Sequence Analysis 3B:

Sequence motif profiles

Katherine Dougan, PhD

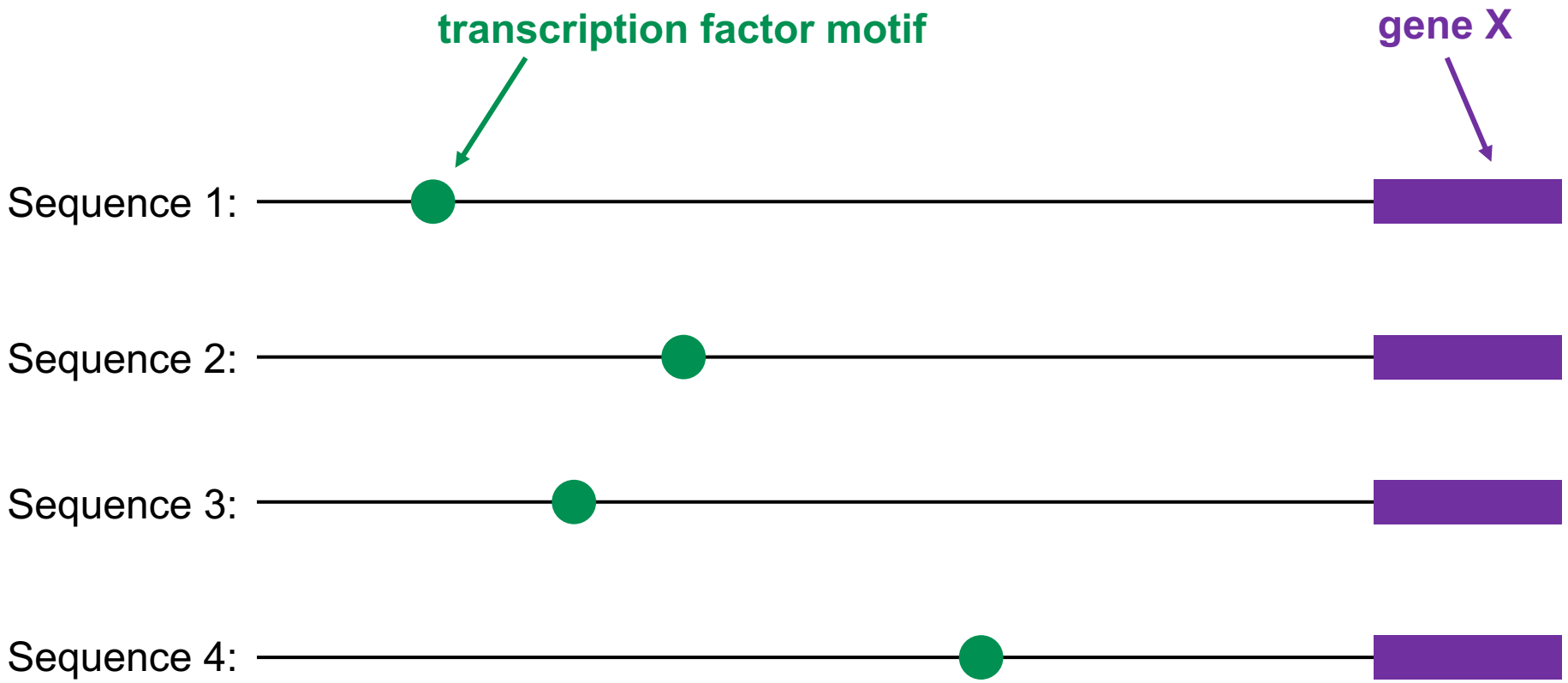
Australian Centre for Ecogenomics
School of Chemistry & Molecular Biosciences
The University of Queensland

SCIE2100 | BINF6000 | Bioinformatics I - Introduction

Outline

- **Identifying motifs in a sequence**
  - Position Specific Scoring Matrix
  - Scoring sequences with profiles
- **Constructing motif profiles**
  - Position Frequency Matrix
  - Position Probability Matrix
  - Position Weight Matrix
  - Finding the most likely start
- **Visualizing motifs with sequence logos**
  - Representing uncertainty with entropy
  - Determining Information Content

Identifying motifs in a sequence



Let's say we want to find a transcription factor motif for gene X, but its location and sequence is highly variable... ***how could we identify the motif better accounting for variability?***

Position Specific Scoring Matrix

We can use a numeric representation of a motif called a ***profile*** to identify the best starting position by calculating a similarity score between the profile and a sequence

For example, we could describe motif M of length K by the probability of encountering different nucleotides at each position of the sequence in profile f …

$$f = \begin{bmatrix} f_{1A} & f_{2A} & f_{3A} & f_{4A} & \cdots & f_{KA} \\ f_{1C} & f_{2C} & f_{3C} & f_{4C} & \cdots & f_{KC} \\ f_{1G} & f_{2G} & f_{3G} & f_{4G} & \cdots & f_{KG} \\ f_{1T} & f_{2T} & f_{3T} & f_{4T} & \cdots & f_{KT} \end{bmatrix}$$

and calculate the ***probability*** of a sequence x being an instance of motif M:

$$P(x|M) = \prod_{i=1}^{K} f_{i,x}$$

Scoring a sequence using a profile

The profile f below was generated from an alignment of motif M to the right

$$P(x|M) = \prod_{i=1}^{K} f_{i,x}$$

	1	2	3	4	5	6	7	8
A	0.29	0.21	0.07	0.07	0.07	0.07	0.07	0.14
C	0.21	0.14	0.07	0.07	0.14	0.07	0.07	0.57
G	0.43	0.57	0.07	0.07	0.07	0.79	0.79	0.21
T	0.07	0.07	0.79	0.79	0.71	0.07	0.07	0.07

Alignment:

AATGCGGA
AATGTGGC
ACTGTGGC
CGTGTGGC
CGTGTGGC
GGTGTGGC
GGTGTGGC
GGTGTGGC
GGTGTGGG
GGTGTGGG

What is the probability of Sequence X being an instance of M?

Sequence X: CCTGCGGC

$$P(X|M) = 0.21 * 0.14 * 0.79 * 0.07 * 0.14 * 0.79 * 0.79 * 0.57 = 0.000081$$

Usually if $P(x|M) \geq threshold$ when compared to the highest possible likelihood (i.e. consensus sequence) then it is a match

What if we do not know where the motif starts?

Finding the most likely start for a motif

We can use a sliding window and calculate the score for each set to find the one with the greatest score

	1	2	3	4	5	6	7	8
A	0.29	0.21	0.07	0.07	0.07	0.07	0.07	0.14
C	0.21	0.14	0.07	0.07	0.14	0.07	0.07	0.57
G	0.43	0.57	0.07	0.07	0.07	0.79	0.79	0.21
T	0.07	0.07	0.79	0.79	0.71	0.07	0.07	0.07

S(1) ATGCGATGACCTGC

S(2) ATGCGATGACCTGC

S(3) ATGCGATGACCTGC

$$P(X|M) = f_{1x} * f_{2x} * f_{3x} * f_{4x} * f_{5x} * f_{6x} * f_{7x} * f_{8x}$$

Let's get into how you construct a Position Specific Scoring Matrix…

Finding the most likely start for a motif

We can use a sliding window and calculate the score for each set to find the one with the greatest score

	1	2	3	4	5	6	7	8
A	0.29	0.21	0.07	0.07	0.07	0.07	0.07	0.14
C	0.21	0.14	0.07	0.07	0.14	0.07	0.07	0.57
G	0.43	0.57	0.07	0.07	0.07	0.79	0.79	0.21
T	0.07	0.07	0.79	0.79	0.71	0.07	0.07	0.07

S(1) ATGCGATGACCTGC

S(2) ATGCGATGACCTGC

S(3) ATGCGATGACCTGC

$$P(X|M) = f_{1x} * f_{2x} * f_{3x} * f_{4x} * f_{5x} * f_{6x} * f_{7x} * f_{8x}$$

Let's get into how you construct a Position Specific Scoring Matrix…

Finding the most likely start for a motif

We can use a sliding window and calculate the score for each set to find the one with the greatest score

	1	2	3	4	5	6	7	8
A	0.29	0.21	0.07	0.07	0.07	0.07	0.07	0.14
C	0.21	0.14	0.07	0.07	0.14	0.07	0.07	0.57
G	0.43	0.57	0.07	0.07	0.07	0.79	0.79	0.21
T	0.07	0.07	0.79	0.79	0.71	0.07	0.07	0.07

S(1) ATGCGATGACCTGC

S(2) ATGCGATGACCTGC

S(3) ATGCGATGACCTGC

$$P(X|M) = f_{1x} * f_{2x} * f_{3x} * f_{4x} * f_{5x} * f_{6x} * f_{7x} * f_{8x}$$

Let's get into how you construct a Position Specific Scoring Matrix…

The different kinds of PSSMs or profile matrices

Let's first discuss the different kinds of profile matrices:

Position Frequency Matrix (PFM) – the position-dependent frequency f or how often each letter (i.e. nucleotide or amino acid) occurs at a given position in N sequences

Position Probability Matrix (PPM) – the probability of each letter at a given position by normalizing the PFM values by the total number of sequences, N

Position Weight Matrix (PWM) – the log likelihood ratios of the PPM

PSSM – Step 1: Position Frequency Matrix (PFM)

$n_{i,a}$ is the count of residue a in column i

PFM

$n_{i,a}$

A	3	2	0	0	0	0	0	1
C	2	1	0	0	1	0	0	7
G	5	7	0	0	0	10	10	2
T	0	0	10	10	9	0	0	0

```
AATGCGGA
AATGTGGC
ACTGTGGC
CGTGTGGC
CGTGTGGC
GGTGTGGC
GGTGTGGC
GGTGTGGC
GGTGTGGG
GGTGTGGG
```

PSSM – Step 2: Add pseudocounts to PFM

$n_{i,a}$ is the count of residue a in column i

PFM

$n_{i,a}$

A	3	2	0	0	0	0	0	1
C	2	1	0	0	1	0	0	7
G	5	7	0	0	0	10	10	2
T	0	0	10	10	9	0	0	0

If we have zeros in our profile, then we need to add **pseudocounts**.

Later on, we will be doing log calculations, and you can't calculate the log of 0…

Pseudocounts

(if zeros in PFM)

A	4	3	1	1	1	1	1	2
C	3	2	1	1	2	1	1	8
G	6	8	1	1	1	11	11	3
T	1	1	11	11	10	1	1	1

And it lets us account for scenarios not currently captured in our model by representing them at very low frequencies

PSSM – Step 3: Position Probability Matrix (PPM)

$n_{i,a}$ is the count of residue a in column i

Pseudocounts
(if zeros in PFM)
$n_{i,a}$

A	4	3	1	1	1	1	1	2
C	3	2	1	1	2	1	1	8
G	6	8	1	1	1	11	11	3
T	1	1	11	11	10	1	1	1

PPM

$$f_{i,a} = \frac{n_{i,a}}{N_{seq}}$$

A	0.29	0.21	0.07	0.07	0.07	0.07	0.07	0.14
C	0.21	0.14	0.07	0.07	0.14	0.07	0.07	0.57
G	0.43	0.57	0.07	0.07	0.07	0.79	0.79	0.21
T	0.07	0.07	0.79	0.79	0.71	0.07	0.07	0.07

Using our PFM that has been adjusted with pseudocounts, we now calculate the probability of each character at each position

PSSM – Step 4: Adjust PPM by background probability

$n_{i,a}$ is the count of residue a in column i

PPM

$$f_{i,a} = \frac{n_{i,a}}{N_{seq}}$$

A	0.29	0.21	0.07	0.07	0.07	0.07	0.07	0.14
C	0.21	0.14	0.07	0.07	0.14	0.07	0.07	0.57
G	0.43	0.57	0.07	0.07	0.07	0.79	0.79	0.21
T	0.07	0.07	0.79	0.79	0.71	0.07	0.07	0.07

Human p_a

```
A: 0.233
C: 0.268
G: 0.267
T: 0.231
```

$$\frac{f_{i,a}}{p_a}$$

A	1.24	0.90	0.30	0.30	0.30	0.30	0.30	0.60
C	0.78	0.52	0.26	0.26	0.52	0.26	0.26	2.13
G	1.61	2.13	0.26	0.26	0.26	2.96	2.96	0.79
T	0.30	0.30	3.42	3.42	3.07	0.30	0.30	0.30

Some organisms do not have the four bases occurring at similar frequencies.

For example, GC content is 38% in yeasts and 19% in Plasmodium falciparum.

PSSM – Step 5: Position Weight Matrix (PWM)

$n_{i,a}$ is the count of residue a in column i

PPM

$$\frac{f_{i,a}}{p_a}$$

A	1.24	0.90	0.30	0.30	0.30	0.30	0.30	0.60
C	0.78	0.52	0.26	0.26	0.52	0.26	0.26	2.13
G	1.61	2.13	0.26	0.26	0.26	2.96	2.96	0.79
T	0.30	0.30	3.42	3.42	3.07	0.30	0.30	0.30

PWM

$$W_{i,a} = \log_2\left(\frac{f_{i,a}}{p_a}\right)$$

A	0.32	-0.15	-1.73	-1.73	-1.73	-1.73	-1.73	-0.73
C	-0.35	-0.94	-1.94	-1.94	-0.94	-1.94	-1.94	1.09
G	0.69	1.09	-1.93	-1.93	-1.93	1.57	1.57	-0.35
T	-1.72	-1.72	1.77	1.77	1.62	-1.72	-1.72	-1.72

Finally, we can calculate the log-likelihood ratios

Finding the most likely start for a motif

Going back to our original example…

	1	2	3	4	5	6	7	8
A	0.29	0.21	0.07	0.07	0.07	0.07	0.07	0.14
C	0.21	0.14	0.07	0.07	0.14	0.07	0.07	0.57
G	0.43	0.57	0.07	0.07	0.07	0.79	0.79	0.21
T	0.07	0.07	0.79	0.79	0.71	0.07	0.07	0.07

ATGCGATGACCTGCGGC

ATGCGATGACCTGCGGC

ATGCGATGACCTGCGGC

Alignment:

AATGCGGA

AATGTGGC

ACTGTGGC

CGTGTGGC

CGTGTGGC

GGTGTGGC

GGTGTGGC

GGTGTGGC

GGTGTGGG

GGTGTGGG

Here we used the **Position Probability Matrix**, and calculated the **PRODUCT** of the values with:

$$P(x|M) = \prod_{i=1}^{K} f_{i,x}$$

Finding the most likely start for a motif

We can also use the *Position Weight Matrix* for this…

	1	2	3	4	5	6	7	8
A	0.32	-0.15	-1.73	-1.73	-1.73	-1.73	-1.73	-0.73
C	-0.35	-0.94	-1.94	-1.94	-0.94	-1.94	-1.94	1.09
G	0.69	1.09	-1.93	-1.93	-1.93	1.57	1.57	-0.35
T	-1.72	-1.72	1.77	1.77	1.62	-1.72	-1.72	-1.72

ATGCGATGACCTGCGGC

ATGCGATGACCTGCGGC

ATGCGATGACCTGCGGC

Alignment:

AATGCGGA

AATGTGGC

ACTGTGGC

CGTGTGGC

CGTGTGGC

GGTGTGGC

GGTGTGGC

GGTGTGGC

GGTGTGGG

GGTGTGGG

If we use the **Position Weight Matrix**, then we instead calculate the **SUMS** of the values with:

$$P(x|M) = \sum_{i=1}^{K} w_{i,x}$$

Finding the most likely start for a motif

When comparing a sequence to the profile for a motif…

If we use the **Position Probability Matrix**, then we calculate the **PRODUCT** of the values with:

$$P(x|M) = \prod_{i=1}^{K} f_{i,x}$$

If we use the **Position Weight Matrix**, then we instead calculate the **SUM** of the values with:

$$P(x|M) = \sum_{i=1}^{K} w_{i,x}$$

Finding the most likely start for a motif

We can also use the *Position Weight Matrix* for this…

	1	2	3	4	5	6	7	8
A	0.32	-0.15	-1.73	-1.73	-1.73	-1.73	-1.73	-0.73
C	-0.35	-0.94	-1.94	-1.94	-0.94	-1.94	-1.94	1.09
G	0.69	1.09	-1.93	-1.93	-1.93	1.57	1.57	-0.35
T	-1.72	-1.72	1.77	1.77	1.62	-1.72	-1.72	-1.72

$$S(x|M) = \sum_{i=1}^{K} w_{i,x}$$

S(1)= GATGACCTGCGGC = -6.95
S(2)= GATGACCTGCGGC = -10.01
S(3)= GATGACCTGCGGC = -4.3
S(4)= GATGACCTGCGGC = -2.44
S(5)= GATGACCTGCGGC = -3.44
S(6)= GATGACCTGCGGC = 1.84

The last option is the most likely start for the motif

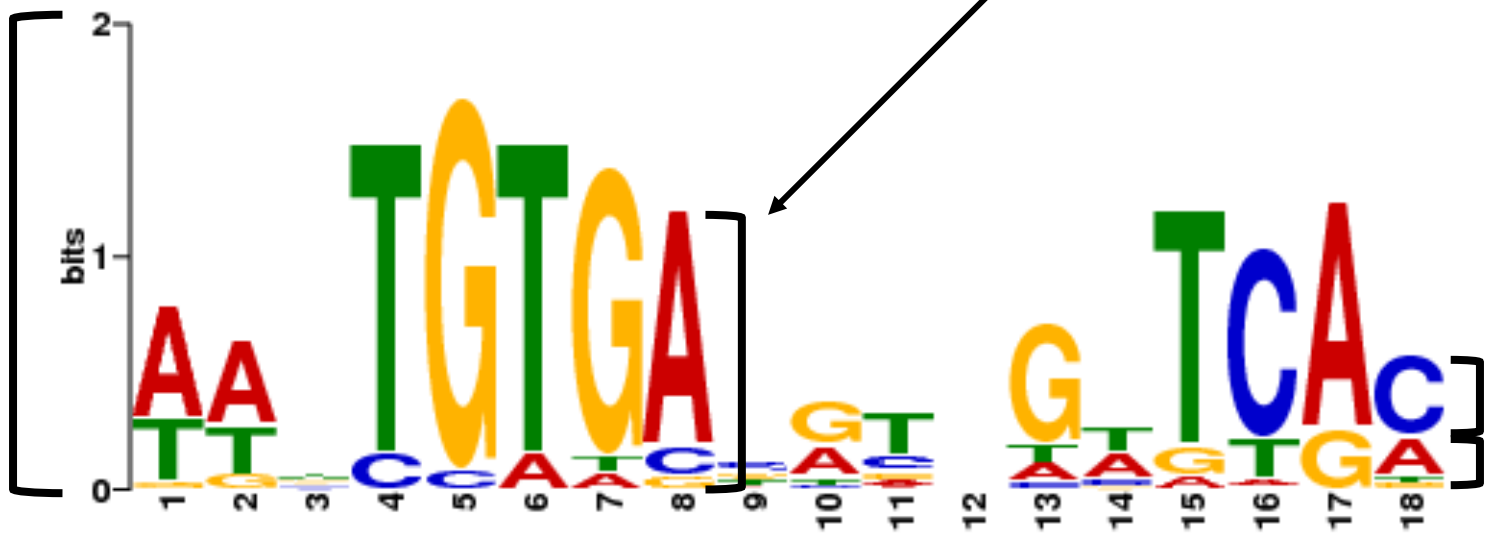
Visualizing motifs with Sequence Logos

Sequence logos are a great way to visualize not only the probability of different characters at each position in a motif, but also the ***uncertainty*** in the motif and the ***information content*** for each position

Entropy is represented by the total height of a column in *bits.*

The total height of characters in a column is the **Information Content.**

The height of an individual character is the **frequency of the residue.**



Let's explore these parameters and how they're calculated…

Representing *uncertainty* in our model with Entropy

Problem: Different positions in a motif will exhibit ***varying levels of conservation.***

We can use **Shannon entropy** to represent uncertainty, or how unpredictable sequences generated from the profile can be.

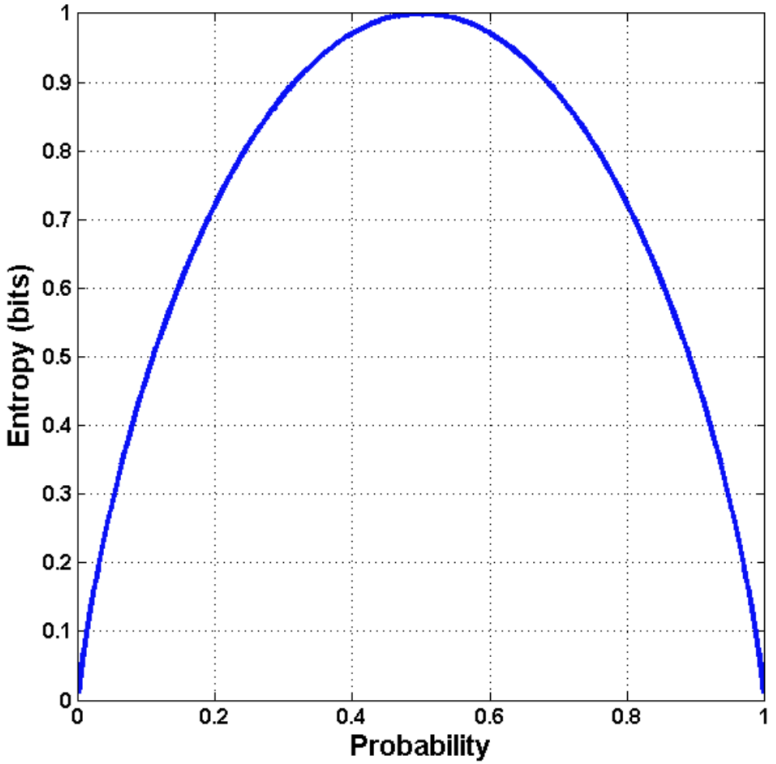
$$H_i = -\sum f_{i,a} \log_2 f_{i,a}$$

$$H = -\left(\frac{1}{100}\log_2\left(\frac{1}{100}\right) + \frac{1}{100}\log_2\left(\frac{1}{100}\right) + \frac{1}{100}\log_2\left(\frac{1}{100}\right) + \frac{97}{100}\log_2\left(\frac{97}{100}\right)\right)$$

$$\boldsymbol{H = 0.25}$$

$$H = -\left(\frac{1}{10}\log_2\left(\frac{1}{10}\right) + \frac{1}{10}\log_2\left(\frac{1}{10}\right) + \frac{1}{10}\log_2\left(\frac{1}{10}\right) + \frac{7}{10}\log_2\left(\frac{7}{10}\right)\right)$$

$$\boldsymbol{H = 1.36}$$



The less equal the probability of different outcomes are , the lower the entropy

Representing *uncertainty* in our model with Entropy

Problem: Different positions in a motif will exhibit ***varying levels of conservation.***

We can use **Shannon entropy** to represent uncertainty, or how unpredictable sequences generated from the profile can be.

$$H_i = -\sum f_{i,a} \log_2 f_{i,a}$$

If each nucleotide is equally likely, then that is the scenario of the highest uncertainty…

$f_{i,a} = \frac{1}{4}$ for all $a \in \{A, C, G, T\}$

$$H_{max} = -\left(\frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right)$$

$\log_2 \frac{1}{4} = -2$

$\boldsymbol{H_{max} = 2\ bits}$ ⟵ The maximum information content of any position is 2 bits

Determining the *Information Content* of a site

The **Information Content** (I) at each site is the ***reduction in entropy***, or how much it reduces uncertainty compared to the background model.

The difference in what we know now compared to what we knew before

$$H_i = -\sum f_{i,a} \log_2 f_{i,a}$$

$$I_i = H_{max} - H_i$$

	1	2	3	4	5	6	7	8
A	0.29	0.21	0.07	0.07	0.07	0.07	0.07	0.14
C	0.21	0.14	0.07	0.07	0.14	0.07	0.07	0.57
G	0.43	0.57	0.07	0.07	0.07	0.79	0.79	0.21
T	0.07	0.07	0.79	0.79	0.71	0.07	0.07	0.07

Let's calculate the *entropy* and *information content* for position 3 in our PPM…

$$H_{max} = -(0.233 * \log_2(0.233) + 0.268 * \log_2(0.268) + 0.267 * \log_2(0.267) + 0.231 \log_2(0.231))$$

$$H_3 = -(0.07 * \log_2(0.07) + 0.07 * \log_2(0.07) + 0.07 * \log_2(0.07) + 0.79 \log_2(0.79))$$

$$I_3 = H_{max} - H_3 = 1.996 - 1.074 = 0.992$$

```
A: 0.233
C: 0.268
G: 0.267
T: 0.231
```

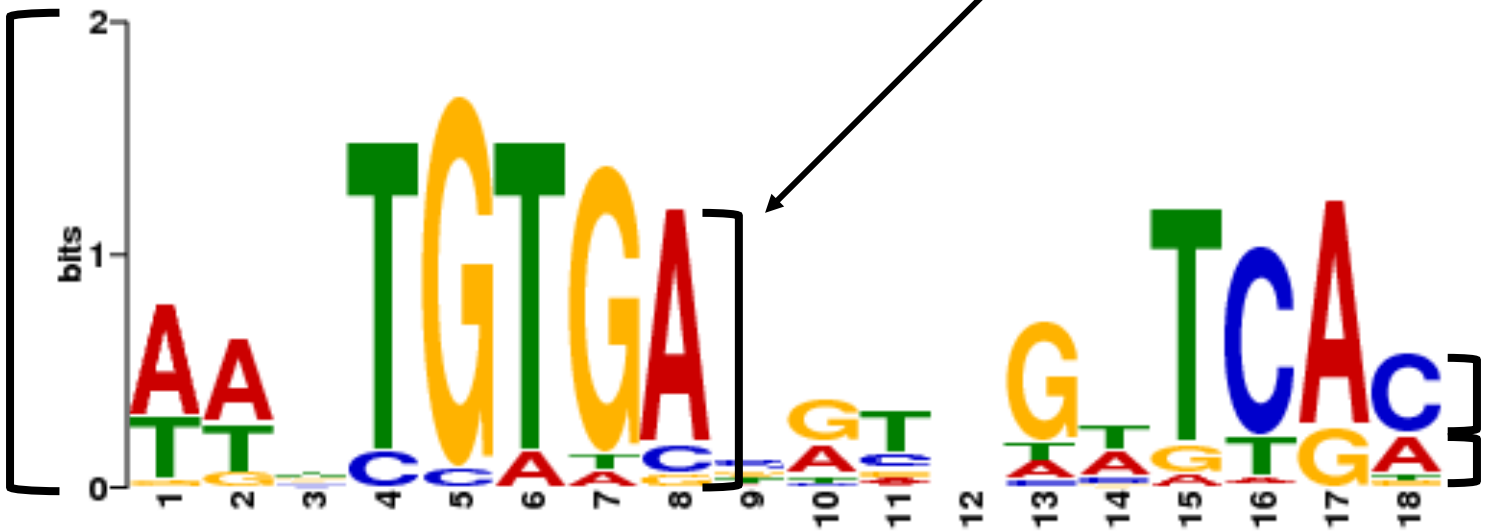
Visualizing motifs with Sequence Logos

Sequence logos are a great way to visualize not only the probability of different characters at each position in a motif, but also the ***uncertainty*** in the motif and the ***information content*** for each position

Entropy is represented by the total height of a column in *bits.*

The total height of characters in a column is the **Information Content.**

The height of an individual character is the **frequency of the residue.**



Reflection

- *What are the strengths of profiles in describing motifs compared to discrete representations?*
- *What two types of PSSM can you apply to a sequence to determine if it is an instance of a motif and how is it calculated?*
- *What are the calculations and steps involved in converting a sequence alignment to a Position Weight Matrix?*
- *How can you represent uncertainty in a profile?*
- *What is the Information Content of a profile and how is it calculated?*
- *What is a sequence logo and what parameters are needed to construct one?*