morgan, PathLength...> [default: Morgan]

Fingerprints to use for calculating similarity/distance between molecules. Supported values: AtomPairs, MACCS166Keys, Morgan, MorganFeatures, PathLength, TopologicalTorsions. The PathLength fingerprints are Daylight like fingerprints. The Morgan and MorganFeature fingerprints are circular fingerprints, corresponding Scitegic's Extended Connectivity Fingerprints (ECFP) and Features Connectivity Fingerprints (FCFP). The values of default parameters for generating fingerprints can be modified using '-p, --paramsFingerprints' option.

--fingerprintsType <IntVect, BitVect, or auto> [default: auto]

Fingerprints type to generate for calculating similarity. Supported values: IntVect, BitVect, Auto.

The following default fingerprints type are automatically generated for available fingerprints, based on the value of similarty metric:

AtomPairs Tanimoto|Dice: IntVect All Others: BitVect MACCS166Keys All: BitVect Morgan Tanimoto|Dice: IntVect All Others: BitVect MorganFeatures Tanimoto|Dice: IntVect All Others: BitVect PathLength All: BitVect TopologicalTorsions Tanimoto|Dice: IntVect All Others: BitVect

The Dice and Tanimoto similarity functions available in RDKit are able to handle fingerprints corresponding to both IntVect and BitVect. All other similarity functions, however, expect BitVect fingerprints to calculate pairwise similarity. Consequently, BitVect fingerprints, instead of default IntVect fingerprints, are generated for AtomPairs, Morgan, MorganFeatures, and TopologicalTorsions during the calculation of similarity using all other similarity functions.

The IntVect fingerprints type is not available for MACCS166Keys and Pathlength fingerprints. In addition, IntVect fingerprints type is only valid for Tanimoto or Dice value of ' -s, --similarityMetric' option. The BitVect fingerprints type is valid for all values of '' -s, --similarityMetric' option.

-e, --examples

Print examples.

-h, --help

Print this help message.

-i, --infile <infile>

Input file name.

--infileParams <Name,Value,...> [default: auto]

A comma delimited list of parameter name and value pairs for reading molecules from files. The supported parameter names for different file formats, along with their default values, are shown below:

```
SD, MOL: removeHydrogens,yes,sanitize,yes,strictParsing,yes
SMILES: smilesColumn,1,smilesNameColumn,2,smilesDelimiter,space,
    smilesTitleLine,auto,sanitize,yes
```

Possible values for smilesDelimiter: space, comma or tab.

-n, --numClusters <number> [default: 10]

Number of clusters to generate during hierarchical clustering. This option is ignored for 'Butina' value of '-c, --clusteringMethod' option.

-o, --outfile <outfile>

Output file name.

--outfileMode <SingleFile or MultipleFiles> [default: SingleFile]

Write out a single file containing molecule clusters or generate an individual file for each cluster. Possible values: SingleFile or MultipleFiles. The molecules are grouped for each cluster before they are written to output file(s) along with appropriate cluster numbers. The cluster number is also appended to output file names during generation of multiple output files.

--outfileParams <Name,Value,...> [default: auto]

A comma delimited list of parameter name and value pairs for writing molecules to files. The supported parameter names for different file formats, along with their default values, are shown below:

```
SD: compute2DCoords,auto,kekulize,no
SMILES: kekulize,no,smilesDelimiter,space, smilesIsomeric,yes,
    smilesTitleLine,yes
```

Default value for compute2DCoords: yes for SMILES input file; no for all other file types. The kekulize and smilesIsomeric parameters are also used during generation of SMILES strings for CSV/TSV files.

--overwrite

Overwrite existing files.

-p, --paramsFingerprints <Name,Value,...> [default: auto]

Parameter values to use for generating fingerprints. The default values are dependent on the value of '-f, --fingerprints' option. In general, it is a comma delimited list of parameter name and value pairs for the name of fingerprints specified using '-f, --fingerprints' option. The supported parameter names along with their default values for valid fingerprints names are shown below:

```
AtomPairs: minLength,1 ,maxLength,30, useChirality,No,
    fpSize, 2048, bitsPerHash,4
Morgan: radius,2, useChirality,No, fpSize, 2048
MorganFeatures:   radius,2, useChirality,No, fpSize, 2048
PathLength: minPath,1, maxPath,7, fpSize, 2048, bitsPerHash,2
TopologicalTorsions: useChirality,No, fpSize, 2048, bitsPerHash,4
```

The fpSize and bitsPerHash are only used for BitVect fingerprints type specified using '--fingerprintsType' option.

-s, --similarityMetric <Dice, Tanimoto...> [default: Tanimoto]

Similarity metric to use for calculating similarity/distance between molecules. Possible values: BraunBlanquet, Cosine, Dice, Kulczynski, RogotGoldberg, Russel, Sokal, Tanimoto.

-w, --workingdir <dir>

Location of working directory which defaults to the current directory.

EXAMPLES

To cluster molecules using Butina methodology at a similarity cutoff of 0.55 with automatic determination of number of clusters, Tanimoto similarity metric corresponding to Morgan fingerprints with radius of 2, and write out a single SMILES file containing clustered molecules along with cluster number for each molecule, type:

```
    % RDKitClusterMolecules.py  -i Sample.smi -o SampleOut.smi
```

To cluster molecules using Butina methodology at a similarity cutoff of 0.55 with automatic determination of number of clusters, Tanimoto similarity metric corresponding to Morgan fingerprints with radius of 2 and type BitVect, fingerprint BitVect size of 4096, and write out a single SMILES file containing clustered molecules along with cluster number for each molecule, type:

```
    % RDKitClusterMolecules.py  -f Morgan  --fingerprintsType  BitVect
      -p "fpSize,4096" -s Tanimoto -i Sample.smi -o SampleOut.smi
```

To cluster molecules using Butina methodology at similarity cutoff of 0.45 with automatic determination of number of clusters, Dice similarity metric corresponding to Morgan fingerprints with radius of 2, and write out multiple SD files containing clustered molecules for each cluster, type:

```
    % RDKitClusterMolecules.py  -b 0.45 -s Dice --outfileMode MultipleFiles
      -i Sample.smi -o SampleOut.sdf
```

To cluster molecules using Ward hierarchical methodology to generate 15 clusters, Dice similarity metric corresponding to Pathlength fingerprints with path length between 1 and 7, and write out a single TSV file for clustered molecules along with cluster numner for each molecule, type:

```
    % RDKitClusterMolecules.py  -c Ward -f PathLength -n 15
      -p ’minPath,1, maxPath,7’ -i Sample.sdf -o SampleOut.tsv
```

To cluster molecules using Centroid hierarchical methodology to generate 5 clusters, Dice similarity metric corresponding to MACCS166Keys fingerprints for molecules in a SMILES CSV file, SMILES strings in column 1, name in column 2, and write out a single SD file for clustered molecules along with cluster numner for each molecule, type:

```
    % RDKitClusterMolecules.py  -c Centroid -f MACCS166Keys --infileParams
      "smilesDelimiter,comma,smilesTitleLine,yes,smilesColumn,1,
      smilesNameColumn,2" --outfileParams "compute2DCoords,yes"
      -i SampleSMILES.csv -o SampleOut.sdf
```

AUTHOR

Manish Sud(msud@san.rr.com)

SEE ALSO

RDKitConvertFileFormat.py, RDKitPickDiverseMolecules.py, RDKitSearchFunctionalGroups.py, RDKitSearchSMARTS.py

COPYRIGHT

The functionality available in this script is implemented using RDKit, an open source toolkit for cheminformatics developed by Greg Landrum.

This file is part of MayaChemTools.