Seminar III: R/Bioconductor

August-December 2009

Bachelor in Genomic Sciences LCG
UNAM - Cuernavaca - Mexico

Biostrings

Isaac F. López Moyado

Electronic mail:
ilopez@lcg.unam.mx

September 17, 2009

Abstract

Here I show an introduction to the package Biostrings from Bioconductor.

1 Author

Author H. Pagès, P. Aboyoun, R. Gentleman, and S. DebRoy

Maintainer Hervé Pagès

2 What is Biostrings?

As described on the Bioconductor page:

> Memory efficient **string containers** **string matching algorithms** and other utilities, for fast **manipulation of large biological sequences** or sets of sequences.

3 What is it used for?

Some of its uses include[1]:

- Pairwise Sequence Alignment Functions
- Evolutionary Models in Protein Alignments
- Removing Adapters from Sequence Reads
- Quality Assurance in Sequencing Experiments
- Computation Profiling
- Computing alignment consensus matrices

4 Relations with other packages

Depends

R , methods , IRanges

Imports

methods , utils , IRanges , Biobase

Depends On Biostrings

BSgenome , BiostringsCinterfaceDemo , ChIPpeakAnno , GGtools , GeneRegionScan , ShortRead , altcdfenvs , matchprobes , microRNA

Imports Biostrings

AffyCompatible , BCRANK , BiostringsCinterfaceDemo , ChIPpeakAnno , GeneRegionScan , MEDME , Rolexa , ShortRead , biocDatasets , gcrma , oligoClasses , pdInfoBuilder , rtracklayer

Suggests Biostrings

SLGI , annotate , oneChannelGUI

[1] You can find more information here

5 Examples

Here are some things you can do with Biostrings. You can find advanced examples here:

```
> library(Biostrings)
```

1. Forensic example (for more information go to this page)

```
> library("BSgenome.Hsapiens.UCSC.hg18")

> Hsapiens

> chr18NoN <- mask(Hsapiens$chr18, "N")
> alphabetFrequency(Hsapiens$chr18, as.prob = TRUE)["N"]

N
0

> matchPattern("GAGCCATGTTCATGCCACTG", chr18NoN)

  Views on a 76117153-letter DNAString subject
subject: CCCTAACCCTAACCCTAACCCTTACCCCTAACCC...GGTCTCTTGCCTCGGCAAAGATTAGATTAGGG
views:
       start      end width
[1] 59099824 59099843    20 [GAGCCATGTTCATGCCACTG]
[2] 65528339 65528358    20 [GAGCCATGTTCATGCCACTG]
[3] 72568199 72568218    20 [GAGCCATGTTCATGCCACTG]
[4] 74769361 74769380    20 [GAGCCATGTTCATGCCACTG]

> xsw <- reverseComplement(DNAString("CAAACCCGACTACCAGCAAC"))
> matchPattern(xsw, chr18NoN)

  Views on a 76117153-letter DNAString subject
subject: CCCTAACCCTAACCCTAACCCTTACCCCTAACCC...GGTCTCTTGCCTCGGCAAAGATTAGATTAGGG
views:
       start      end width
[1] 59100110 59100129    20 [GTTGCTGGTAGTCGGGTTTG]

> GAAA <- paste(rep("GAAA", 21), collapse = "")
> mT <- matchPattern(GAAA, chr18NoN)
> countPattern(GAAA, chr18NoN)
```

```
[1] 7

> length(mT)

[1] 7

> mT

  Views on a 76117153-letter DNAString subject
subject: CCCTAACCCTAACCCTAACCCTTACCCCTAACCC...GGTCTCTTGCCTCGGCAAAGATTAGATTAGGG
views:
       start      end width
[1]  2604564  2604647    84 [GAAAGAAAGAAAGAAAGAAAGAAA...AAAGAAAGAAAGAAAGAAAGAA
[2]  2604568  2604651    84 [GAAAGAAAGAAAGAAAGAAAGAAA...AAAGAAAGAAAGAAAGAAAGAA
[3]  2604572  2604655    84 [GAAAGAAAGAAAGAAAGAAAGAAA...AAAGAAAGAAAGAAAGAAAGAA
[4] 49915245 49915328    84 [GAAAGAAAGAAAGAAAGAAAGAAA...AAAGAAAGAAAGAAAGAAAGAA
[5] 49915249 49915332    84 [GAAAGAAAGAAAGAAAGAAAGAAA...AAAGAAAGAAAGAAAGAAAGAA
[6] 49915253 49915336    84 [GAAAGAAAGAAAGAAAGAAAGAAA...AAAGAAAGAAAGAAAGAAAGAA
[7] 49915257 49915340    84 [GAAAGAAAGAAAGAAAGAAAGAAA...AAAGAAAGAAAGAAAGAAAGAA

> GAAA.x <- paste(rep("GAAA", 18), collapse = "")
> mT <- matchPattern(GAAA.x, chr18NoN)
> countPattern(GAAA.x, chr18NoN)

[1] 19

> length(mT)

[1] 19

> mT

  Views on a 76117153-letter DNAString subject
subject: CCCTAACCCTAACCCTAACCCTTACCCCTAACCC...GGTCTCTTGCCTCGGCAAAGATTAGATTAGGG
views:
        start      end width
 [1]  2604564  2604635    72 [GAAAGAAAGAAAGAAAGAAAGAA...AAAGAAAGAAAGAAAGAAAGAA
 [2]  2604568  2604639    72 [GAAAGAAAGAAAGAAAGAAAGAA...AAAGAAAGAAAGAAAGAAAGAA
 [3]  2604572  2604643    72 [GAAAGAAAGAAAGAAAGAAAGAA...AAAGAAAGAAAGAAAGAAAGAA
 [4]  2604576  2604647    72 [GAAAGAAAGAAAGAAAGAAAGAA...AAAGAAAGAAAGAAAGAAAGAA
 [5]  2604580  2604651    72 [GAAAGAAAGAAAGAAAGAAAGAA...AAAGAAAGAAAGAAAGAAAGAA
 [6]  2604584  2604655    72 [GAAAGAAAGAAAGAAAGAAAGAA...AAAGAAAGAAAGAAAGAAAGAA
 [7] 19831616 19831687    72 [GAAAGAAAGAAAGAAAGAAAGAA...AAAGAAAGAAAGAAAGAAAGAA
 [8] 30239572 30239643    72 [GAAAGAAAGAAAGAAAGAAAGAA...AAAGAAAGAAAGAAAGAAAGAA
 [9] 49915245 49915316    72 [GAAAGAAAGAAAGAAAGAAAGAA...AAAGAAAGAAAGAAAGAAAGAA
[10] 49915249 49915320    72 [GAAAGAAAGAAAGAAAGAAAGAA...AAAGAAAGAAAGAAAGAAAGAA
[11] 49915253 49915324    72 [GAAAGAAAGAAAGAAAGAAAGAA...AAAGAAAGAAAGAAAGAAAGAA
[12] 49915257 49915328    72 [GAAAGAAAGAAAGAAAGAAAGAA...AAAGAAAGAAAGAAAGAAAGAA
[13] 49915261 49915332    72 [GAAAGAAAGAAAGAAAGAAAGAA...AAAGAAAGAAAGAAAGAAAGAA
[14] 49915265 49915336    72 [GAAAGAAAGAAAGAAAGAAAGAA...AAAGAAAGAAAGAAAGAAAGAA
[15] 49915269 49915340    72 [GAAAGAAAGAAAGAAAGAAAGAA...AAAGAAAGAAAGAAAGAAAGAA
[16] 59099881 59099952    72 [GAAAGAAAGAAAGAAAGAAAGAA...AAAGAAAGAAAGAAAGAAAGAA
[17] 61328762 61328833    72 [GAAAGAAAGAAAGAAAGAAAGAA...AAAGAAAGAAAGAAAGAAAGAA
[18] 61328766 61328837    72 [GAAAGAAAGAAAGAAAGAAAGAA...AAAGAAAGAAAGAAAGAAAGAA
[19] 61711107 61711178    72 [GAAAGAAAGAAAGAAAGAAAGAA...AAAGAAAGAAAGAAAGAAAGAA
```

2.
```
> seqR1 <- RNAString("UCUUCCGAGACGAUGCUAGCAGCUAGCUAG")
> seqD1 <- cDNA(seqR1)
> seqD1

  30-letter "DNAString" instance
seq: AGAAGGCTCTGCTACGATCGTCGATCGATC

> reverse(seqD1)

  30-letter "DNAString" instance
seq: CTAGCTAGCTGCTAGCATCGTCTCGGAAGA

> translate(seqD1)

  10-letter "AAString" instance
seq: RRLCYDRRSI
```

3.
```
> f1 <- system.file("extdata", "someORF.fa", package = "Biostrings")
> file.info(f1)
> f1
> ff <- readFASTA(f1, strip.descs = TRUE)
> writeFASTA(ff, file = "", append = FALSE, width = 80)
```

4.
```
> pairwiseAlignment(pattern = c("superman"), subject = "supercalifragilistic

Global PairwiseAlignedFixedSubject (1 of 1)
pattern: [1] superman
subject: [1] supercal
score: -78.72394
```

```
> pairwiseAlignment(pattern = c("batman"), subject = "supercalifragilisticex

Global PairwiseAlignedFixedSubject (1 of 1)
pattern: [1] b-----a----------t-----man
subject: [1] supercalifragilisticexpial
score: -146.9763

> pairwiseAlignment("spiderman", "humptydumpty", type = "overlap",
+     gapOpening = -2, gapExtension = -1)

Overlap PairwiseAlignedFixedSubject (1 of 1)
pattern: [1] sp--id-erman
subject: [4] -pty-du--m--
score: 2.945211
```

5.

```
> data(BLOSUM62)
> pairwiseAlignment(AAString("RRLCYDRRSI"), AAString("HAQTYVALKYDRRSIERWW"),
+     substitutionMatrix = BLOSUM62, gapOpening = -12, gapExtension = -4)

Global PairwiseAlignedFixedSubject (1 of 1)
pattern: [1] RR-----LCYDRRSI
subject: [1] HAQTYVALKYDRRSI
score: -29
```

6. This example is taken from this document, page 21.

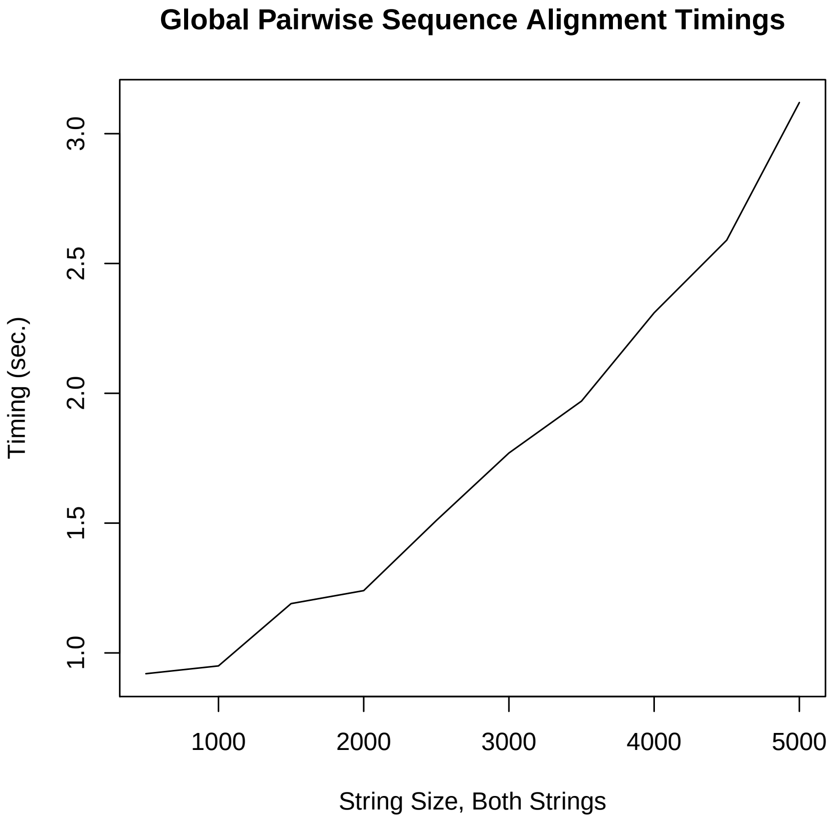
```
> N <- as.integer(seq(500, 5000, by = 500))
> timings <- rep(0, length(N))
> names(timings) <- as.character(N)
> for (i in seq_len(length(N))) {
+     string1 <- DNAString(paste(sample(DNA_ALPHABET[1:4], N[i],
+         replace = TRUE), collapse = ""))
+     string2 <- DNAString(paste(sample(DNA_ALPHABET[1:4], N[i],
+         replace = TRUE), collapse = ""))
+     timings[i] <- system.time(pairwiseAlignment(string1, string2,
+         type = "global"))[["user.self"]]
+ }
> timings

 500 1000 1500 2000 2500 3000 3500 4000 4500 5000
0.98 1.11 1.11 1.35 1.43 1.92 2.05 2.23 2.64 3.33
```

```
> coef(summary(lm(timings ~ poly(N, 2))))

             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 1.8150000  0.0372618 48.709405 4.022581e-10
poly(N, 2)1 2.2046799  0.1178322 18.710343 3.093480e-07
poly(N, 2)2 0.5587893  0.1178322  4.742248 2.102371e-03

> plot(N, timings, xlab = "String Size, Both Strings", ylab = "Timing (sec.)
+     type = "l", main = "Global Pairwise Sequence Alignment Timings")
```



6 Why Biostrings?

What I find interesting about Biostrings is that it allows me to work with sequences, a must in genomics, and it also enriches the things that can be done with R. With such uses, you can imagine how we can apply this program not only to this subject but also to filogenetics and our lab work. I am interested in Biostrings because I consider it to be an alternative to other tools and programming languages.

7 Extra Information

- As of today, they have released the version 2.13.39.
- You can download this package with the next R code:

  ```
  > source("http://bioconductor.org/biocLite.R")
  > biocLite("Biostrings")
  ```

```
> sessionInfo()

R version 2.10.0 Under development (unstable) (2009-08-15 r49252)
i386-pc-mingw32

locale:
[1] LC_COLLATE=Spanish_Mexico.1252  LC_CTYPE=Spanish_Mexico.1252
[3] LC_MONETARY=Spanish_Mexico.1252 LC_NUMERIC=C
[5] LC_TIME=Spanish_Mexico.1252

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] BSgenome.Hsapiens.UCSC.hg18_1.3.11 BSgenome_1.13.11
[3] Biostrings_2.13.39                 IRanges_1.3.60

loaded via a namespace (and not attached):
[1] Biobase_2.5.5 tools_2.10.0
```