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、
4.名前
5.1-1000のスコア
6.リードや遺伝子の向き（+/-)
7.CDSのスタートポジション。リードなら２の座標と同じになる。
8.CDSのエンドポジション。
9.exonの数
10.各exonのサイズ（数値をコンマで区切り全て記載する）
11.exonのスタート位置。
となっている。

ちなみに、ENCODEにあるbigwigファイルはwiggle形式にみかけたbed形式が多い。紛らわしい。

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。

bed形式

```
#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
```

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。

なお、ENCODEのデータファイルのダウンロード方法は、
1.「Files」にある「File details」のタブを選択
2.「Processed data」から「File type」が「bigwig」になっているものをダウンロード
(Safariだと設定によってはダウンロード中にスリープになると中断されることがあるので設定を確認するかwgetコマンドやcurlコマンドでダウンロードすること)

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、スコアが0のものは、マッピング可能な遺伝子座であるがシーケンスのリードシグナルがないことを意味している。
したがって、スコア0の位置をヌクレオソーム排他的な領域であると見做すことにする。

上記の例だと、1bpごとのスコアが出ているので領域が見やすくなっているが、

```
#bedGraph section chr1:0-840194
chr1    0       56898   0
chr1    56898   56998   1.02889
chr1    56998   91465   0
chr1    91465   91468   1.02889
chr1    91468   91469   2.05778
```

```
chr1    91469   91476   3.08639
chr1    91476   91487   4.11528
chr1    91487   91565   1.32123
chr1    91565   91568   0.9909
```

のような形の場合、スコアが記録されている遺伝子座の長さが一定でない。
したがって、スタート位置と終了位置が繋がっているのか、繋がっていない場合はどう処理するのかということを念頭に置いて処理をする必要がある。

いくつか方針があるが、最も簡単なものはwiggleデータ(今回はbedデータだが)をわかりやすい形式に書き換える方法である。
つまり、上記のデータを次にように書き換えるのである。(ヘッダーの下の行から６、７行目)

```
#bedGraph section chr1:91469-91487
chr1    91470   3.08639
chr1    91471   3.08639
chr1    91472   3.08639
chr1    91473   3.08639
chr1    91474   3.08639
chr1    91475   3.08639
chr1    91476   4.11528
chr1    91477   4.11528
chr1    91478   4.11528
chr1    91479   4.11528
chr1    91480   4.11528
chr1    91481   4.11528
chr1    91482   4.11528
chr1    91483   4.11528
chr1    91484   4.11528
chr1    91485   4.11528
chr1    91486   4.11528
```

このように書き換えることによって1bp単位でのデータの処理ができるようになる。
しかし、一般的にデータ容量が増加するためストレージの空き具合と相談する必要がある。

このような処理はshellやawkを使用するのが早いかもしれない。いわゆるシェル芸というやつである。

```
cat test.wig | grep -Gv section | awk '{for (i = 0; i < $3-$2; i++) print
$1" "$2+i" "$4;}' > test.data
```

(こういう作業がサクッとできるようになると色々と手間が省けて効率が上がる。perlワンライナー、sed、awk、grepといったshellコマンド系をいじっておくと良い)

ここから数行連続してスコアが特定の範囲内にある箇所を抜き出せばいい。 これに関しては課題として残しておこう。 抜き出した場所はbed形式で管理できるようにすると後々便利である。

特定した領域の周辺分布の計算

特別難しい作業ではない。前の章の課題がクリアできたらできるはず。
bed形式で保存した遺伝子座の上流下流数bp〜数kbpの領域における、別のbed形式におけるスコアの平均を出したりすればいい。