NAME

InfoFingerprintsFiles.pl - List information about fingerprints data in SD, FP and CSV/TSV text file(s)

SYNOPSIS

InfoFingerprintsFiles.pl SDFile(s) FPFile(s) TextFile(s)...

InfoFingerprintsFiles.pl [-a, --all] [--AverageBitDensity] [--BitDensity] [-c, --count] [-c, --ColMode *ColNum | ColLabel*] [--DataCheck] [-d, --detail *InfoLevel*] [-e, --empty] [--FingerprintsCol *col number | col name*] [--FingerprintsField *FieldLabel*] [--FingerprintsType] [--FingerprintsDescription] [--FingerprintsSize] [--FingerprintsBitStringFormat] [--FingerprintsBitOrder] [--FingerprintsVectorValuesType] [--FingerprintsVectorValuesFormat] [-h, --help] [--InDelim *comma | semicolon*] [--NumOfOnBits] [--NumOfNonZeroValues] [-w, --WorkingDir dirname] SDFile(s) FPFile(s) TextFile(s)...

DESCRIPTION

List information about fingerprints data in *SD*, *FP* and *CSV/TSV* text file(s): number of rows containing fingerprints data, type of fingerprints vector, description and size of fingerprints, bit density and average bit density for bit-vector fingerprints strings, and so on.

The scripts InfoFingerprintsSDFiles.pl and InfoFingerprintsTextFiles.pl have been removed from the current release of MayaChemTools and their functionality merged with this script.

The valid *SDFile* extensions are .sdf and .sd. All SD files in a current directory can be specified either by *.sdf or the current directory name.

The valid *FPFile* extensions are .fpf and .fp. All FP files in a current directory can be specified either by *.fpf or the current directory name.

The valid *TextFile* extensions are .csv and .tsv for comma/semicolon and tab delimited text files respectively. All other file names are ignored. All 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.

Format of fingerprint strings data in SDFile(s), FPFile(s) and TextFile(s) is automatically detected.

Example of FP file containing fingerprints bit-vector string data:

Example of FP file containing fingerprints vector string data:

```
# Package = MayaChemTools 7.4
# ReleaseDate = Oct 21, 2010
# TimeStamp = Mon Mar 7 15:14:01 2011
# FingerprintsStringType = FingerprintsVector
# Description = PathLengthBits:AtomicInvariantsAtomTypes:MinLength1:...
# VectorStringFormat = IDsAndValuesString
# VectorValuesType = NumericalValues
Cmpd1 338;C F N O C:C C:N C=O CC CF CN CO C:C:C C:C:N C:CC C:CF C:CN C:
N:C C:NC CC:N CC=O CCC CCN CCO CNC NC=O O=CO C:C:C:C C:C:C:N C:C:CC...;
33 1 2 5 21 2 2 12 1 3 3 20 2 10 2 2 1 2 2 2 8 2 5 1 1 1 1 19 2 8 2 2 2 2
6 2 2 2 2 2 2 2 2 3 2 2 1 4 1 5 1 1 18 6 2 2 1 2 10 2 1 2 1 2 2 2 2 2 . . .
Cmpd2 103;C N O C=N C=O CC CN CO CC=O CCC CCN CCO CNC N=CN NC=O NCN O=C
O C CC=O CCCC CCCN CCCO CCNC CNC=N CNC=O CNCN CCCC=O CCCCC CCCCN CC...;
15 4 4 1 2 13 5 2 2 15 5 3 2 2 1 1 1 2 17 7 6 5 1 1 1 2 15 8 5 7 2 2 2 2
1 2 1 1 3 15 7 6 8 3 4 4 3 2 2 1 2 3 14 2 4 7 4 4 4 4 1 1 1 2 1 1 1 ...
. . . . . .
```

Example of SD file containing fingerprints bit-vector string data:

> <PathLengthFingerprints>

FingerprintsBitVector; PathLengthBits: AtomicInvariantsAtomTypes: MinLength1: MaxLength8; 1024; HexadecimalString; Ascending; 9c8460989ec8a49913991a6603130b0a19e8051c89184414953800cc2151082844a201042800130860308e8204d402800831048940e44281c00060449a5000ac80c894114e006321264401600846c05016446208190410805000304a10205b0100e04c0038ba0fad0209c0ca8b1200012268b61c0026aa0660a11014a011d46

\$\$\$\$

Example of CSV Text file containing fingerprints bit-vector string data:

```
"CompoundID", "PathLengthFingerprints"

"Cmpd1", "FingerprintsBitVector; PathLengthBits: AtomicInvariantsAtomTypes
:MinLength1: MaxLength8; 1024; HexadecimalString; Ascending; 9c8460989ec8a4
9913991a6603130b0a19e8051c89184414953800cc2151082844a20104280013086030
8e8204d402800831048940e44281c00060449a5000ac80c894114e006321264401..."
......
```

The current release of MayaChemTools supports the following types of fingerprint bit-vector and vector strings:

FingerprintsVector;AtomNeighborhoods:AtomicInvariantsAtomTypes:MinRadius0:MaxRadius2;41;AlphaNumericalValues;ValuesString;NR0-C.X1.B01.H3-ATC1:NR1-C.X3.B03.H1-ATC1:NR2-C.X1.B01.H3-ATC1:NR2-C.X3.B04-ATC1 NR0-C.X 1.B01.H3-ATC1:NR2-C.X3.B03.H1-ATC1:NR2-C.X1.B01.H3-ATC1:NR2-C.X3.B04-ATC1 NR0-C.X2.B02.H2-ATC1:NR1-C.X2.B02.H2-ATC1:NR1-C.X3.B03.H1-ATC1:NR2-C.X3.B03.H1-ATC1:NR2-C.X2.B02.H2-ATC1:NR2-C.X3.B03.H1-ATC1:NR2-C.X2.B02.H2-ATC1:NR2-C.X2.B03.H1-ATC1.NR2-C.X2.B03.H1-ATC1:NR2-C.X2.B03.H1-ATC1:NR2-C.X2.B03.H1-ATC1.NR2-C.X2.B03.H1-AT

FingerprintsVector;AtomTypesCount:AtomicInvariantsAtomTypes:ArbitrarySize;10;NumericalValues;IDsAndValuesString;C.X1.B01.H3 C.X2.B02.H2 C.X2.B03.H1 C.X3.B03.H1 C.X3.B04 F.X1.B01 N.X2.B02.H1 N.X3.B03 O.X1.B01.H1 O.X1.B02;2 4 14 3 10 1 1 1 3 2

FingerprintsVector;AtomTypesCount:SLogPAtomTypes:ArbitrarySize;16;NumericalValues;IDsAndValuesString;C1 C10 C11 C14 C18 C20 C21 C22 C5 CS F N11 N4 O10 O2 O9;5 1 1 1 14 4 2 1 2 2 1 1 1 1 3 1

FingerprintsVector; EStateIndicies: ArbitrarySize; 11; NumericalValues; IDs AndValuesString; SaaCH SaasC SaasN SdO SdssC SsCH3 SsF SsOH SssCH2 SssN H SsssCH; 24.778 4.387 1.993 25.023 -1.435 3.975 14.006 29.759 -0.073 3 .024 -2.270

FingerprintsVector; ExtendedConnectivity: AtomicInvariantsAtomTypes: Radius2;60;AlphaNumericalValues; ValuesString;73555770 333564680 352413391 666191900 1001270906 1371674323 1481469939 1977749791 2006158649 21414 08799 49532520 64643108 79385615 96062769 273726379 564565671 85514103 5 906706094 988546669 1018231313 1032696425 1197507444 1331250018 1338 532734 1455473691 1607485225 1609687129 1631614296 1670251330 17303...

FingerprintsVector; ExtendedConnectivity: FunctionalClassAtomTypes: Radiu s2;57; AlphaNumericalValues; ValuesString; 24769214 508787397 850393286 8 62102353 981185303 1231636850 1649386610 1941540674 263599683 32920567 1 571109041 639579325 683993318 723853089 810600886 885767127 90326012 7 958841485 981022393 1126908698 1152248391 1317567065 1421489994 1455 632544 1557272891 1826413669 1983319256 2015750777 2029559552 20404...

FingerprintsVector; ExtendedConnectivity: EStateAtomTypes: Radius2;62;Alp haNumericalValues; ValuesString; 25189973 528584866 662581668 671034184 926543080 1347067490 1738510057 1759600920 2034425745 2097234755 21450 44754 96779665 180364292 341712110 345278822 386540408 387387308 50430 1706 617094135 771528807 957666640 997798220 1158349170 1291258082 134 1138533 1395329837 1420277211 1479584608 1486476397 1487556246 1566...

FingerprintsVector; PathLengthCount: AtomicInvariantsAtomTypes: MinLength 1: MaxLength8; 432; NumericalValues; IDsAndValuesPairsString; C.X1.BO1.H3 2 C.X2.BO2.H2 4 C.X2.BO3.H1 14 C.X3.BO3.H1 3 C.X3.BO4 10 F.X1.BO1 1 N.X 2.BO2.H1 1 N.X3.BO3 1 O.X1.BO1.H1 3 O.X1.BO2 2 C.X1.BO1.H3C.X3.BO3.H1 2 C.X2.BO2.H2C.X2.BO2.H2 1 C.X2.BO2.H2C.X3.BO3.H1 4 C.X2.BO2.H2C.X3.BO 4 1 C.X2.BO2.H2N.X3.BO3 1 C.X2.BO3.H1:C.X2.BO3.H1 10 C.X2.BO3.H1:C....

FingerprintsVector; PathLengthCount: MMFF94AtomTypes: MinLength1: MaxLength8;463; NumericalValues; IDsAndValuesPairsString; C5A 2 C5B 2 C=ON 1 CB 1 8 COO 1 CR 9 F 1 N5 1 NC=O 1 O=CN 1 O=CO 1 OC=O 1 OR 2 C5A:C5B 2 C5A:N 5 2 C5ACB 1 C5ACR 1 C5B:C5B 1 C5BC=ON 1 C5BCB 1 C=ON=O=CN 1 C=ONNC=O 1 CB:CB 18 CBF 1 CBNC=O 1 COO=O=CO 1 COOCR 1 COOCC=O 1 CRCR 7 CRN5 1 CR OR 2 C5A:C5B:C5B 2 C5A:C5BC=ON 1 C5A:C5BCB 1 C5A:N5:C5A 1 C5A:N5CR ...

FingerprintsVector; TopologicalAtomPairs: AtomicInvariantsAtomTypes: MinD istancel: MaxDistancel0; 223; NumericalValues; IDsAndValuesString; C.X1.BO1.H3-D1-C.X3.BO3.H1 C.X2.BO2.H2-D1-C.X2.BO2.H2 C.X2.BO2.H2-D1-C.X3.BO3.H1 C.X2.BO2.H2-D1-C.X3.BO3.C.X2.BO3.H1-D1-...; 2 1 4 1 1 10 8 1 2 6 1 2 2 1 2 1 2 1 2 1 2 1 5 1 10 12 2 2 1 2 1 9 1 3 1 1 1 2 2 1 3 6 1 6 14 2 2 2 3 1 3 1 8 2 2 1 3 2 6 1 2 2 5 1 3 1 23 1...

FingerprintsVector;TopologicalAtomPairs:FunctionalClassAtomTypes:MinDistancel:MaxDistancel0;144;NumericalValues;IDsAndValuesString;Ar-D1-Ar Ar-D1-Ar.HBA Ar-D1-HBD Ar-D1-Hal Ar-D1-None Ar.HBA-D1-None HBA-D1-NI HBA-D1-None HBA-D1-NI HBA-HBD-D1-None HBD-D1-None NI-D1-None No...; 23 2 1 1 2 1 1 1 1 2 1 1 7 28 3 1 3 2 8 2 1 1 1 5 1 5 24 3 3 4 2 13 4 1 1 4 1 5 22 4 4 3 1 19 1 1 1 1 1 2 2 3 1 1 8 25 4 5 2 3 1 26 1 4 1 ...

FingerprintsVector; TopologicalAtomTorsions: AtomicInvariantsAtomTypes; 3
3; NumericalValues; IDsAndValuesString; C.X1.B01.H3-C.X3.B03.H1-C.X3.B04-C.X3.B04 C.X1.B01.H3-C.X3.B03.H1-C.X3.B04-N.X3.B03 C.X2.B02.H2-C.X2.B0
2.H2-C.X3.B03.H1-C.X2.B02.H2 C.X2.B02.H2-C.X2.B02.H2-C.X3.B03.H1-O...;
2 2 1 1 2 2 1 1 3 4 4 8 4 2 2 6 2 2 1 2 1 1 2 1 1 2 6 2 4 2 1 3 1

FingerprintsVector;TopologicalAtomTriplets:SYBYLAtomTypes:MinDistance1:MaxDistance10;2332;NumericalValues;IDsAndValuesString;C.2-D1-C.2-D9-C.3-D10 C.2-D1-C.2-D9-C.ar-D10 C.2-D1-C.3-D1-C.3-D2 C.2-D1-C.3-D10-C.3-D9 C.2-D1-C.3-D2-C.3-D3 C.2-D1-C.3-D2-C.ar-D3 C.2-D1-C.3-D3-C.3-D4 C.2-D1-C.3-D3-N.ar-D4 C.2-D1-C.3-D3-O.3-D2 C.2-D1-C.3-D4-C.3-D5 C.2-D1-C.3-D5-C.3-D6 C.2-D1-C.3-D5-O.3-D4 C.2-D1-C.3-D6-C.3-D7 C.2-D1-C.3-D7...

FingerprintsVector;TopologicalPharmacophoreAtomPairs:ArbitrarySize:Min Distancel:MaxDistancel0;54;NumericalValues;IDsAndValuesString;H-D1-H H -D1-NI HBA-D1-NI HBD-D1-NI H-D2-H H-D2-HBA H-D2-HBD HBA-D2-HBA H-D3-HBA H-D3-HBD H-D3-NI HBD-D3-NI HBD-D3-NI H-D4-H H-D4-H BA H-D4-HBD HBA-D4-HBA HBA-D4-HBD HBD-D4-HBD H-D5-H H-D5-HBA H-D5-...; 18 1 2 1 22 12 8 1 2 18 6 3 1 1 1 22 13 6 5 7 2 28 9 5 1 1 1 36 16 10 3 4 1 37 10 8 1 35 10 9 3 3 1 28 7 7 4 18 16 12 5 1 2 1

FingerprintsVector; TopologicalPharmacophoreAtomTriplets: ArbitrarySize: MinDistance1: MaxDistance10;696; NumericalValues; IDsAndValuesString; Arl-Arl-Arl Arl-Arl-Hl Arl-Arl-HBAl Arl-Arl-HBDl Arl-Hl-Hl Arl-Hl-HBAl Arl-Hl-HBDl Arl-HBAl-HBAl Hl-HB-HBDl Hl-HBAl-HBDl Hl-HBAl-HBAl Hl-HBAl-HBDl Hl-HBAl-HBDl Hl-HBAl-HBAl Hl-HBAl-HBDl Hl-HBAl-NII Hl-HBDl-NII HBAl-HBDl-NII HBAl-HBDl-NII Arl-...; 46 106 8 3 83 11 4 1 21 5 3 1 2 2 1 1 1 100 101 18 11 145 132 26 14 23 28 3 3 5 4 61 45 10 4 16 20 7 5 1 3 4 5 3 1 1 1 1 5 4 2 1 2 2 2 1 1 1 119 123 24 15 185 202 41 25 22 17 3 5 85 95 18 11 23 17 3 1 1 6 4 ...

FingerprintsVector;TopologicalPharmacophoreAtomTriplets:FixedSize:MinDistance1:MaxDistance10;2692;OrderedNumericalValues;ValuesString;46 106

OPTIONS

-a, --all

List all the available information.

--AverageBitDensity

List average bit density of fingerprint bit-vector strings.

--BitDensity

List bit density of fingerprints bit-vector strings data in each row.

--count

List number of data entries containing fingerprints bit-vector or vector strings data. This is default behavior.

-c, --ColMode ColNum | ColLabel

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

-d, --detail InfoLevel

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

--DataCheck

Validate fingerprints data specified using --FingerprintsCoI and list information about missing and invalid data.

-e, --empty

List number of rows containing no fingerprints data.

--FingerprintsCol col number | col name

This value is -c, --colmode specific. It corresponds to column in CSV/TSV *TextFile(s)* containing fingerprints data. Possible values: *col number or col label*. Default value: *first column containing the word Fingerprints in its column label*.

--FingerprintsField *FieldLabel*

Fingerprints field label to use during listing of fingerprints information for *SDFile(s)*. Default value: *first data field label containing the word Fingerprints in its label*.

--FingerprintsType

List types of fingerprint strings: FingerprintsBitVector or FingerprintsVector.

--FingerprintsDescription

List types of fingerprints: PathLengthBits, PathLengthCount, MACCSKeyCount, ExtendedConnectivity and so on.

--FingerprintsSize

List size of fingerprints.

--FingerprintsBitStringFormat

List format of fingerprint bit-vector strings: BinaryString or HexadecimalString.

--FingerprintsBitOrder

List order of bits data in fingerprint bit-vector bit strings: Ascending or Descending.

--FingerprintsVectorValuesType

List type of values in fingerprint vector strings: OrderedNumericalValues, NumericalValues or AlphaNumericalValues.

--FingerprintsVectorValuesFormat

List format of values in fingerprint vector strings: ValuesString, IDsAndValuesString, IDsAndValuesPairsString, ValuesAndIDsString or ValuesAndIDsPairsString.

-h, --help

Print this help message.

--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.

--NumOfOnBits

List number of on bits in fingerprints bit-vector strings data in each row.

--NumOfNonZeroValues

List number of non-zero values in fingerprints vector strings data in each row.

-w, --WorkingDir DirName

Location of working directory. Default: current directory.

EXAMPLES

To count number of lines containing fingerprints bit-vector or vector strings data present in FP file, in a column name containing Fingerprint substring in text file, and in a data field with Fingerprint substring in its label, type:

- % InfoFingerprintsFiles.pl SampleFPBin.csv
- % InfoFingerprintsFiles.pl SampleFPBin.sdf SampleFPBin.fpf SampleFPBin.csv
- % InfoFingerprintsFiles.pl SampleFPHex.sdf SampleFPHex.fpf SampleFPHex.csv
- % InfoFingerprintsFiles.pl SampleFPcount.sdf SampleFPcount.fpf SampleFPcount.csv

To list all available information about fingerprints bit-vector or vector strings data present in FP file, in a column name containing Fingerprint substring in text file, and in a data field with Fingerprint substring in its label, type:

- % InfoFingerprintsFiles.pl -a SampleFPHex.sdf SampleFPHex.fpf SampleFPHex.csv
- % InfoFingerprintsFiles.pl -a SampleFPcount.sdf SampleFPcount.fpf SampleFPcount.csv

To list all available information about fingerprints bit-vector or vector strings data present in a column named Fingerprints in text file, type:

- % InfoFingerprintsFiles.pl -a --ColMode ColLabel --FingerprintsCol Fingerprints SampleFPHex.sdf
- % InfoFingerprintsFiles.pl -a --ColMode ColLabel --FingerprintsCol Fingerprints SampleFPcount.csv

To list all available information about fingerprints bit-vector or vector strings data present in a data field names Fingerprints in SD file, type:

- % InfoFingerprintsFiles.pl -a --FingerprintsField Fingerprints SampleFPHex.sdf
- % InfoFingerprintsFiles.pl -a --FingerprintsField Fingerprints SampleFPcount.sdf

To list bit density, average bit density, and number of on bits for fingerprints bit-vector strings data present in FP file, in a column name containing Fingerprint substring in text file, and in a data field with Fingerprint substring in its label, type:

```
% InfoFingerprintsFiles.pl --BitDensity --AverageBitDensity
--NumOfOnBits SampleFPBin.csv SampleFPBin.sdf SampleFPBin.fpf
```

To list vector values type, format and number of non-zero values for fingerprints vector strings data present in FP file, in a column name containing Fingerprint substring in text file, and in a data field with Fingerprint substring in its label along with fingerprints type and description, type:

```
% InfoFingerprintsFiles.pl --FingerprintsType --FingerprintsDescription
--FingerprintsVectorValuesType --FingerprintsVectorValuesFormat
--NumOfNonZeroValues SampleFPcount.csv SampleFPcount.sdf
SampleFPcount.fpf
```

AUTHOR

Manish Sud <msud@san.rr.com>

SEE ALSO

SimilarityMatricesFingerprints.pl, SimilaritySearchingFingerprints.pl, AtomNeighborhoodsFingerprints.pl, AtomNeighborhoodsFingerprints.pl, ExtendedConnectivityFingerprints.pl, MACCSKeysFingerprints.pl, PathLengthFingerprints.pl, TopologicalAtomPairsFingerprints.pl, TopologicalAtomTorsionsFingerprints.pl, TopologicalPharmacophoreAtomPairsFingerprints.pl, TopologicalPharmacophoreAtomTripletsFingerprints.pl

COPYRIGHT

Copyright (C) 2017 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.

www.MayaChemTools.org Page 7