NAME

>   RDKitPickDiverseMolecules.py - Pick a diverse subset of molecules

SYNOPSIS

>   RDKitPickDiverseMolecules.py [--clusteringMethod <Centroid, CLink...>] [--fingerprints <MACCS166Keys,
>   Morgan, PathLength...>] [--infileParams <Name,Value,...>] [--mode <MaxMin or HierarchicalClustering>] [
>   --numMols <number>] [--outfileParams <Name,Value,...>] [--overwrite] [--paramsFingerprints
>   <Name,Value,...>] [--similarityMetric <Dice, Tanimoto...>] [-w <dir>] -i <infile> -o <outfile>
>
>   RDKitPickDiverseMolecules.py -h | --help | -e | --examples

DESCRIPTION

>   Pick a subset of diverse molecules based on a variety of 2D fingerprints using MaxMin [Ref 135] or an
>   available hierarchical clustering methodology and write them to a file.
>
>   The default fingerprints types for various fingerprints are shown below:

```
    AtomPairs              IntSparseIntVect
    MACCS166Keys           ExplicitBitVect
    Morgan                 UIntSparseIntVect
    MorganFeatures         UIntSparseIntVect
    PathLength             ExplicitBitVect
    TopologicalTorsions    LongSparseIntVect
```

>   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, ExplicitBitVect fingerprints are generated for AtomPairs, Morgan,
>   MorganFeatures, and TopologicalTorsions for similarity calculations instead of default IntVect fingerprints.
>
>   The supported input file formats are: SD (.sdf, .sd), SMILES (.smi, .csv, .tsv, .txt)
>
>   The supported output file formats are: SD (.sdf, .sd), SMILES (.smi)

OPTIONS

>   -c, --clusteringMethod <Centroid, CLink...> [default: Centroid]
>
>>   Clustering method to use for picking a subset of diverse molecules during hierarchical clustering.
>>   Supported values: Centroid, CLink, Gower, McQuitty, SLink, UPGMA, Ward. This option is ignored for
>>   'MaxMin' value of '-m, --mode' option. The Clink and SLink corresponding to CompleteLink and
>>   SingleLink cluster method.
>
>   -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.
>
>   -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.

-m, --mode <MaxMin or HierarchicalClustering> [default: MaxMin]

    Pick a diverse subset of molecules using MaxMin or hierarchical clustering methodology.

-n, --numMols <number> [default: 25]

    Number of diverse molecules to pick.

-o, --outfile <outfile>

    Output file name.

--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.

--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 the 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
Morgan:   radius,2, useChirality,No
MorganFeatures:   radius,2, useChirality,No
PathLength: minPath,1, maxPath,7, fpSize, 2048, bitsPerHash,2
TopologicalTorsions: useChirality,No
```

-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 pick 25 diverse molecules using MaxMin methodology, Tanimoto similarity metric corresponding to Morgan fingerprints with radius of 2, and write out a SMILES file, type:

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

    To pick 50 diverse molecules using MaxMin methodology, Dice similarity metric corresponding to PathLength fingerprints with max path length of 6, and write out a SD file, type:

```
% RDKitPickDiverseMolecules.py  -m MaxMin -f PathLength -s Dice -n 50
  -p 'maxPath,6' -i Sample.sdf -o SampleOut.sdf
```

    To pick 25 diverse molecules using Centroid hierarchical clustering methodology, Tanimoto similarity metric corresponding to Morgan fingerprints with radius of 2, and write out a SMILES file, type:

```
% RDKitPickDiverseMolecules.py  -m HierarchicalClustering -i Sample.smi
  -o SampleOut.smi
```

    To pick 50 diverse molecules using Ward hierarchical methodology methodology, Dice similarity metric corresponding to MorganFeatures fingerprints with radius of 2 along with deploying chirality, and write out a SD file, type:

```
% RDKitPickDiverseMolecules.py  -m HierarchicalClustering -c Ward -n 50
  -f MorganFeatures -p 'radius,2,useChirality,No' -i Sample.sdf -o
  SampleOut.sdf
```

To pick 25 diverse molecules using MaxMin methodology, Tanimoto similarity metric corresponding to Morgan fingerprints with radius of 2 from a CSV SMIKES file , SMILES strings in column 1, name in olumn 2, and write out a SD file, type:

```
% RDKitPickDiverseMolecules.py  --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

RDKitClusterMolecules.py, RDKitConvertFileFormat.py, RDKitSearchFunctionalGroups.py, RDKitSearchSMARTS.py

COPYRIGHT

Copyright (C) 2019 Manish Sud. All rights reserved.

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.

MayaChemTools is free software; you can redistribute it and/or modify it under the terms of the GNU Lesser General Public License as published by the Free Software Foundation; either version 3 of the License, or (at your option) any later version.