ゲノム情報解析基礎（2014 年 4 月 30 日）課題の回答とコメント

課題 4：TAIR から取得した転写開始点上流 500bp の multi-FASTA ファイル（TAIR10_upstream_500_20101028.fa）中には 500 bp でない配列が 2 つ含まれている。

1. この 2 つの配列のみを抽出するためには以下のテンプレートコードのどこをどう変更すればよいか示せ（ここをこう変えるとよい、などでよい）。

```
in_f <- "TAIR10_upstream_500_20101028.fa"#入力ファイル名を指定してin_fに格納
out_f <- "hoge5.png"                     #出力ファイル名を指定してout_fに格納
param1 <- c(461, 500)                    #抽出したい範囲の始点と終点を指定
param2 <-  500                           #配列長を指定
param_fig <- c(700, 400)                 #ファイル出力時の横幅と縦幅を指定(単位はピクセル

#必要なパッケージをロード
library(Biostrings)                      #パッケージの読み込み
library(seqLogo)                         #パッケージの読み込み

#入力ファイルの読み込み
fasta <- readDNAStringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み
fasta                                    #確認してるだけです

#前処理(配列長が500bpのもののみフィルタリング後、解析したいサブセットを抽出)
obj <- as.logical(width(fasta) == param2)#条件を満たすかどうかを判定した結果をobjに格納
fasta <- fasta[obj]                      #objがTRUEとなるもののみ抽出した結果をfastaに格
fasta                                    #確認してるだけです
fasta <- subseq(fasta, param1[1], param1[2])#param1で指定した始点と終点の範囲の配列を
fasta                                    #確認してるだけです

#本番(sequence logoを実行)
```

ここを!=にすればいいのです。「!==」は減点。「<」は実質的に目的を達成可能なので OK。

2. 500 bp でない 2 つの配列の配列長を示せ。

   26 と 224 です。

3. このコードは、上流 500bp のファイルを読み込んで 461 番目から 500 番目の範囲を切り出して、sequence logos を行っている。この目的を達成するために、subseq 関数のデフォルトオプションの「start=461 と end=500」を利用している。一方、これは上流 40bp、つまり最後から 40bp を切り出していることと同義であるため、「end=500 と**何か**」というオプションで表現することもできる。この**何か**を示せ。

   「width=40」です。もちろん「param1 <- 40 と width=param1」でも OK です。ただし、「width=…」という記述がない場合は減点。「width=param1」だけの場合も減点。「40」のような数値だけの場合も減点。