Exercises - Week 13

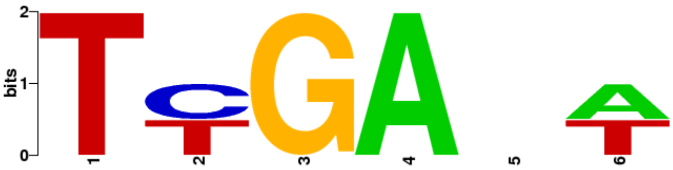
Genomics and bioinformatics

1 Motif model

The consensus is T {C,T} GA {A,C,G,T} {A,T}, or TYGANW using the IUPAC convention. The matrix M is

	1	2	3	4	5	6
A	0	0	0	1	1/4	1/2
C	0	1/2	0	0	1/4	0
G	0	0	1	0	1/4	0
T	1	1/2	0	0	1/4	1/2

The information content is $I = (2, 1, 2, 2, 0, 1)$. The non-zero logo heights are: $\text{Height}_{T1} = 2$, $\text{Height}_{C2} = \text{Height}_{T2} = 1/2$, $\text{Height}_{G3} = 2$, $\text{Height}_{A4} = 2$, $\text{Height}_{A6} = \text{Height}_{T6} = 1/2$. The corresponding logo is[1]



2 Motif finding

1) The $N = 10$ possible substrings of length $L = 6$ are

```
ATTGAC
TTGACA
TGACAC
CCTTGA
CTTGAC
TTGACA
TTGACA
ATTGAC
TTGACA
TGACAC
```

[1] One can use http://weblogo.berkeley.edu/logo.cgi without the "Small Sample Correction" option.

2) The initial $10 \cdot M$ is

	1	2	3	4	5	6
A	2	0	2	4	5	5
C	2	1	0	2	4	5
G	0	2	4	3	1	0
T	6	7	4	1	0	0

3) The $N = 10$ probabilities are

$$\begin{aligned} p_1 = p_5 = p_8 = 2 \cdot 7 \cdot 4 \cdot 3 \cdot 5 \cdot 5 \cdot 1/10^6 &= 4.2 \cdot 10^{-3} \\ p_2 = p_6 = p_7 = p_9 = 6 \cdot 7 \cdot 4 \cdot 4 \cdot 4 \cdot 5 \cdot 1/10^6 &= 1.344 \cdot 10^{-2} \\ p_3 = p_{10} = 6 \cdot 2 \cdot 2 \cdot 2 \cdot 5 \cdot 5 \cdot 1/10^6 &= 1.2 \cdot 10^{-3} \\ p_4 = 2 \cdot 1 \cdot 4 \cdot 1 \cdot 1 \cdot 5 \cdot 1/10^6 &= 4 \cdot 10^{-5} \end{aligned}$$

4) We have

$$Const = \sum_{k=1}^{N} p_k = 3 \cdot 4.2 \cdot 10^{-3} + 4 \cdot 1.344 \cdot 10^{-2} + 2 \cdot 1.2 \cdot 10^{-3} + 4 \cdot 10^{-5} = 0.0688 \ .$$

The updated $Const \cdot M$ is

	1	2	3	4	5	6
A	$p_1 + p_8$	0	$p_3 + p_{10}$	$p_2 + p_6 + p_7 + p_9$	$p_1 + p_3 + p_5 + p_8 + p_{10}$	$p_2 + p_4 + p_6 + p_7 + p_9$
C	$p_4 + p_5$	p_4	0	$p_3 + p_{10}$	$p_2 + p_6 + p_7 + p_9$	$p_1 + p_3 + p_5 + p_8 + p_{10}$
G	0	$p_3 + p_{10}$	$p_2 + p_6 + p_7 + p_9$	$p_1 + p_5 + p_8$	p_4	0
T	$p_2 + p_3 + p_6 + p_7 + p_9 + p_{10}$	$p_1 + p_2 + p_5 + p_6 + p_7 + p_8 + p_9$	$p_1 + p_4 + p_5 + p_8$	p_4	0	0

In summary, the updated M is approximately

	1	2	3	4	5	6
A	0.12	0	0.03	0.78	0.22	0.78
C	0.06	0	0	0.03	0.78	0.22
G	0	0.03	0.78	0.18	0	0
T	0.82	0.96	0.18	0	0	0

5) The consensus was TT {G,T} AA {A,C} and is now TTGACA, which was expected since TTGACA appears in each of the four binding sites.