Package 'PEAm5C'

January 6, 2018

Type Package

Title An integrated R toolkit for plant m5C analysis.

Version 0.1.3

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 PEA-m5C is designed for transcriptome-wide m5C annotation. This package contains a set of functions such as feature encoding, cross validation, transcriptome-wide m5C annotation and user-specific model training.

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 datasets for cross validation.

Usage

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

Arguments

data	A string vector representing the names of training samples or a string vector recording the names of training samples.
cvnum	which specifies the number of fold in cross-validation.
seed	A numeric value used for generating random seed.

Author(s)

Jingjing Zhai, Chuang Ma, Jie Song.

Examples

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

extra_model *Extracting model from PEA_ml*

Description

Extracting the user-specific model from function "PEA_ml"

Usage

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

Arguments

res	A list, which can be obtained from function "PEA_ml"
ignum	a integer number,which indicates the number of features

Value

models and selected features

Examples

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

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

Description

For a given motif, all transcript sequences will be scanned according to the user-specific flanking sequence length (upstream/downstream), then the fixed length of sequences centered on motif will be returned.

Usage

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

Arguments

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

Value

A list of sequences around motif.

Author(s)

Jie Song, Jingjing Zhai, Chuang Ma

Examples

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

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 *Transcriptome-wide m5C predictor training under machine learning framework.*

Description

This function used for transcriptome-wide m5C predictor construction. First, the fixed number (parameter "independent_num") of independent samples (positive and negative samples) are randomly sampled from training samples. The the k-fold cross-validation would be performed based on the training samples but excluding the independent samples. Finally, the m5C predictor, performance evaluation on independent test datasets and cross-validation results will be returned.

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
feature_num	A numeric value, 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.
...	optional parameters to be passed to the low level function randomForest.default

Value

An object of result, which is a list with the components including cross validation detailed information , feature weights, and the AUC value of independent set.

Author(s)

Jie Song, Jingjing Zhai, Chuang Ma

Examples

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

---

predict_m5c *Predicting m5C sites by PEA-m5C*

Description

Predicting m5C modification sites based on PEA-m5C. PEA-m5C provide four threshold (VH-mode: very high confidence with specificity of 99 precent; HMode: high confidence mode with specificity of 95 precent; NMode: normal confidence with specificity of 90 precent; LMode: low confidence mode with specificity of 85 precent) to meet different requirements.

Usage

```
predict_m5c(sample_feature,mode=NULL)
```

Arguments

sample_feature	A dataframe or list of RNA sequence.
mode	A string character of "VH","H","N","L",it means different mode.

Value

A matrix with 4 columns including transcripts ID, candidate m5C position, probabilistic scores for being m5C and which mode of m5C.

Author(s)

Jie Song, Jingjing Zhai, Chuang Ma

Examples

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

predict_self_model *Predicting CMRs sites by user-specific model*

Description

Predicting the m5C through user-specific squences.

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"))
seq <- PEA_ml(pos_sample = pos_sample,neg_sample = neg_sample)
model <- extra_model(res = seq)
model
#
res <- predict_self_model(models = model,sequence_dir = paste0(system.file(package = "PEAm5c"),"/data/cdna.fa"))
table(res[,4])
```

Index