(BioC2011) Biostrings lab - Exercises

Hervé Pagès*

July 28, 2011

Exercise 1

a. *Generate a random DNAString of length 2000. (You will need:* `sample()`, `DNA_BASES`, `paste()`, `DNAString()`.*)*

b. *Create views on it.*

c. *Invert the views.*

d. *Count the frequencies of the DNA letters: (a) in the DNAString object, (b) inside the views, (c) outside the views. Do a sanity check.*

Exercise 2

a. *Load Affymetrix hgu95av2 probe sequences into a DNAStringSet object.*

b. *Remove the first 10 probes.*

c. *Which probes contain more than 16 A's?*

d. *Reverse complement the probes.*

e. *Trim the first (5') and last (3') two bases.*

f. *Generate the sequences of the mismatch probes (MM probes) by replacing the middle nucleotide of each PM probe by its reverse complement.*

g. *Which probes contain more than 9 consecutive A's? (You can use* `vcountPattern()` *for this.) Display their sequences.*

Exercise 3

a. *Load BSgenome data package for hg19.*

b. *Count the number of times each Affymetrix hgu95av2 probe hits Human chr22. (You will need:* `PDict()`, `countPDict()`.*)*

c. *Which probes have more than 2000 hits? Display their sequences.*

*Fred Hutchinson Cancer Research Center, Seattle, WA 98008

Exercise 4

The goal of this exercise is to count the nb of times each Human transcript is hit by a hgu95av2 probe. We use the TxDb.Hsapiens.UCSC.hg19.knownGene package for the locations of the transcripts and their exons.

a. *Extract the Human transcriptome with* `extractTranscriptsFromGenome()` *(defined in the GenomicFeatures package).*

b. *Use* `vcountPDict()` *(with* `'collapse=2'`*) to count the nb of hits per transcript.*