Bioconductor for Sequence Analysis

Martin Morgan

2012-07-02 Mon

Contents

1 Packages

1.1 A workflow – RNA seq

- Experimental design
- Sample prep, library generation
- Quality assessment / exploration / remediation
- Alignment
- Counting
- Statistical assessment
- Biological context (annotation)

1.1.1 Quality assessment / exploration / remediation

```
library(Biostrings)  # pattern matching, DNA sequence manipulation
library(ShortRead)   # fastq files

fl <- dir(pattern=".fastq$")
fq <- readFastq(fl)
abc <- alphabetByCycle(sread(fq))
abc[1:4, 1:5]
matplot(t(abc[1:4,]), type="l", lwd=2, lty=1)
```

1.1.2 Counting

```
library(TxDb.Hsapiens.UCSC.hg19.knownGene)
genes <- exonsBy(TxDb.Hsapiens.UCSC.hg19.knownGene, "gene")

library(GenomicRanges)

fls <- dir("/path/to", ".*bam$")
counts <- summarizeOverlaps(fls, genes)
```

1.1.3 Statistical assessment

```
library(DESeq)
## ...
```

1.1.4 Biological context

```
library(org.Hs.eg.db)                 # info about genes
library(GO.db)                        # gene ontology categories
library(BSgenome.Hsapiens.UCSC.hg19)  # whole-genome sequence
library(SNPlocs.Hsapiens.dbSNP.20101109) # known SNPs
library(SIFT.Hsapiens.dbSNP132)       # SNP consequences
library(biomaRt)                      # online data resources

select(org.Hs.eg.db, entrezIds, c("GENENAME", "SYMBOL", "GO"))
```

- Solution 1: bad

```
system.time({
    x <- numeric()
    for (i in 1:50000)
        x[i] <- log(i)
})
```

- Solution 2: good

```
system.time({
    y <- log(1:50000)
})

identical(x, y)    ## same answer!
```

1.2 Survey

Selected Packages
BiocViews

2 Performance: dealing with big data

2.1 Iterate

```
library(Rsamtools)          # work with bam (aligned read) files
fl <- open(BamFile("/path/to/file.bam"))
while (length(x <- yield(fl))) {
    ## do work
}
```

2.2 Sample

```
library(ShortRead)          # work with fastq files
fl <- FastqSampler("/path/to/file.fastq")
sample <- yield(fl)
```

2.3 Select

```
library(VariantAnnotation) # work with 'variant call format' files
which <- GRanges("chr1", IRanges(100000, width=10000))
param <- ScanVcfParam(which=which)
fl <- open(VcfFile("/path/to/file.vcf"))
vcf <- readVcf(fl, "hg19", param=param)
```

2.4 Tune

Task: calculate log of numbers 1 through 50000

2.5 Parallelize

2.5.1 Demo

```
library(parallel)
options(mc.cores=8)

## fun demo
sleeper <- function(i) {
    Sys.sleep(1)
    i
}

system.time(res0 <- lapply(1:8, sleeper))
system.time(res1 <- mclapply(1:8, sleeper))

res0
identical(res0, res1)
```

2.5.2 Set-up

```
## regions of interest
data("ex", package="SequenceAnalysisData") # from another course
```

```
# aligned reads
fls <- dir(pattern=".bam$", full=TRUE)

## count reads aligning to 'regions of interest'
counter <-
    function(filePath, roi)
{
    aln <- readGappedAlignments(filePath)
    strand(aln) <- "*"
    hits <- countOverlaps(aln, roi)
    countOverlaps(roi, aln[hits==1])
}
```

2.5.3 Sequential

```
system.time({
    counts0 <- sapply(fls, counter, ex)
})
counts0
```

2.5.4 Parallel

```
sapply <- function(...)
   simplify2array(mclapply(...))

system.time({
    counts1 <- sapply(fls, counter, ex)
})
identical(counts0, counts1)
```

2.6 Reduce

Big data becomes small as analysis progresses

3 Classes

3.1 Benefits

- Coordinate complicated data –> fewer 'clerical' errors
- Ensure data integrity, e.g., alignment and annotation using same genome
- Use appropriate methods

3.2 Working with classes

Define classes to represent complicated data

- Two different object systems: S3, S4
- Bioconductor uses primarily S4 objects

Work with classes similarly

- 'accessors' to extract data
- 'Generic' functions with specialized 'methods' for each class

Inheritance of classes and methods

3.3 Common classes in sequence analysis

Derived from eSet
DNAStringSet and friends
GRanges, GRangesList
SummarizedExperiment

4 Help

4.1 Html / static

Vignettes

- also:

```
vignette(package="VariantAnnotation")
vignette(package="VariantAnnotation", "VariantAnnotation")
```

Manual pages

4.2 Interactive

```
showMethods("translate")
showMethods(class="DNAStringSet", where="package:Biostrings")

class ? DNAStringSet

method ? "translate,DNAStringSet"
```

4.3 Community

Web site: http://bioconductor.org

- Work flows
- Packages and vignettes
- Conferences and Courses

Mailing lists
Twitter – @Bioconductor; #Bioc