# Package 'ENGEP'

September 20, 2023

**Title** Predicting expression levels of spatially unmeasured genes in a query spatial dataset based on ensemble learning

**Version** 1.0

**Description**
ENGEP is an ensemble learning tool for spatially unmeasured genes expression prediction.
It integrates the results of different reference datasets and prediction methods,
instead of relying on a single reference dataset or method.

**License** GPL (>= 2)

**Encoding** UTF-8

**LazyData** true

**Roxygen** list(markdown = TRUE)

**RoxygenNote** 7.1.1

**Imports** caret,
Matrix,
parallel,
propr,
proxy,
proxyC,
Seurat,
stats

**Depends** R (>= 2.10)

## R topics documented:

compute_simi *compute_simi*

## Description

Function to caculate similarity matrix between spatial and reference data. Distance measures (such as Manhattan distance, Canberra distance, Euclidean distance, and phi_s proportionality measure) are transformed into similarity scores using $s = 1/(1+d)$. The codes are modified from the calculateSimilarity function scclassify.

## Usage

```
compute_simi(spcom, sccom, similarity)
```

## Arguments

| | |
|---|---|
| spcom | common gene expression matrix of spatial data. |
| sccom | common gene expression matrix of reference data. |
| similarity | a vector indicates the ten similarity measure that are used. |

## Value

similarity matrix between spatial and reference data.

engep_predict *engep_predict*

## Description

Function to get expression levels of unmeasured genes predicted by ENGEP.

## Usage

```
engep_predict(
  spa_counts,
  ref_list,
  pre_genes,
  nCpus = 6,
  simi_list = NULL,
  n0 = NULL,
  parallel = TRUE,
  k_list = NULL,
  get_baes = FALSE,
  get_weight = FALSE
)
```

### Arguments

| | |
|---|---|
| spa_counts | gene expression matrix of spatial dataset (gene by cell). |
| ref_list | each element in the list is a gene expression matrix (gene by cell) of an original single reference dataset. |
| pre_genes | an array contains names of genes to be predicted, if is NULL, ENGEP will predict the intersection of unique genes of each references. |
| nCpus | number of (maximum) cores to use for parallel execution, default to 6. |
| simi_list | a vector indicates the ten similarity measures that are used, note that the similarity measures used should be chosen from the ten similarity measures mentioned in the document. Default is c("pearson", "spearman","cosine", "jaccard","weighted_rank", "manhattan","canberra","euclidean", "phi_s","rho_p") |
| n0 | the number of cells of sub-reference dataset, default is 8000. |
| parallel | a logical value to indicate if the process should be run parallelly in multiple threads, default to TURE. |
| k_list | a list contains different values of $k$ (number of neighbors in knn), default is (20,30,40,50). |
| get_baes | a logical value to indicate whethe to return base results or not, default is FAISE. |
| get_weight | a logical value to indicate whethe to return weights of different references, default is FAISE. |

### Value

the predicted expression levels of unmeasured genes.

---

## ensemble_result *ensemble_result*

---

### Description

Function to combine base results by a weighted average ensemble method.

### Usage

```
ensemble_result(result_single, get_weight = FALSE)
```

### Arguments

| | |
|---|---|
| result_single | a list contains base results predicted by different reference datasets, different similarity measures and different values of k. |
| get_weight | a logical value to indicate whethe to return weights of different references, default is FAISE. |

### Value

ensembel result.

gene_dataliat *gene_dataliat*

## Description

Function to randomly partition large reference into small reference. equal-sized sub-datasets.

## Usage

```
gene_dataliat(spa_counts, ref_list, pre_genes, n0)
```

## Arguments

| | |
|---|---|
| spa_counts | gene expression matrix of spatial dataset (gene by cell). |
| ref_list | each element in the list is a gene expression matrix (gene by cell) of an original single reference dataset. |
| pre_genes | an array contains names of genes to be predicted, if is NULL, ENGEP will predict the intersection of unique genes of each references. If you let pre_genes = NULL, we suggest you to use reference datasets with high variable genes. |
| n0 | the number of cells of sub-reference dataset, default is 8000 |

## Value

a list contains equal-sized reference sub-datasets with common genes and predicted genes, spatial dataset with common genes, and unique genes in references.

imp_new *imp_new*

## Description

Function to generate base results by using different reference datasets, different similarity measures and different values of k.

## Usage

```
imp_new(i, spcom, sccomlist, similist, k.list, sc_implist)
```

## Arguments

| | |
|---|---|
| i | each time one reference expression matrix in the list is used to predict the expression. |
| spcom | expression matrix of spatial data with common genes. |
| sccomlist | a list contains expression matrices of different reference datasets with common genes. |
| similist | a vector indicates the ten similarity measure that are used. |
| k.list | a list contains different values of $k$ (number of neighbors in knn), default is (20,30,40,50). |
| sc_implist | a list contains expression matrices of different reference datasets with genes to be predicted. |

## Value

the predicted expression of unmeasured genes and the R-squared sore between the predicted expression and the measured expression of training genes representing predictive power of reference.

---

`weight_r2` *weight_r2*

---

## Description

Function to convert the predictive power to weight, with a range of values between 0.1 and 0.9.

## Usage

```
weight_r2(r2_vec)
```

## Arguments

| | |
|---|---|
| `r2_vec` | R-squared sore which presents the predictive power of reference data. |

## Value

weight converted from R-squared sore.

# Index