

モチーフ解析

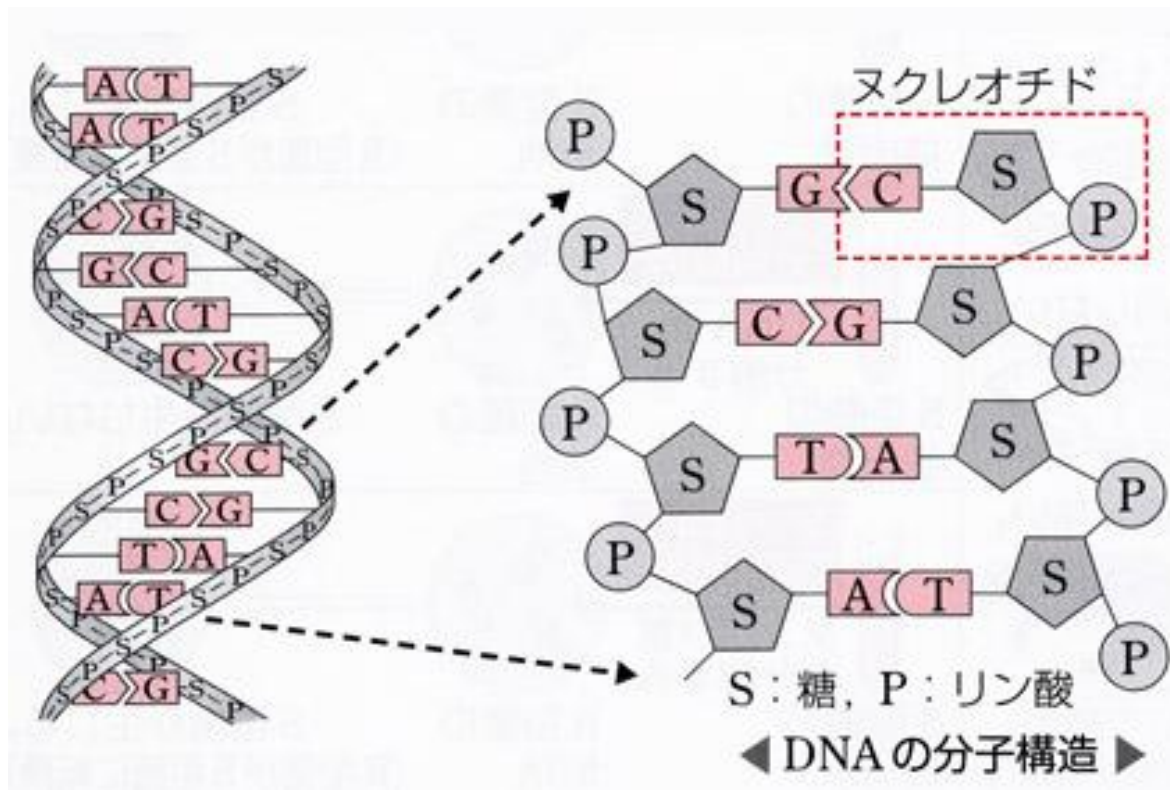
東京科学大学 ILA国府台
中林 潤

JH人材育成課 バイオインフォマティクソン育成講座



⑧

塩基配列



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

though yet of hamlet our dear brothers death the memory be green and that it
is thus be fitted to bear our hearts in grief and our whole kingdom to be contra
cted in one brow of woe yet so far hath discretion fought with nature that w
e with wisest sorrow think on him together with remembrance of ourselves
therefore our sometime sister now our queen the imperial jointress to th
is warlike state have we eastward with a defeated joy with a auspicious an
d a dropping eye with a mirth in funeral and with a dirge in marriage in equals
cal weighing delight and do let a Kent to wife nor have we here in barr'd your
better wisdoms which have freely gone with this affair along for all our
r thanks



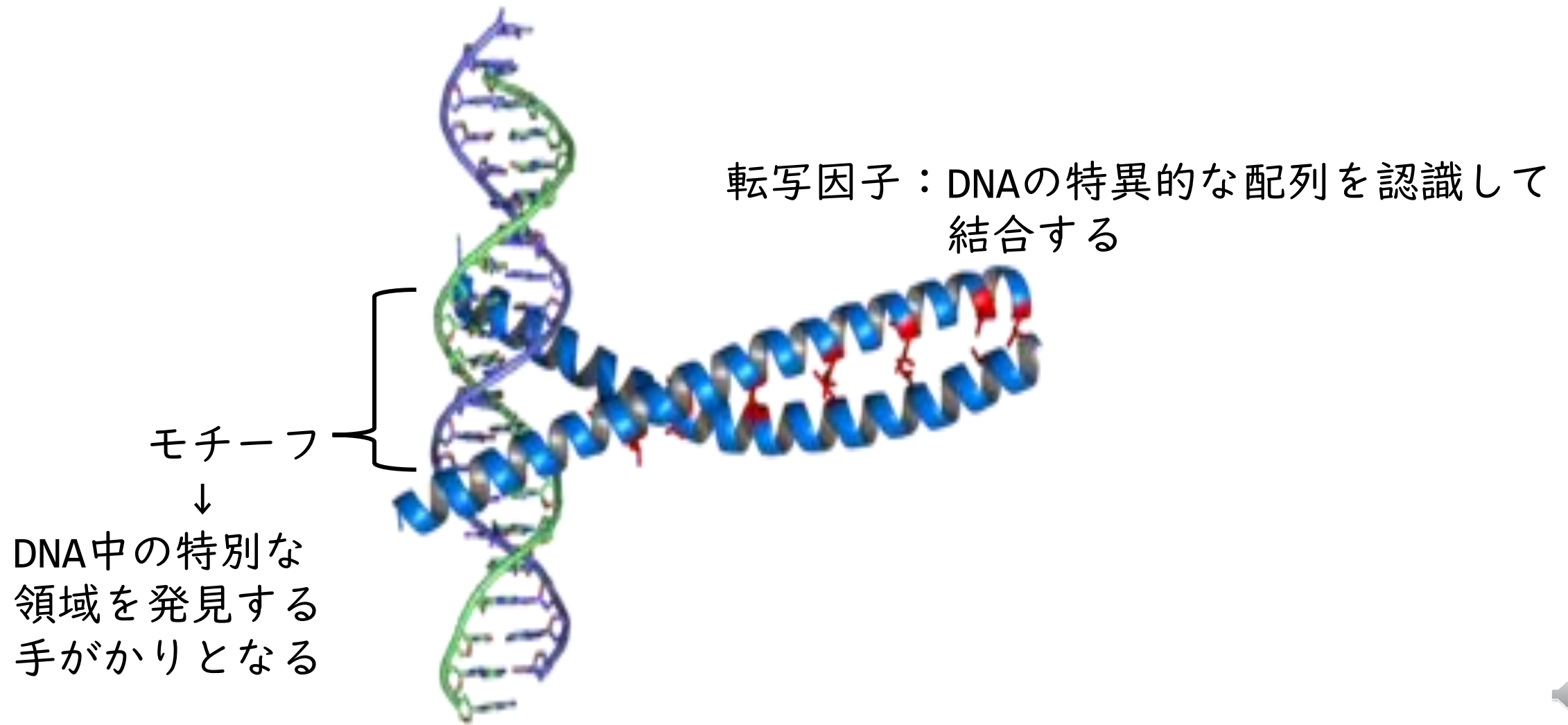
モチーフ解析

though yet of hamlet our dear brothers death the memory be green and that it
tus be fitted to bear our hearts in grief and our whole kingdom to be contra
cted in one brow of woe yet so far hath discretion fought **with** nature that w
e with wisest sorrow think on him together **with** remembrance of ourselves
therefore our sometimes sister now our queen the imperial jointress to th
is warlike state have we east were **with** a defeated joy **with** an auspicious an
dad dropping eye **with** mirth in funeral and **with** dirge in marriage in equals
cal weighing delight and do let a Kent to wife nor have we here in barr'd your
better wisdoms which have freely gone 15 with this affair along for all ou
r thanks

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



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 = $-\log_2(\text{事象の起こる確率})$

- コイン投げ

$$\begin{aligned}\text{情報量} &= -\log_2(\text{表が出る確率}) \\ &= -\log_2 \frac{1}{2} = 1 \text{ (bit)}\end{aligned}$$



- トランプ

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



モチーフ “AAAA” が持つ情報量

#1 AAAAAAAAAAAAAAAAAA
#2 AAAAAAAAAAAAAAAAAA
#3 AAAAAAAAAAAAAAAAAA
#4 AAAAAAAAAAAAAAAAAA

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 GT**AAAA**TCCGTCGTG
#2 GTTGTTC**CCAAAA**GC
#3 GTCTGT**AAAA**GGCTC
#4 G**AAAA**TTGCCTCCGT

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 GTATGTTTCATGTCTGATACT
#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 GTAAGTACAG**GAT**AGCCACAG
#2 GTATGTTCCC**GATA**AGTTTA
#3 GTATGTTTCATGTCT**GATA**CT
#4 GCAT**GAT**AGCTGCCCAAGTT
#5 GTATGTTGAC**GATA**TTACTT
#6 GTAAGTATCCAG**GATA**TTACT



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

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

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  GTATGTTTCATGTCTGATACT  
#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

```
#1          GTAAGTACAGATAGCCACAG
#2          GTATGTTCCCGATAAGTTTA
#3          GTATGTTCA TGTCTGATACT
#4 GCATGATAGCTGC CCAAGTT
#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

#1 GTA**AGT**ACAGATAGCCACAG
#2 GTATG**TCCC**GATAAGTTTA
#3 GTATGTTCA**TGTC**TGATACT
#4 GCATGATAGCTGC**CCA**AGTT
#6 GTAAG**TATCC**AGATATTACT



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

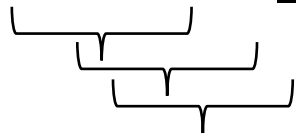
	1	2	3	4
A	0.4	0	0.2	0.4
T	0.4	0.2	0.4	0
G	0	0.4	0	0
C	0.2	0.4	0.4	0.6



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 GTATGTTTGACGATATTACTT



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

GTAT : $-\infty$

TATG : $-\infty$

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

```
#1          GTAAGTACAGATAGCCACAG
#2          GTATGTTCCCGATAAGTTTA
#3          GTATGTTTCATGTCTGATACT
#4 GCATGATAGCTGCCCAGTT
#5          GTATGTTGACGATATTACTT
#6          GTAAGTATCCAGATATTACT
```



PSSMが高い配列が選択される

#1 GTAAGTACAG**GAT**AGCCACAG
#2 GTATGTTCCC**GATA**AGTTTA
#3 GTATGTTTCATGTCT**GATA**CT
#4 GCATGATAGCTG**CCA**AGTT
#5 GTATGTTG**CGAT**ATTACTT
#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

```
GTAAGTACAGGATAGCCACAG
GTATGTTCCCATAAGTTTA
GTATGTTTCATGTCTGATACT
          GCATGATAGCTGCCCAAGTT
GTAAGTATCCAATAATTACT
```



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
- 求め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

The screenshot shows the MEME Suite website interface. At the top, the browser address bar displays <https://meme-suite.org/meme/>. The main heading is "The MEME Suite" with the subtitle "Motif-based sequence analysis tools".

On the left, there is a sidebar menu for "MEME Suite 5.5.8" with the following items: Queue status: OPEN, Jobs running: 1, Jobs waiting to run: 0; Motif Discovery; Motif Enrichment; Motif Scanning; Motif Comparison; Gene Regulation; Utilities; Manual; Guides & Tutorials; Sample Outputs; File Format Reference; Databases; Download & Install; Help; Alternate Servers; Authors & Citing; Recent Jobs; and a link to "Previous version 5.5.7".

The central workflow diagram illustrates the process: "Your DNA, RNA or protein sequences" and "Motif databases" feed into "Motif Discovery" (tools: MEME, STREME, XSTREME, MEME-CHIP, GLAM2, MoMo) and "Motif Enrichment" (tools: SEA, CentriMo, AME, SpaMo, GOMo). "GO databases" also feed into "Motif Enrichment". "Discovered motifs (de novo)" and "Enriched motifs" lead to "Annotated motifs". "Sequence databases" feed into "Motif Scanning" (tools: FIMO, MAST, MCAST, GLAM2SCAN), which produces "Annotated sequences". "Motif Comparison" (tool: Tomtom) takes "Annotated motifs" and "Your BED file of genomic loci" as input to produce "Aligned motifs". "T-Gene" takes "Your BED file of genomic loci" as input to produce "Regulatory gene targets". A note says: "Mouse-over for information on each software tool or resource. Click to submit a job to the tool or to view database details."

Below the diagram is a grid of tool logos and descriptions:

- MEME**: Multiple Em for Motif Elicitation
- SEA**: Simple Enrichment Analysis
- FIMO**: Find Individual Motif Occurrences
- STREME**: Sensitive, Thorough, Rapid, Enriched Motif Elicitation
- AME**: Analysis of Motif Enrichment
- MAST**: Motif Alignment & Search Tool
- XSTREME**: Motif Discovery and Enrichment Analysis
- CentriMo**: Local Motif Enrichment Analysis
- MCAST**: Motif Cluster Alignment and Search Tool
- MEME-CHIP**: Motif Analysis of Large Nucleotide Datasets
- SpaMo**: Spaced Motif Analysis Tool
- GLAM2Scan**: Scanning with Gapped Motifs
- GLAM2**: Gapped Local Alignment of Motifs
- GOMo**: Gene Ontology for Motifs
- T-Gene**: Predicting Target Genes
- MoMo**: Modification Motifs
- Tomtom**: Motif Comparison Tool
- GT-Scan**: Identifying Unique Genomic Targets
- BED2FASTA**: Convert BED file to FASTA
- DREME**: Discriminative Regular Expression Motif Elicitation

At the bottom, a text block states: "Development of the MEME Suite was funded by grant R01 GM103544 from the National Institutes of Health."



MEMEの実行

MEME Suite 5.5.8
Queue status: OPEN
Jobs running: 1
Jobs waiting to run: 0

MEME
Multiple Em for Motif Elicitation
Version 5.5.8

MEME discovers novel, **ungapped** motifs (recurring, fixed-length patterns) in your sequences (sample output from sequences). MEME splits variable-length patterns into two or more separate motifs. See this [Manual](#) for more information.

Data Submission Form

Perform motif discovery on DNA, RNA, protein or custom alphabet datasets.

Select the motif discovery mode [?](#)

Classic mode Discriminative mode Differential Enrichment mode

Select the sequence alphabet

Use sequences with a standard alphabet or specify a custom alphabet. [?](#)

DNA, RNA or Protein Custom ファイルの選択 ファイルが選択されていません

Input the primary sequences

Enter sequences in which you want to find motifs. [?](#)

Upload BED file [?](#)

Specify the genome your BED file is based on.

UCSC Mammal Genomes

Mouse [?](#)

mm10 [?](#)

Select the BED file to upload. [?](#)

ファイルの選択 ファイルが選択されていません

Select the site distribution

How do you expect motif sites to be distributed in sequences? [?](#)

Zero or One Occurrence Per Sequence (zoops) [?](#)

Select the number of motifs

How many motifs should MEME find? [?](#)

3

Input job details

(Optional) Enter your email address. [?](#)

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



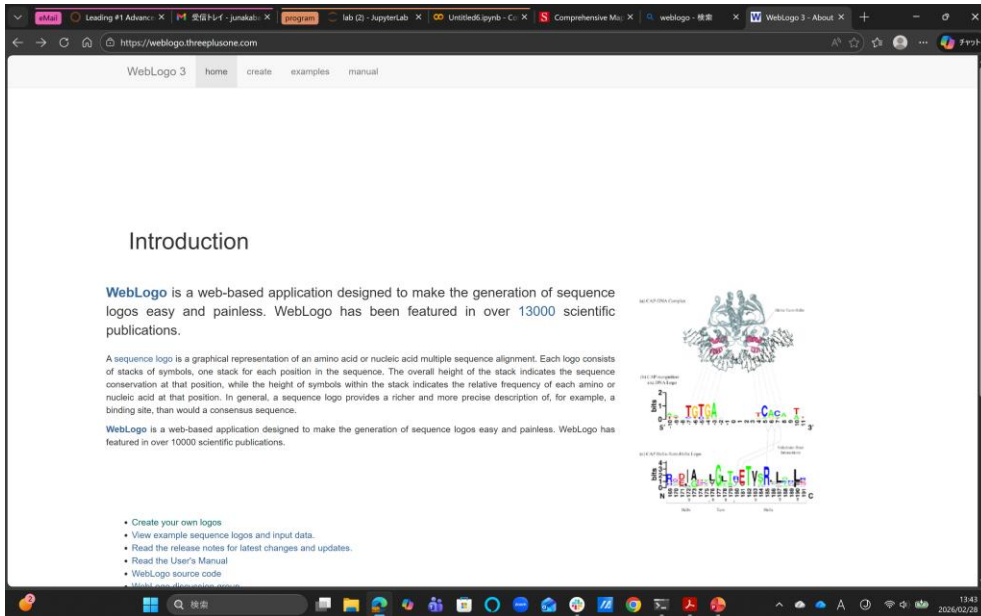
シーケンスロゴ



Reverse Opposite:



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



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.threepiusone.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。

