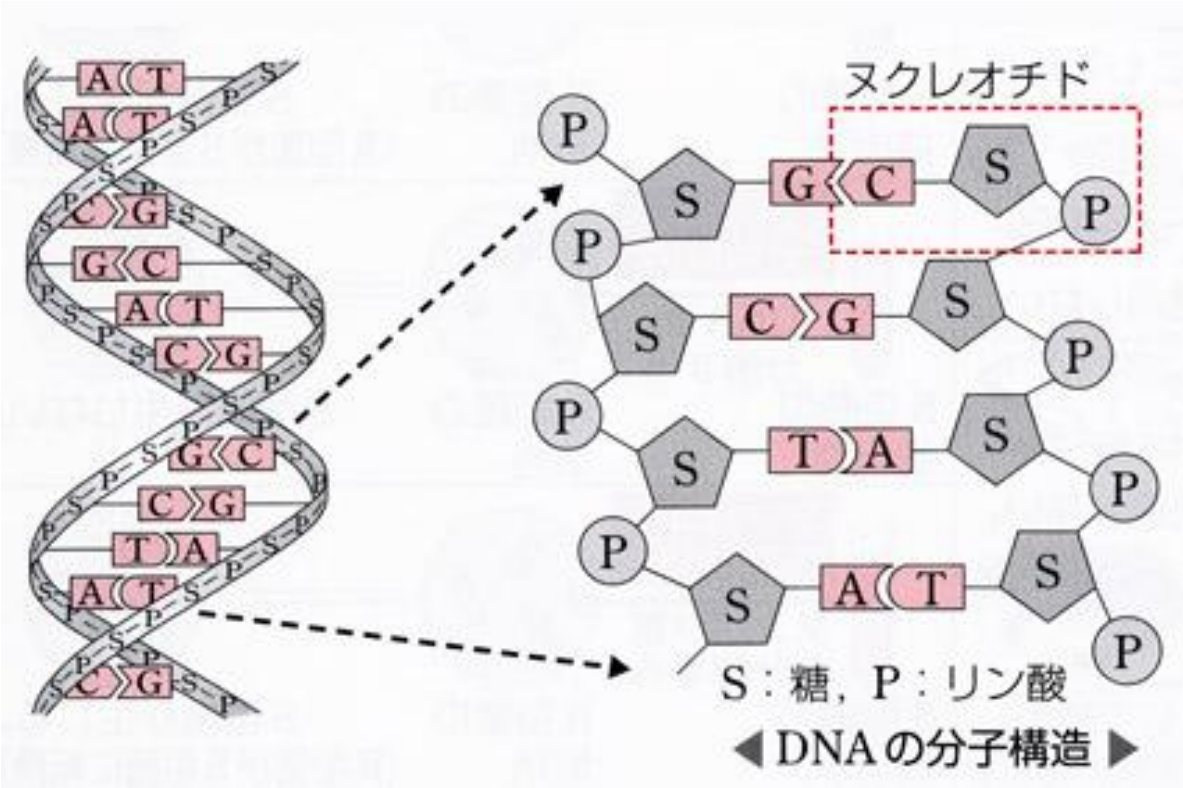
モチーフ解析

東京科学大学　ILA国府台
中林　潤

JH人材育成課　バイオインフォマティシャン育成講座　⑧

塩基配列



◀DNAの分子構造▶

base A : adenin
G : guanine
T : thymine
C : cytosine

生物種が持つ遺伝子の総体：ゲノム
ヒトゲノム：30億塩基対

モチーフ解析

- 塩基配列中から特定の配列を検索する

モチーフ解析

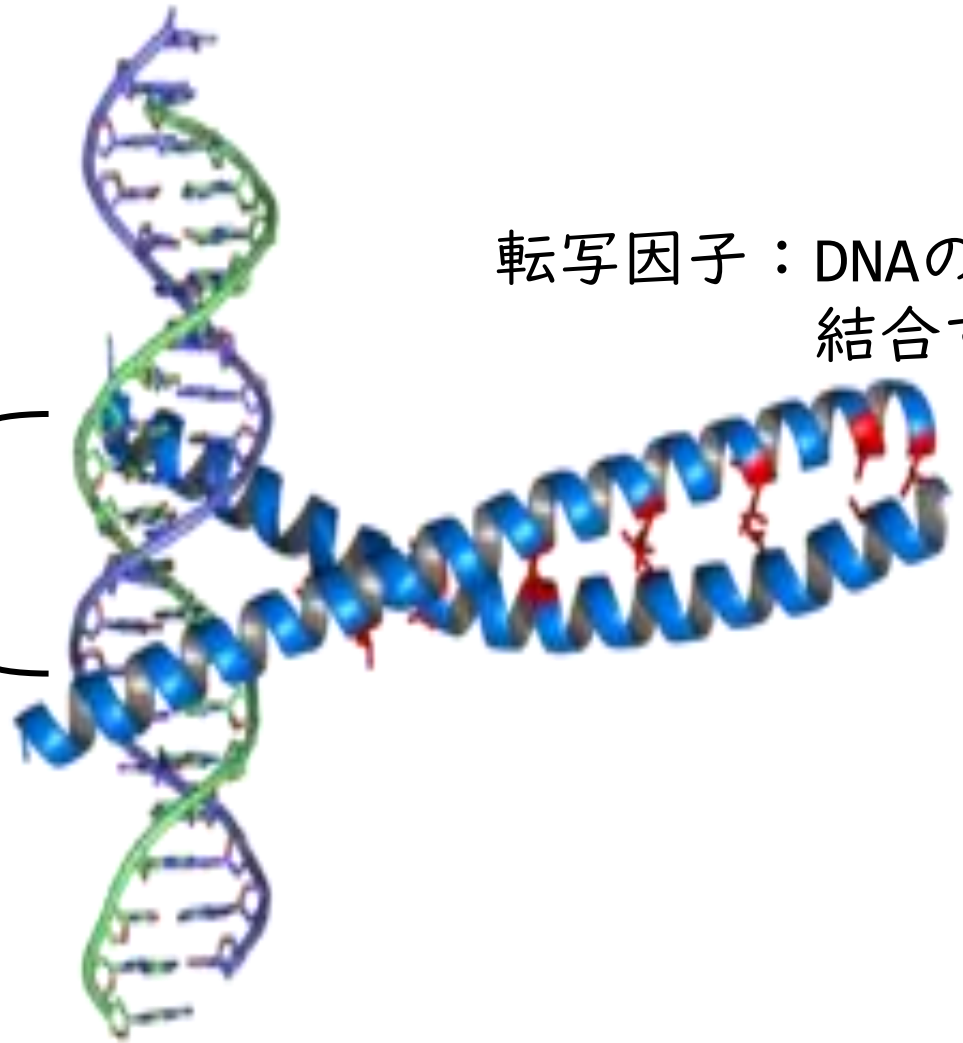
thoughyetofhamletourdearbrothersdeaththememorybegreenandthati
tusbefittedtobearourheartsingriefandourwholekingdomtobecontra
ctedinonebrowofwoeyetsofarhathdiscretionfoughtwithnaturethatw
ewithwisestsorrowthinkonhimtogetherwithremembranceofourselves
thereforeoursometimesisternowourqueentheimperialjointresstoth
iswarlikestatehaveweastwerewithadefeatedjoywithanauspiciousan
dadroppingeyewithmirthinfuneralandwithdirgeinmarriageinequals
caleweighingdelightanddoletakentowifenorhavewehereinbarrdyour
betterwisdomswhichhavefreelygone15withthisaffairalongforallou
rthanks

モチーフ解析

thoughyetofhamletourdearbrothersdeaththememorybegreenandthati
tusbefittedtobearourheartsingriefandourwholekingdomtobecontra
ctedinonebrowofwoeyetsofarhathdiscretionfoughtwithnaturethatw
ewithwisestsorrowthinkonhimtogetherwithremembranceofourselves
thereforeoursometimesisternowourqueentheimperialjointresstoth
iswarlikestatehaveweastwerewithadefeatedjoywithanauspiciousan
dadroppingeyewithmirthinfuneralandwithdirgeinmarriageinequals
caleweighingdelightanddoletakentowifenorhavewehereinbarrdyour
betterwisdomswhichhavefreelygone15withthisaffairalongforallou
rthanks

偶然では起こりえない配列→機能的な意義があるのではないか

転写因子の結合モチーフ



転写因子：DNAの特異的な配列を認識して
結合する

モチーフ
↓
DNA中の特別な
領域を発見する
手がかりとなる

モチーフ解析

- 30億塩基対の中から特定の配列を探す。

- 全ての組み合わせを総当たりで調べる。

  30億^{2}の計算時間が必要

  不可能

発見的（ヒューリスティックな）解析法

- モンテカルロ法

  乱数を使った計算を繰り返すことで近似解を求める方法。正確な解を求めるのが困難な問題にも適用できる。

- モチーフ解析

  配列の持つ情報量を手掛かりにして、試行錯誤を繰り返しモチーフを検索する。

情報量

- 情報量 $= -\log_2($事象の起こる確率 $)$

- コイン投げ

  情報量$= -\log_2$（表が出る確率）

  $= -\log_2 \frac{1}{2} = 1$ (bit)

- トランプ

  情報量$= -\log_2 \frac{1}{53} = 5.72792$

モチーフ“AAAA”が持つ情報量

#1 AAAAAAAAAAAAAAAA
#2 AAAAAAAAAAAAAAAA
#3 AAAAAAAAAAAAAAAA
#4 AAAAAAAAAAAAAAAA

周囲もAのみ
情報として意味を持たない

#1 GTCACATCAGTCGTG
#2 GTTGTTCACAGAAGC
#3 GTCTGTACATGGCAC
#4 GTAGATAGCCTCCGT

AAAAが含まれない
情報として意味を持たない

#1 GTAAAATCCGTCGTG
#2 GTTGTTCCCAAAAGC
#3 GTCTGTAAAAGGCTC
#4 GAAAATTGCCTCCGT

AAAAのみが周囲と比較して高頻度に出現する
情報として意味を持つ

20塩基×6配列の中から4塩基のモチーフを探索する

#1 GTAAGTACAGATAGCCACAG
#2 GTATGTTCCCGATAAGTTTA
#3 GTATGTTCATGTCTGATACT
#4 GCATGATAGCTGCCCAAGTT
#5 GTATGTTGACGATATTACTT
#6 GTAAGTATCCAGATATTACT

20塩基×6配列の中から4塩基のモチーフを探索する

#1 GTAAGTACAGATAGCCACAG
#2 GTATGTTCCCGATAAGTTTA
#3 GTATGTTCATGTCTGATACT
#4 GCATGATAGCTGCCCAAGTT
#5 GTATGTTGACGATATTACTT
#6 GTAAGTATCCAGATATTACT

位置特異的スコア行列

position-specific scoring matrix (PSSMs)

長さL塩基のモチーフ

1	2	3	4	…	L
A	G	G	C		T

$$\mathrm{PSSMs} = \sum_{j=1}^{L} \log_2 \left(\frac{P_{\{A,T,G,C\},j}}{P_{\{A,T,G,C\}}} \right)$$

p_{ij}： j番目の塩基が{A,T,G,C}のいずれかである確率
p_i：塩基{A,T,G,C}の頻度

長さLの塩基配列が実現する確率

ステップ1

ランダムに1配列を選択し、抜き取る

#1 GTAAGTACAGATAGCCACAG
#2 GTATGTTCCCGATAAGTTTA
#3 GTATGTTCATGTCTGATACT
#4 GCATGATAGCTGCCCAAGTT
#5 GTATGTTGACGATATTACTT
#6 GTAAGTATCCAGATATTACT

ステップ2

残った配列中から4文字の配列をランダムに選択

```
                    GTAAGTACAGATAGCCACAG  #1
               GTATGTTCCCGATAAGTTTA       #2
          GTATGTTCATGTCTGATACT            #3
GCATGATAGCTGCCCAAGTT                      #4
               GTAAGTATCCAGATATTACT       #6
```

ステップ2

残った配列中から4文字の配列をランダムに選択

```
#1                GTAAGTACAGATAGCCACAG
#2              GTATGTTCCCGATAAGTTTA
#3           GTATGTTCATGTCTGATACT
#4 GCATGATAGCTGCCCAAGTT
#6              GTAAGTATCCAGATATTACT
```

```
1 2 3 4
A G T A
T C C C
T G T C
C C A A
A T C C
```

	1	2	3	4
A	2	0	1	2
T	2	1	2	0
G	0	2	0	0
C	1	2	2	3

ステップ2

塩基の出現頻度を計算

4文字選んだ時に1番目がAの確率

T
G
C

4文字選んだ時に2番目がAの確率

T
G
C

	1	2	3	4
A	0.4	0	0.2	0.4
T	0.4	0.2	0.4	0
G	0	0.4	0	0
C	0.2	0.4	0.4	0.6

ステップ3

抜き取っておいた配列で4文字の配列が実現する確率を計算する

#5 GTATGTTGACGATATTACTT

$$F = \sum_{i=1}^{K} \sum_{j=1}^{M} 4 \log_2 \frac{p_{ij}}{p_i}$$

GTAT : -∞
TATG : -∞

TGAC : 10.055

	1	2	3	4	p_i
A	0.4	0	0.2	0.4	34/120=0.283
T	0.4	0.2	0.4	0	41/120=0.342
G	0	0.4	0	0	25/120=0.208
C	0.2	0.4	0.4	0.6	20/120=0.167

ステップ4

Fが最大となる4文字を選択して、ステップ1から繰り返し

```
                  GTAAGTACAGATAGCCACAG
#1
            GTATGTTCCCGATAAGTTTA
#2
      GTATGTTCATGTCTGATACT
#3
GCATGATAGCTGCCCAAGTT
#4
            GTATGTTGACGATATTACTT
#5
            GTAAGTATCCAGATATTACT
#6
```

PSSMが高い配列が選択される

```
#1          GTAAGTACAGATAGCCACAG
#2        GTATGTTCCCGATAAGTTTA
#3  GTATGTTCATGTCTGATACT
#4   GCATGATAGCTGCCCAAGTT
#5          GTATGTTGACGATATTACTT
#6          GTAAGTATCCAGATATTACT
                      21.35
```

	1	2	3	4	p_i
A	0	0.8	0.4	0.8	34/120=0.283
T	0	0	0.6	0.2	41/120=0.342
G	0.6	0.2	0	0	25/120=0.208
C	0.4	0	0	0	20/120=0.167

結果

最終的にモチーフが見つかる

GTAAGTACAGATAGCCACAG

GTATGTTCCCGATAAGTTTA

GTATGTTCATGTCTGATACT

GCATGATAGCTGCCCAAGTT

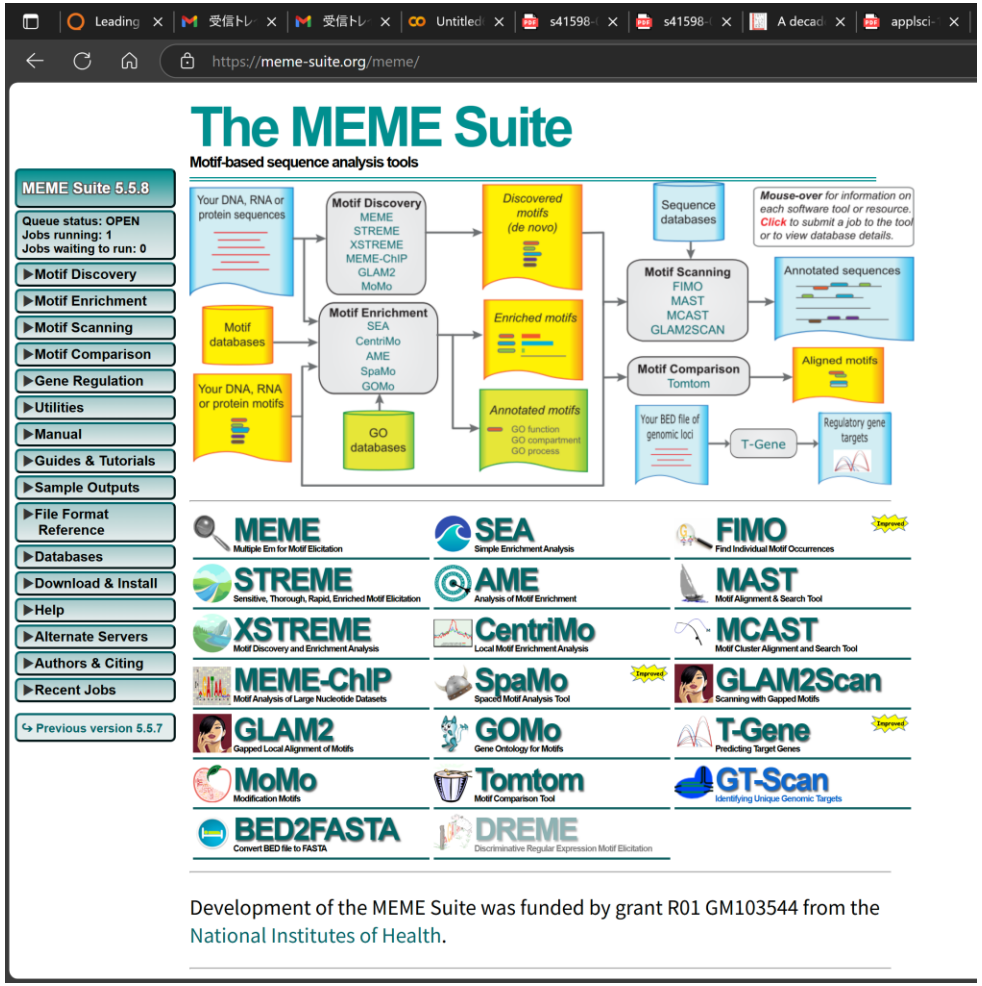
GTAAGTATCCAGATATTACT

発見的な（ヒューリスティック）解析手法

- 試行錯誤を繰り返して、徐々に正解に近づいていく解析方法
- 正解が求まる保証がない
- 求まった解が正しい保証がない

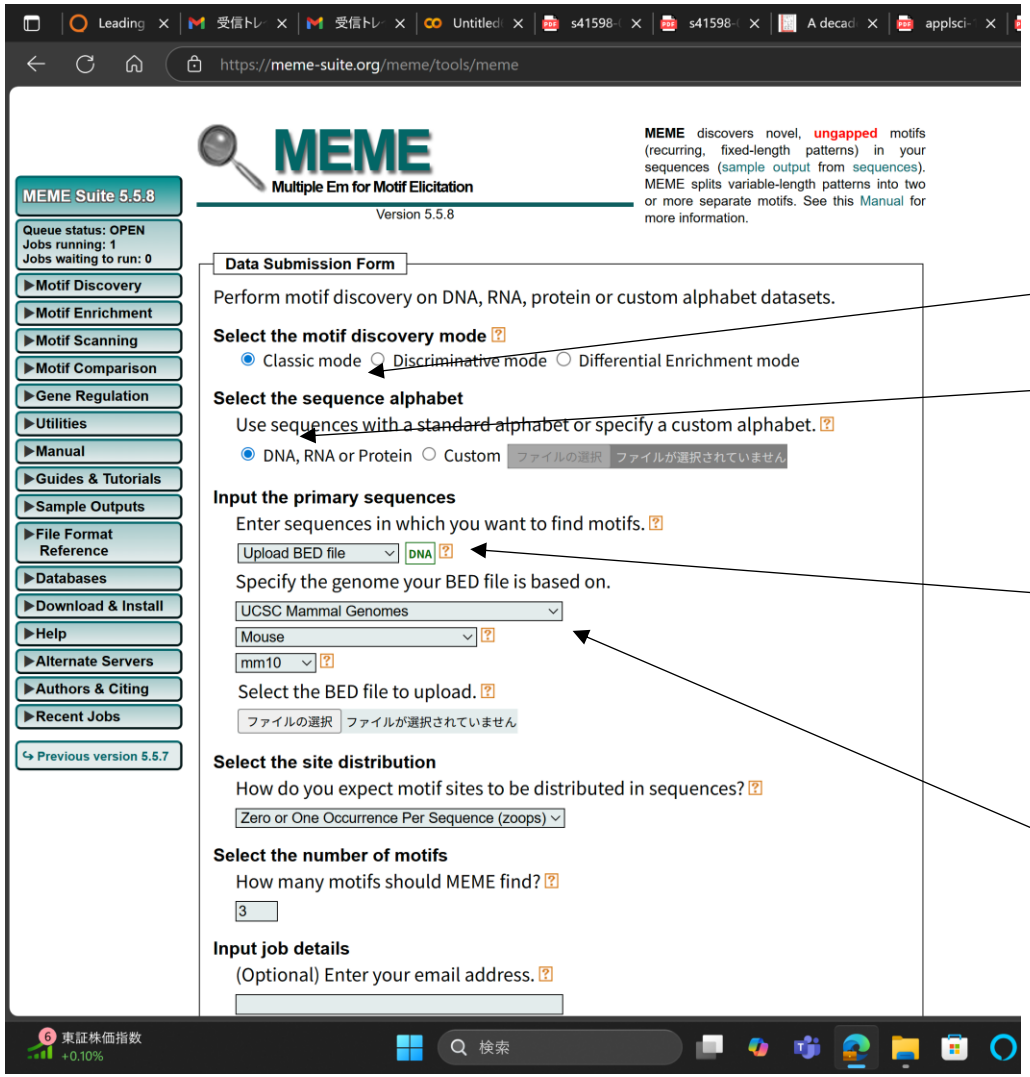
MEME Suite

- Webベースのモチーフ解析アプリケーション集



https://meme-suite.org/meme/

MEMEの実行



classic mode

DNA

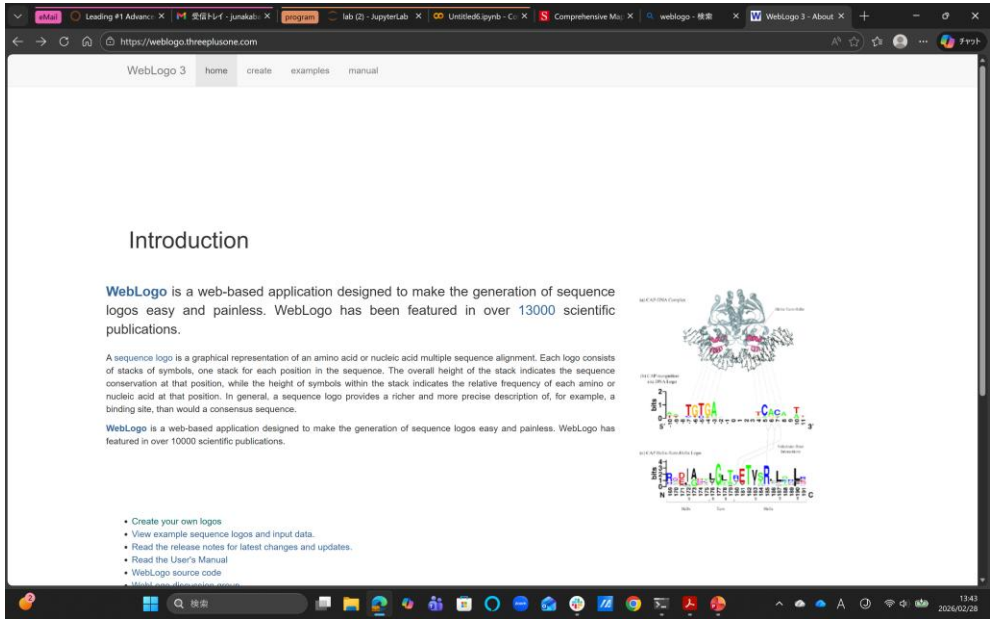
配列
FASTA形式
BEDフォーマット
今回はBEDフォーマットを選択

参照ゲノム情報
UCSC mammal genome
Mouse
mm10

シーケンスロゴ

$$R_i = \log_2(|C|) - \left(-\sum_{i=\{A,T,G,C\}} P_i \log_2 P_i + \frac{|C|-1}{2 \log_e 2}\right)$$

$$h_i(X) = P_i(X) R_i$$



WebLogo3：Webベースのシーケンスロゴ作成用アプリケーション

https://weblogo.threeplusone.com/

まとめ

- モチーフ解析
  塩基配列の中から特定の配列を検索する。

- Position Specific Scoring Matrix
  配列が持つ情報量を表す。

- ヒューリスティック（発見的）な解析方法
  試行錯誤を繰り返して正解に近づいていく解析手法
  正解が得られる保証がない。
  得られた解が正しい保証がない。