Series 2

Genomics and bioinformatics - Week 2

September 27, 2011

1 Description

In today's session you will retrieve publicly available genome sequence and annotation data for a particular species and use it to extract some biological information about that species.

2 Downloading the genome data

From the UCSC Genome Browser select the latest assembly of the *Mus musculus* genome.
Scroll down to "Assembly details" and click on the "Downloads" link.
From "Data set by chromosome" download the `.fa.gz` file for any chromosome.
From "Annotation database" download the `knownGene.txt.gz` and `kgAlias.txt.gz` files.

3 Programming exercise

3.1 Using Python

1. Read the `.fa` file and extract the chromosome sequence
2. Determine the length of the sequence
3. Calculate the number of As, Gs, Cs and Ts in the sequence
4. Compute the GC-content of the chromosome
5. Plot GC content vs chromosome length (use `matplotlib`)

Questions

1. What does the GC-content tell us about a genome?
2. Using data for all chromosomes, calculate the average GC-content of the mouse genome. Does this compare to the value mentioned in your lecture slides?

3.2 Using R

1. Read the `knownGene.txt` and `kgAlias.txt` files
2. Extract the annotation corresponding to your chromosome (`chrom` column) from `knownGene.txt`
3. Extract a list all the exons in the chromosome (`exonStarts`, `exonEnds` columns)

4. Compute exon sizes. Create a table as shown below,

geneID	exonNum	exonStart	exonEnd	exonSize
uc007aet.1	1	3195984	3197398	1415
uc007aet.1	2	3203519	3205713	2195
uc007aeu.1	1	3204562	3207049	2488
uc007aeu.1	2	3411782	3411982	201
uc007aeu.1	3	3660632	3661579	948

5. Plot the exon size distribution

Note

The `knownGene.txt` and `kgAlias.txt` files are both tab-delimited files.
Column headers for the `knownGene.txt` file are,
`geneID chrom strand txStart txEnd cdsStart cdsEnd exonCount exonStarts exonEnds proteinID alignID`
Column headers for the `kgAlias.txt` file are,
`geneID geneName`

Questions

1. List names of genes with,

   a) the longest exon, b) the shortest exon and c) most number of exons.

2. List all the intronless genes in the chromosome.

3. What can you tell about the exon size distribution across different mouse chromosomes?

Note

Cross `knownGene.txt` and `kgAlias.txt` to obtain `geneNames` corresponding to each `geneID`.

geneID	geneName
uc007aet.1	AK135172, mKIAA1889, uc007aet.1
uc007aeu.1	NM_001011874, NP_001011874, Q5GH67, XKR4_MOUSE, Xkr4, Xrg4, uc007aeu.1

3.3 Reference documentation

For R - `http://cran.r-project.org/doc/manuals/refman.pdf`
For Python - `http://docs.python.org/tutorial/`

If you need help:

1. Go through last week's exercise session for examples

2. Use Google

3. Use the ? or help() with R commands

4. Ask us