

NAME

AtomTypesFingerprints.pl - Generate atom types fingerprints for SD files

SYNOPSIS

AtomTypesFingerprints.pl SDFfile(s)...

```
AtomTypesFingerprints.pl [-a, --AtomIdentifierType AtomicInvariantsAtomTypes | DREIDINGAtomTypes |
EStateAtomTypes | MMFF94AtomTypes | SLogPAtomTypes | SYBYLAtomTypes | TPSAAAtomTypes | UFFAtomTypes] [
--AtomicInvariantsToUse "AtomicInvariant, AtomicInvariant..."] [--FunctionalClassesToUse
"FunctionalClass1, FunctionalClass2..."] [--AtomTypesSetToUse ArbitrarySize | FixedSize] [--BitsOrder Ascending |
Descending] [-b, --BitStringFormat BinaryString | HexadecimalString] [--CompoundID DataFieldName or
LabelPrefixString] [--CompoundIDLabel text] [--CompoundIDMode DataField | MolName | LabelPrefix |
MolNameOrLabelPrefix] [--DataFields "FieldLabel1, FieldLabel2..."] [-d, --DataFieldsMode All | Common | Specify |
CompoundID] [-f, --Filter Yes | No] [--FingerprintsLabelMode FingerprintsLabelOnly | FingerprintsLabelWithIDs] [
--FingerprintsLabel text] [-h, --help] [-k, --KeepLargestComponent Yes | No] [-m, --mode AtomTypesCount |
AtomTypesBits] [-i, --IgnoreHydrogens Yes | No] [--OutDelim comma | tab | semicolon] [--output SD | FP | text | all] [
-o, --overwrite] [-q, --quote Yes | No] [-r, --root RootName] [-s, --size number] [--ValuesPrecision number] [
-v, --VectorStringFormat IDsAndValuesString | IDsAndValuesPairsString | ValuesAndIDsString | ValuesAndIDsPairsString] [
-w, --WorkingDir DirName]
```

DESCRIPTION

Generate atom types fingerprints for *SDFfile(s)* and create appropriate SD, FP or CSV/TSV text file(s) containing fingerprints bit-vector or vector strings corresponding to molecular fingerprints.

Multiple SDFfile 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.

The current release of MayaChemTools supports generation of atom types fingerprints corresponding to following -a, --AtomIdentifierTypes:

```
AtomicInvariantsAtomTypes, DREIDINGAtomTypes, EStateAtomTypes,
FunctionalClassAtomTypes, MMFF94AtomTypes, SLogPAtomTypes,
SYBYLAtomTypes, TPSAAAtomTypes, UFFAtomTypes
```

Based on the values specified for -a, --AtomIdentifierType along with other specified parameters such as --AtomicInvariantsToUse and --FunctionalClassesToUse, initial atom types are assigned to all non-hydrogen atoms or all atoms in a molecule

Using the assigned atom types and specified -m, --Mode, one of the following types of fingerprints are generated:

```
AtomTypesCount - A vector containing count of atom types
AtomTypesBits - A bit vector indicating presence/absence of atom types
```

For *AtomTypesCount* fingerprints, two types of atom types set size are allowed as value of --AtomTypesSetToUse option:

```
ArbitrarySize - Corresponds to only atom types detected in molecule
FixedSize - Corresponds to fixed number of atom types previously defined
```

For *AtomTypesBits* fingerprints, only *FixedSize* atom type set is allowed.

ArbitrarySize corresponds to atom types detected in a molecule where as *FixedSize* implies a fix number of all possible atom types previously defined for a specific -a, --AtomIdentifierType.

Fix number of all possible atom types for supported *AtomIdentifierTypes* in current release of MayaChemTools are:

AtomIdentifier	Total	TotalWithoutHydrogens
DREIDINGAtomTypes	37	34
EStateAtomTypes	109	87
MMFF94AtomTypes	212	171
SLogPAtomTypes	72	67
SYBYLAtomTypes	45	44

The functional class abbreviations correspond to:

```
HBD: HydrogenBondDonor
HBA: HydrogenBondAcceptor
PI : PositivelyIonizable
NI : NegativelyIonizable
Ar : Aromatic
Hal : Halogen
H : Hydrophobic
RA : RingAtom
CA : ChainAtom
```

Functional class atom type specification for an atom corresponds to:

```
Ar.CA.H.HBA.HBD.Hal.NI.PI.RA
```

AtomTypes::FunctionalClassAtomTypes module is used to assign functional class atom types. It uses following definitions [Ref 60-61, Ref 65-66]:

```
HydrogenBondDonor: NH, NH2, OH
HydrogenBondAcceptor: N[!H], O
PositivelyIonizable: +, NH2
NegativelyIonizable: -, C(=O)OH, S(=O)OH, P(=O)OH
```

--AtomTypesSetToUse *ArbitrarySize* | *FixedSize*

Atom types set size to use during generation of atom types fingerprints.

Possible values for *AtomTypesCount* values of -m, --mode option: *ArbitrarySize* | *FixedSize*; Default value: *ArbitrarySize*.

Possible values for *AtomTypesBits* value of -m, --mode option: *FixedSize*; Default value: *FixedSize*.

FixedSize value is not supported for *AtomicInvariantsAtomTypes* value of -a, --AtomI identifierType option.

ArbitrarySize corresponds to only atom types detected in molecule; *FixedSize* corresponds to fixed number of previously defined atom types for specified -a, --AtomI identifierType.

--BitsOrder *Ascending* | *Descending*

Bits order to use during generation of fingerprints bit-vector string for *AtomTypesBits* value of =item

--BitsOrder *Ascending* | *Descending*

Bits order to use during generation of fingerprints bit-vector string for *AtomTypesBits* value of -m, --mode option. Possible values: *Ascending*, *Descending*. Default: *Ascending*.

Ascending bit order which corresponds to first bit in each byte as the lowest bit as opposed to the highest bit.

Internally, bits are stored in *Ascending* order using Perl vec function. Regardless of machine order, big-endian or little-endian, vec function always considers first string byte as the lowest byte and first bit within each byte as the lowest bit.

-b, --BitStringFormat *BinaryString* | *HexadecimalString*

Format of fingerprints bit-vector string data in output SD, FP or CSV/TSV text file(s) specified by --output used during *AtomTypesBits* value of -m, --mode option. Possible values: *BinaryString*, *HexadecimalString*.

Default value: *BinaryString*.

BinaryString corresponds to an ASCII string containing 1s and 0s. *HexadecimalString* contains bit values in ASCII hexadecimal format.

Examples:

```
FingerprintsBitVector;AtomTypesBits:DREIDINGAtomTypes;34;BinaryString;
Ascending;001010101010100000000000000000000000000000000000000
```

```
FingerprintsBitVector;AtomTypesBits:MMFF94AtomTypes;171;BinaryString;
Ascending;10000101010000000000001100000000000000001010000101000000000000
0000000000000000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000000000000000
```

--CompoundID *DataFieldName* or *LabelPrefixString*

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

For *DataField* value of --CompoundID Mode option, it 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 --CompoundID Mode:

```
MolID
ExtReg
```

Examples for *LabelPrefix* or *MolNameOrLabelPrefix* value of --CompoundID Mode:

```
Compound
```

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

--CompoundIDLabel *text*

Specify compound ID column label for FP or CSV/TSV text file(s) used during *CompoundID* value of --DataFieldsMode option. Default: *CompoundID*.

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

Specify how to generate compound IDs and write to FP or CSV/TSV text file(s) along with generated fingerprints for *FP* | *text* | *all* values of --output option: 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 --CompoundID Mode, 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.

This is only used for *CompoundID* value of --DataFieldsMode option.

--DataFields "*FieldLabel1,FieldLabel2,...*"

Comma delimited list of *SDFFile(s)* data fields to extract and write to CSV/TSV text file(s) along with generated fingerprints for *text* | *all* values of --output option.

This is only used for *Specify* value of --DataFieldsMode option.

Examples:

```
Extreg
MolID,CompoundName
```

-d, --DataFieldsMode *All* | *Common* | *Specify* | *CompoundID*

Specify how data fields in *SDFFile(s)* are transferred to output CSV/TSV text file(s) along with generated fingerprints for *text* | *all* values of --output option: transfer all SD data field; transfer SD data files common to all compounds; extract specified data fields; generate a compound ID using molname line, a compound prefix, or a combination of both. Possible values: *All* | *Common* | *specify* | *CompoundID*. Default value: *CompoundID*.

-f, --Filter *Yes* | *No*

Specify whether to check and filter compound data in *SDFFile(s)*. Possible values: *Yes* or *No*. Default value: *Yes*.

By default, compound data is checked before calculating fingerprints and compounds containing atom data corresponding to non-element symbols or no atom data are ignored.

--FingerprintsLabelMode *FingerprintsLabelOnly* | *FingerprintsLabelWithIDs*

Specify how fingerprints label is generated in conjunction with --FingerprintsLabel option value: use fingerprints label generated only by --FingerprintsLabel option value or append atom type value IDs to --FingerprintsLabel option value.

Possible values: *FingerprintsLabelOnly* | *FingerprintsLabelWithIDs*. Default value: *FingerprintsLabelOnly*.

This option is only used for *FixedSize* value of *-e*, *--AtomTypesSetToUse* option during generation of *AtomTypesCount* fingerprints and ignored for *AtomTypesBits*.

Atom type IDs appended to *--FingerprintsLabel* value during *FingerprintsLabelWithIDs* values of *--FingerprintsLabelMode* correspond to fixed number of previously defined atom types.

--FingerprintsLabel text

SD data label or text file column label to use for fingerprints string in output SD or CSV/TSV text file(s) specified by *--output*. Default value: *AtomTypesFingerprints*.

-h, --help

Print this help message.

-i, --IgnoreHydrogens Yes | No

Ignore hydrogens during fingerprints generation. Possible values: *Yes* or *No*. Default value: *Yes*.

For *yes* value of *-i, --IgnoreHydrogens*, any explicit hydrogens are also used for generation of atom type fingerprints; implicit hydrogens are still ignored.

-k, --KeepLargestComponent Yes | No

Generate fingerprints for only the largest component in molecule. Possible values: *Yes* or *No*. Default value: *Yes*.

For molecules containing multiple connected components, fingerprints can be generated in two different ways: use all connected components or just the largest connected component. By default, all atoms except for the largest connected component are deleted before generation of fingerprints.

-m, --mode AtomTypesCount | AtomTypesBits

Specify type of atom types fingerprints to generate for molecules in *SDFFile(s)*. Possible values: *AtomTypesCount* or *AtomTypesBits*. Default value: *AtomTypesCount*.

For *AtomTypesCount* values of *-m, --mode* option, a fingerprint vector string is generated. The vector string corresponding to *AtomTypesCount* contains count of atom types.

For *AtomTypesBits* value of *-m, --mode* option, a fingerprint bit-vector string containing zeros and ones indicating presence or absence of atom types is generated.

For *AtomTypesCount* atom types fingerprints, two types of atom types set size can be specified using *-a, --AtomTypesSetToUse* option: *ArbitrarySize* or *FixedSize*. *ArbitrarySize* corresponds to only atom types detected in molecule; *FixedSize* corresponds to fixed number of atom types previously defined.

For *AtomTypesBits* atom types fingerprints, only *FixedSize* is allowed.

Combination of *-m, --Mode* and *--AtomTypesSetToUse* along with *-a, --AtomIdentifierType* allows generation of following different atom types fingerprints:

Mode	AtomIdentifierType	AtomTypesSetToUse
AtomTypesCount	AtomicInvariantsAtomTypes	ArbitrarySize [Default]
AtomTypesCount	DREIDINGAtomTypes	ArbitrarySize
AtomTypesCount	DREIDINGAtomTypes	FixedSize
AtomTypesBits	DREIDINGAtomTypes	FixedSize
AtomTypesCount	EStateAtomTypes	ArbitrarySize
AtomTypesCount	EStateAtomTypes	FixedSize
AtomTypesBits	EStateAtomTypes	FixedSize
AtomTypesCount	FunctionalClassAtomTypes	ArbitrarySize
AtomTypesCount	MMFF94AtomTypes	ArbitrarySize
AtomTypesCount	MMFF94AtomTypes	FixedSize
AtomTypesBits	MMFF94AtomTypes	FixedSize
AtomTypesCount	SLogPAtomTypes	ArbitrarySize
AtomTypesCount	SLogPAtomTypes	FixedSize
AtomTypesBits	SLogPAtomTypes	FixedSize

AtomTypesCount	SYBYLAtomTypes	ArbitrarySize
AtomTypesCount	SYBYLAtomTypes	FixedSize
AtomTypesBits	SYBYLAtomTypes	FixedSize
AtomTypesCount	TPSAAAtomTypes	FixedSize
AtomTypesBits	TPSAAAtomTypes	FixedSize
AtomTypesCount	UFFAtomTypes	ArbitrarySize
AtomTypesCount	UFFAtomTypes	FixedSize
AtomTypesBits	UFFAtomTypes	FixedSize

The default is to generate *AtomicInvariantAtomTypes* fingerprints corresponding to *ArbitrarySize* as value of `--AtomTypesSetToUse` option.

`--OutDelim` *comma | tab | semicolon*

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

`--output` *SD | FP | text | all*

Type of output files to generate. Possible values: *SD, FP, text, or all*. Default value: *text*.

`-o, --overwrite`

Overwrite existing files.

`-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: `<Root>.<Ext>`. Default for new file names: `<SDFileName><AtomTypesFP>.<Ext>`. The file type determines `<Ext>` value. The `sdf, fpf, csv, and tsv` `<Ext>` values are used for SD, FP, comma/semicolon, and tab delimited text files, respectively. This option is ignored for multiple input files.

`-v, --VectorStringFormat` *ValuesString | IDsAndValuesString | IDsAndValuesPairsString | ValuesAndIDsString | ValuesAndIDsPairsString*

Format of fingerprints vector string data in output SD, FP or CSV/TSV text file(s) specified by `--output` used during `<AtomTypesCount>` value of `-m, --mode` option. Possible values: *ValuesString, IDsAndValuesString | IDsAndValuesPairsString | ValuesAndIDsString | ValuesAndIDsPairsString*.

Default value during *ArbitrarySize* value of `-e, --AtomTypesSetToUse` option: *IDsAndValuesString*. Default value during *FixedSize* value of `-e, --AtomTypesSetToUse` option: *ValuesString*.

Example of *SD* file containing atom types fingerprints string data:

```

... ..
... ..
$$$$
... ..
... ..
... ..
41 44 0 0 0 0 0 0 0 0 0999 V2000
-3.3652 1.4499 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
... ..
2 3 1 0 0 0 0
... ..
M END
> <CmpdID>
Cmpd1

> <AtomTypesFingerprints>
FingerprintsVector;AtomTypesCount;AtomicInvariantsAtomTypes;ArbitrarySi
ze;10;NumericalValues;IDsAndValuesString;C.X1.BO1.H3 C.X2.BO2.H2 C.X2.B
O3.H1 C.X3.BO3.H1 C.X3.BO4 F.X1.BO1 N.X2.BO2.H1 N.X3.BO3 O.X1.BO1.H1 O.
X1.BO2;2 4 14 3 10 1 1 1 3 2

```

```
-r SampleATFP -o Sample.sdf
```

To generate E-state atom types count fingerprints of arbitrary size in vector string format and create a SampleATFP.csv file containing sequential compound IDs along with fingerprints vector strings data, type:

```
% AtomTypesFingerprints.pl -m AtomTypesCount -a EStateAtomTypes
--AtomTypesSetToUse ArbitrarySize -r SampleATFP -o Sample.sdf
```

To generate E-state atom types count fingerprints of fixed size in vector string with IDsAndValues format and create a SampleATFP.csv file containing sequential compound IDs along with fingerprints vector strings data, type:

```
% AtomTypesFingerprints.pl -m AtomTypesCount -a EStateAtomTypes
--AtomTypesSetToUse FixedSize -v IDsAndValuesString
-r SampleATFP -o Sample.sdf
```

To generate E-state atom types bits fingerprints of fixed size in bit-vector string format and create a SampleATFP.csv file containing sequential compound IDs along with fingerprints vector strings data, type:

```
% AtomTypesFingerprints.pl -m AtomTypesBits -a EStateAtomTypes
--AtomTypesSetToUse FixedSize -r SampleATFP -o Sample.sdf
```

To generate MMFF94 atom types count fingerprints of arbitrary size in vector string format and create a SampleATFP.csv file containing sequential compound IDs along with fingerprints vector strings data, type:

```
% AtomTypesFingerprints.pl -m AtomTypesCount -a MMFF94AtomTypes
--AtomTypesSetToUse ArbitrarySize -r SampleATFP -o Sample.sdf
```

To generate MMFF94 atom types count fingerprints of fixed size in vector string format and create a SampleATFP.csv file containing sequential compound IDs along with fingerprints vector strings data, type:

```
% AtomTypesFingerprints.pl -m AtomTypesCount -a MMFF94AtomTypes
--AtomTypesSetToUse FixedSize -r SampleATFP -o Sample.sdf
```

To generate MMFF94 atom types count fingerprints of fixed size in vector string with IDsAndValues format and create a SampleATFP.csv file containing sequential compound IDs along with fingerprints vector strings data, type:

```
% AtomTypesFingerprints.pl -m AtomTypesCount -a MMFF94AtomTypes
--AtomTypesSetToUse FixedSize -v IDsAndValuesString
-r SampleATFP -o Sample.sdf
```

To generate MMFF94 atom types bits fingerprints of fixed size in bit-vector string format and create a SampleATFP.csv file containing sequential compound IDs along with fingerprints vector strings data, type:

```
% AtomTypesFingerprints.pl -m AtomTypesBits -a MMFF94AtomTypes
--AtomTypesSetToUse FixedSize -r SampleATFP -o Sample.sdf
```

To generate MMFF94 atom types count fingerprints of arbitrary size in vector string format and create a SampleATFP.csv file containing compound ID from molecule name line along with fingerprints vector strings data, type:

```
% AtomTypesFingerprints.pl -m AtomTypesCount -a MMFF94AtomTypes
--DataFieldsMode CompoundID --CompoundIDMode MolName
-r SampleATFP -o Sample.sdf
```

To generate MMFF94 atom types count fingerprints of arbitrary size in vector string format and create a SampleATFP.csv file containing compound IDs using specified data field along with fingerprints vector strings data, type:

```
% AtomTypesFingerprints.pl -m AtomTypesCount -a MMFF94AtomTypes
--DataFieldsMode CompoundID --CompoundIDMode DataField --CompoundID
Mol_ID -r SampleATFP -o Sample.sdf
```

To generate MMFF94 atom types count fingerprints of arbitrary size in vector string format and create a SampleATFP.csv file containing compound ID using combination of molecule name line and an explicit compound prefix along with fingerprints vector strings data, type:

```
% AtomTypesFingerprints.pl -m AtomTypesCount -a MMFF94AtomTypes
--DataFieldsMode CompoundID --CompoundIDMode MolnameOrLabelPrefix
--CompoundID Cmpd --CompoundIDLabel MolID -r SampleATFP -o Sample.sdf
```

To generate MMFF94 atom types count fingerprints of arbitrary size in vector string format and create a SampleATFP.csv file containing specific data fields columns along with fingerprints vector strings data, type:

```
% AtomTypesFingerprints.pl -m AtomTypesCount -a MMFF94AtomTypes
--DataFieldsMode Specify --DataFields Mol_ID -r SampleATFP
-o Sample.sdf
```

To generate MMFF94 atom types count fingerprints of arbitrary size in vector string format and create a SampleATFP.csv file containing common data fields columns along with fingerprints vector strings data, type:

```
% AtomTypesFingerprints.pl -m AtomTypesCount -a MMFF94AtomTypes
--DataFieldsMode Common -r SampleATFP -o Sample.sdf
```

To generate MMFF94 atom types count fingerprints of arbitrary size in vector string format and create SampleATFP.sdf, SampleATFP.fpf and SampleATFP.csv files containing all data fields columns in CSV file along with fingerprints vector strings data, type:

```
% AtomTypesFingerprints.pl -m AtomTypesCount -a MMFF94AtomTypes
--DataFieldsMode All --output all -r SampleATFP -o Sample.sdf
```

AUTHOR

Manish Sud <msud@san.rr.com>

SEE ALSO

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

COPYRIGHT

Copyright (C) 2004-2012 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.