

NAME

SimilarityMatrixSDFiles.pl - Calculate similarity matrices using fingerprints data in SDFFile(s)

SYNOPSIS

SimilarityMatrixSDFiles.pl SDFFile(s)...

SimilarityMatrixSDFiles.pl [-a, --alpha *number*] [b, --beta *number*] [--CompoundID *DataFieldName* or *LabelPrefixString*] [--CompoundIDMode *DataField* | *MolName* | *LabelPrefix* | *MolNameOrLabelPrefix*] [-d, --detail *InfoLevel*] [-f, --fast] [--FingerprintsField *FieldLabel*] [--FingerprintsFormatMode *Internal* | *Specify*] [--FingerprintsString *Hexadecimal* | *Binary* | *RawBinary*] [-h, --help] [-m, --mode *All* | "*Tanimoto*,*Tversky*,..."] [--OutDelim *comma* | *tab* | *semicolon*] [-o, --overwrite] [-p, --precision *number*] [-q, --quote *Yes* | *No*] [-r, --root *RootName*] [-w, --WorkingDir *dirname*] SDFFile(s)...

DESCRIPTION

Calculate similarity matrices using fingerprints data field in *SDFFile(s)* and generate CSV/TSV text files containing values for specified similarity coefficients.

Multiple SDF file names are separated by spaces. The valid file extensions are *.sdf* and *.sd*. All other file names are ignored. All the SD files in a current directory can be specified either by **.sdf* or the current directory name.

OPTIONS

-a, --alpha *number*

Value of alpha parameter for calculating *Tversky* similarity coefficient specified for -m, --mode option. It corresponds to weights assigned for bits set to "1" in a pair of fingerprint bit vectors during the calculation of similarity coefficient. Possible values: *0 to 1*. Default value: *0.5*.

b, --beta *number*

Value of alpha parameter for calculating *WeightedTanimoto* and *WeightedTversky* similarity coefficients specified for -m, --mode option. It is used to weight the contributions of bits set to "0" during the calculation of similarity coefficients. Possible values: *0 to 1*. Default value of *1* makes *WeightedTanimoto* and *WeightedTversky* equivalent to *Tanimoto* and *Tversky*.

--CompoundID *DataFieldName* or *LabelPrefixString*

This value is --CompoundIDMode specific and indicates how compound ID is generated.

For *DataField* value of --CompoundIDMode option, this option corresponds to datafield label name whose value is used as compound ID; otherwise, it's a prefix string used for generating compound IDs like *LabelPrefixString*<Number>. Default value, *Cmpd*, generates compound IDs which look like *Cmpd*<Number>.

Examples for *DataField* value of --CompoundIDMode:

```
MolID
ExtReg
```

Examples for *LabelPrefix* or *MolNameOrLabelPrefix* value of --CompoundIDMode:

```
Compound
```

The values specified above generates compound IDs which correspond to *Compound*<Number> instead of default value of *Cmpd*<Number>.

--CompoundIDMode *DataField* | *MolName* | *LabelPrefix* | *MolNameOrLabelPrefix*

Specify how to generate compound IDs for similarity matrix CSV/TSV text file(s): use a *SDFFile(s)* datafield value; use molname line from *SDFFile(s)*; generate a sequential ID with specific prefix; use combination of both *MolName* and *LabelPrefix* with usage of *LabelPrefix* values for empty molname lines.

Possible values: *DataField* | *MolName* | *LabelPrefix* | *MolNameOrLabelPrefix*. Default: *LabelPrefix*.

For *MolNameAndLabelPrefix* value of --CompoundIDMode, molname line in *SDFFile(s)* takes precedence over sequential compound IDs generated using *LabelPrefix* and only empty molname values are replaced with sequential compound IDs.

-d, --detail *InfoLevel*

Level of information to print about lines being ignored. Default: *1*. Possible values: *1, 2* or *3*

-f, --fast

In this mode, fingerprints field specified using --FingerprintsField is assumed to contain valid fingerprints data and no checking is performed before calculating similarity matrices. By default,

fingerprints data is validated before computing pairwise similarity coefficients.

--FingerprintsField *FieldLabel*

Fingerprints field label to use during calculation similarity matrices for *SDFFile(s)*. Default value: *first data field label containing the word Fingerprints in its label*

--FingerprintsFormatMode *Internal | Specify*

Specify format of fingerprints data in *SDFFile(s)*: use default format which MayaChemTools fingerprint generation scripts use to write out fingerprints data or explicitly specify format of fingerprints. Possible values: *Internal | Specify*. Default value: *Internal*.

Internal fingerprints string format consists of four parts delimited by semicolon:

<Type:StringType:Size:String>. For example:

```
"PathLength:Binary:512:010011..."
"MDLKeys166FP:Binary:166:010011..."
"MDLKeys166Count:Vector:166:0 1 2..."
```

For *Specify* value of --FingerprintsFormatMode option, --FingerprintsString is used to interpret fingerprints string.

--FingerprintsString *Hexadecimal | Binary | RawBinary*

Format of fingerprints string during *Specify* value of --FingerprintsFormatMode option. Possible values: *Hexadecimal, Binary, or RawBinary*. Default value: *none*; its value must be explicitly specified.

-h, --help

Print this help message

-m, --mode *All | "Tanimoto,[Tversky,...]"*

Specify what similarity coefficients to use for calculating similarity matrices for fingerprints data values in *TextFile(s)*: calculate similarity matrices for all supported similarity coefficients or specify a comma delimited list of similarity coefficients. Possible values: *All | "Tanimoto,[Tversky,...]"*. Default: *Tanimoto*

All uses complete list of supported similarity coefficients: *BaroniUrbani, Buser, Cosine, Dice, Dennis, Euclid, Forbes, Fossum, Hamann, Jacard, Kulczynski1, Kulczynski2, Manhattan, Matching, McConnaughey, Ochiai, Pearson, RogersTanimoto, RussellRao, Simpson, SkoalSneath1, SkoalSneath2, SkoalSneath3, Tanimoto, Tversky, Yule, WeightedTanimoto, WeightedTversky*. These similarity coefficients are described below.

For two fingerprints bit vectors A and B of same size, let:

```
Na = Number of bits set to "1" in A
Nb = Number of bits set to "1" in B
Nc = Number of bits set to "1" in both A and B
Nd = Number of bits set to "0" in both A and B

Nt = Number of bits set to "1" or "0" in A or B (Size of A or B)
Nt = Na + Nb - Nc + Nd

Na - Nc = Number of bits set to "1" in A but not in B
Nb - Nc = Number of bits set to "1" in B but not in A
```

Then, various similarity coefficients [Ref. 40 - 42] for a pair of bit vectors A and B are defined as follows:

BaroniUrbani: $(\sqrt{Nc * Nd} + Nc) / (\sqrt{Nc * Nd} + Nc + (Na - Nc) + (Nb - Nc))$ (same as Buser)

Buser: $(\sqrt{Nc * Nd} + Nc) / (\sqrt{Nc * Nd} + Nc + (Na - Nc) + (Nb - Nc))$ (same as BaroniUrbani)

Cosine: $Nc / \sqrt{Na * Nb}$ (same as Ochiai)

Dice: $(2 * Nc) / (Na + Nb)$

Dennis: $(Nc * Nd - ((Na - Nc) * (Nb - Nc))) / \sqrt{Nt * Na * Nb}$

Euclid: $\sqrt{(Nc + Nd) / Nt}$

Forbes: $(Nt * Nc) / (Na * Nb)$

Fossum: $(Nt * ((Nc - 1/2) ** 2)) / (Na * Nb)$

Hamann: $((Nc + Nd) - (Na - Nc) - (Nb - Nc)) / Nt$

Jaccard: $Nc / ((Na - Nc) + (Nb - Nc) + Nc) = Nc / (Na + Nb - Nc)$ (same as Tanimoto)

Kulczynski1: $Nc / ((Na - Nc) + (Nb - Nc)) = Nc / (Na + Nb - 2Nc)$

Kulczynski2: $((Nc / 2) * (2 * Nc + (Na - Nc) + (Nb - Nc))) / ((Nc + (Na - Nc)) * (Nc + (Nb - Nc))) = 0.5 * (Nc / Na + Nc / Nb)$

Manhattan: $((Na - Nc) + (Nb - Nc)) / Nt = (Na + Nb - 2Nc) / Nt$

Matching: $(Nc + Nd) / Nt$

McConnaughey: $(Nc ** 2 - (Na - Nc) * (Nb - Nc)) / (Na * Nb)$

Ochiai: $Nc / \text{SQRT}(Na * Nb)$ (same as Cosine)

Pearson: $((Nc * Nd) - ((Na - Nc) * (Nb - Nc))) / \text{SQRT}(Na * Nb * ((Na - Nc + Nd) * (Nb - Nc + Nd)))$

RogersTanimoto: $(Nc + Nd) / ((Na - Nc) + (Nb - Nc) + Nt) = (Nc + Nd) / (Na + Nb - 2Nc + Nt)$

RussellRao: Nc / Nt

Simpson: $Nc / \text{MIN}(Na, Nb)$

SkoalSneath1: $Nc / (Nc + 2 * (Na - Nc) + 2 * (Nb - Nc)) = Nc / (2 * Na + 2 * Nb - 3 * Nc)$

SkoalSneath2: $(2 * Nc + 2 * Nd) / (Nc + Nd + Nt)$

SkoalSneath3: $(Nc + Nd) / ((Na - Nc) + (Nb - Nc)) = (Nc + Nd) / (Na + Nb - 2 * Nc)$

Tanimoto: $Nc / ((Na - Nc) + (Nb - Nc) + Nc) = Nc / (Na + Nb - Nc)$ (same as Jaccard)

Tversky: $Nc / (\alpha * (Na - Nc) + (1 - \alpha) * (Nb - Nc) + Nc) = Nc / (\alpha * (Na - Nb) + Nb)$

Yule: $((Nc * Nd) - ((Na - Nc) * (Nb - Nc))) / ((Nc * Nd) + ((Na - Nc) * (Nb - Nc)))$

Values of Tanimoto/Jaccard and Tversky coefficients are dependent on only those bit which are set to "1" in both A and B. In order to take into account all bit positions, modified versions of Tanimoto [Ref. 42] and Tversky [Ref. 43] have been developed.

Let:

Na' = Number of bits set to "0" in A

Nb' = Number of bits set to "0" in B

Nc' = Number of bits set to "0" in both A and B

Tanimoto': $Nc' / ((Na' - Nc') + (Nb' - Nc') + Nc') = Nc' / (Na' + Nb' - Nc')$

Tversky': $Nc' / (\alpha * (Na' - Nc') + (1 - \alpha) * (Nb' - Nc') + Nc') = Nc' / (\alpha * (Na' - Nb') + Nb')$

Then:

WeightedTanimoto = $\beta * \text{Tanimoto} + (1 - \beta) * \text{Tanimoto}'$

WeightedTversky = $\beta * \text{Tversky} + (1 - \beta) * \text{Tversky}'$

--OutDelim *comma | tab | semicolon*

Delimiter for output CSV/TSV text file(s). Possible values: *comma, tab, or semicolon* Default value: *comma*

-o, --overwrite

Overwrite existing files

-p, --precision *number*

Precision of calculated values in the output file. Default: up to 2 decimal places. Valid values: positive integers

-q, --quote *Yes | No*

Put quote around column values in output CSV/TSV text file(s). Possible values: *Yes or No*. Default value: *Yes*

-r, --root *RootName*

New file name is generated using the root: $\langle \text{Root} \rangle \langle \text{Mode} \rangle . \langle \text{Ext} \rangle$. Default for new file names: $\langle \text{TextFileName} \rangle \langle \text{Mode} \rangle . \langle \text{Ext} \rangle$. The csv, and tsv $\langle \text{Ext} \rangle$ values are used for comma/semicolon, and tab delimited text files respectively. This option is ignored for multiple input files.

-w, --WorkingDir *DirName*

Location of working directory. Default: current directory

EXAMPLES

To generate a similarity matrix corresponding to Tanimoto similarity coefficient for fingerprints data in any internal fingerprint format present in a data field with Fingerprint substring in its label and create a SampleFPTanimoto.csv file containing sequentially generated compound IDs with Cmpd prefix, type:

```
% SimilarityMatrixSDFFiles.pl -o SampleFP.sdf
```

To generate similarity matrices corresponding to all supported similarity coefficient for fingerprints data in any internal fingerprint format present in a data field with Fingerprint substring in its label and create a SampleFPTanimoto.csv file containing sequentially generated compound IDs with Cmpd prefix, type:

```
% SimilarityMatrixSDFFiles.pl -o -m All SampleFP.sdf
```

To generate a similarity matrix corresponding to Buser, Dice and Tanimoto similarity coefficient for fingerprints data in any internal fingerprint format present in a data field with Fingerprint substring in its label and create a SampleFPTanimoto.csv file containing sequentially generated compound IDs with Cmpd prefix, type:

```
% SimilarityMatrixSDFFiles.pl -o -m "Buser,Dice,Tanimoto" SampleFP.sdf
```

To generate a similarity matrix corresponding to Tanimoto similarity coefficient for fingerprints data in any internal fingerprint format present in a data field name PathLengthFingerprints and create a SampleFPTanimoto.csv file containing compound IDs present in data field name Cmpd_ID with Cmpd prefix, type:

```
% SimilarityMatrixSDFFiles.pl -o --FingerprintsField PathLengthFingerprints  
--CompoundIDMode DataField --CompoundID Cmpd_ID SampleFP.sdf
```

To generate a similarity matrix corresponding to Tanimoto similarity coefficient for fingerprints data in any binary bit-string format present in a data field with Fingerprint substring in its label and create a SampleFPTanimoto.csv file containing sequentially generated compound IDs with Cmpd prefix, type:

```
% SimilarityMatrixSDFFiles.pl -o --FingerprintsFormatMode Specify  
--FingerprintsString Binary SampleFP.sdf
```

To generate a similarity matrix corresponding to Tanimoto similarity coefficient for fingerprints data in any internal fingerprint format present in a data field with Fingerprint substring in its label and create a SampleFPTanimoto.csv file containing compound IDs from molname line or sequentially generated compound IDs with Mol prefix, type generated compound IDs with Cmpd prefix, type:

```
% SimilarityMatrixSDFFiles.pl -o --CompoundIDMode MolnameOrLabelPrefix  
--CompoundID Mol SampleFP.sdf
```

To generate a similarity matrix corresponding to Tanimoto similarity coefficient for fingerprints data in any internal fingerprint format present in a data field with Fingerprint substring in its label and create a SampleFPTanimoto.tsv file without any quotes around values along with sequentially generated compound IDs with Cmpd prefix, type:

```
% SimilarityMatrixSDFFiles.pl -o --OutDelim Tab --quote No SampleFP.sdf
```

AUTHOR

Manish Sud <msud@san.rr.com>

SEE ALSO

InfoFingerprintsSDFFiles.pl, PathLengthFingerprints.pl, SimilarityMatrixTextFiles.pl

COPYRIGHT

Copyright (C) 2004-2008 Manish Sud. All rights reserved.

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.