Package 'eiR'

October 9, 2013

Type Package

Title Accelerated similarity searching of small molecules

Version 1.0.3

Date 2013-04-30

Author Kevin Horan

Maintainer Kevin Horan <khoran@cs.ucr.edu>

Suggests RCurl,snow

Description The eiR package provides utilities for accelerated structure similarity searching of very large small molecule data sets using an embedding and indexing approach.

License Artistic-2.0

System Requirements GSL (>=1.14) http://www.gnu.org/software/gsl/

Depends R (>= 2.10.0), ChemmineR (>= 2.11.18), methods

biocViews Infrastructure, DataImport, Clustering, Bioinformatics,Proteomics

Imports snow, tools, snowfall, RUnit, methods,ChemmineR,RCurl,digest,BiocGenerics

R topics documented:

addTransform *Add Transform*

Description

New descriptor types can be added using the addTransform function. These transforms are basically just ways to read descriptors from compound definitions, and to convert descriptors between string and object form. This conversion is required because descriptors are stored as strings in the SQL database, but are used by the rest of the program as objects.

There are two main components that need to be added. The addTransform function takes the name of the transform and two functions, toString, and toObject. These have slightly different meanings depending on the component you are adding. The first component to add is a transform from a chemical compound format, such as SDF, to a descriptor format, such as atom pair (AP), in either string or object form. The toString function should take any kind of chemical compound source, such an SDF file, an SDF object or an SDFset, and output a string representation of the descriptors. Since this function can be written in terms of other functions that will be defined, you can usually accept the default value of this function. The toObject function should take the same kind of input, but output the descriptors as an object. The actual return value is a list containing the names of the compounds (in the names field), and the actual descriptor objects (in the descriptors field).

The second component to add is a transform that converts between string and object representations of descriptors. In this case the toString function takes descriptors in object form and returns a string representation for each. The toObject function performs the inverse operation. It takes descriptors in string form and returns them as objects. The objects returned by this function will be exactly what is handed to the distance function, so you need to make sure that the two match each other.

Usage

```
addTransform(descriptorType, compoundFormat = NULL, toString = NULL, toObject)
```

Arguments

descriptorType	The name of the type of the descritor being added.
compoundFormat	The format of the compound data the descriptor will be extracted from.
toString	See description. If this parameter is NULL and compoundFormat is not NULL, then a default function will be used for this value.
toObject	See description. If compoundFormat is not NULL, then the return value of this function should be a list with the fields "names" and "descriptors", containing the compound names and descriptor objects, respectivly. If compoundFormat is NULL, then the return value should be a collection of descriptor objects, in whatever format the distance function for this descrptor type requires.

Value

No value returned.

Author(s)

Kevin Horan

See Also

setDefaultDistance

Examples

```
# adding support for atompair (ap) descriptors extracted from
# sdf formmatted data.

    #first component
    addTransform("ap","sdf",
        # Any sdf source -> APset
        toObject = function(input,dir="."){
            sdfset=if(is.character(input) && file.exists(input)){
                read.SDFset(input)
            }else if(inherits(input,"SDFset")){
                input
            }else{
           stop(paste("unknown type for 'input', or filename does not exist. type found:",class(input)))
            }
            list(names=sdfid(sdfset),descriptors=sdf2ap(sdfset))
        }
    )

    #second component
    addTransform("ap",
        # APset -> string,
        toString = function(apset,dir="."){
            unlist(lapply(ap(apset), function(x) paste(x,collapse=", ")))
        },
        # string or list -> AP set list
        toObject= function(v,dir="."){
            if(inherits(v,"list") || length(v)==0)
                return(v)

            as( if(!inherits(v,"APset")){
                    names(v)=as.character(1:length(v));
                    read.AP(v,type="ap",isFile=FALSE)
                } else v,
                "list")
        }
    )
```

---

eiAdd *Add new compounds*

Description

Add additional compounds to and existing database

Usage

```
eiAdd(r,d,refIddb,additions,dir=".",format="sdf",
descriptorType="ap",distance=getDefaultDist(descriptorType))
```

Arguments

r	The number of references used to build the database you wish to query agains.
d	The number of dimensions used to build the database you wish to query agains.
refIddb	An Iddb formatted file containing the reference IDs of the database you wish to append to. This should almost always be the file: run-r-d/<long random string>.cdb. The refIddb value should also be returned by eiMakeDb.
additions	The compounds to add. This can be either a file in sdf format, or and SDFset object.
dir	The directory where the "data" directory lives. Defaults to the current directory.
format	The format of the data given in additions. Currently only "sdf" is supported.
descriptorType	The format of the descriptor. Currently supported values are "ap" for atom pair, and "fp" for fingerprint.
distance	The distance function to be used to compute the distance between two descriptors. A default function is provided for "ap" and "fp" descriptors.

Details

New Compounds can be added to an existing database, however, the reference compounds cannot be changed. This will also update the matrix file in the run/job directory with the new compounds.

Author(s)

Kevin Horan

See Also

eiMakeDb eiPerformanceTest eiQuery

Examples

```
library(snow)
r<- 50
d<- 40

#initialize
data(sdfsample)
dir=file.path(tempdir(),"add")
```

```
dir.create(dir)
eiInit(sdfsample[1:99],dir=dir)

#create compound db
refIddb=eiMakeDb(r,d,numSamples=20,dir=dir,
   cl=makeCluster(1,type="SOCK",outfile=""))

#find compounds similar two each query
eiAdd(r,d,refIddb,sdfsample[100],dir=dir)
```

`eiCluster` *Cluster compounds*

Description

Uses Jarvis-Patrick clustering to cluster the compound database using the LSH algorithm to quickly find nearest neighbors.

Usage

```
eiCluster(r,d,K,minNbrs, dir=".",cutoff=NULL,
 descriptorType="ap",distance=getDefaultDist(descriptorType),
  W = 1.39564, M=19,L=10,T=30,type="cluster",linkage="single")
```

Arguments

`r`	The number of references used to build the database you wish to query against.
`d`	The number of dimensions used to build the database you wish to query against.
`K`	The number of neighbors to consider for each compound.
`minNbrs`	The minimum number of neighbors that two comopunds must have in common in order to be joined.
`dir`	The directory where the "data" directory lives. Defaults to the current directory.
`descriptorType`	The format of the descriptor. Currently supported values are "ap" for atom pair, and "fp" for fingerprint.
`distance`	The distance function to be used to compute the distance between two descriptors. A default function is provided for "ap" and "fp" descriptors.
`cutoff`	Distance cutoff value. Compounds having a distance larger this this value will not be included in the nearest neighbor table. Note that this is a distance value, not a similarity value, as is often used in other ChemmineR functions.
`W`	Tunable LSH parameter. See LSHKIT page for details. http://lshkit.sourceforge.net/dd/d2a/mplsh-tune_8cpp.html
`M`	Tunable LSH parameter. See LSHKIT page for details. http://lshkit.sourceforge.net/dd/d2a/mplsh-tune_8cpp.html

L	Number of hash tables
T	Number of probes
type	If "cluster", returns a clustering, else, if "matrix", returns a nearest neighbor matrix.
linkage	Can be one of "single", "average", or "complete", for single linkage, average linkage and complete linkage merge requirements, respectively. In the context of Jarvis-Patrick, average linkage means that at least half of the pairs between the clusters under consideration must pass the merge requirement. Similarly, for complete linkage, all pairs must pass the merge requirement. Single linkage is the normal case for Jarvis-Patrick and just means that at least one pair must meet the requirement.

Details

The jarvis patrick clustering algorithm takes a set of items, a distance function, and two parameters, K, and minNbrs. For each item, it find the K nearest neighbors of that item. Normally this requires computing the distance between every pair of items. However, using Locality Sensative Hashing (LSH), the set of nearst neighbors can be found in near constant time. Once the nearest neighbor matrix is computed, the algorithm makes one pass through the items and merges all pairs that have at least minNbrs neighbors in common.

Although not required, it is avisable to specify a cutoff value. This is the maximum distance two items can have from each other and still be considered to be neighbors. It is thus possible for an item to end up with less than K neighbors if less than K items are close enough to it. If a cutoff is not specified, it is possible for highly un-related items to be listed as neighbors of another item simply because nothing else was nearby. This can lead to items being joined into clusters with which they have no true connection.

Value

If type is "cluster", returns a clustering. This will be a vector in which the names are the compound names, and the values are the cluster labels. Otherwise, if type is "matrix", returns a nearest neighbor matrix. This will be a matrix with a row for each compound. Each row will contain the index value of the neibhboring compounds. If there are not K neibhbors for a compound, that row will be padded with NAs.

Author(s)

Kevin Horan

Examples

```
library(snow)
   r<- 50
   d<- 40

   #initialize
   data(sdfsample)
   dir=file.path(tempdir(),"cluster")
```

```
    dir.create(dir)
    eiInit(sdfsample,dir=dir)

    #create compound db
    eiMakeDb(r,d,numSamples=20,dir=dir, cl=makeCluster(1,type="SOCK",outfile=""))

print(dir())
print(dir(dir))
print(dir(file.path(dir,"data")))
print(dir(file.path(dir,"run-50-40")))
eiCluster(r,d,K=5,minNbrs=2,cutoff=0.5,dir=dir)
```

`eiInit` *Initialize a compound database*

Description

Takes the raw compound database in whatever format the given measure supports and creates a "data" directory.

Usage

```
eiInit(compoundDb,dir=".",format="sdf",descriptorType="ap",append=FALSE)
```

Arguments

`compoundDb`	Either a filename of and SDF file, or an SDFset.
`dir`	The directory where the "data" directory lives. Defaults to the current directory.
`format`	The format of the data in compoundDb. Currenly only "sdf" is supported.
`descriptorType`	The format of the descriptor. Currently supported values are "ap" for atom pair, and "fp" for fingerprint.
`append`	If true the given compounds will be added to an existing database and the <data-dir>/Main.iddb file will be udpated with the new compound id numbers. This should not normally be used directly, use `eiAdd` instead to add new compounds to a database.

Details

EiInit can take either an SDFset, or a filename. SDF is supported by default. It might complain if your SDF file does not follow the SDF specification. If this happens, you can create an SDFset with the `read.SDFset` command and then use that instead of the filename.

EiInit will create a folder called 'data'. Commands should always be executed in the folder containing this directory (ie, the parent directory of "data"), or else specify the location of that directory with the `dir` option.

Value

A directory called "data" will have been created in the current working directory. The generated compound ids of the given compounds will be returned. These can be used to reference a compound or set of compounds in other functions, such as eiQuery.

Author(s)

Kevin Horan

See Also

eiMakeDb eiPerformanceTest eiQuery

Examples

```
data(sdfsample)
dir=file.path(tempdir(),"init")
dir.create(dir)
eiInit(sdfsample,dir=dir)
```

---

eiMakeDb *Create an embedded database*

Description

Uses the initalized compound data to create an embedded compound databbase with r reference compounds in d dimensions.

Usage

```
eiMakeDb(refs,d,descriptorType="ap",distance=getDefaultDist(descriptorType),
dir=".",numSamples=cdbSize(dir)*0.1,
cl=makeCluster(1,type="SOCK"))
```

Arguments

refs	The reference compounds to use to build the database you wish to query against. Refs can be one of three things. It can be a filename of an iddb file giving the index values of the reference compounds to use, it can be vector of index values, or it can be a scalar value giving the number of randomly selected references to use.
d	The number of dimensions used to build the database you wish to query against.
descriptorType	The format of the descriptor. Currently supported values are "ap" for atom pair, and "fp" for fingerprint.
distance	The distance function to be used to compute the distance between two descriptors. A default function is provided for "ap" and "fp" descriptors.

dir	The directory where the "data" directory lives. Defaults to the current directory.
numSamples	The number of non-reference samples to be chosen now to be used later by the eiPerformanceTest function.
cl	A SNOW cluster can be given here to run this function in parrallel.

Details

This function will embedd compounds from the data directory in another space which allows for more efficient searching. The main two parameters are r and d. r is the number of reference compounds to use and d is the dimension of the embedding space. We have found in practice that setting d to around 100 works well. r should be large enough to "represent" the full compound database. Note that an r by r matrix will be constructed during the course of execution, so r should be less than about 46,000 to avoid overflowing an integer. Since this is the longest running step, a SNOW cluster can be provided to parallelize the task.

To help tune these values, `eiMakeDb` will pick `numSamples` non-reference samples which can later be used by the `eiPerformanceTest` function.

`eiMakdDb` does its job in a job folder, named after the number of reference compounds and the number of embedding dimensions. For example, using 300 reference compounds to generate a 100-dimensional embedding (r=300, d=100) will result in a job folder called run-300-100. The embedding result is the file matrix.<r>.<d>. In the above example, the output would be run-300-100/matrix.300.100.

Value

Creates files in `dir` ("run-r-d" by default). The return value is the name of the refIddb file, which needs to be given to other functions such as eiQuery or eiAdd.

Author(s)

Kevin Horan

See Also

`eiInit` `eiPerformanceTest` `eiQuery`

Examples

```
library(snow)

r<- 50
d<- 40

#initialize
data(sdfsample)
dir=file.path(tempdir(),"makedb")
dir.create(dir)
eiInit(sdfsample,dir=dir)

#create compound db
```

```
refIddb=eiMakeDb(r,d,numSamples=20,dir=dir,
   cl=makeCluster(1,type="SOCK",outfile=""))
```

`eiPerformanceTest` *Test the performance of LSH search*

Description

Tests the performance of embedding and LSH.

Usage

```
eiPerformanceTest(r,d,distance=getDefaultDist(descriptorType),descriptorType="ap",
dir=".",K=200, W = 1.39564, M=19,L=10,T=30)
```

Arguments

`r`	The number of references used to build the database you wish to query agains.
`d`	The number of dimensions used to build the database you wish to query agains.
`distance`	The distance function to be used to compute the distance between two descriptors. A default function is provided for "ap" and "fp" descriptors.
`descriptorType`	The format of the descriptor. Currently supported values are "ap" for atom pair, and "fp" for fingerprint.
`dir`	The directory where the "data" directory lives. Defaults to the current directory.
`K`	Number of search results to use for LSH performance test.
`W`	Tunable LSH parameter. See LSHKIT page for details. http://lshkit.sourceforge.net/dd/d2a/mplsh-tune_8cpp.html
`M`	Tunable LSH parameter. See LSHKIT page for details. http://lshkit.sourceforge.net/dd/d2a/mplsh-tune_8cpp.html
`L`	Number of hash tables
`T`	Number of probes

Details

This will perform two different tests. The first tests the embedding results in similarity search. The way this works is by approximating 1,000 random similarity searches (determined by data/test_queries.iddb) by nearest neighbor search using the coordinates from the embedding results. The search results are then compared to the reference search results (chemical-search.results.gz).

The comparison results are summarized in two types of files. The first type lists the recall for different k values, k being the number of numbers to retrieve. These files are named as “recall-ratio-k”. For example, if the recall is 70 compound search - 70 of the 100 results are among the real top-100 compounds - then the value at line 100 is 0.7. Several relaxation ration are used, each generating a file in this form. For instance, recall.ratio-10 is the file listing the recalls when relaxation ratio is 10. The other file, recall.csv, lists recalls of different relaxation ratios in one file

by limiting to selected k value. In this CSV file, the rows correspond to different relaxation ratios, and the columns are different k values. You will be able to pick an appropriate relaxation ratio for the k values you are interested in.

The second test measures the performance of the Locality Sensitive Hash (LSH). The results for lsh-assisted search will be in run-r-d/indexed.performance. It's a 1,000-line files of recall values. Each line corresponds to one test query. LSH search performance is highly sensitive to your LSH parameters (K, W, M, L, T). The default parameters are listed in the man page for `eiPerformanceTest`. When you have your embedding result in a matrix file, you should follow instruction on http://lshkit.sourceforge.net/dd/d2a/mplsh-tune_8cpp.html to find the best values for these parameters.

Value

No value is returned. Creates files in `dir`/run-r-d.

Author(s)

Kevin Horan

See Also

`eiInit` `eiMakeDb` `eiQuery`

Examples

```
library(snow)

r<- 50
d<- 40

#initialize
data(sdfsample)
dir=file.path(tempdir(),"perf")
dir.create(dir)
eiInit(sdfsample,dir=dir)

#create compound db
eiMakeDb(r,d,numSamples=20,dir=dir,
   cl=makeCluster(1,type="SOCK",outfile=""))

eiPerformanceTest(r,d,dir=dir,K=22)
```

---

eiQuery *Perform a query on an embedded database*

Description

Finds similar compounds for each query.

Usage

```
eiQuery(r,d,refIddb,queries,format="sdf",
dir=".",descriptorType="ap",distance=getDefaultDist(descriptorType),
K=200, W = 1.39564, M=19,L=10,T=30)
```

Arguments

r	The number of references used to build the database you wish to query against.
d	The number of dimensions used to build the database you wish to query against.
refIddb	An Iddb formatted file containing the reference IDs of the database you wish to query against. This should almost always be the file: run-r-d/<long random string>.cdb. The refIddb value should also be returned by eiMakeDb.
queries	This can be either an SDFset, or a file containg 1 or more query compounds.
format	The format in which the queries are given. Valid values are: "sdf" when queries is either a filename of an sdf file, or and SDFset object; "compound_id" when queries is a list of id numbers; and "name", when queries is a list of compound names, as returned by cid(apset).
dir	The directory where the "data" directory lives. Defaults to the current directory.
descriptorType	The format of the descriptor. Currently supported values are "ap" for atom pair, and "fp" for fingerprint.
distance	The distance function to be used to compute the distance between two descriptors. A default function is provided for "ap" and "fp" descriptors. The Tanimoto function is used by default.
K	The number of results to return.
W	Tunable LSH parameter. See LSHKIT page for details. http://lshkit.sourceforge.net/dd/d2a/mplsh-tune_8cpp.html
M	Tunable LSH parameter. See LSHKIT page for details. http://lshkit.sourceforge.net/dd/d2a/mplsh-tune_8cpp.html
L	Number of hash tables
T	Number of probes

Details

This function identifies the database by the r, d, and refIddb parameters. The queries can be given in a few different formats, see the queries parameter for details. The LSH algorithm is used to quickly identify compounds similar to the queries.

Value

Returns a data frame with columns 'query', 'target', 'target_ids', and 'distance'. 'query' and 'target' are the compound names and distance is the distance between them, as computed by the given distance function.'target_ids' is the compound id of the target. Query namess are repeated for each matching target found.

Author(s)

Kevin Horan

See Also

eiInit eiMakeDb eiPerformanceTest

Examples

```
library(snow)
r<- 50
d<- 40

#initialize
data(sdfsample)
dir=file.path(tempdir(),"query")
dir.create(dir)
eiInit(sdfsample,dir=dir)

#create compound db
refIddb=eiMakeDb(r,d,numSamples=20,dir=dir,
   cl=makeCluster(1,type="SOCK",outfile=""))

#find compounds similar two each query
results = eiQuery(r,d,refIddb,sdfsample[1:2],K=15,dir=dir)
```

---

example_compounds *Example Compounds*

Description

122 compounds in SDF format, stored as a list. Each element of the list is one line of text. This is just used in some unit tests.

Format

The format is: chr [1:12222] "3540" " OpenBabel06051210572D" "" ...

`setDefaultDistance` *Set the default distance function for a descriptor type*

Description

Set the default distance function for a descriptor type. This is the distance function that will be used if none is given for a particular function call.

Usage

```
setDefaultDistance(descriptorType, distance)
```

Arguments

`descriptorType`	The type of the descriptor to set a distance function for. Built-in values are "ap" and "fp". Additional values can be set as well.
`distance`	A distance function taking two descriptor objects (as returned by toObject in a descriptor transform, see `\ ink{addTransform}` for details), and returning a distance value.

Value

No return value.

Author(s)

Kevin Horan

See Also

`addTransform`

Examples

```
setDefaultDistance("ap", function(d1,d2) 1-cmp.similarity(d1,d2) )
```

Index