main.R

xiucheng quek

22 April 2016

This is written to document the analysis and the generation pertaining to identifying the optimized parameters for dot-aligner

# Introduction

Previously, to optimize parameters for dot-aligner a total of 8764 combinations of parameters were used on a total of 3064 pairwise alignments. We then conducted RNAalifold and RNAdistance on the results and compared the differences between scores derived from structural alignments via dot aligner and reference alignment from brailbase alignments are considered "gold standards".

To choose the optimal parameters, for a given pairwise alignment, we ranked the differences between the SCI and RNAdistance score (the structure conservation index ) produced using dot aligner and Brial.

In the ranking of the differences in SCI (sci\_diff) generated between dotaligner and brailbase ( dotaligner - brailbase ), 1. Parameters that gives the largest sci\_diff are ranked top and given a score of 0. T 2. Parameters with The lowest sci\_diff are given the score based on the position of the rank. 3. Parameters with the same score are given the same rank and thus have the same score 4. The sum of rank score will help determine which parameters are optimial 5. The parameters which gives the lowest

To ensure that we are selecting for parameters that produce consistent results across various sequence pair, we calculated the CoV, mean and median of the score produced across the 3096 pairs

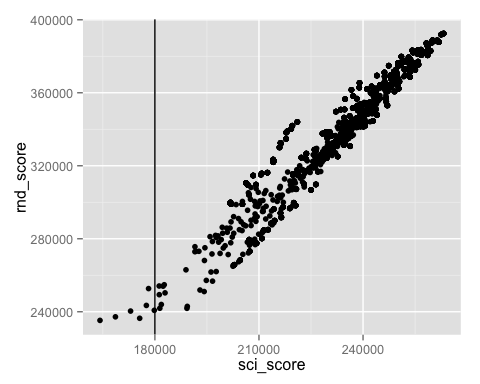
## Dependencies

library(reshape2)  
library(ggplot2)

# Figure 1 : Cumulative ranked score of SCI-diff and RNAdistance difference (RND-diff)

The ideal combinations should have both low sci-diff and differences in rnadistance score from the original sequence

source('figure1A.R')  
pFigure1A <- generateFigure1A('data/rnd\_ranking/cumulative\_score\_rnd\_2104/part-00000', 'data/sci\_ranking/cumulative\_score\_2104/part-00000')  
  
pFigure1A + geom\_vline(xintercept=180000)



Based on the distribution of SCI vs RND, I selected a cut off of 180000 cumulative socre for sci

figure1AData <- pFigure1A$data  
top\_parameters <- figure1AData[figure1AData$sci\_score < 180000 ,]  
combinations <- read.table('data/combination\_identifier\_2104/part-00000')  
colnames(combinations) <- c('e','o','k', 't', 'comb\_id')  
top\_parameters\_combination <- merge(combinations, top\_parameters, by.x= 'comb\_id', by.y='combination')  
top\_parameters\_combination

## comb\_id e o k t sci\_score rnd\_score  
## 1 12 0.8 1.0 0 1 177546 243464  
## 2 1503 0.2 0.8 0 1 178182 252692  
## 3 1779 1.0 1.0 0 1 175620 236416  
## 4 1809 0.4 1.0 0 1 168631 237212  
## 5 1943 0.6 1.0 0 1 172994 240371  
## 6 2104 1.0 0.6 0 1 179830 240777  
## 7 2676 0.2 1.0 0 1 164180 235260

I also did a rank product to get the top 20 score

figure1AData$ranked\_score<- (figure1AData$sci\_score /100000) \* (figure1AData$rnd\_score /100000)  
ordered\_figure1A <- figure1AData[order(figure1AData$ranked\_score),]  
top\_20 <- head(ordered\_figure1A, n = 20)   
  
pFigure1A + geom\_point(data=top\_20, colour ='red')



top\_20\_parameters <- merge(combinations, top\_20, by.x= 'comb\_id', by.y='combination')  
write.table(top\_20\_parameters, sep='\t', file='top\_20\_parameters.tsv', row.names = F)  
top\_20\_parameters

## comb\_id e o k t sci\_score rnd\_score ranked\_score  
## 1 12 0.8 1.0 0 1 177546 243464 4.322606  
## 2 72 1.0 0.4 0 1 193002 251899 4.861701  
## 3 398 0.6 0.6 0 1 181268 249346 4.519845  
## 4 687 0.8 0.2 0 1 189345 242940 4.599947  
## 5 796 0.8 0.6 0 1 181424 241908 4.388792  
## 6 1223 0.4 0.6 0 1 182364 254013 4.632283  
## 7 1503 0.2 0.8 0 1 178182 252692 4.502517  
## 8 1779 1.0 1.0 0 1 175620 236416 4.151938  
## 9 1809 0.4 1.0 0 1 168631 237212 4.000130  
## 10 1852 0.6 0.8 0 1 182699 254794 4.655061  
## 11 1943 0.6 1.0 0 1 172994 240371 4.158274  
## 12 2032 1.0 0.2 0 1 189240 241876 4.577261  
## 13 2104 1.0 0.6 0 1 179830 240777 4.329893  
## 14 2425 0.8 0.8 0 1 182930 250343 4.579524  
## 15 2511 0.6 0.2 0 1 194209 251065 4.875908  
## 16 2544 0.4 0.8 0 1 181238 254106 4.605366  
## 17 2662 1.0 0.8 0 1 181856 243978 4.436886  
## 18 2676 0.2 1.0 0 1 164180 235260 3.862499  
## 19 2783 0.8 0.4 0 1 194824 257177 5.010425  
## 20 2834 0.2 0.6 0 1 188961 262969 4.969089

# Figure 1B plot the mean of SCI and mean RDD of ( in progress )

# Figure 1C : ( code not evaluted)

Check markdown document for code

# 