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。

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