生信编程直播课程优秀学员作业展示1

Original x2yline 生信技能树 2017-03-15 08:47

收录于合集
#学徒作业

117个

题目 人类基因组外显子区域长度

学员：x2yline

具体题目详情请参考生信技能树论坛

题目数据来源为：ftp://ftp.ncbi.nlm.nih.gov/pub/CCDS/current_human/CCDS.current.txt

解题思路（比较适合R语言）如下

R语言实现

第一次完成的代码是**未考虑外显子overlap情况**（只去重了完全相同的外显子）写的

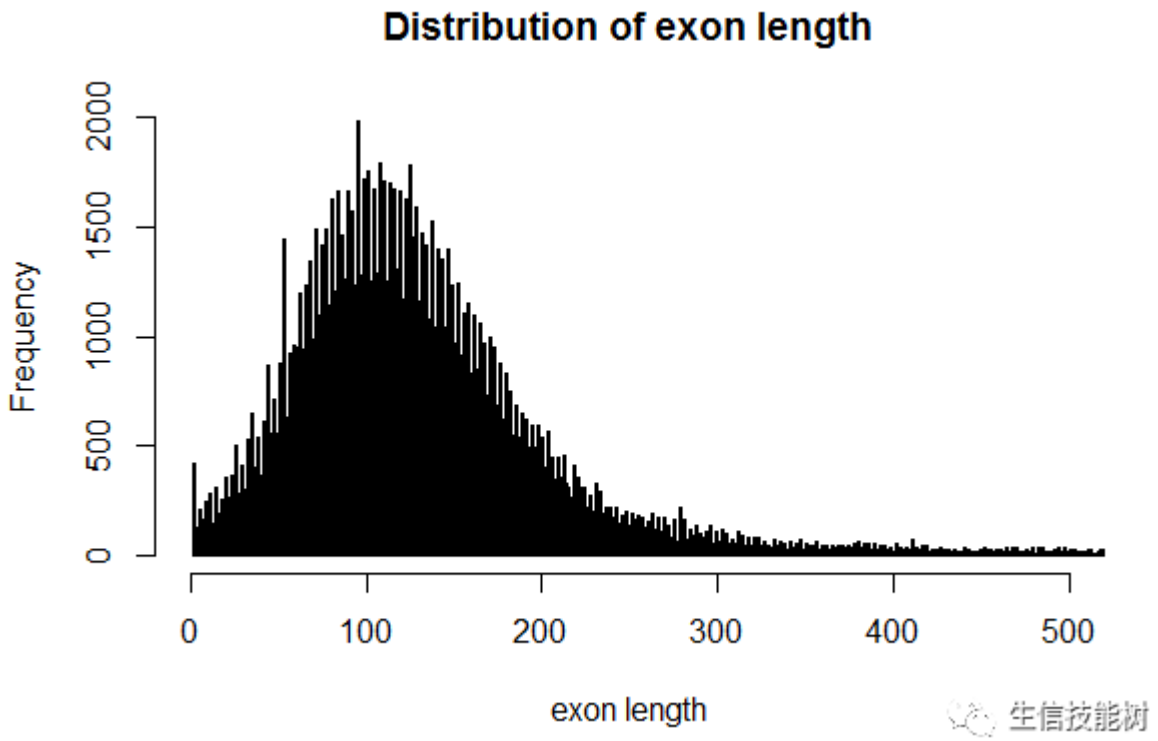
运行计算时间：14.74084 secs

最后运行结果：36048075

第一版代码如下：

```
setwd('E:\\r\\biotrainee_demo\\class1')#修改工作路径
t1 <- Sys.time()#把程序运行之前的时间赋值给t1
directory = 'CCDS.current.txt'#把文件名赋值给directory

data <- read.table(directory, sep='\t',
                   stringsAsFactors=F, header=T)[c(1,10)]#读取数据并提取出第一和第十列

get_gene <-function(data_item){
  # 该函数用于apply执行
  # 输入的数据为仅含原始数据第1列和第10列的dataframe
  # 用apply函数执行后输出的数据为每个基因外显子的坐标，
  # 一个基因的所有外显子以逗号分隔组成一个string，所有基因的string组成一个vector
  # 用apply函数执行后，最后格式为c('111-112, 115-135, 125-138', '254-258',...)
  if (!data_item[2] =='-'){
    exon_ranges <- data_item[2]
    exon_ranges <- substr(exon_ranges, start=2, stop=nchar(exon_ranges)-1)# 去除首尾的中括号符号
```

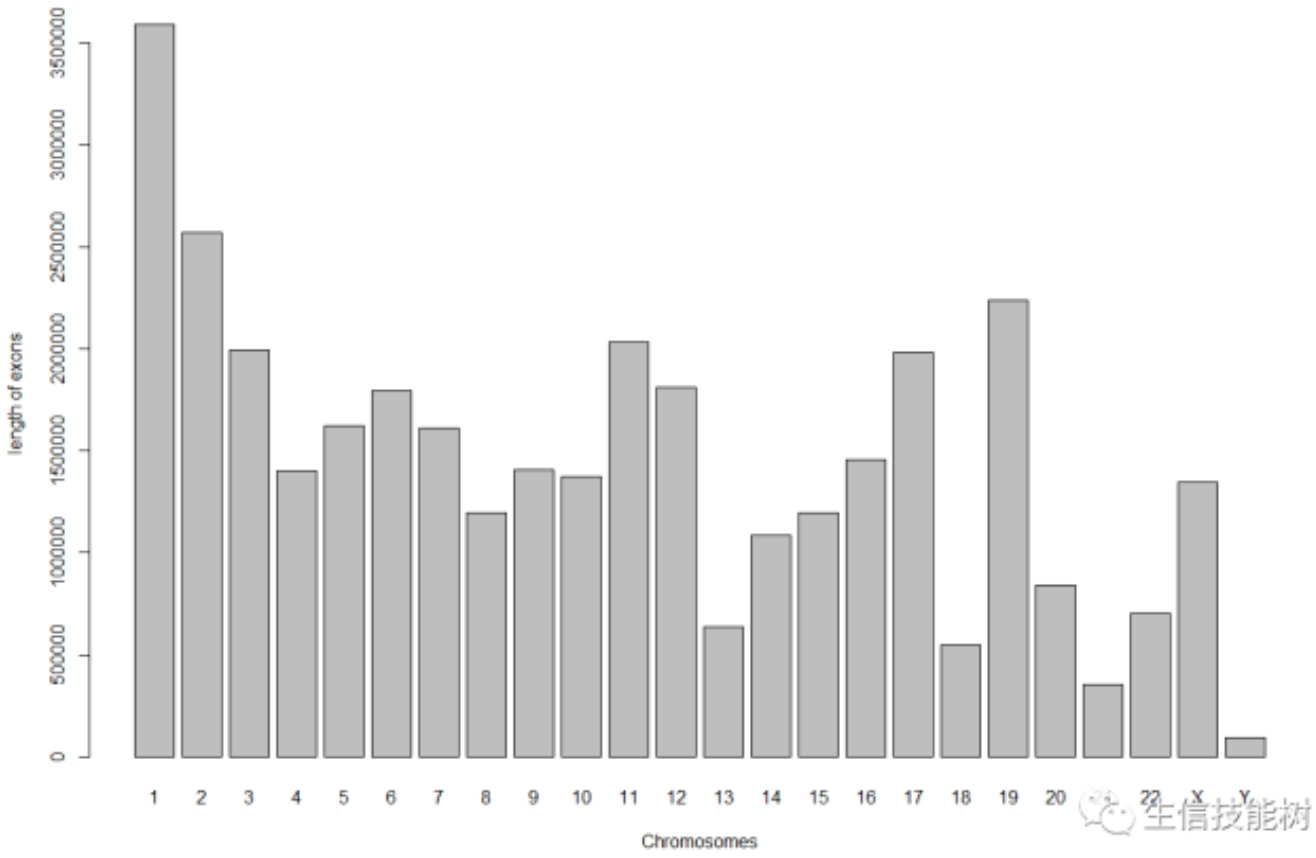
```
  }
}

get_exon <- function(gene){
  # 输入的数据为c('111-112, 115-135', '125-138', '254-258,...')
  # 把i号染色体上的所有外显子后在一起，并去除完全相同的外显子
  # 输出的数据为c('111-112','115-135', '125-138', '254-258', ...)
  exon <- unique(strsplit(gene,", ")[[1]])
}

get_length <- function(exon){
  # 输入的数据为lapply(c('111-112','115-135', '125-138', '254-258', ...),fun)
  # 输出结果为左右两坐标之差+1即外显子的长度
  loc <- strsplit(exon,"-")[[1]]
  a <- as.numeric(loc[2])-as.numeric(loc[1]) +1 #每个外显子的碱基数目
  a
}

exon_length = 0
exon_length_items = NULL
for (i in unique(data[,1])){
  gene_i <- paste(apply(data[which(data[1]==i & data[2] != '-'),], 1, get_gene),collapse=', ')
  exon_i <-  get_exon(gene_i)
  exon_i_length <- sapply(exon_i, get_length)
  exon_length <- exon_length + sum(exon_i_length)
  exon_length_items <- c(exon_i_length, exon_length_items)
  names(exon_length_items)[1:length(exon_i_length)] <- i
}

hist(exon_length_items,xlim=c(0,500),breaks = 20000,
     main='Distribution of exon length', xlab='exon length')

difftime(Sys.time(), t1, units = 'secs')# 计算执行完成后时间与t1的间隔

print(paste('all exons length is',exon_length))
```



第二版代码如下

```
setwd('E:\\r\\biotrainee_demo1')
t1 <- Sys.time()
directory = 'CCDS.current.txt'
# 读取数据并提取第1列和第10列
data <- read.table(directory, sep='\t',
```

```
                    stringsAsFactors=F, header=T)[c(1,10)]

get_gene <-function(data_item){
  # 用apply执行该函数
  # 输入的数据为仅含原始数据第1列和第10列的dataframe
  # 输出的数据为c('111-112, 115-135, 125-138', '254-258',...)
  if (!data_item[2] =='-'){
    exon_ranges <- data_item[2]
    exon_ranges <- substr(exon_ranges, start=2, stop=nchar(exon_ranges)-1)
  }
}

get_exon <- function(gene){
  # 输入的数据为c('111-112, 115-135, 125-138, 254-258,...')
  # 输出的数据为c('111-112','115-135', '125-138', '254-258', ...)
  exon <- unique(strsplit(gene,", ")[[1]])# 注: strsplit的输出结果为列表
}

get_length <- function(exon){
  # 输入的数据为lapply(c('111-112','115-135', '125-138', '254-258', ...),fun)
  # 输出结果为两坐标值和左右两坐标之差
  loc <- strsplit(exon,"-")[[1]]
  a <- c(as.numeric(loc[1]), as.numeric(loc[2])-as.numeric(loc[1]), as.numeric(loc[2]))
  #if (a==0){
  #print(loc)
  #}
  a
}

exon_length = NULL
for (i in unique(data[,1])){
  # paste 函数把i号染色体的所有外显子的坐标合并为一个character对象
  # gene_i的格式为'111-112, 115-135, 125-138, 254-258,...'
  gene_i <- paste(apply(data[which(data[1]==i & data[2] != '-'),], 1, get_gene),collapse=', ')
  # exon_i的格式为c('111-112','115-135', '125-138', '254-258', ...)
  exon_i <-  lapply(get_exon(gene_i), get_length)
  mat <- matrix(unlist(exon_i), ncol=3, byrow = T)
  #mat <- mat[order(mat[,2], decreasing = F),]
  #mat <- mat[order(mat[,1], decreasing = F),]

  # 使用matrix 是因为vector太长会报错
  #R memory management / cannot allocate vector of size n MB
  base_loc <- matrix(unique(unlist(apply(mat, 1, function(x) c(x[1]:x[3])))))

  exon_length <- c(exon_length , dim(base_loc)[1] * dim(base_loc)[2])
}

# 耗时长度
difftime(Sys.time(), t1, units = 'secs')
chrs <- unique(data[,1])
barplot(exon_length,names.arg=chrs,xlab='Chromosomes',ylab='length of exons')
print(paste('all exons length is',sum(exon_length)))
```



python实现

- jupyter编辑器太强大了，非常好用，但是没有查看当前变量的功能，所以最终还是选择spyder作为python编写平台（有shift+enter键相当于Rstudiod的ctr+r键，也有查看当前已有变量数值的功能）
- 关于open(file, 'rt')的解释

w,r,wt,rt都是python里面文件操作的模式。w是写模式，r是读模式。

t是windows平台特有的所谓text mode(文本模式）,区别在于会自动识别windows平台的换行符。

类Unix平台的换行符是\n，而windows平台用的是\r\n两个ASCII字符来表示换行，python内部采用的是\n来表示换行符。

rt模式下，python在读取文本时会自动把\r\n转换成\n. wt模式下，Python写文件时会用\r\n来表示换行。

python代码实现（第一次写的与老师的代码大致相同，用for循环即可，不推荐用以下的方法做）

```python
import pandas as pd
import numpy as np
file = r'E:\r\biotrainee_demo1\CCDS.current.txt'


def calculate_exon(file):
  data = pd.read_csv(file, sep='\t',\
    usecols=[0,9])

#data.loc[1:10,:]
#  data[0:3]
#  data.iloc[1:3]
#  data.iloc[3]
  all_length = 0

  for i in data.iloc[:,0].unique():
    # get the data of chrosome i
    # iloc[row_vector,col_vect]
    # iloc[row_vector]
    data_i = data.loc[data.iloc[:,0] == i]
    type(data_i)
    type(data_i.iloc[:,1])
    # remove the '[]' in column2
    data_j = data_i.iloc[:,1].apply(lambda x: x[1:-1])
    data_p = data_j.apply(lambda x: x.split(', '))
    data_g = data_p.apply(lambda x: pd.Series(x))
    # 把nan填充为 0-0
    data_f = np.array(data_g.fillna('0-0'))
    # 去除重复的外显子
    data_f = np.unique(data_f.reshape((data_f.shape[0]*data_f.shape[1], 1)))
    data_f = pd.DataFrame(data_f)
    data_m = data_f.apply(lambda x: \
      x.apply(lambda y: (y.split('-')[0])))
    data_n = data_f.apply(lambda x: \
      x.apply(lambda y: (y.split('-')[-1])))
    # pd.to_numeric can only apply to a 1-d array
    data_mi = data_m.apply(lambda x: pd.to_numeric(x, downcast='float'))
    data_ni = data_n.apply(lambda x: pd.to_numeric(x, downcast='float'))
    all_length += (data_ni - data_mi).sum().sum()
  return(all_length)

length = calculate_exon(file)
print(length)
```

运算速度有点慢，因为是临时学的pandas和numpy，很多步骤还没有优化

未去重overlap结果为：36046283

编程感悟

由于开始R是没有基础的，用通过R包swirl学习了一下lapply，apply和sapply函数的使用，对于迭代数目比较多的循环来说，R语言的for循环效率远远不如apply系列函数，应该尽量避免for循环处理，而python的for循环运算速度较快，可以使用for循环处理一下比较大的数据。

本文编辑：思考问题的熊

收录于合集 #学徒作业 117

上一篇 · 生信编程直播课程优秀学员学习心得及作业展示3

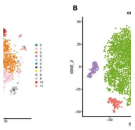
People who liked this content also liked

TLC点板，产物极性忽大忽小，咋回事！

有机合成路线

单细胞专题 | 9.如何人工注释单细胞类群？

生物信息云



生物材料家族盘点：水凝胶有多火？扒一扒近些年用它发的正刊

EngineeringForLife