

第1部の掟

①今年度のリンク先の、②第1部の該当項目をクリックして、そこを利用してください。例えば、③統計解析の初日(2016.07.20)はこちらをクリック

(Rで)塩基配列解析

～NGS, RNA-seq, ゲノム, トランスクリプトーム, 正規化, 発現変動, 統計, モデル, バイオインフォマティクス～
(last modified 2016/06/13, since 2011)

- [基本的な利用法](#) (last modified 2015/04/03)
- [サンプルデータ](#) (last modified 2016/05/13)
- [バイオインフォマティクス人材育成カリキュラム\(次世代シーケンサ\)](#) | [NGSハンズオン講習会2016](#) (last modified 2016/06/13)
- [バイオインフォマティクス人材育成カリキュラム\(次世代シーケンサ\)](#) | [NGSハンズオン講習会2015](#) (last modified 2015/04/03)
- [バイオインフォマティクス人材育成カリキュラム\(次世代シーケンサ\)](#) | [速習コース2014](#) (last modified 2014/05/12)
- [書籍 | トランスクリプトーム解析 | について](#) (last modified 2014/05/12)



バイオインフォマティクス人材育成カリキュラム(次世代シーケンサ) | NGSハンズオン講習会2016

平成28年度[NGSハンズオン講習会](#)の実習で用いるリンク先、データファイル、コピペ用コード集などはここで示します。

- [はじめに\(講習会参加者必読\)](#) (last modified 2016/06/13)
- 事前準備 | [Bio-Linux 8とRのインストール状況確認\(2016.07.19\)](#) (last modified 2016/06/07)
- [第1部 | 統計解析 | について](#) (last modified 2016/06/02)
- [第1部 | 統計解析 | **ゲノム解析、塩基配列解析\(2016.07.20\)**](#) (last modified 2016/05/07)
- [第1部 | 統計解析 | **トランスクリプトーム解析1\(2016.07.21\)**](#) (last modified 2016/05/23)
- [第1部 | 統計解析 | **トランスクリプトーム解析2\(2016.07.22\)**](#) (last modified 2016/06/02)
- [第2部 | NGS解析\(初～中級\) | について](#) (last modified 2016/05/26)
- [第2部 | NGS解析\(初～中級\) | \[NGS解析基礎\\(2016.07.25\\)\]\(#\)](#) (last modified 2016/04/26)
- [第2部 | NGS解析\(初～中級\) | \[ゲノムReseq、変異解析\\(2016.07.26\\)\]\(#\)](#) (last modified 2016/04/26)
- [第2部 | NGS解析\(初～中級\) | \[RNA-seq\\(2016.07.27\\)\]\(#\)](#) (last modified 2016/04/26)
- [第2部 | NGS解析\(初～中級\) | \[ChIP-seq\\(2016.07.28\\)\]\(#\)](#) (last modified 2016/04/26)
- [第3部 | NGS解析\(中～上級\) | について](#) (last modified 2016/06/13)
- [第3部 | NGS解析\(中～上級\) | \[Linux環境でのデータ解析: JavaやRの利用法\\(2016.08.01\\)\]\(#\)](#) (last modified 2016/06/13)
- [第3部 | NGS解析\(中～上級\) | \[Linux環境でのデータ解析: マッピング、トリミング、アセンブリ\\(2016.08.02\\)\]\(#\)](#) (last modified 2016/06/13)
- [第3部 | NGS解析\(中～上級\) | \[クラウド環境との連携、ロングリードデータの解析\\(2016.08.03\\)\]\(#\)](#) (last modified 2016/04/26)
- [第3部 | NGS解析\(中～上級\) | \[トランスクリプトームアセンブリ、発現量推定\\(2016.08.04\\)\]\(#\)](#) (last modified 2016/04/26)



What's new
このウェブページは基本的に体系的に更新されています。私(門田)の演習依頼・研究系workshop

第1部の掟

①この日の講義資料や解析データ(hoge.zip)はここにあります。適宜アップデートしているので、最終更新日に気をつけてください

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料のスライド25あたりからやる予定です。スライド24あたりまでは自習。

- [講義資料PDF](#)(2016.06.08版; 約5MB)
 - [hoge.zip](#)(2016.05.13版; 約2MB)
- 
- NGSデータ解析戦略
 - [DDBJ Pipeline: Nagasaki et al., DNA Res., 2013](#)
 - [Illumina BaseSpace](#)
 - [Galaxy: Goecks et al., Genome Biol., 2010](#)
 - NGSデータ
 - *Lactobacillus hokkaidonensis* LOOC260(T): [Tanizawa et al., BMC Genomics, 2015](#)
 - [DRR024500*](#) (PacBioデータ; 後に問題が判明し削除済み)
 - [DRR024501](#) (paired-end MiSeqデータ)
 - [DRR054113](#) (PacBioデータ; 再登録後のもの)
 - [DRR054114](#) (PacBioデータ; 再登録後のもの)
 - [DRR054115](#) (PacBioデータ; 再登録後のもの)
 - [DRR054116](#) (PacBioデータ; 再登録後のもの)
 - de novoアセンブリ
 - [Velvet: Zerbino and Birney, Genome Res., 2008](#)
 - [Platanus: Kajitani et al., Genome Res., 2014](#)
 - [日本乳酸菌学会誌のNGS連載第6回ゲノムアセンブリ](#)
 - [原稿PDF](#)
 - ウェブ資料PDF
 - [Windows用](#)(2016.03.29版; 約25MB)
 - [Macintosh用](#)
 - DDBJ PipelineでPlatanus
 - Platanus ver. 1.2.2を実行したzip圧縮ファイル([platanusResult.zip](#); 約2.2MB)
 - k-mer解析(k=1)

第1部の掟

①この部分からページ下部に移動することが、②のスライドをめくっていくことに相当します



第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料の②(スライド25あたりからやる予定です。スライド24あたりまでは自習。)

- [講義資料PDF](#)(2016.06.08版; 約5MB)
- [hoge.zip](#)(2016.05.13版; 約2MB)
- NGSデータ解析戦略
 - [DDBJ Pipeline: Nagasaki et al., DNA Res., 2013](#)
 - [Illumina BaseSpace](#)
 - [Galaxy: Goecks et al., Genome Biol., 2010](#)
- NGSデータ
 - *Lactobacillus hokkaidonensis* LOOC260(T): [Tanizawa et al., BMC Genomics, 2015](#)
 - [DRR024500*](#) (PacBioデータ; 後に問題が判明し削除済み)
 - [DRR024501](#) (paired-end MiSeqデータ)
 - [DRR054113](#) (PacBioデータ; 再登録後のもの)
 - [DRR054114](#) (PacBioデータ; 再登録後のもの)
 - [DRR054115](#) (PacBioデータ; 再登録後のもの)
 - [DRR054116](#) (PacBioデータ; 再登録後のもの)
- de novoアセンブリ
 - [Velvet: Zerbino and Birney, Genome Res., 2008](#)
 - [Platanus: Kajitani et al., Genome Res., 2014](#)
- [日本乳酸菌学会誌のNGS連載第6回ゲノムアセンブリ](#)
 - [原稿PDF](#)
 - ウェブ資料PDF
 - [Windows用](#)(2016.03.29版; 約25MB)
 - [Macintosh用](#)
- DDBJ PipelineでPlatanus
 - Platanus ver. 1.2.2を実行したzip圧縮ファイル([platanusResult.zip](#); 約2.2MB)
- k-mer解析(k=1)

第1部の掟

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料の①(スライド25あたりからやる予定です。スライド24あたりまでは自習。

- [講義資料PDF](#)(2016.06.08版; 約5MB)
- [hoge.zip](#)(2016.05.13版; 約2MB)
- NGSデータ解析戦略
 - [DDBJ Pipeline: Nagasaki et al., DNA Res., 2013](#)
 - [Illumina BaseSpace](#)
 - [Galaxy: Goecks et al., Genome Biol., 2010](#)
- NGSデータ
 - *Lactobacillus hokkaidonensis* LOOC260(T): [Tanizawa et al., BMC Genomics, 2015](#)
 - [DRR024500*](#) (PacBioデータ; 後に問題が判明し削除済み)
 - [DRR024501](#) (paired-end MiSeqデータ)
 - [DRR054113](#) (PacBioデータ; 再登録後のもの)
 - [DRR054114](#) (PacBioデータ; 再登録後のもの)
 - [DRR054115](#) (PacBioデータ; 再登録後のもの)
 - [DRR054116](#) (PacBioデータ; 再登録後のもの)
- de novoアセンブリ
 - [Velvet: Zerbino and Birney, Genome Res., 2008](#)
 - [Platanus: Kajitani et al., Genome Res., 2014](#)
- [日本乳酸菌学会誌のNGS連載第6回ゲノムアセンブリ](#)
 - [原稿PDF](#)
 - ウェブ資料PDF
 - [Windows用](#)(2016.03.29版; 約25MB)
 - [Macintosh用](#)
- DDBJ PipelineでPlatanus
 - Platanus ver. 1.2.2を実行したzip圧縮ファイル([platanusResult.zip](#); 約2.2MB)
- k-mer解析(k=1)

例えば②は、①のスライド6の補足情報です。そして③は、①のスライド24の補足情報となります。ウェブ資料中でスライド番号を明記していないのは、スライドの挿入・追加・削除による番号のずれの修正で疲弊するのを避けるためで…した

第1部の掟

ここが最も重要！①のスライド35-36に相当するのが、②の部分



第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料の①(スライド25あたりからやる予定です。スライド24あたりまでは自習。

- [講義資料PDF](#)(2016.06.08版; 約5MB)
- [hoge.zip](#)(2016.05.13版; 約2MB)
- NGSデータ解析戦略
 - [DDBJ Pipeline: Nagasaki et al., DNA Res.](#)
 - [Illumina BaseSpace](#)
 - [Galaxy: Goecks et al., Genome Biol., 2010](#)
- NGSデータ
 - Lactobacillus hokkaidonensis LOOC260(T)
 - [DRR024500*](#) (PacBioデータ; 後に)
 - [DRR024501](#) (paired-end MiSeqデータ)
 - [DRR054113](#) (PacBioデータ; 再登録)
 - [DRR054114](#) (PacBioデータ; 再登録)
 - [DRR054115](#) (PacBioデータ; 再登録)
 - [DRR054116](#) (PacBioデータ; 再登録)
- de novoアセンブリ
 - [Velvet: Zerbino and Birney, Genome Res.](#)
 - [Platanus: Kajitani et al., Genome Res., 2011](#)
- [日本乳酸菌学会誌のNGS連載第6回ゲノムアセンブリ](#)
 - [原稿PDF](#)
 - ウェブ資料PDF
 - [Windows用](#)(2016.03.29版; 約25MB)
 - [Macintosh用](#)
- DDBJ PipelineでPlatanus
 - Platanus ver. 1.2.2を実行したzip圧縮ファイル
- k-mer解析(k=1)

```
• DDBJ PipelineでPlatanus
  ◦ Platanus ver. 1.2.2を実行したzip圧縮ファイル(platanusResult.zip; 約2.2MB)
• k-mer解析(k=1)
  塩基ごとの出現頻度解析は、k-mer解析のk=1の場合に相当します。「イントロ | 一般 | k-mer解析 | k=1(塩基ごとの出現頻度解析) | Biostrings」の例題7は、以下と同じ。入力ファイルは out\_gapClosed.fa (約2.4MB)

  in_f <- "out_gapClosed.fa"           #入力ファイル名を指定してin_fに格納
  out_f <- "hoge7.txt"                 #出力ファイル名を指定してout_fに格納
  param_base <- c("A", "C", "G", "T", "N") #出力させたい塩基を指定

  #必要なパッケージをロード
  library(Biostrings)                 #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

  #本番
  hoge <- alphabetFrequency(fasta)     #A,C,G,T,..の数を各配列ごとにカウントした結果をhogeに格納
  obj <- is.element(colnames(hoge), param_base) #条件を満たすかどうかを判定した結果をobjに格納
  #out <- colSums(hoge[, obj])         #列ごとの総和をoutに格納
  out <- apply(as.matrix(hoge[, obj]), 2, sum) #列ごとの総和をoutに格納

  #ファイルに保存
  write.table(out, out_f, sep="\t", append=F, quote=F, row.names=T, col.names=F) #tmp
```

一気に結果を得る
実際の利用時は、[rcodel.txt](#)のような無駄なコメントを除いてスリムにした一連のスク립トを作成しておき、一気にコピペする。ファイルの中身は以下と同じ。

```
param_base <- c("A", "C", "G", "T", "N") #出力させたい塩基を指定
library(Biostrings)                       #パッケージの読み込み
```

第1部の掟

①赤下線部分にも書いてあります。何が言いたいかというところ…



第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料のスライド25あたりからやる予定です。スライド24あたりまでは自習。

- [講義資料PDF](#)(2016.06.08版; 約5MB)
- [hoge.zip](#)(2016.05.13版; 約2MB)
- NGSデータ解析戦略
 - [DDBJ Pipeline: Nagasaki et al., DNA Res.](#)
 - [Illumina BaseSpace](#)
 - [Galaxy: Goecks et al., Genome Biol., 2010](#)
- NGSデータ
 - Lactobacillus hokkaidonensis LOOC260(T)
 - [DRR024500*](#) (PacBioデータ; 後に)
 - [DRR024501](#) (paired-end MiSeqデータ)
 - [DRR054113](#) (PacBioデータ; 再登録)
 - [DRR054114](#) (PacBioデータ; 再登録)
 - [DRR054115](#) (PacBioデータ; 再登録)
 - [DRR054116](#) (PacBioデータ; 再登録)
 - de novoアセンブリ
 - [Velvet: Zerbino and Birney, Genome Res.](#)
 - [Platanus: Kajitani et al., Genome Res., 2011](#)
 - [日本乳酸菌学会誌のNGS連載第6回ゲノムアセンブリ](#)
 - [原稿PDF](#)
 - [ウェブ資料PDF](#)
 - [Windows用](#)(2016.03.29版; 約25MB)
 - [Macintosh用](#)
- DDBJ PipelineでPlatanus
 - Platanus ver. 1.2.2を実行したzip圧縮ファイル([platanusResult.zip](#); 約2.2MB)
- k-mer解析(k=1)

- DDBJ PipelineでPlatanus
 - Platanus ver. 1.2.2を実行したzip圧縮ファイル([platanusResult.zip](#); 約2.2MB)
- k-mer解析(k=1)

塩基ごとの出現頻度解析は、k-mer解析のk=1の場合に相当します。「[イントロ](#) | [一般](#) | [k-mer解析](#) | [k=1\(塩基ごとの出現頻度解析\)](#) | [Biostrings](#)」の例題7は、以下と同じ。入力ファイルは[out_gapClosed.fa](#) (約2.4MB)

```
in_f <- "out_gapClosed.fa" #入力ファイル名を指定してin_fに格納
out_f <- "hoge7.txt" #出力ファイル名を指定してout_fに格納
param_base <- c("A", "C", "G", "T", "N") #出力させたい塩基を指定

#必要なパッケージをロード
library(Biostrings) #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

#本番
hoge <- alphabetFrequency(fasta) #A,C,G,T,..の数を各配列ごとにカウントした結果をhogeに格納
obj <- is.element(colnames(hoge), param_base) #条件を満たすかどうかを判定した結果をobjに格納
#out <- colSums(hoge[, obj]) #列ごとの総和をoutに格納
out <- apply(as.matrix(hoge[, obj]), 2, sum) #列ごとの総和をoutに格納

#ファイルに保存
write.table(out, out_f, sep="\t", append=F, quote=F, row.names=T, col.names=F) #tmp
```

一気に結果を得る
実際の利用時は、[rcodel.txt](#)のような無駄なコメントを除いてスリムにした一連のスク립トを作成しておき、一気にコピペする。ファイルの中身は以下と同じ。

```
param_base <- c("A", "C", "G", "T", "N") #出力させたい塩基を指定
library(Biostrings) #パッケージの読み込み
```

①のスライド35-36がそれぞれ②と③ですが、ここで書いてある通りにやらなくてよい(やるな)ということです!

第1部の掟

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料の、スライド25あたりからやる予定です。スライド24あたりまでは自習。

- 講義資料PDF(2016.06.08版; 約5MB)
- hoge.zip(2016.05.13版; 約2MB)



イントロ | 一般 | k-mer解析 | k=1(塩基ごとの出現頻度解析) | Biostrings

k-mer解析(k=1)

(Rで)塩基配列解析

①(アセンブリ実行結果の)multi-FASTAファイルを読み込んで、塩基ごとの出現頻度解析を行う項目

```

in_f <- "hoge1.fa"
out_f <- "hoge1.txt"

#必要なパッケージをロード
library(Biostrings)

#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

#本番
out <- alphabetFrequency(fasta)

```

Jul 20 2016, NGS/ハンズオン講習会 35

イントロ | 一般 | k-mer解析 | k=1(塩基ごとの出現頻度解析) | Biostrings

k-mer解析(k=1)

①例題7が、PlatanusのStep3実行後のファイル(②out_gapClosed.fa)を入力とするものなので、そのままコピペできて便利。これを実行します

```

in_f <- "out_gapClosed.fa"
out_f <- "hoge1.txt"

#必要なパッケージをロード
library(Biostrings)

#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

#本番
hoge <- alphabetFrequency(fasta)
obj <- is.element(colnames(hoge), param_base)#条件を満たすかどうかを判定した結果をobj
out <- apply(as.matrix(hoge[, 2, sum]), 2, sum)#列ごとの総和をoutに格納

#ファイルに保存
write.table(out, out_f, sep="\t", append=F, quote=F, row.names=T, col.names=F)#ts

```

Jul 20 2016, NGS/ハンズオン講習会 36



第1部の掟

①のスライド35-36は、やったふりをする。実際には②を眺め、③の(特に例題番号)記述内容を確認し…

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料の①(スライド25あたりからやる予定です。スライド24あたりまでは自習。

- [講義資料PDF](#)(2016.06.08版; 約5MB)
- [hoge.zip](#)(2016.05.13版; 約2MB)
- NGSデータ解析戦略
 - [DDBJ Pipeline: Nagasaki et al., DNA Res.](#)
 - [Illumina BaseSpace](#)
 - [Galaxy: Goecks et al., Genome Biol., 2010](#)
- NGSデータ
 - Lactobacillus hokkaidonensis LOOC260(T)
 - [DRR024500*](#) (PacBioデータ; 後に)
 - [DRR024501](#) (paired-end MiSeqデータ)
 - [DRR054113](#) (PacBioデータ; 再登録)
 - [DRR054114](#) (PacBioデータ; 再登録)
 - [DRR054115](#) (PacBioデータ; 再登録)
 - [DRR054116](#) (PacBioデータ; 再登録)
 - de novoアセンブリ
 - [Velvet: Zerbino and Birney, Genome Res.](#)
 - [Platanus: Kajitani et al., Genome Res., 2011](#)
 - [日本乳酸菌学会誌のNGS連載第6回ゲノムアセンブリ](#)
 - [原稿PDF](#)
 - ウェブ資料PDF
 - [Windows用](#)(2016.03.29版; 約25MB)
 - [Macintosh用](#)
- DDBJ PipelineでPlatanus
 - Platanus ver. 1.2.2を実行したzip圧縮ファイル
- k-mer解析(k=1)

②

- DDBJ PipelineでPlatanus
 - Platanus ver. 1.2.2を実行したzip圧縮ファイル([platanusResult.zip](#); 約2.2MB)
- k-mer解析(k=1)

塩基ごとの出現頻度解析は、k-mer解析のk=1の場合に相当します。「[イントロ](#) | [一般](#) | [k-mer解析](#) | [k=1\(塩基ごとの出現頻度解析\)](#)」の例題7は、以下と同じ。入力ファイルは [out_gapClosed.fa](#) (約2.4MB)

```
in_f <- "out_gapClosed.fa" #入力ファイル名を指定してin_fに格納
out_f <- "hoge7.txt" #出力ファイル名を指定してout_fに格納
param_base <- c("A", "C", "G", "T", "N") #出力させたい塩基を指定

#必要なパッケージをロード
library(Biostrings) #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

#本番
hoge <- alphabetFrequency(fasta) #A,C,G,T,..の数を各配列ごとにカウントした結果を
obj <- is.element(colnames(hoge), param_base) #条件を満たすかどうかを判定した結果をobj
#out <- colSums(hoge[, obj]) #列ごとの総和をoutに格納
out <- apply(as.matrix(hoge[, obj]), 2, sum) #列ごとの総和をoutに格納

#ファイルに保存
write.table(out, out_f, sep="\t", append=F, quote=F, row.names=T, col.names=F) #tmp
```

③

一気に結果を得る
実際の利用時は、[rcodel.txt](#)のような無駄なコメントを除いてスリムにした一連のスクリプトを作成しておき、一気にコピーする。ファイルの中身は以下と同じ。

```
param_base <- c("A", "C", "G", "T", "N") #出力させたい塩基を指定
library(Biostrings) #パッケージの読み込み
```


第1部の掟

①のスライド35-36は、やったふりをする。実際には②を眺め、③の(特に例題番号)記述内容を確認し…④の赤枠内をコピーして効率的に進めていきましょう

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料の①(スライド25あたりからやる予定です。スライド24あたりまでは自習。

- [講義資料PDF](#)(2016.06.08版; 約5MB)
- [hoge.zip](#)(2016.05.13版; 約2MB)
- NGSデータ解析戦略
 - [DDBJ Pipeline: Nagasaki et al., DNA Res.](#)
 - [Illumina BaseSpace](#)
 - [Galaxy: Goecks et al., Genome Biol., 2010](#)
- NGSデータ
 - Lactobacillus hokkaidonensis LOOC260(T)
 - [DRR024500*](#) (PacBioデータ; 後に)
 - [DRR024501](#) (paired-end MiSeqデータ)
 - [DRR054113](#) (PacBioデータ; 再登録)
 - [DRR054114](#) (PacBioデータ; 再登録)
 - [DRR054115](#) (PacBioデータ; 再登録)
 - [DRR054116](#) (PacBioデータ; 再登録)
- de novoアセンブリ
 - [Velvet: Zerbino and Birney, Genome Res.](#)
 - [Platanus: Kajitani et al., Genome Res., 2011](#)
- [日本乳酸菌学会誌のNGS連載第6回ゲノムアセンブリ](#)
 - [原稿PDF](#)
 - ウェブ資料PDF
 - [Windows用](#)(2016.03.29版; 約25MB)
 - [Macintosh用](#)
- DDBJ PipelineでPlatanus
 - Platanus ver. 1.2.2を実行したzip圧縮ファイル
- k-mer解析(k=1)

②

- DDBJ PipelineでPlatanus
 - Platanus ver. 1.2.2を実行したzip圧縮ファイル([platanusResult.zip](#); 約2.2MB)
- k-mer解析(k=1)

塩基ごとの出現頻度解析は、k-mer解析のk=1の場合に相当します。「[イントロ](#) | [一般](#) | [k-mer解析](#) | [k=1\(塩基ごとの出現頻度解析\)](#) | [Biostrings](#)」の例題7は、以下と同じ。入力ファイルは [out_gapClosed.fa](#) (約2.4MB)

```
in_f <- "out_gapClosed.fa" #入力ファイル名を指定してin_fに格納
out_f <- "hoge7.txt" #出力ファイル名を指定してout_fに格納
param_base <- c("A", "C", "G", "T", "N") #出力させたい塩基を指定

#必要なパッケージをロード
library(Biostrings) #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

#本番
hoge <- alphabetFrequency(fasta) #A,C,G,T,..の数を各配列ごとにカウントした結果をhogeに格納
obj <- is.element(colnames(hoge), param_base) #条件を満たすかどうかを判定した結果をobjに格納
#out <- colSums(hoge[, obj]) #列ごとの総和をoutに格納
out <- apply(as.matrix(hoge[, obj]), 2, sum) #列ごとの総和をoutに格納

#ファイルに保存
write.table(out, out_f, sep="\t", append=F, quote=F, row.names=T, col.names=F) #tmp
```

③

④

- 一気に結果を得る

実際の利用時は、[rcodel.txt](#)のような無駄なコメントを除いてスリムにした一連のスク립トを作成しておき、一気にコピーする。ファイルの中身は以下と同じ。

```
param_base <- c("A", "C", "G", "T", "N") #出力させたい塩基を指定
library(Biostrings) #パッケージの読み込み
```

第1部の掟

尚、①の例題番号と、②出力ファイルの hogeX.txtの番号は(基本的に)同じです。これも確認手段として有効利用しましょう

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料のスライド25あたりからやる予定です。スライド24あたりまでは自習。

- [講義資料PDF](#)(2016.06.08版; 約5MB)
- [hoge.zip](#)(2016.05.13版; 約2MB)
- NGSデータ解析戦略
 - [DDBJ Pipeline: Nagasaki et al., DNA Res.](#)
 - [Illumina BaseSpace](#)
 - [Galaxy: Goecks et al., Genome Biol., 2010](#)
- NGSデータ
 - Lactobacillus hokkaidonensis LOOC260(T)
 - [DRR024500*](#) (PacBioデータ; 後に)
 - [DRR024501](#) (paired-end MiSeqデータ)
 - [DRR054113](#) (PacBioデータ; 再登録)
 - [DRR054114](#) (PacBioデータ; 再登録)
 - [DRR054115](#) (PacBioデータ; 再登録)
 - [DRR054116](#) (PacBioデータ; 再登録)
- de novoアセンブリ
 - [Velvet: Zerbino and Birney, Genome Res.](#)
 - [Platanus: Kajitani et al., Genome Res., 2011](#)
- [日本乳酸菌学会誌のNGS連載第6回ゲノムアセンブリ](#)
 - [原稿PDF](#)
 - [ウェブ資料PDF](#)
 - [Windows用](#)(2016.03.29版; 約25MB)
 - [Macintosh用](#)
- DDBJ PipelineでPlatanus
 - Platanus ver. 1.2.2を実行したzip圧縮ファイル
- k-mer解析(k=1)

- DDBJ PipelineでPlatanus
 - Platanus ver. 1.2.2を実行したzip圧縮ファイル([platanusResult.zip](#); 約2.2MB)
- k-mer解析(k=1)

塩基ごとの出現頻度解析は、k-mer解析のk=1の場合に相当します。「イントロ | 一般 | k-mer解析 | k=1(塩基ごとの出現頻度解析) | [Biostrings](#)」の例題7は、以下と同じ。入力ファイルは [out_gapClosed.fa](#) (約2.4MB)

```
in_f <- "out_gapClosed.fa" #① 入力ファイル名を指定してin_fに格納
out_f <- "hoge7.txt" #出力ファイル名を指定してout_fに格納
param_base <- c("A", "C", "G", "T", "N") #② 出力させたい塩基を指定

#必要なパッケージをロード
library(Biostrings) #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

#本番
hoge <- alphabetFrequency(fasta) #A,C,G,T,..の数を各配列ごとにカウントした結果をhogeに格納
obj <- is.element(colnames(hoge), param_base) #条件を満たすかどうかを判定した結果をobjに格納
#out <- colSums(hoge[, obj]) #列ごとの総和をoutに格納
out <- apply(as.matrix(hoge[, obj]), 2, sum) #列ごとの総和をoutに格納

#ファイルに保存
write.table(out, out_f, sep="\t", append=F, quote=F, row.names=T, col.names=F) #tmp
```

一気に結果を得る
実際の利用時は、[rcodel.txt](#)のような無駄なコメントを除いてスリムにした一連のスク립トを作成しておき、一気にコピペする。ファイルの中身は以下と同じ。

```
param_base <- c("A", "C", "G", "T", "N") #出力させたい塩基を指定
library(Biostrings) #パッケージの読み込み
```

第1部の掟

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料の①スライド25あたりからやる予定です。スライド24あたりまで

- [講義資料PDF](#)(2016.06.08版; 約5MB)
- [hoge.zip](#)(2016.05.13版; 約2MB)

ここはそれほど重要ではありませんが念のため。①の②スライド45では、rcode1.txtというテキストファイルを開いてもらっています。このとき、エディタによっては改行コードが崩れていたり、OSの違いによってコピペ実行時に不具合が生じるかもしれませんので…



実際の利用時は、hogeフォルダ直下にある①rcode1.txtのように、無駄なコメントを除いてスリムにした一連のスクリプトを作成しておき、一気にコピペ

一気に結果を得る

```
rcode1.txt
param_base = c("A", "C", "G", "T", "N") #出力させたい塩基を指定
library(Biostrings) #パッケージの読み込み
#####
### Step 1:
#####
in_f <- "out_contig.fa" #入力ファイル名を指定してin_fに格納
out_f <- "result_step1.txt" #出力ファイル名を指定してout_fに格納
#####
fasta <- readDNASTringSet(in_f, format="fasta") #in_fで指定したファイルの読み込み
hoge <- alphabetFrequency(fasta) #A,C,G,T,...の数を各配列ごとにカウントした
obj <- is.element(colnames(hoge), param_base) #条件を満たすかどうかを判定した結果
out <- apply(as.matrix(hoge[, obj]), 2, sum) #列ごとの総和をoutに格納
write.table(out, out_f, sep="\t", append=F, quote=F, row.names=T, col.names=F) #
#####
### Step 2:
#####
in_f <- "out_scaffold.fa" #入力ファイル名を指定してin_fに格納
out_f <- "result_step2.txt" #出力ファイル名を指定してout_fに格納
#####
fasta <- readDNASTringSet(in_f, format="fasta") #in_fで指定したファイルの読み込み
hoge <- alphabetFrequency(fasta) #A,C,G,T,...の数を各配列ごとにカウントした
obj <- is.element(colnames(hoge), param_base) #条件を満たすかどうかを判定した結果
out <- apply(as.matrix(hoge[, obj]), 2, sum) #列ごとの総和をoutに格納
write.table(out, out_f, sep="\t", append=F, quote=F, row.names=T, col.names=F) #
#####
### Step 3:
#####
```

45

②

第1部の掟

念のため、②のところでもコピー用コードを提供していますので、必要に応じてご利用ください



第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料の①(スライド25あたりからやる予定です。スライド24あたりまでは自習。

- [講義資料PDF](#)(2016.06.08版; 約5MB)
- [hoge.zip](#)(2016.05.13版; 約2MB)

②

- 一気に結果を得る
実際の利用時は、[rcode1.txt](#)のような無駄なコメントを除いてスリムにした一連の скриптを作成しておき、一気にコピーする。ファイルの中身は以下と同じ。

実際の利用時は、[rcode1.txt](#)のようにした一連の скрипт

一気に結果を得る

```
rcode1.txt
param_base <- c("A", "C", "G", "T", "N") #出力させたい塩基を指定
library(Biostrings) #パッケージの読み込み

#####
### Step 1
#####
in_f <- "out_contig.fa" #入力ファイル名を指定してin_fに格納
out_f <- "result_step1.txt" #出力ファイル名を指定してout_fに格納

fasta <- readDNAStringSet(in_f, format="fasta") #in_fで指定したファイルの読み込み
hoge <- alphabetFrequency(fasta) #A,C,G,T,..の数を各配列ごとにカウントした結果
obj <- is.element(colnames(hoge), param_base) #条件を満たすかどうかを判定した結果
out <- apply(as.matrix(hoge[, obj]), 2, sum) #列ごとの総和をoutに格納
write.table(out, out_f, sep="\t", append=F, quote=F, row.names=T, col.names=F)

#####
### Step 2
#####
in_f <- "out_scaffold.fa" #入力ファイル名を指定してin_fに格納
out_f <- "result_step2.txt" #出力ファイル名を指定してout_fに格納

fasta <- readDNAStringSet(in_f, format="fasta") #in_fで指定したファイルの読み込み
hoge <- alphabetFrequency(fasta) #A,C,G,T,..の数を各配列ごとにカウントした結果
obj <- is.element(colnames(hoge), param_base) #条件を満たすかどうかを判定した結果
out <- apply(as.matrix(hoge[, obj]), 2, sum) #列ごとの総和をoutに格納
write.table(out, out_f, sep="\t", append=F, quote=F, row.names=T, col.names=F)

#####
### Step 3
#####
```

```
param_base <- c("A", "C", "G", "T", "N") #出力させたい塩基を指定
library(Biostrings) #パッケージの読み込み

#####
### Step 1
#####
in_f <- "out_contig.fa" #入力ファイル名を指定してin_fに格納
out_f <- "result_step1.txt" #出力ファイル名を指定してout_fに格納

fasta <- readDNAStringSet(in_f, format="fasta") #in_fで指定したファイルの読み込み
hoge <- alphabetFrequency(fasta) #A,C,G,T,..の数を各配列ごとにカウントした結果
obj <- is.element(colnames(hoge), param_base) #条件を満たすかどうかを判定した結果をo
out <- apply(as.matrix(hoge[, obj]), 2, sum) #列ごとの総和をoutに格納
write.table(out, out_f, sep="\t", append=F, quote=F, row.names=T, col.names=F)

#####
### Step 2
#####
in_f <- "out_scaffold.fa" #入力ファイル名を指定してin_fに格納
out_f <- "result_step2.txt" #出力ファイル名を指定してout_fに格納

fasta <- readDNAStringSet(in_f, format="fasta") #in_fで指定したファイルの読み込み
```

- 配列数
[rcode2.txt](#)は、配列数をカウントする必要最小限のコード。ファイルの中身は以下と同じ。

```
library(Biostrings) #パッケージの読み込み
```

第1部の掟

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料のスライド25あたりからやる予定です。スライド24あたりまでは自習。

- 講義資料PDF(2016.06.08版; 約5MB)
- hoge.zip(2016.05.13版; 約2MB)

赤枠で示すように、①スライドのタイトルで対応づけをある程度行えます。場所がよくわからなくなったら、「(Rで)塩基配列解析」のウェブサイト内で文字列検索してみてもいいでしょう

一気に結果を得る ①

実際の利用時は、rcode1.txtのような無駄なコメントを除いた一連のスクリプト

一気に結果を得る ①

実際の利用時は、rcode1.txtのような無駄なコメントを除いてスリムにした一連のスクリプトを作成しておき、一気にコピーする。ファイルの中身は以下と同じ。

```
param_base <- c("A", "C", "G", "T", "N")#出力させたい塩基を指定
library(Biostrings) #パッケージの読み込み

#####
### Step 1
#####
in_f <- "out_contig.fa" #入力ファイル名を指定してin_fに格納
out_f <- "result_step1.txt" #出力ファイル名を指定してout_fに格納

fasta <- readDNASTringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み
hoge <- alphabetFrequency(fasta) #A,C,G,T,..の数を各配列ごとにカウントした結果
obj <- is.element(colnames(hoge), param_base)#条件を満たすかどうかを判定した結果
out <- apply(as.matrix(hoge[, obj]), 2, sum)#列ごとの総和をoutに格納
write.table(out, out_f, sep="\t", append=F, quote=F, row.names=T, col.names=F)#o

#####
### Step 2
#####
in_f <- "out_scaffold.fa" #入力ファイル名を指定してin_fに格納
out_f <- "result_step2.txt" #出力ファイル名を指定してout_fに格納

fasta <- readDNASTringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み
hoge <- alphabetFrequency(fasta) #A,C,G,T,..の数を各配列ごとにカウントした結果
obj <- is.element(colnames(hoge), param_base)#条件を満たすかどうかを判定した結果
out <- apply(as.matrix(hoge[, obj]), 2, sum)#列ごとの総和をoutに格納
write.table(out, out_f, sep="\t", append=F, quote=F, row.names=T, col.names=F)#o

#####
### Step 3
#####
in_f <- "out_scaffold.fa" #入力ファイル名を指定してin_fに格納
out_f <- "result_step2.txt" #出力ファイル名を指定してout_fに格納

fasta <- readDNASTringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み
```

配列数

rcode2.txtは、配列数をカウントする必要最小限のコード。ファイルの中身は以下と同じ。

```
library(Biostrings) #パッケージの読み込み
```

第1部の掟

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料のスライド25あたりからやる予定です。スライド24あたりまでは自習。

- 講義資料PDF(2016.06.08版; 約5MB)
- hoge.zip(2016.05.13版; 約2MB)

①スライド50についても、②でコピー用コードを提供しています。この場合、タイトルが「配列数」と短いので、キーワード検索は厳しいかもしれませんが、③の次の内容なので、普通にページを下にスクロールすれば見つけれられます

③ 一気に結果を得る
実際の利用時は、rcode2.txtのような無駄なコメントを除いてスリムにした一連の скриプトを作成しておき、一気にコピーする。ファイルの中身は以下と同じ。

```
param_base <- c("A", "C", "G", "T", "N")#出力させたい塩基を指定
library(Biostrings) #パッケージの読み込み

#####
### Step 1
#####
in_f <- "out_contig.fa" #入力ファイル名を指定してin_fに格納
out_f <- "result_step1.txt" #出力ファイル名を指定してout_fに格納

fasta <- readDNASTringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み
hoge <- alphabetFrequency(fasta) #A,C,G,T,..の数を各配列ごとにカウントした結果
obj <- is.element(colnames(hoge), param_base)#条件を満たすかどうかを判定した結果をo
out <- apply(as.matrix(hoge[, obj]), 2, sum)#列ごとの総和をoutに格納
write.table(out, out_f, sep="\t", append=F, quote=F, row.names=T, col.names=F)#o

#####
### Step 2
#####
in_f <- "out_scaffold.fa" #入力ファイル名を指定してin_fに格納
out_f <- "result_step2.txt" #出力ファイル名を指定してout_fに格納

fasta <- readDNASTringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み
```

② 配列数
rcode2.txtは、配列数をカウントする必要最小限のコード。ファイルの中身は以下と同じ。

```
library(Biostrings) #パッケージの読み込み
```

配列数は、①Step1 → Step2で減らそうと予想。③(hogeフォルダ直下とする必要最小限のコード。349-

① 配列数

```
library(Biostrings) #パッケージの読み込み

### Step 1 ###
in_f <- "out_contig.fa" #入力ファイル名を指定してin_fに格納
fasta <- readDNASTringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み
length(fasta) #配列数をカウント

### Step 2 ###
in_f <- "out_scaffold.fa" #入力ファイル名を指定してin_fに格納
fasta <- readDNASTringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み
length(fasta) #配列数をカウント

### Step 3 ###
in_f <- "out_gapClosed.fa" #入力ファイル名を指定してin_fに格納
fasta <- readDNASTringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み
length(fasta) #配列数をカウント
```

```
> getwd()
[1] "C:/Users/kad"
> list.files(patt
[1] "out_contig.f
[3] "out_gapClose
[5] "out_scaffold
> ### Step 1 ##
> in_f <- "out_co
> fasta <- readDN
> length(fasta)
[1] 349
>
> ### Step 2 ##
> in_f <- "out_sc
> fasta <- readDN
> length(fasta)
[1] 117
>
> ### Step 3 ##
> in_f <- "out_ga
> fasta <- readDN
> length(fasta)
[1] 117
```

第1部の掟

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料の①(1)スライド25あたりからやる予定です。スライド24あたりまでは自習。

- [講義資料PDF](#)(2016.06.08版; 約5MB)
- [hoge.zip](#)(2016.05.13版; 約2MB)

掟に従って、ページをスクロールしていかないと…、①の②スライド120のところで必ずつまづきます。なぜならここでコピペしてもらったコードは、どこか他のところに例題があるわけではなく

k=6, 8, 10, 12

k値の違いによるk-mer出現頻度分布の傾向を把握。①k=6のときの②k-mer出現回数の中央値(median)は8。③最多出現回数は39回

```
k=6, 8, 10, 12
必要最小限のコードにして、k値の違い(k=6, 8, 10, and 12)による影響の全体像を把握。

in_f <- "sample34.ngs.fasta" #入力ファイル名を指定してin_fに格納
library(Biostrings) #パッケージの読み込み
fasta <- readDNAStringSet(in_f, format="fasta") #in_fで指定したファイルの読み込み

param_kmer <- 6
hoge <- oligonucleotideFrequency(fasta, width=param_k$)
out <- colSums(hoge)
kmer <- out[out > 0]
length(kmer)
table(kmer)
median(kmer)

param_kmer <- 8
hoge <- oligonucleotideFrequency(fasta, width=param_k$)
out <- colSums(hoge)
kmer <- out[out > 0]
length(kmer)
table(kmer)
median(kmer)

param_kmer <- 10
hoge <- oligonucleotideFrequency(fasta, width=param_k$)
out <- colSums(hoge)
kmer <- out[out > 0]
length(kmer)
table(kmer)
median(kmer)

R Console
> in_f <- "sample34.ngs.fasta" #入力ファイル名
> library(Biostrings) #パッケージの読み込み
> fasta <- readDNAStringSet(in_f, format="fasta") #in_f$
>
> param_kmer <- 6 #k-merのkの値
> hoge <- oligonucleotideFrequency(fasta, width=param_k$) #列ごとの総和
> out <- colSums(hoge) #1回以上出現し
> kmer <- out[out > 0] #1回以上出現し
> length(kmer)
[1] 862
> table(kmer) #k-merの種類
kmer
 2  3  4  5  6  7  8  9 10 11 12 13 14 15
 5 10 36 74 109 134 146 112 68 43 29 11 17 11
16 17 18 19 20 21 22 24 26 27 28 29 30 39
13 7 10 4 6 4 4 1 2 2 1 1 1 1
> median(kmer) #出現回数の中
[1] 8
```

120



第1部の掟

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料のダウンロード25あたりからやる予定です。

- [講義資料PDF](#)(2016.06.08版; 約5MB)
- [hoge.zip](#)(2016.05.13版; 約2MB)



- k=6, 8, 10, 12
必要最小限のコードにして、k値の違い(k=6, 8, 10, and 12)による影響の全体像を把握。

k=6, 8, 10, 12

k値の違いを把握。①中央値(median)

```

k=6, 8, 10, 12
必要最小限のコードにして、k値の違い(k=6, 8, 10, and 12)による影響の全体像を把握。

in_f <- "sample34_ngs.fasta" #入力ファイル名を指定してin_fに格納
library(Biostrings) #パッケージの読み込み
fasta <- readDNAStringSet(in_f, format="fasta") #in_fで指定したファイルの読み込み

param_kmer <- 6
hoge <- oligonucleotideFrequency(fasta, width=param_kmer) #k連続塩基の出現頻度情報
out <- colSums(hoge) #列ごとの総和をoutに格納
kmer <- out[out > 0] #1回以上出現したk-merのみをkmerに格納
length(kmer) #1回以上出現したk-merの種類数を表示
table(kmer) # (k-merの種類は問わずに) k-merの出現回数分布
median(kmer) #出現回数の中央値(median)を表示

param_kmer <- 8
hoge <- oligonucleotideFrequency(fasta, width=param_kmer) #k連続塩基の出現頻度情報
out <- colSums(hoge) #列ごとの総和をoutに格納
kmer <- out[out > 0] #1回以上出現したk-merのみをkmerに格納
length(kmer) #1回以上出現したk-merの種類数を表示
table(kmer) # (k-merの種類は問わずに) k-merの出現回数分布
median(kmer) #出現回数の中央値(median)を表示

param_kmer <- 10
hoge <- oligonucleotideFrequency(fasta, width=param_kmer) #k連続塩基の出現頻度情報
out <- colSums(hoge) #列ごとの総和をoutに格納
kmer <- out[out > 0] #1回以上出現したk-merのみをkmerに格納
length(kmer) #1回以上出現したk-merの種類数を表示
table(kmer) # (k-merの種類は問わずに) k-merの出現回数分布
median(kmer) #出現回数の中央値(median)を表示
    
```

Jul 20 2016, NGS/ハンズオン講習会

```

in_f <- "sample34_ngs.fasta" #入力ファイル名を指定してin_fに格納
library(Biostrings) #パッケージの読み込み
fasta <- readDNAStringSet(in_f, format="fasta") #in_fで指定したファイルの読み込み

param_kmer <- 6 #k-merのkの値を指定
hoge <- oligonucleotideFrequency(fasta, width=param_kmer) #k連続塩基の出現頻度情報
out <- colSums(hoge) #列ごとの総和をoutに格納
kmer <- out[out > 0] #1回以上出現したk-merのみをkmerに格納
length(kmer) #1回以上出現したk-merの種類数を表示
table(kmer) # (k-merの種類は問わずに) k-merの出現回数分布
median(kmer) #出現回数の中央値(median)を表示

param_kmer <- 8 #k-merのkの値を指定
hoge <- oligonucleotideFrequency(fasta, width=param_kmer) #k連続塩基の出現頻度情報
out <- colSums(hoge) #列ごとの総和をoutに格納
kmer <- out[out > 0] #1回以上出現したk-merのみをkmerに格納
length(kmer) #1回以上出現したk-merの種類数を表示
table(kmer) # (k-merの種類は問わずに) k-merの出現回数分布
median(kmer) #出現回数の中央値(median)を表示

param_kmer <- 10 #k-merのkの値を指定
hoge <- oligonucleotideFrequency(fasta, width=param_kmer) #k連続塩基の出現頻度情報
out <- colSums(hoge) #列ごとの総和をoutに格納
kmer <- out[out > 0] #1回以上出現したk-merのみをkmerに格納
length(kmer) #1回以上出現したk-merの種類数を表示
table(kmer) # (k-merの種類は問わずに) k-merの出現回数分布
median(kmer) #出現回数の中央値(median)を表示
    
```

- 作図(k=6, 8, 10)
上記のk=6, 8, 10の結果より、横軸の最大値は39であることがわかっているが、ヒストグラム全体の変遷が把握しづらいので1から20の範囲を指定している。このあたりは結果を眺めながら自分の好みに合わせて微調整する。

```

in_f <- "sample34_ngs.fasta" #入力ファイル名を指定してin_fに格納
param_fig <- c(150, 210) #ヒストグラム描画時の横幅と縦幅を指定(単位はpx)
    
```


第1部の掟

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料の①スライド25あたりからやる予定です。スライド24あたりまでは自習。

- [講義資料PDF](#)(2016.06.08版; 約5MB)
- [hoge.zip](#)(2016.05.13版; 約2MB)

次は、作業ディレクトリについてです。②スライド54からは、どこで作業するかは明記していません。しかし、`hoge4.fa`がどこに存在するかについては、自分で適切に判断しましょう。デフォルトはhogeフォルダ直下にありますので、作業ディレクトリの変更をするなり、現在の作業ディレクトリ中に`hoge4.fa`をコピーするなりしましょう。例えばhogeフォルダ内にあるplatanusResultフォルダ中で作業を行いたいのなら、そこにはデフォルトでは`hoge4.fa`はないので、自分でコピーすればいいでしょ、ということです。

②平成27年度NGSハンズオン講習会(2015.07.29)のスライド94~)で用いたhoge4.faを入力として、配列ごと(この場合コンティグごと)に16種類の2連続塩基の出現頻度解析の概念を説明します

イントロ 一般 塩基配列(reverse.complement)を取得 (last modified 2013/06/14)
イントロ 一般 塩基(reverse)を取得 (last modified 2013/06/14)
イントロ 一般 k-mer解析(k=1)(塩基ごとの出現頻度解析) Biostrings (last modified 2016/04/27) NEW
イントロ 一般 k-mer解析(k=2)(2連続塩基の出現頻度解析) Biostrings (last modified 2016/01/28)
イントロ 一般 k-mer解析(k=3)(3連続塩基の出現頻度解析) Biostrings (last modified 2016/01/28)
イントロ 一般 k-mer解析(k=n)(n連続塩基の出現頻度解析) Biostrings (last modified 2016/01/28)
(毎時予定)イントロ 一般 2連続塩基の出現頻度解析 Biostrings (last modified 2016/01/28)
(毎時予定)イントロ 一般 塩基配列解析 Biostrings (last modified 2016/01/28)
(毎時予定)イントロ 一般 塩基配列解析 Biostrings (last modified 2016/01/28)

イントロ | 一般 | [k-mer解析 | k=2\(2連続塩基の出現頻度解析\)](#) | [Biostrings](#)

[Biostrings](#) パッケージを用いて、multi-FASTA形式ファイルを読み込んで、"AA", "AC", "AG", "AT", "CA", "CC", "CG", "CT", "GA", "GC", "GG", "GT", "TA", "TC", "TG", "TT" の計4 = 16通りの2連続塩基の出現頻度を調べるやり方を示します。k-mer解析のk=2の場合に相当します。ヒトゲノムでCGの割合が割合よりも低い(Lander et al., 2001; Saxena et al., 2006)ですが、それを簡単に検証できます。

「ファイル」ディレクトリの変更で解析したいファイルを置いてあるディレクトリに移動し以下をコピー。

```
1. イントロ | 一般 | ランダムな塩基配列を作成の4を実行して得られたmulti-FASTAファイル(hoge4.fa)の場合:  
タイトル通りの出現頻度です。  
  
in_f <- "hoge4.fa"  
out_f <- "hoge1.txt"  
  
#必要なパッケージをロード  
library(Biostrings)  
  
#入力ファイルの読み込み  
fasta <- readDNAStringSet(in_f, format="fasta")  
  
#本番  
out <- dinucleotideFrequency(fasta)  
  
#ファイルに保存  
tmp <- cbind(names(fasta), out)  
write.table(tmp, out_f, sep="\t", append=TRUE)
```

```
>contig_1  
CGGACAGCTCCTCGGCATCCGGAT  
>contig_2  
GTCTGCCTCAAGCGCCCCAAGTGGGTTTGGAGGCTAACATCGCAAGTGG  
ACACTCAGTCCGGCCGTCTGGTTGGCAGGGGCAGAGACCCAGCACACCTT  
GTC  
>contig_3  
TGTAAGAGAAGGCGGTATCAGCGTCCACTTACACGATCCGTTACTAATT  
GTATGAGGTCGGSCA  
>contig_4  
CGTGCATGATCCACACAGCAGTAAACCGGACCTCTACCTATGAACATG
```



54

第1部の掟

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料の①スライド25あたりからやる予定です。スライド24あたりまでは自習。

- [講義資料PDF](#)(2016.06.08版; 約5MB)
- [hoge.zip](#)(2016.05.13版; 約2MB)

続、作業ディレクトリ。②スライド72のところは、入力ファイルはありません。出力ファイルが生成されるだけです。つまり自分がどこで作業を行っているかを適切に把握できていればそれでいいのです

ランダムデータ生成

この後に用いるランダムな塩基配列、および仮想NGSデータの作成について解説します。①サンプルデータの、②例題32

サンプルデータ

1. Illumina 36bp single-end human (SRA000289) data (Marioni et al., Genome Res., 2008)

「Kidney 7 samples vs Liver 7 samples」のRNA-seqの塩基配列データ(Supplementary Table 1.txt)です。サンプルは二つの条件(1.5 nM and 1 nM)でシーケンスされており、15 nMのものか5 samples vs 5 samples、1.5 nMのものか5 samples vs 5 samplesのいずれかです。

32. k-mer解析用のランダム配列から生成したFASTA形式ファイル(sample32_ref.fastaとsample32_nes.fasta)です。k-merの長さのリファレンス配列を生成したのち、k-mer長の部分配列を1000回ランダム抽出したものです。塩基の存在比はAが22%、Cが38%、Gが38%、Tが22%になっています。リファレンス配列(仮想ゲノム配列)がsample32_ref.fastaで、1000回ランダム抽出した仮想NGSデータがsample32_nes.fastaです。リード長20塩基で1000回抽出したデータは200塩基となり、50塩基からなる元のゲノム配列の4倍シーケンスしていることになります(4x coverageに相当)。入力したNGS配列形式はシングルシーケンスデータ(ランダムな塩基配列)の生成から基本的に同じです。

```
out_f1 <- "sample32_ref.fasta" #出力ファイル名を指定してout_f1に格納
out_f2 <- "sample32_nes.fasta" #出力ファイル名を指定してout_f2に格納
param_len_ref <- 50 #リファレンス配列の長さを指定
naranbi <- c("A", "C", "G", "T") #以下の数値指定時にACGTの並びを間違えないようにする
param_composition <- c(22, 38, 38, 22) #("A", "C", "G", "T"の並びで)各塩基の存在比率を指定
param_len_ngs <- 20 #リード長を指定
param_num_ngs <- 10 #リード数を指定
param_desc <- "kkk" #FASTA形式ファイルのdescription行に記載する内容
```

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

#本番(リファレンス配列生成)
set.seed(1010) #おまじない(同じ乱数になるようにするため)
ACGTset <- rep(naranbi, param_composition) #naranbi中の塩基がparam_compositionで指定した数
```

72

