NAME

EStateIndiciesFingerprints.pl - Generate E-state indicies fingerprints for SD files

SYNOPSIS

EStateIndiciesFingerprints.pl SDFile(s)...

EStateIndiciesFingerprints.pl [--AromaticityModel *AromaticityModelType*] [--CompoundID *DataFieldName or LabelPrefixString*] [--CompoundIDLabel *text*] [--CompoundIDMode *DataField* | *MolName* | *LabelPrefix* | *MolNameOrLabelPrefix*] [--DataFields *"FieldLabel1,FieldLabel2,..."*] [-d, --DataFieldsMode *All* | *Common* | *Specify* | *CompoundID*] [-e, --EStateAtomTypesSetToUse *ArbitrarySize or FixedSize*] [-f, --Filter *Yes* | *No*] [--FingerprintsLabelMode *FingerprintsLabelOnly* | *FingerprintsLabelWithIDs*] [--FingerprintsLabel *text*] [-h, --help] [-k, --KeepLargestComponent *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 E-state indicies fingerprints [Ref 75-78] for *SDFile(s)* and create appropriate SD, FP, or CSV/TSV text file(s) containing fingerprints bit-vector or vector strings corresponding to molecular fingerprints.

Multiple SDFile 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.

E-state atom types are assigned to all non-hydrogen atoms in a molecule using module AtomTypes::EStateAtomTypes.pm and E-state values are calculated using module AtomicDescriptors::EStateValues.pm. Using E-state atom types and E-state values, EStateIndiciesFingerprints constituting sum of E-state values for E-sate atom types is generated.

Two types of E-state atom types set size are allowed:

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

Module AtomTypes::EStateAtomTypes.pm, used to assign E-state atom types to non-hydrogen atoms in the molecule, is able to assign atom types to any valid atom group. However, for *FixedSize* value of EStateAtomTypesSetToUse, only a fixed set of E-state atom types corresponding to specific atom groups [Appendix III in Ref 77] are used for fingerprints.

The fixed size E-state atom type set size used during generation of fingerprints contains 87 E-state non-hydrogen atom types in EStateAtomTypes.csv data file distributed with MayaChemTools.

Combination of Type and EStateAtomTypesSetToUse allow generation of 2 different types of E-state indicies fingerprints:

```
Type                        EStateAtomTypesSetToUse

EStateIndicies              ArbitrarySize      [ default fingerprints ]
EStateIndicies              FixedSize
```

Example of *SD* file containing E-state indicies fingerprints string data:

```
... ...
... ...
$$$$
... ...
... ...
... ...
41 44  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
... ...
2  3  1  0  0  0  0
... ...
M  END
>  <CmpdID>
Cmpd1

>  <EStateIndiciesFingerprints>
FingerprintsVector;EStateIndicies:ArbitrarySize;11;NumericalValues;IDsA
ndValuesString;SaaCH SaasC SaasN SdO SdssC SsCH3 SsF SsOH SssCH2 SssNH
SssCH;24.778 4.387 1.993 25.023 -1.435 3.975 14.006 29.759 -0.073 3.02
4 -2.270

$$$$
```

```
... ...
... ...
```

Example of *FP* file containing E-state indicies fingerprints string data:

```
#
# Package = MayaChemTools 7.4
# Release Date = Oct 21, 2010
#
# TimeStamp = Fri Mar 11 14:35:11 2011
#
# FingerprintsStringType = FingerprintsVector
#
# Description = EStateIndicies:ArbitrarySize
# VectorStringFormat = IDsAndValuesString
# VectorValuesType = NumericalValues
#
Cmpd1 11;SaaCH SaasC SaasN SdO SdssC...;24.778 4.387 1.993 25.023 -1...
Cmpd2 9;SdNH SdO SdssC SsCH3 SsNH...;7.418 22.984 -1.583 5.387 5.400...
... ...
... ..
```

Example of CSV *Text* file containing E-state indicies fingerprints string data:

```
"CompoundID","EStateIndiciesFingerprints"
"Cmpd1","FingerprintsVector;EStateIndicies:ArbitrarySize;11;NumericalVa
lues;IDsAndValuesString;SaaCH SaasC SaasN SdO SdssC SsCH3 SsF SsOH SssC
H2 SssNH SsssCH;24.778 4.387 1.993 25.023 -1.435 3.975 14.006 29.759 -0
.073 3.024 -2.270"
"Cmpd2","FingerprintsVector;EStateIndicies:ArbitrarySize;9;NumericalVal
ues;IDsAndValuesString;SdNH SdO SdssC SsCH3 SsNH2 SsOH SssCH2 SssNH Sss
sCH;7.418 22.984 -1.583 5.387 5.400 19.852 1.737 5.624 -3.319"
... ...
... ...
```

The current release of MayaChemTools generates the following types of E-state fingerprints vector strings:

```
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;EStateIndicies:FixedSize;87;OrderedNumericalValues;
ValuesString;0 0 0 0 0 0 0 3.975 0 -0.073 0 0 24.778 -2.270 0 0 -1.435
4.387 0 0 0 0 0 0 3.024 0 0 0 0 0 0 1.993 0 29.759 25.023 0 0 0 0 1
4.006 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0


FingerprintsVector;EStateIndicies:FixedSize;87;OrderedNumericalValues;
IDsAndValuesString;SsLi SssBe SssssBem SsBH2 SssBH SssssB SssssBm SsCH3
SdCH2 SssCH2 StCH SdsCH SaaCH SssssCH SddC StsC SdssC SaasC SaaaC Sssss
C SsNH3p SsNH2 SssNH2p SdNH SssNH SaaNH StN SssssNHp SdsN SaaN SssssN Sd
0 0 0 0 0 0 0 3.975 0 -0.073 0 0 24.778 -2.270 0 0 -1.435 4.387 0 0 0
0 0 0 3.024 0 0 0 0 0 0 0 1.993 0 29.759 25.023 0 0 0 0 14.006 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0...
```

OPTIONS

--AromaticityModel *MDLAromaticityModel* | *TriposAromaticityModel* | *MMFFAromaticityModel* | *ChemAxonBasicAromaticityModel* | *ChemAxonGeneralAromaticityModel* | *DaylightAromaticityModel* | *MayaChemToolsAromaticityModel*

Specify aromaticity model to use during detection of aromaticity. Possible values in the current release are: *MDLAromaticityModel, TriposAromaticityModel, MMFFAromaticityModel, ChemAxonBasicAromaticityModel, ChemAxonGeneralAromaticityModel, DaylightAromaticityModel or MayaChemToolsAromaticityModel*. Default value: *MayaChemToolsAromaticityModel*.

The supported aromaticity model names along with model specific control parameters are defined in AromaticityModelsData.csv, which is distributed with the current release and is available under lib/data directory. Molecule.pm module retrieves data from this file during class instantiation and makes it available to method DetectAromaticity for detecting aromaticity corresponding to a specific model.

--CompoundID *DataFieldName or LabelPrefixString*

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

For *DataField* value of --CompoundIDMode 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 --CompoundIDMode:

```
MolID
ExtReg
```

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

```
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 *SDFile(s)* datafield value; use molname line from *SDFile(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 *SDFile(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 *SDFiles(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 *SDFile(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*.

-e, --EStateAtomTypesSetToUse *ArbitrarySize* | *FixedSize*

E-state atom types set size to use during generation of E-state indicies fingerprints. Possible values: *ArbitrarySize* | *FixedSize*; Default value: *ArbitrarySize*.

ArbitrarySize corrresponds to only E-state atom types detected in molecule; *FixedSize* corresponds to fixed number of previously defined E-state atom types.

For *EStateIndicies*, a fingerprint vector string is generated. The vector string corresponding to *EStateIndicies* contains sum of E-state values for E-state atom types.

Module AtomTypes::EStateAtomTypes.pm is used to assign E-state atom types to non-hydrogen atoms in the molecule which is able to assign atom types to any valid atom group. However, for *FixedSize* value of EStateAtomTypesSetToUse, only a fixed set of E-state atom types corresponding to specific atom groups [Appendix III in Ref 77] are used for fingerprints.

The fixed size E-state atom type set size used during generation of fingerprints contains 87 E-state non-hydrogen atom types in EStateAtomTypes.csv data file distributed with MayaChemTools.

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

Specify whether to check and filter compound data in SDFile(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 E-state 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, --EStateAtomTypesSetToUse option during generation of *EStateIndicies* E-state fingerprints.

E-state atom type IDs appended to --FingerprintsLabel value during *FingerprintsLabelWithIDs* values of --FingerprintsLabelMode correspond to fixed number of previously defined E-state 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: *EStateIndiciesFingerprints*.

-h, --help

Print this help message.

-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.

--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><EStateIndiciesFP>.<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.

--ValuesPrecision *number*

Precision of values for E-state indicies option. Default value: up to *3* decimal places. Valid values: positive integers.

-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 for *EStateIndicies*. Possible values: *ValuesString, IDsAndValuesString, IDsAndValuesPairsString, ValuesAndIDsString, ValuesAndIDsPairsString*.

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

Examples:

```
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
```

-w, --WorkingDir *DirName*

Location of working directory. Default: current directory.

EXAMPLES

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

```
% EStateIndiciesFingerprints.pl -r SampleESFP -o Sample.sdf
```

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

```
% EStateIndiciesFingerprints.pl -e FixedSize -r SampleESFP
  -o Sample.sdf
```

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

```
% EStateIndiciesFingerprints.pl -e FixedSize -v IDsAndValuesString
  -r SampleESFP -o Sample.sdf
```

To generate E-state fingerprints of fixed size in vector string format and create a SampleESFP.csv file containing compound ID from molecule name line along with fingerprints vector strings data, type

```
% EStateIndiciesFingerprints.pl -e FixedSize
  --DataFieldsMode CompoundID --CompoundIDMode MolName
  -r SampleESFP -o Sample.sdf
```

To generate E-state fingerprints of fixed size in vector string format and create a SampleESFP.csv file containing compound IDs using specified data field along with fingerprints vector strings data, type:

```
% EStateIndiciesFingerprints.pl -e FixedSize
  --DataFieldsMode CompoundID --CompoundIDMode DataField --CompoundID
  Mol_ID -r SampleESFP -o Sample.sdf
```

To generate E-state fingerprints of fixed size in vector string format and create a SampleESFP.csv file containing compound ID using combination of molecule name line and an explicit compound prefix along with fingerprints vector strings data, type:

```
% EStateIndiciesFingerprints.pl -e FixedSize
  --DataFieldsMode CompoundID --CompoundIDMode MolnameOrLabelPrefix
  --CompoundID Cmpd --CompoundIDLabel MolID -r SampleESFP -o Sample.sdf
```

To generate E-state fingerprints of fixed size in vector string format and create a SampleESFP.csv file containing specific data fields columns along with fingerprints vector strings data, type:

```
% EStateIndiciesFingerprints.pl -e FixedSize
  --DataFieldsMode Specify --DataFields Mol_ID -r SampleESFP
  -o Sample.sdf
```

To generate E-state fingerprints of fixed size in vector string format and create a SampleESFP.csv file containing common data fields columns along with fingerprints vector strings data, type:

```
% EStateIndiciesFingerprints.pl -e FixedSize
 --DataFieldsMode Common -r SampleESFP -o Sample.sdf
```

To generate E-state fingerprints of fixed size in vector string format and create SampleESFP.sdf, SampleESFP.fpf, and SampleESFP.csv files containing all data fields columns in CSV file along with fingerprints vector strings data, type:

```
% EStateIndiciesFingerprints.pl -e FixedSize
 --DataFieldsMode All --output all -r SampleESFP -o Sample.sdf
```

AUTHOR

Manish Sud <msud@san.rr.com>

SEE ALSO

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

COPYRIGHT