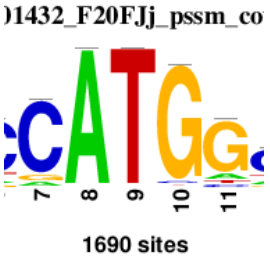
Mar 31, 2017

Genome-wide Kozak Sequence Over-represented Motif Analysis



DOI
dx.doi.org/10.17504/protocols.io.hikb4cw

Mariana Rius[1], Joshua Rest[1]

[1]Stony Brook University

Protist Research to Opti... Collier Lab

Mariana Rius
Stony B

OPEN ACCESS

DOI: dx.doi.org/10.17504/protocols.io.hikb4cw

Protocol Citation: Mariana Rius, Joshua Rest 2017. Genome-wide Kozak Sequence Over-represented Motif Analysis. **protocols.io https://dx.doi.org/10.17504/protocols.io.hikb4cw**

Protocol status: Working

Created: March 31, 2017

Last Modified: March 08, 2018

Protocol Integer ID: 5420

Abstract

Bioinformatic approach to identifying over-represented motifs in the region framing the start codon (25 bp up and downstream) for genes annotated in the three sequenced Labyrinthulomycete genomes (*Aurantiochytrium limacinum, Schizochytrium aggregatum, and Aplanochytrium kergulense*).

Download gene annotation (gff) file and fasta file for species of interest

1 *Schizochytrium aggregatum*
Schag1_GeneCatalog_genes_20121220.gff
Schag1_AssemblyScaffolds.fasta from
http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=Schag1

Aurantiochytrium limacinum
Aurli1_GeneCatalog_genes_20120618.gff
Aurli1_AssemblyScaffolds.fasta from
http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=Aurli1

Aplanchytrium kergulense
Aplke1_GeneCatalog_genes_20121220.gff
Aplke1_AssemblyScaffolds.fasta from
http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=Aplke

Note

Using R version 3.3.2 and the following packages:
doBy (doBy_4.5-15)
data.table (data.table_1.10.0)
seqinr (seqinr_3.3-3)

Command

Create working gene catalog for organism of interest. Schizochytrium aggregatum (Schag1) code provided herein as an example. (R 3.3.2)

```
ShGeneCat <- read.delim(
```

OPTIONAL: Create .rda file to facilitate access to annotations

2 Create subset of annotation file.

Command

Example of ShGeneCat. (R 3.3.2)

```
colnames(ShGeneCat) <- c(
```

Identify the coordinates of 25 base pairs up and downstream of all annotated coding start sites

3 Retain only genes with a protein ID

Command

```
ShGeneCat <- ShGeneCat[!(is.na(ShGeneCat$PID)),]
```

4 Identify species and term

Command

```
term <-
```

5 Create new destination for identified coordinates

Command

```
ShGeneWg <- ShGeneCat[]
```

6 Write table with coordinates of region of interest for each gene. Here 25 bases up and downstream were isolated as region of interest.

Command

```
promC <- do.call(
```

7 Change any negative start sites to 1

Command

```
promC[promC[,'start'] < 1, 'start'] <- 1
write.table(promC,file=paste(species,term,
```

Create FASTA file containing region of interest

8 Using FASTA files previously downloaded:

Schag1_AssemblyScaffolds.fasta
Aurli1_AssemblyScaffolds.fasta
Aplke1_AssemblyScaffolds.fasta

Run bedtools command to retrieve sequence data.

Command

bedtools 2.15.0

```
bedtools getfasta -s -fi Schag1_AssemblyScaffolds.fasta -bed
Sh.wg.promC.gff -fo Sh.wg.promC.fasta -name
```

9 Use bioawk to discard any sequences not containing an 'ATG' as the start codon.

Command

```
bioawk -c fastx 'substr($seq,26,3) ~ /ATG/ { print
```

Use RSATprotist to identify over-represented motifs in sequences

10 Use RSATprotist online in the web interface
http://rsat01.biologie.ens.fr/rsa-tools/

Input FASTA file:
Sh.wg.promC.ATG26.fasta

1 - Choose your type of data to analyse
ChIP-seq
List of gene names
Sequences
Matrices (PSSM)
Coordinates (BED)
List of variants

2 - Choose your biological question/ analysis to perform
Are there over-represented motifs in these sequences?
I want to scan these sequences with a motif

3 - Relevant RSAT programs
oligo-analysis (words)
dyad-analysis (spaced pairs)

http://rsat01.biologie.ens.fr/rsa-tools/