Introduction to Scientific Computing for Biologists

ISCB20.09 - R for Bioinformatics

An Introduction to `seqinr` and `Biconductor`

Md. Jubayer Hossain
https://jhossain.me/
jubayer@hdrobd.org

Founder
Health Data Research Organization
Lead Organizer
Scientific Computing for Biologists

11 February 2021

IRanges

```r
# Load IRanges
library(IRanges)
```

```
Loading required package: BiocGenerics

Loading required package: parallel


Attaching package: ’BiocGenerics’

The following objects are masked from ’package:parallel’:

    clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
    clusterExport, clusterMap, parApply, parCapply, parLapply,
    parLapplyLB, parRapply, parSapply, parSapplyLB

The following objects are masked from ’package:stats’:
```

IRanges

A range is defined by start and end.

```r
# IRanges with numeric arguments.
ir1 <- IRanges(start = 20, end = 30)
ir1
```

```
IRanges object with 1 range and 0 metadata columns:
          start       end     width
      <integer> <integer> <integer>
  [1]        20        30        11
```

IRanges

```r
# IRanges with numeric arguments.
ir2 <- IRanges(start = c(1, 3, 5), end = c(7, 9, 11))
ir2
```

```
IRanges object with 3 ranges and 0 metadata columns:
          start       end     width
      <integer> <integer> <integer>
  [1]         1         7         7
  [2]         3         9         7
  [3]         5        11         7
```

Equation: width = end - start + 1

IRanges

```r
# IRanges with logical vector
ir3 <- IRanges(start = c(TRUE, FALSE, T, F))
ir3
```

```
IRanges object with 2 ranges and 0 metadata columns:
          start       end     width
      <integer> <integer> <integer>
  [1]         1         1         1
  [2]         3         3         1
```

Rle - run length encoding

- Rle stands for Run length encoding
- Computes and stores the lengths and values of a vector or factor.
- Rle is general S4 container used to save long repetitive vectors efficiently.

```r
num <- c(3, 2, 1, 5, 6, 7, 8)
Rle(num)
```

```
numeric-Rle of length 7 with 7 runs
  Lengths: 1 1 1 1 1 1 1
  Values : 3 2 1 5 6 7 8
```

GRanges

- A GRanges is a data structure for storing genomic intervals.
- They are fast and efficient.

```r
# Load GenomicRanges
library(GenomicRanges)
```

```
Loading required package: GenomeInfoDb
```

GRanges

```
# Create GRanges object
gr1 <- GRanges("chr1:200-300")
gr1
```

```
GRanges object with 1 range and 0 metadata columns:
      seqnames    ranges strand
         <Rle> <IRanges>  <Rle>
  [1]     chr1   200-300      *
  -------
  seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

- GRanges class is a container to save genomic intervals by chromosome.
- Minimum arguments `chr1:200-300`
- GRanges `seqnames` and `seqinfo`.

GRanges

```r
gr2 <- GRanges(seqnames = "chr1",
               strand = c("+", "-", "+"),
        ranges = IRanges(start = c(1, 3, 5), width = 3))
gr2
```

```
GRanges object with 3 ranges and 0 metadata columns:
      seqnames    ranges strand
         <Rle> <IRanges>  <Rle>
  [1]     chr1       1-3      +
  [2]     chr1       3-5      -
  [3]     chr1       5-7      +
  -------
  seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

Patterns Finding

- Gene start
- Protein end
- Regions that enhance or silence gene expression
- Conserved regions between organisms
- Genetic variation

Pattern Matching

- `matchPattern(pattern, subject)`
  - 1 string to 1 string
- `vmatchPattern(pattern, subject)`
  - 1 set of strings to 1 string
  - 1 string to a set of strings
- `findPalindromes()` - Find palindromic regions in a single sequence

Introduction to ShortRead

ShortRead: a Bioconductor package for input, quality assessment and exploration of high-throughput sequence data.

Sequence Data Handling with `ShortRead`

- Reading and Writing `FASTA` File
- Reading and Writing `FASTQ` File
- `FASTQ` Sampling

Quality Control(QC) with `ShortRead`

- Quality scores - Phred table
- Encoding - Phred +33
- `fastq` quality
- Quality Assessment

Match and Filter

- Duplicate sequences
  - Biological sequence duplicates occur in nature.
  - Amplification from the steps in library preparation (PCR)
  - Sequencing the sample more than once
- Remove Duplicates
  - Whole genome sequencing or exome sequencing
  - Mark duplicates using a threshold.
  - RNA-seq and ChIP-seq