

## 解析方針

### 転写効率におけるコアプロモータの重要性

- ・コアプロモータを改良することで、従来の高発現プロモータをさらに効率化することができる。
- ・コアプロモータは転写開始点を決定することで、転写されるmRNAの配列を決定する。
- ・導入遺伝子高発現系を最適化していく場合、1塩基のずれもなく意図した通りにmRNAが転写される必要があるが、コアプロモーターの配列によっては、転写開始点が複数に分散し、意図したものとは異なるmRNAが一定の比率で転写される。
- ・従来の高発現プロモーターでは、このように高度な設計が施されることを想定していないため、転写開始点の分散には注意を払う必要がある。
- ・そもそも分散型のプロモーターを使用している場合や、本来は収束型であるにも関わらず、発現力セットの操作を行う際に、適切なコアプロモーターを破壊してしまった場合などでは、転写開始点の分散や発現能力の低下が生じる。

### 内在性遺伝子の多くは、分散型のコアプロモーター

収束に関わるコアプロモーター要素を持たない場合や、複数個所に収束に関わるコアプロモーター要素が存在する場合などでは、基本転写因子群が結合する位置が1か所に限定されないため、認識される様々な位置から転写が行われる。

転写効率が高く、転写開始点が1塩基単位で厳密に収束しているコアプロモータに共通する配列の特徴を探る

#### 参考文献 :

<https://bsw3.naist.jp/ko-kato/research.html>

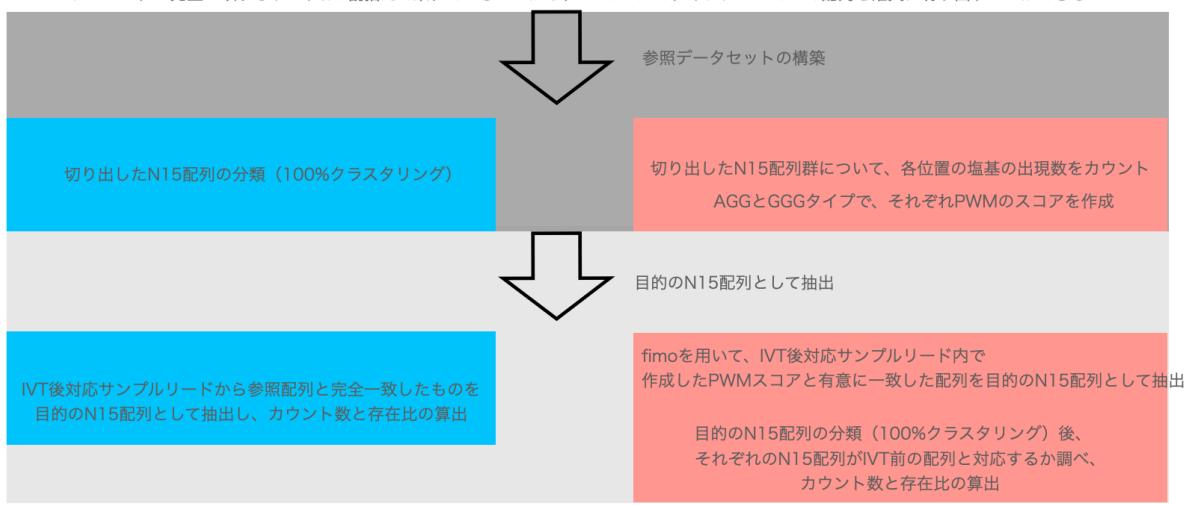
### (1) リード処理

リードトリムの方法

IVT前後の対応関係に着目したリード処理

- ・IVT前のN15配列を、IVT後のサンプルリードにおけるN15配列の切り出し時の参照配列とする
- ・この時、n=2間の和集合を考える

IVT前（サンプルAとB）のリードにおいて、  
T7プロモータと完全一致するリードは9割拾えて来れていることから、AGGとGGGタイプについてN15配列を確実に切り出すことができる



## (1.5) k-merの範囲の絞り込み

5-mer -> 8 merと変化させた時にIVT前後で分布がズれていく様子が観察できれば良い

この時、IVT前で正規分布がかけることが前提条件となる

まずは、IVT前でJellyfish countによる最適なk-merを以下の方法で探索

(1) X回出現するユニークなk-mer数の最大値を求める

(2) 異なるk-mer間の最大値同士を比較し、その中でも最大なる時のk-merをベストk-merと定義

## k-merプロファイリングの活用

(1) 配列検索のフィルタリング

(2) de Bruijn グラフによるゲノムアセンブリ

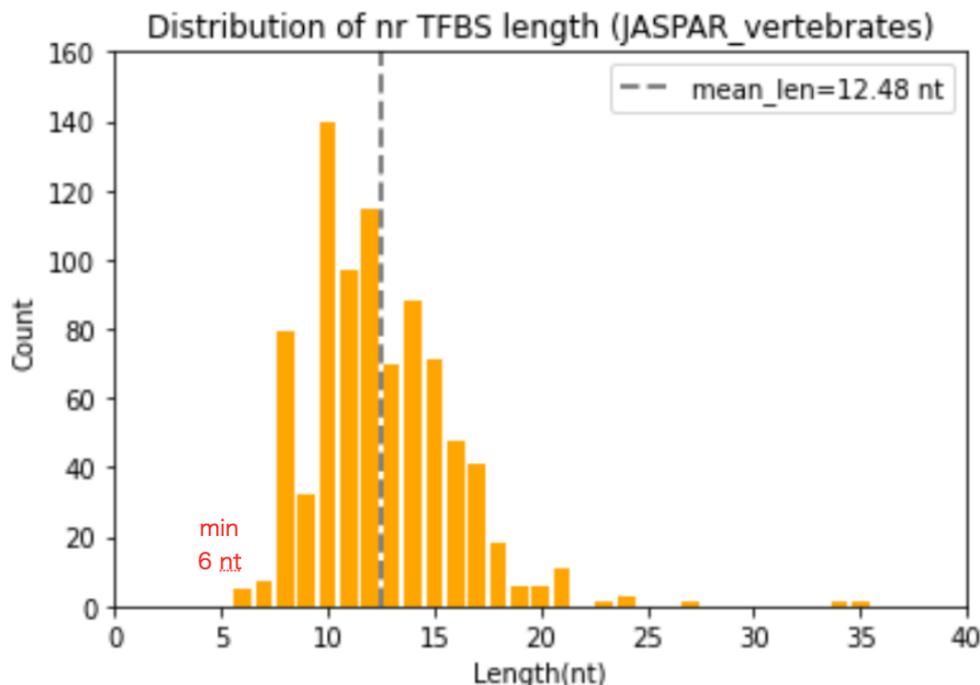
(3) 種の判別

## k-mer解析ツール

・k-merカウントツール：Squeakr

・kat hist ・jellyfish count、histo

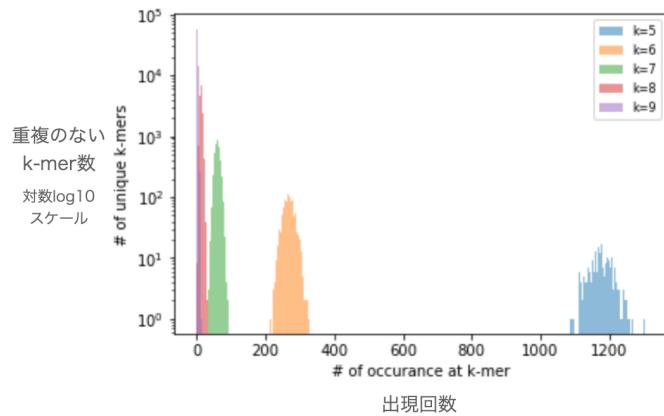
## 参考：TFBS長の分布



## 参考：k-mer頻度の理論分布

## ・データセットn=100,000の時

k-mer頻度解析における理論分布 (n = 100,000)



各配列に対して、  
k=5, 6, 7, 8, 9と変化させた時の  
k-mer頻度を調べた



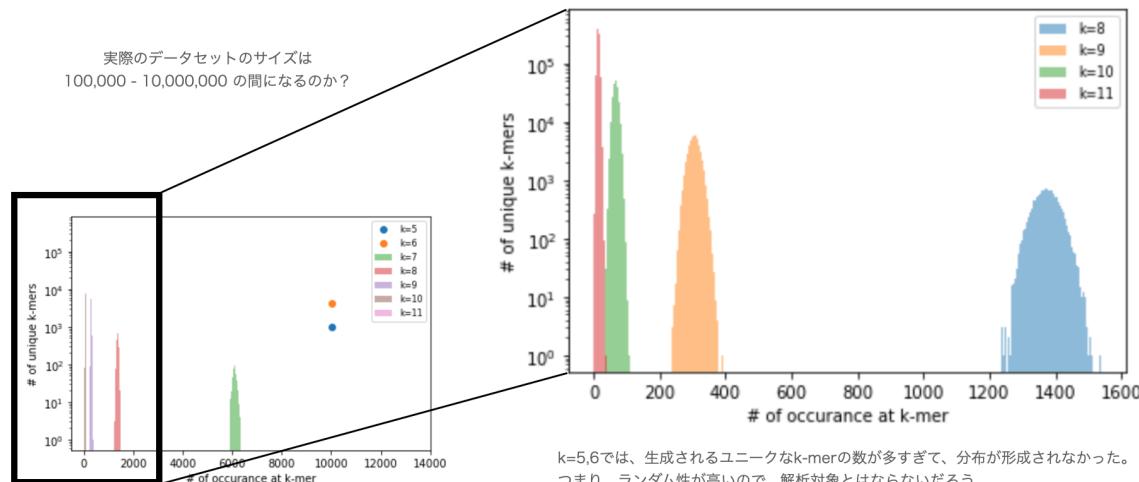
kの値が小さくなるにつれて、出現回数が増加 (理論分布)

ATGC組成に偏りのある配列が複数存在するならば、  
出現回数の分布が右側にシフト？

実際のデータセットのサイズは100,000 - 10,000,000 の間になるのか？

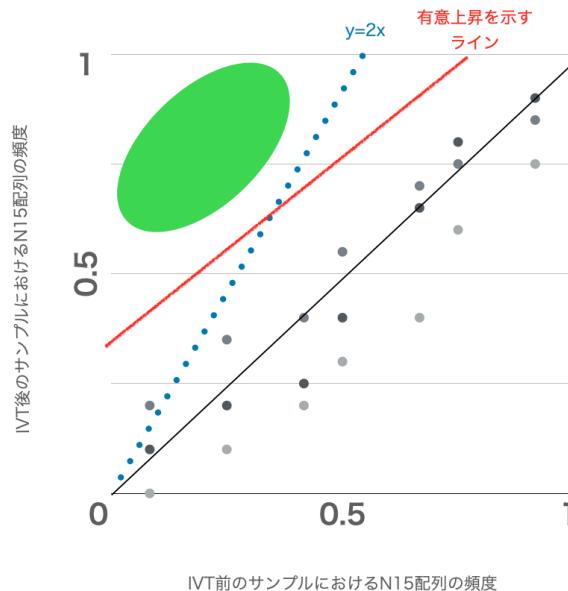
## ・データセットn=10,000,000の時

k-mer頻度解析における理論分布 (n = 10,000,000)



## (2) 目的配列の抽出

高い効率をもたらす転写制御配列グループの決定



カウント数において

IVT前 : 3 → IVT後 : 6は確かに2倍の増加を示すが、前後で総カウント数が異なるれば、増加しているかどうかはわからない

例えば、前後で総カウント数が異なる時、

IVT前 : 3/100 → IVT後 : 6/200は同じ

ただし、IVT後で、12/200ならば、2倍の増加を示す

→ yとx軸は配列の出現頻度とする

クラスター距離間の乖離から境界線を決定するようなSVMは、生物学的な説明がしづらいのでSVMの使用はできるだけ避けたい

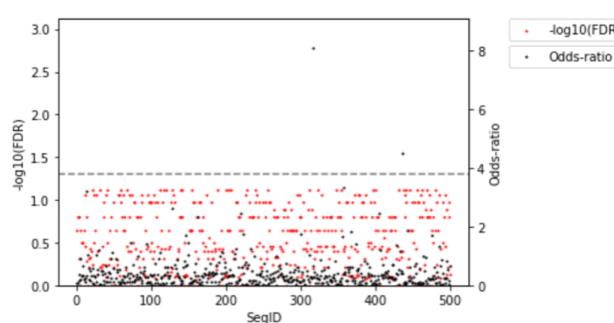
高い効率をもたらす転写制御配列の決定境界領域を  
2群の比率の差の検定による有意水準ライン(上昇)の上と定義

または

高い効率をもたらす転写制御配列の決定境界領域を  
2群の比率の差の検定による有意水準ラインと変動2倍のラインで  
囲まれた領域内と定義

散布図は以下のように修正

x軸: SeqID、y1軸:-log10(FDR)、y2軸:オッズ比とした時の2軸プロット



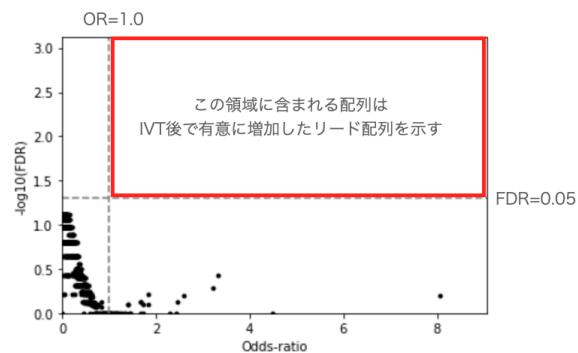
メリット:

x軸はクラスタリング順に配列がソートされているので、  
配列内の塩基組成の傾向がイメージできる

デメリット:

複雑

x軸: オッズ比、y軸:  $-\log_{10}(FDR)$ とした時の散布図



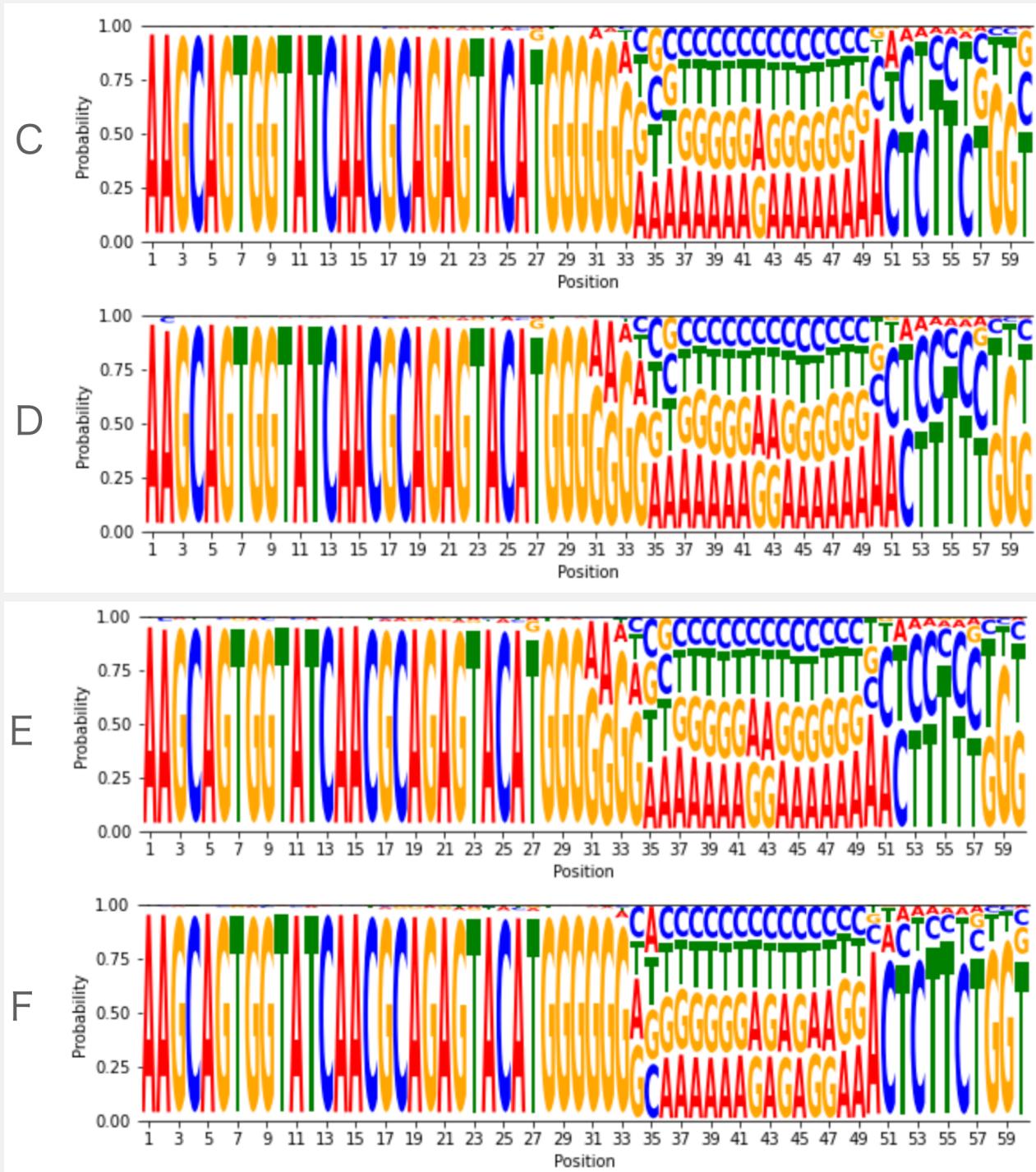
メリット:

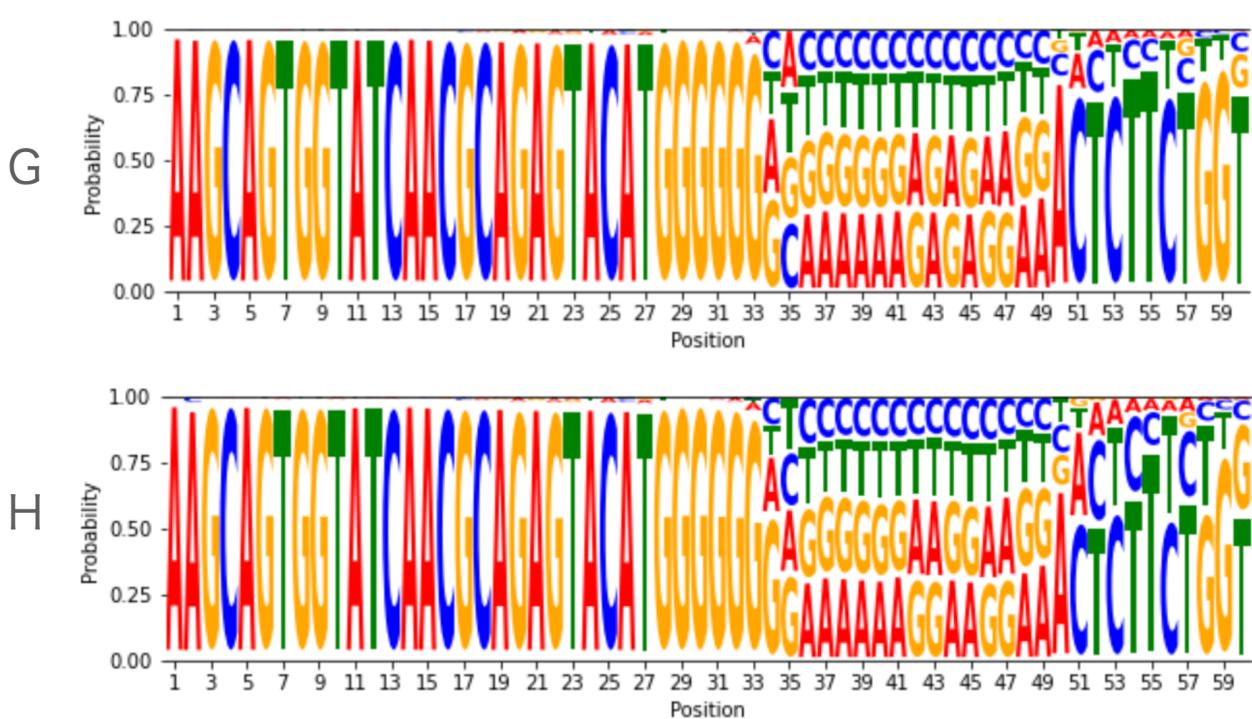
散布図として成立しており、判断基準を可視化できる

デメリット:

配列内の塩基組成がイメージできない

## ライブラリー構造





懸念点: 転写開始点のズレが、タンパク質翻訳に影響を与える

読み枠のずれでNMDを与えるような異常なタンパク質を産出することを考えると、5'UTR開始点のズれを考慮すべきだ

NGSライブリヤーの各塩基位置で出現頻度が最大値をとる塩基から構成されるDNA断片を土台として、5'UTR領域の開始位置を一時的に決定

IVTにおけるtemplate switchで塩基の挿入、また欠失?が生じると仮定して、一時的に決定した5'UTR開始位置から、±3 bpのズれを考えてみる

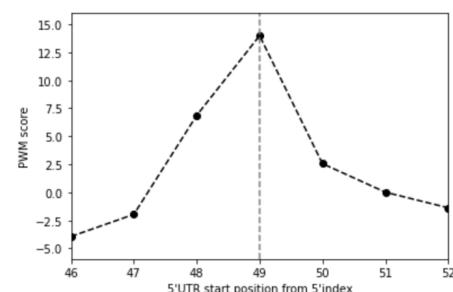
前後の6通りのPWMスコアを調べてみる

Refとずれても、共通する塩基のポジションを把握する

共通する塩基の出現頻度で、できる限りリードをrescueできるかも？

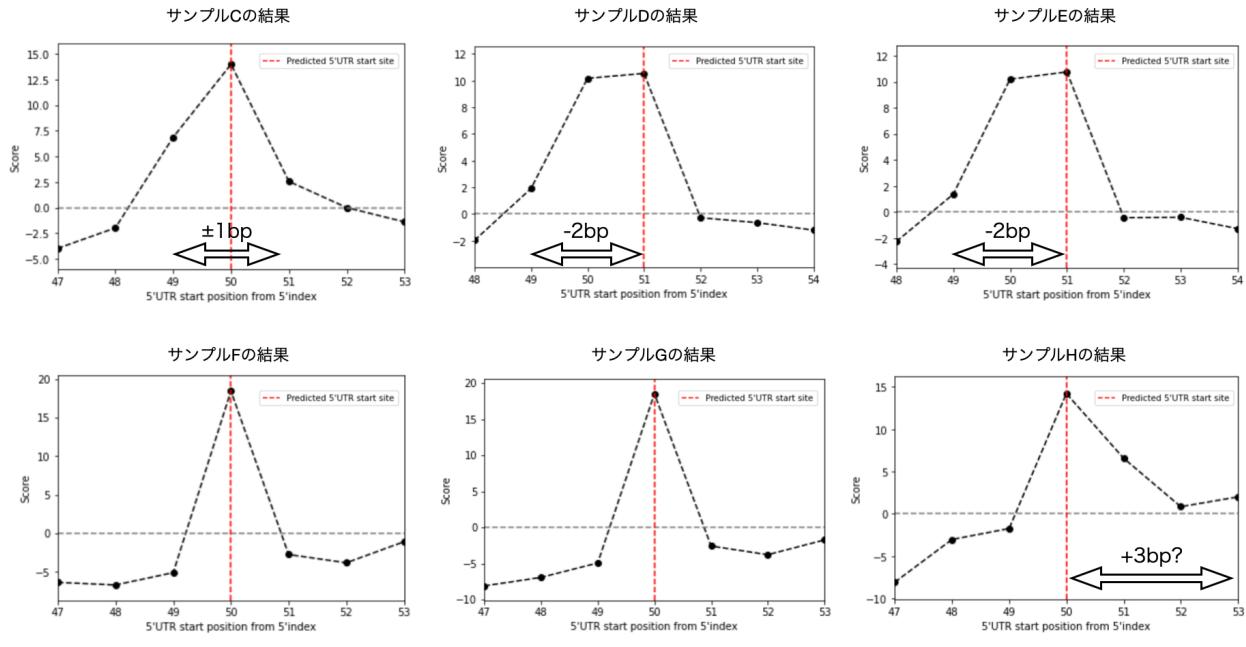
サンプルCの結果

対象となる配列アラインメント中の各位置における文字の出現頻度に基づいて類似性スコアを定義。  
スコアは、目的の配列がランダム配列との程度異なるかを示す。  
スコアの意味  
・ 目的の配列である確率とランダムサイトである確率が同じである場合、0。  
・ ランダムサイトよりも目的サイトである可能性が高い場合、0より大きい。  
・ 目的サイトよりもランダムサイトである可能性が高い場合、0よりも小さい。



TODO: +側で、ライブラリー構造のATGC比率の調査領域を+3以上伸ばす。

修正点：+側で、ライブラリー構造のATGC比率の調査領域を+3以上伸ばす。+側では、score算出時の要素数が2-3個少なく、過大・小評価している可能性がある



### (3) 目的配列内でのAGCT比率に偏りがある領域の同定

高い効率をもたらす転写制御配列内で、AGCT比率に偏りのある領域の同定

配列数が少ない時

- 配列グループ内で、エンリッチメント解析し、シャノンの情報量から同定
- マルチプルアラインメントで配列クラスタリング解析

配列数が多いと少ない時

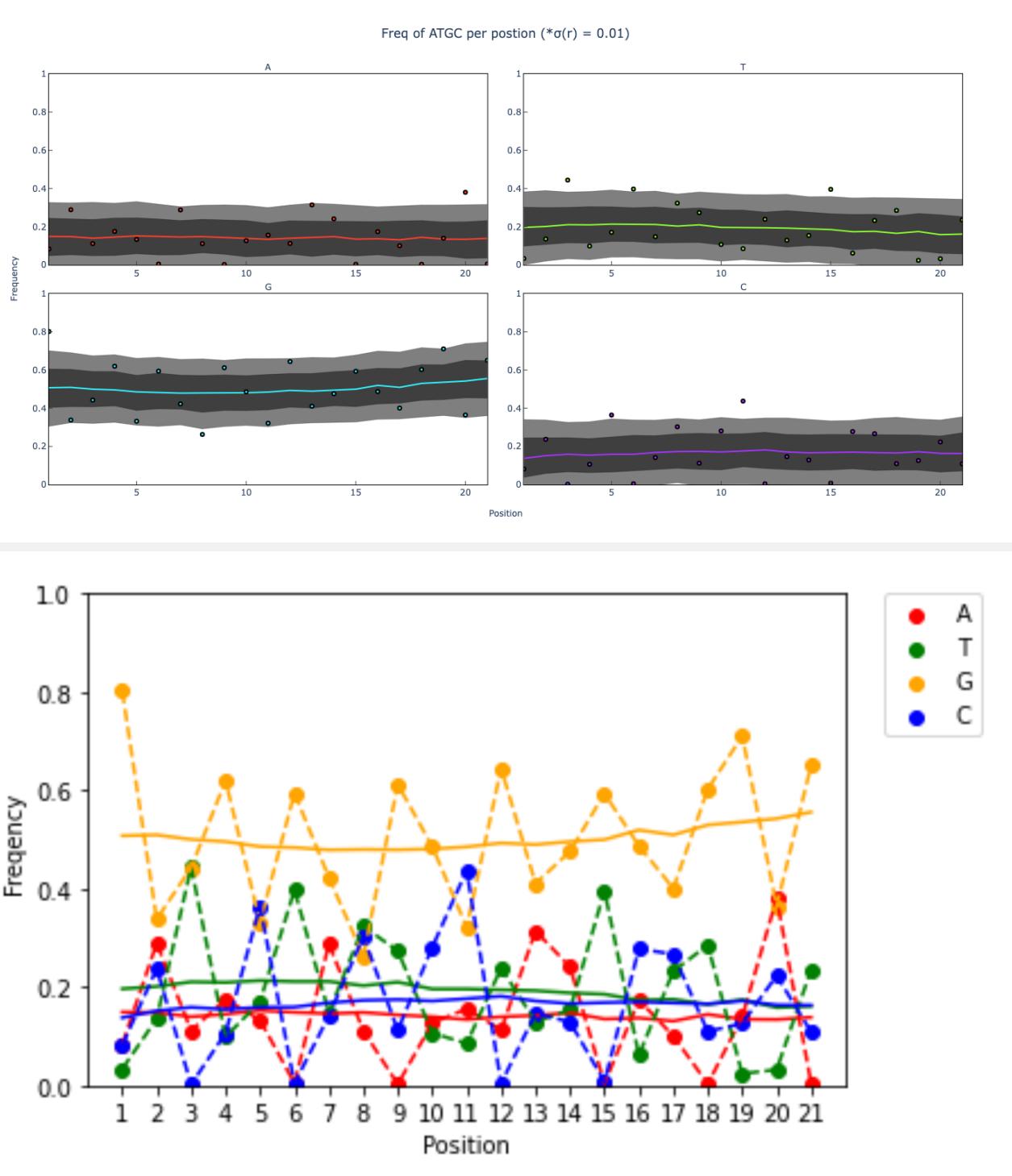
各N15配列について、  
IVT前のデータ、正解データのN15内の各位置におけるATGC頻度分布を理論値とし、サンプル数を全選抜した配列数として、カイ2乗適合度検定

配列グループ内または各N15配列で、  
ATGCの出現頻度を求め、各塩基の出現頻度について1次元の空間構造予測モデルを使ってmotifスコアの予測モデルを構築

出現頻度pについて予測し、 $\log_2(p/bg=0.25)$ でPWMの作成 → motifスコアの算出

<https://biopython.org/docs/1.76/api/Bio.motifs.matrix.html>

### 例) デモデータを用いて、各位置におけるAGCT出現頻度の予測



予測値は、ベイズ予測分布の中央値を示す。

中央値は、各位置でATGCの出現頻度の総和が1になることを確認済。

観測結果から、ある21塩基挿入配列にはAGCT比率に偏りのある領域が存在するだろうと仮定して、配列内におけるAGCT比率について、1次元の空間構造モデルで表現した。上のグラフ結果で対象となった配列は、配列内でAGCT比率に偏りがある領域ではなく、全ての位置でGが0.5%の確率で出現しやすいことが予想された。

このモデルでは、近接領域で応答変数の値に偏りのない、ランダムウォークの状態が見られる時には、グラフは水平線を描く。

逆に、応答変数の値に局所性が見られれば、グラフの概形は激しく変動したものになる。