Package 'PEAm5C'

January 5, 2018

Type Package

Title An integrated R toolkit for plant m5C analysis.

Version 0.1.1

Author Jie Song, Jingjing Zhai amd Chuang Ma

Maintainer Jie Song <q2516581@gmail.com>

Depends R (>= 3.3.2), seqinr (>= 3.3-6), stringr (>= 1.2.0), randomForest (>= 4.6-12),ggplot2 (>= 2.2.1)

Imports pROC (>= 1.5.4), PRROC (>= 1.0-5),FSelector (>= 0.21),e1071 (>= 1.6-8)

Description PEAm5C is an integrated R toolkit designed to identify plants' m5C sites. The toolkit contains functions such as analysis sequence characteristics, m5C prediction from transcriptional group size, and custom build recognition model.

License GPL (>= 2)

Encoding UTF-8

LazyData true

URL http://bioinfo.nwafu.edu.cn/software

NeedsCompilation no

R topics documented:

cvgroup *Cross validation grouping*

Description

Grouping data sets by cross validation

Usage

```
cvgroup(data,cvnum,seed=1234)
```

Arguments

data	A numeric vector, the number of samples (such as 1:1000)
cvnum	A numeric value, the number of Cross validation group (k-fold)
seed	Arguments passed to set.seed()

Author(s)

Jingjing Zhai, Chuang Ma, Jie Song.

Examples

```
cvgroup(1:100,5)
```

extra_model *extracting model from PEA_ml*

Description

Proposing the desired model from the established result information

Usage

```
extra_model(res,ignum)
```

Arguments

res	a list , result obtained from PEA_ml
ignum	a integer number,which indicates the number of features

Value

models and selected feature

Examples

```
load(paste0(system.file(package = "PEAm5c"),"/data/samples.Rds"))
aaa <- PEA_ml(pos_sample = pos_sample,neg_sample = neg_sample)
ddd <- extra_model(res = aaa,ignum=150)
ddd
```

`extra_motif_seq` *Scanning the motifs in the transcripts.*

Description

For a given motif, all transcript sequences will be scanned and then according to the size of "upstream" "downstream", the output fits the sequence of length.

Usage

```
extra_motif_seq(input_seq_dir,text='c',up=5,end=5)
```

Arguments

`input_seq_dir`	A path character, representing the file name of the sequence in FASTA format.
`text`	A string character, which specifies the motif to be searched.
`up`	A integer number, the length of the upstream sequence required.
`end`	A integer number, the length of the downstream sequence required.

Value

A list of sequences around motif.

Author(s)

Jie Song, Jingjing Zhai, Chuang Ma

Examples

```
aaa <- extra_motif_seq(input_seq_dir = paste0(system.file(package = "PEAm5c"),"/data/cdna.fa"),up = 5)
aaa
```

FeatureExtract *Feature encoding*

Description

This function contains three feature encoding scheme, binary, k-mer and PseDNC. For binary encoding scheme, a vector of 404 (4*101) features is generated through assigning 'A', 'C', 'G', 'U' and 'N' with (1,0,0,0), (0,1,0,0), (0,0,1,0), (0,0,0,1) and (0,0,0,0), respectively. Here 'N' is a gap used to ensure the fixed features of each sample, if an m6A/non- m6A site occurs near the initiation or termination of the transcript. For K-mer encoding, the composition of short sequence with different lengths was considered to encoding samples. For PseDNC (pseudo dinucleotide composition) encoding, the local and global sequence-order information along the RNA sequence was used for scoring the each sample.

Usage

```
FeatureExtract(RNAseq, lambda = 6, w = 0.9)
```

Arguments

RNAseq	A list containing the FASTA format sequences.
lambda	The lambda parameter for the PseDNC-related features, default is 6.
w	The weighting parameter for PseDNC-related features, default is 0.9.

Value

A matrix with features.

Author(s)

Jie Song, Jingjing Zhai, Chuang Ma

Examples

```
aaa <- extra_motif_seq(input_seq_dir = paste0(system.file(package = "PEAm5c"),"/data/cdna.fa"),up = 5)
aaa <- lapply(aaa, c2s)
bbb <- FeatureExtract(aaa)
bbb[1:10,]
```

PEA_ml *Machine learning-based Random Forest algorithm and information gain rank for sequences*

Description

First of all, the function separate the positive and negative sample to the independent test set automatically according to setting and training set, concentrated training shall be carried out in accordance with the cross validation in training and assessment, in the results returned to characteristics of the evaluation, prediction score and evaluation as well as the sample group. In this case, the function can define the proportion of positive and negative samples (when negative sample size can be selected to maintain multiple models), and provide two methods of random forest and support vector machine.

Usage

```
PEA_ml(pos_sample,neg_sample,independent_num=100,ig="ALL",
                ratio = 1,modeltype = "RFC",cvnum = 5,repeatTimes = 1, ntree=200,over_sampling = F)
```

Arguments

pos_sample	A numeric matrix recording the features for positive sample.
neg_sample	A numeric matrix recording the features for nagative sample.
independent_num	A numeric value, the number of independent sample
ig	A numeric value,vector or "ALL", the number of selected features based the top of information gain rank, the "ALL" means all features
modeltype	A character string, which specifies machine learing method.
cvnum	An integer value, the number of fold for cross validation.
repeatTimes	An integer value,If the negative sample is larger than the limit of the positive sample, the number of the negative samples and the number of samples of the positive sample is repeated
over_sampling	Logical value, where TRUE represents balance the positive and negative samples according to the ratio based smote simulation
ratio	A numeric value, where 1 represents balance the positive and negative sample.

Value

A list of result.

The first level is used feature num group.

The second level is cross validation group.

The third level is the detail information including positives.test.score.id, negatives.test.score.id, positives.test.score,negatives.test.score, positives.test, negatives.test, auc_test, auc_test_id

Author(s)

Jie Song, Jingjing Zhai, Chuang Ma

Examples

```
load(paste0(system.file(package = "PEAm5c"),"/data/samples.Rds"))
aaa <- PEA_ml(pos_sample = pos_sample,neg_sample = neg_sample)
aaa
```

`predict_m5c` *Predicting m5C sites by PEA-m5C*

Description

Predicting m5C information of the sequences by the proposed m5C models based Random Forest.

Usage

```
predict_m5c(sample_feature)
```

Arguments

`sample_feature` A dataframe or list for target sequence.

Value

A matrix with transcript , its sites, predicted score and m5C level.

Author(s)

Jie Song, Jingjing Zhai, Chuang Ma

Examples

```
aaa <- extra_motif_seq(input_seq_dir = paste0(system.file(package = "PEAm5c"),"/data/cdna.fa"),up = 5)
aaa <- lapply(aaa, c2s)
bbb <- FeatureExtract(aaa)
ccc <- predict_m5c(bbb)
```

predict_self_model *Predict CMRs sites by user-specific models*

Description

The CMRs information of the detection sequences was evaluated by the user-specific models.

Usage

```
predict_self_model(models,sequence_dir,end = 5,up = 5)
```

Arguments

models	A dataframe self models and selected feature based extra_model.
sequence_dir	A path representing the filename of the sequence in FASTA format.
up	A integer number, the length of the upstream sequence required.
end	A integer number, the length of the downstream sequence required.

Examples

```
load(paste0(system.file(package = "PEAm5c"),"/data/samples.Rds"))
aaa <- PEA_ml(pos_sample = pos_sample,neg_sample = neg_sample)
ddd <- extra_model(res = aaa)
ddd
#
eee <- predict_self_model(models = ddd,sequence_dir = paste0(system.file(package = "PEAm5c"),"/data/cdna.fa"))
table(eee[,4])
```

Index