データ解析と可視化

FASTA形式

">"から始まるヘッダーと、その次の行から始まる塩基あるいはアミノ酸配列の 1 文字表記で記されたファイル形式。

wiggle形式

遺伝子座ごとのタンパク質などの結合スコアが記録されているファイルの形式。 variableStepとfixedStepの二種類があり、variableStepは遺伝子座の具体的な位置とスコアが連ねて記録してある。一方、fixedStepは特定の位置から一定の間隔ごとのスコアが記録してある。

bed形式

wiggle形式同様に結合スコアなどが記録されていたり、遺伝子座の位置が記録されている形式。 3またはそれ以上のカラムからなり、

最初の３列に記載する情報は、 1.染色体の名前（chr1など) 2.リードや遺伝子のスタートポジション（ポジションは1でなく0スタート） 3.リードや遺伝子のエンドポジション である。 それ以降は、

例

variableStep

```
#variableStep chrom=chr2
300701 12.5
300702 12.5
300703 12.5
300704 12.5
300705 12.5
```

fixedStep

```
#fixedStep chrom=chr3 start=400601 step=100
11
22
33
```

なお、上記のfixedStepのファイルをvariableStepに変換すると、

```
#variableStep chrom=chr3
400601 11
400701 22
400801 33
```

となる。

Encyclopedia of DNA Elements(ENCODE : https://www.encodeproject.org/)にあるwiggle形式のファイルの多くはbigwigというバイナリ形式に圧縮されているので、BigWigToWig(https://www.encodeproject.org/software/bigwigtowig/)で染色体番号ごとにwiggle形式に変換しなければならない。 BigWigToWigと、以下のシェルスクリプトファイルを同じフォルダ内に用意して１つ目の引数にbigwigファイルの拡張子(.bigwig)を外したものを入れれば変換できる。

```
for i in `seq 1 22`
do
./bigWigToWig -chrom=chr$i $1.bigWig ${1}chr$i.wig
done
```

wiggleファイルの解析

本来、wiggleファイルはIntegrative Genomics Viewer(IGV : http://software.broadinstitute.org/software/igv/)で可視化することでゲノムワイドなスコア分布を見ることができる。 しかし、ここでは数値解析を行おうと思う。

松島が行ったのはある特徴を持つ領域の選出と、その周辺の分布の計算である。

特徴領域の選出

MNaseデータからヌクレオソームが結合していない配列を選出した。 MNaseデータは https://www.encodeproject.org/files/ENCFF000VME/ を用いた。 ここでの作業の目的は150塩基対以上スコアが０の領域を抜き出すことである。 このMNaseデータはbed形式でかかれたものがwiggle形式で保存されたものである。

```
#bedGraph section chr10:60441-79056
chr10   60441   60442   1.3
chr10   60442   60443   1.5
chr10   60443   60444   1.6
chr10   60444   60445   1.8
chr10   60445   60446   1.9
chr10   60446   60447   2.1
chr10   60447   60448   2.2
chr10   60448   60449   2.3
chr10   60449   60450   2.5
```

このような形になっている。 #から始まるヘッダー部分の３列目にデータの内訳が書いてる。 この場合、10番染色体(chr10)の60441塩基から79056塩基まで(60441-79056)における結合スコアが存在していることを表している。 データ部分の1行目は染色体番号、2番目はスコア計測領域のスタート位置３番目はスコア計測領域の終了位置、4番目がスコアである。 実験プロトコルにも書いてあるが、