

NAME  
SimilarityMatrixTextFiles.pl - Calculate similarity matrices using fingerprints data in TextFile(s)

## SYNOPSIS

SimilarityMatrixTextFiles.pl TextFile(s)...

SimilarityMatrixTextFiles.pl [-a, --alpha *number*] [b, --beta *number*] [-c, --ColMode *ColNum* | *ColLabel*] [--CompoundIDCol *col number* | *col name*] [-d, --detail *InfoLevel*] [-f, --fast] [--FingerprintsCol *col number* | *col name*] [--FingerprintsFormatMode *Internal* | *Specify*] [--FingerprintsString *Hexadecimal* | *Binary* | *RawBinary*] [-h, --help] [-m, --mode *All* | "*Tanimoto*,*[Tversky,...]*"] [--InDelim *comma* | *semicolon*] [--OutDelim *comma* | *tab* | *semicolon*] [-o, --overwrite] [-p, --precision *number*] [-q, --quote *Yes* | *No*] [-r, --root *RootName*] [-w, --WorkingDir *dirname*] TextFile(s)...

## DESCRIPTION

Calculate similarity matrices using fingerprints data column specified by a column number or label in *TextFile(s)* and generate CSV/TSV text files containing values for specified similarity coefficients.

The valid file extensions are *.csv* and *.tsv* for comma/semicolon and tab delimited text files respectively. All other file names are ignored. All the text files in a current directory can be specified by *\*.csv*, *\*.tsv*, or the current directory name. The *--indelim* option determines the format of *TextFile(s)*. Any file which doesn't correspond to the format indicated by *--indelim* option is ignored.

## 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*.

-c, --ColMode *ColNum* | *ColLabel*

Specify how columns are identified in *TextFile(s)*: using column number or column label. Possible values: *ColNum* or *ColLabel*. Default value: *ColNum*

--CompoundIDCol *col number* | *col name*

This value is -c, --ColMode mode specific. It specifies column to use to retrieve compound ID for similarity matrices in output *TextFile(s)*. Possible values: *col number* or *col label*. Default value: *first column containing the word compoundID in its column label or sequentially generated 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 column specified using --FingerprintsCol 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.

--FingerprintsCol *col number* | *col name*

This value is -c, --colmode specific. It specifies fingerprints column to use during calculation similarity matrices for *TextFile(s)*. Possible values: *col number* or *col label*. Default value: *first column containing the word Fingerprints in its column label*.

--FingerprintsFormatMode *Internal* | *Specify*

Specify format of fingerprints data in *TextFile(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*, *Jaccard*, *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:  $(\text{SQRT}(\text{Nc} * \text{Nd}) + \text{Nc}) / (\text{SQRT}(\text{Nc} * \text{Nd}) + \text{Nc} + (\text{Na} - \text{Nc}) + (\text{Nb} - \text{Nc}))$  ( same as Buser )

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

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

Dice:  $(2 * \text{Nc}) / (\text{Na} + \text{Nb})$

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

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

Forbes:  $(\text{Nt} * \text{Nc}) / (\text{Na} * \text{Nb})$

Fossum:  $(\text{Nt} * ((\text{Nc} - 1/2) ** 2)) / (\text{Na} * \text{Nb})$

Hamann:  $((\text{Nc} + \text{Nd}) - (\text{Na} - \text{Nc}) - (\text{Nb} - \text{Nc})) / \text{Nt}$

Jaccard:  $\text{Nc} / ((\text{Na} - \text{Nc}) + (\text{Nb} - \text{Nc}) + \text{Nc}) = \text{Nc} / (\text{Na} + \text{Nb} - \text{Nc})$  (same as Tanimoto)

Kulczynski1:  $\text{Nc} / ((\text{Na} - \text{Nc}) + (\text{Nb} - \text{Nc})) = \text{Nc} / (\text{Na} + \text{Nb} - 2\text{Nc})$

Kulczynski2:  $((\text{Nc} / 2) * (2 * \text{Nc} + (\text{Na} - \text{Nc}) + (\text{Nb} - \text{Nc}))) / ((\text{Nc} + (\text{Na} - \text{Nc})) * (\text{Nc} + (\text{Nb} - \text{Nc}))) = 0.5 * (\text{Nc} / \text{Na} + \text{Nc} / \text{Nb})$

Manhattan:  $((\text{Na} - \text{Nc}) + (\text{Nb} - \text{Nc})) / \text{Nt} = (\text{Na} + \text{Nb} - 2\text{Nc}) / \text{Nt}$

Matching:  $(\text{Nc} + \text{Nd}) / \text{Nt}$

McConnaughey:  $(\text{Nc} ** 2 - (\text{Na} - \text{Nc}) * (\text{Nb} - \text{Nc})) / (\text{Na} * \text{Nb})$

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

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

RogersTanimoto:  $(\text{Nc} + \text{Nd}) / ((\text{Na} - \text{Nc}) + (\text{Nb} - \text{Nc}) + \text{Nt}) = (\text{Nc} + \text{Nd}) / (\text{Na} + \text{Nb} - 2\text{Nc} + \text{Nt})$

RussellRao:  $\text{Nc} / \text{Nt}$

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

SkoalSneath1:  $\text{Nc} / (\text{Nc} + 2 * (\text{Na} - \text{Nc}) + 2 * (\text{Nb} - \text{Nc})) = \text{Nc} / (2 * \text{Na} + 2 * \text{Nb} - 3 * \text{Nc})$

SkoalSneath2:  $(2 * \text{Nc} + 2 * \text{Nd}) / (\text{Nc} + \text{Nd} + \text{Nt})$

SkoalSneath3:  $(\text{Nc} + \text{Nd}) / ((\text{Na} - \text{Nc}) + (\text{Nb} - \text{Nc})) = (\text{Nc} + \text{Nd}) / (\text{Na} + \text{Nb} - 2 * \text{Nc})$

Tanimoto:  $\text{Nc} / ((\text{Na} - \text{Nc}) + (\text{Nb} - \text{Nc}) + \text{Nc}) = \text{Nc} / (\text{Na} + \text{Nb} - \text{Nc})$  (same as Jaccard)

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

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

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':  $\text{Nc}' / ((\text{Na}' - \text{Nc}') + (\text{Nb}' - \text{Nc}') + \text{Nc}') = \text{Nc}' / (\text{Na}' + \text{Nb}' - \text{Nc}')$

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

--InDelim *comma | semicolon*

Input delimiter for CSV *TextFile(s)*. Possible values: *comma or semicolon*. Default value: *comma*. For TSV files, this option is ignored and *tab* is used as a delimiter.

--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 column name containing Fingerprint substring and create a SampleFPTanimoto.csv file containing compound IDs retrieved from column name containing CompoundID substring, type:

```
% SimilarityMatrixTextFiles.pl -o SampleFP.csv
```

To generate similarity matrices corresponding to all supported similarity coefficients for fingerprints data in any internal fingerprint format present in a column name containing Fingerprint substring and create SampleFP[CoefficientName].csv files containing compound IDs retrieved from column name containing CompoundID substring, type:

```
% SimilarityMatrixTextFiles.pl -o -m All SampleFP.csv
```

To generate similarity matrices corresponding to Buser, Dice and Tanimoto similarity coefficients for fingerprints data in any internal fingerprint format present in a column name containing Fingerprint substring and create SampleFP[CoefficientName].csv files containing compound IDs retrieved from column name containing CompoundID substring, type:

```
% SimilarityMatrixTextFiles.pl -o -m "Buser,Dice,Tanimoto" SampleFP.csv
```

To generate a similarity matrix corresponding to Tanimoto similarity coefficient for fingerprints data in any intern fingerprint format present in a column number 2 and create a SampleFPTanimoto.csv file containing compound IDs retrieved from column number 1, type:

```
% SimilarityMatrixTextFiles.pl -o --ColMode ColNum --CompoundIDCol 1
--FingerprintsCol 2 SampleFP.csv
```

To generate a similarity matrix corresponding to Tanimoto similarity coefficient for fingerprints data in any internal fingerprint format in column named PathLengthFingerprints and create a SampleFPTanimoto.csv file containing compound IDs retrieved from column named CompoundID, type:

```
% SimilarityMatrixTextFiles.pl -o --ColMode ColLabel --CompoundIDCol
CompoundID --FingerprintsCol PathLengthFingerprints SampleFP.csv
```

To generate a similarity matrix corresponding to Tanimoto similarity coefficient for fingerprints data in as hexadecimal bit-string format in column named Fingerprints and create a SampleFPTanimoto.csv file containing compound IDs retrieved from column named MolID, type:

```
% SimilarityMatrixTextFiles.pl -o --ColMode ColLabel --CompoundIDCol
MolID --FingerprintsCol Fingerprints --FingerprintsFormatMode Specify
--FingerprintsString Hexadecimal SampleFP.csv
```

To generate a similarity matrix corresponding to Tanimoto similarity coefficient for fingerprints data in any internal fingerprint format present in a column name containing Fingerprint substring and create a SampleFPTanimoto.tsv file without any quotes around values along with compound IDs retrieved from column name containing CompoundID substring, type:

```
% SimilarityMatrixTextFiles.pl -o --OutDelim Tab --quote No SampleFP.csv
```

#### AUTHOR

Manish Sud <msud@san.rr.com>

#### SEE ALSO

InfoFingerprintsTextFiles.pl, PathLengthFingerprints.pl, SimilarityMatrixSDFFiles.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.