


Detecting Motifs

A **motif** is a sequence pattern of functional significance.
Example: The **TATA box** is a motif that helps the polymerase find the transcription start site.

Table MM2.1 Nucleotide frequencies in 389 known TATA boxes.

Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	61	16	352	3	354	268	360	222	155	56	83	82	82	68	77
C	145	46	0	10	0	0	3	2	44	135	147	127	118	107	101
G	152	18	2	2	5	0	10	44	157	150	128	128	128	139	140
T	31	309	35	374	30	121	6	121	33	48	31	52	61	75	71

Copyright © 2006 A. Malcolm Campbell ©2018 Sami Khuri



Sequence Motifs

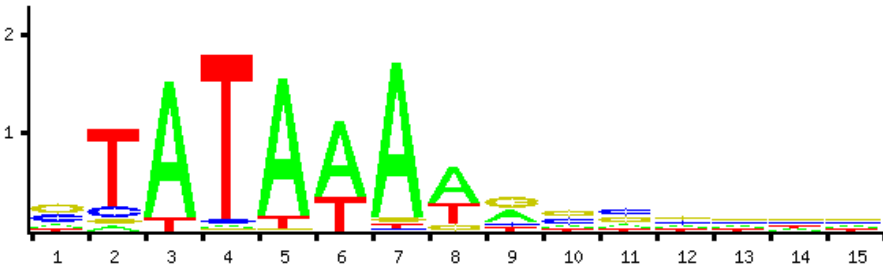



Table MM2.1 Nucleotide frequencies in 389 known TATA boxes.

Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	61	16	352	3	354	268	360	222	155	56	83	82	82	68	77
C	145	46	0	10	0	0	3	2	44	135	147	127	118	107	101
G	152	18	2	2	5	0	10	44	157	150	128	128	128	139	140
T	31	309	35	374	30	121	6	121	33	48	31	52	61	75	71

Copyright © 2006 A. Malcolm Campbell ©2018 Sami Khuri




Creating Tables of Frequencies

The probability of having an A in the first position is: $61/389 = 0.1568$
 The probability of a T in the second position is: $309/389 = 0.7943$
 Similarly for all 4 bases at all 15 positions.
 We can thus create a table of frequencies.

Table MM2.1 Nucleotide frequencies in 389 known TATA boxes.

Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	61	16	352	3	354	268	360	222	155	56	83	82	82	68	77
C	145	46	0	10	0	0	3	2	44	135	147	127	118	107	101
G	152	18	2	2	5	0	10	44	157	150	128	128	128	139	140
T	31	309	35	374	30	121	6	121	33	48	31	52	61	75	71

Copyright © 2006 A. Malcolm Campbell ©2018 Sami Khuri



Creating Log-Odds Tables

Instead of creating a table of frequencies, we create a table of log-odds.
 Suppose that the genome-wide average G and C content is 44%.
 Then the probability of an A is $0.56/2 = 0.28$.

$\log_2(0.1568/0.28) = \log_2(0.56) = -0.84$.
 Note that the base of the logarithm here is 2.
 Similarly, $\log_2(0.7943/0.28) = 1.5$.

Table MM2.1 Nucleotide frequencies in 389 known TATA boxes.

Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	61	16	352	3	354	268	360	222	155	56	83	82	82	68	77
C	145	46	0	10	0	0	3	2	44	135	147	127	118	107	101
G	152	18	2	2	5	0	10	44	157	150	128	128	128	139	140
T	31	309	35	374	30	121	6	121	33	48	31	52	61	75	71

Copyright © 2006 A. Malcolm Campbell ©2018 Sami Khuri

The Log-Odds Tables

Table MM2.1 Nucleotide frequencies in 389 known TATA boxes.

Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	61	16	352	3	354	268	360	222	155	56	83	82	82	68	77
C	145	46	0	10	0	0	3	2	44	135	147	127	118	107	101
G	152	18	2	2	5	0	10	44	157	150	128	128	128	139	140
T	31	309	35	374	30	121	6	121	33	48	31	52	61	75	71

↓

Table MM2.2 Position weight matrix.

A	-0.84	-2.77	1.69	-5.18	1.70	1.30	1.76	1.03	0.51	-0.96	-0.39	-0.41	-0.41	-0.68	-0.50
C	0.76	-0.90	-99.00	-3.10	-99.00	-99.00	-4.80	-5.42	-0.96	0.66	0.78	0.57	0.46	0.32	0.24
G	0.83	-2.25	-5.42	-5.42	-4.10	-99.00	-3.06	-0.96	0.88	0.81	0.58	0.58	0.58	0.70	0.71
T	-1.81	1.50	-1.64	1.78	-1.86	0.15	-4.14	0.15	-1.72	-1.18	-1.81	-1.07	-0.84	-0.54	-0.62


Copyright © 2006 A. Malcolm Campbell ©2018 Sami Khuri

Taking Log-Odds

$$\frac{P(\text{observed})}{P(\text{expected})} \text{ is } \begin{cases} > 1 \\ = 1 \\ < 1 \end{cases}$$

$$\log_b \left(\frac{P(\text{observed})}{P(\text{expected})} \right) \text{ is } \begin{cases} > 0 \\ = 0 \\ < 0 \end{cases}$$


©2018 Sami Khuri



What is the Significance of Log-Odds

- If the nucleotide is **more likely** to occur at a given position than it is to occur overall, the ratio will be **bigger than 1.0** and the **log odds is positive**.
- If the nucleotide is **less likely** to occur at a certain position than it is to occur overall, then the ratio will be **smaller than 1.0** and the **log odds is negative**.

©2018 Sami Khuri



Using Log-Odds Tables (I)

Table MM2.2 Position weight matrix.


A	-0.84	-2.77	1.69	-5.18	1.70	1.30	1.76	1.03	0.51	-0.96	-0.39	-0.41	-0.41	-0.68	-0.50
C	0.76	-0.90	-99.00	-3.10	-99.00	-99.00	-4.80	-5.42	-0.96	0.66	0.78	0.57	0.46	0.32	0.24
G	0.83	-2.25	-5.42	-5.42	-4.10	-99.00	-3.06	-0.96	0.88	0.81	0.58	0.58	0.58	0.70	0.71
T	-1.81	1.50	-1.64	1.78	-1.86	0.15	-4.14	0.15	-1.72	-1.18	-1.81	-1.07	-0.84	-0.54	-0.62

Table MM2.3 PWM score of the 15 bp sequence ACATATATAAGCTGG.

	A	C	A	T	A	T	A	T	A	A	G	C	T	G	G
A	-0.84	-2.77	1.69	-5.18	1.70	1.30	1.76	1.03	0.51	-0.96	-0.39	-0.41	-0.41	-0.68	-0.50
C	0.76	-0.90	-99.00	-3.10	-99.00	-99.00	-4.80	-5.42	-0.96	0.66	0.78	0.57	0.46	0.32	0.24
G	0.83	-2.25	-5.42	-5.42	-4.10	-99.00	-3.06	-0.96	0.88	0.81	0.58	0.58	0.58	0.70	0.71
T	-1.81	1.50	-1.64	1.78	-1.86	0.15	-4.14	0.15	-1.72	-1.18	-1.81	-1.07	-0.84	-0.54	-0.62

Table MM2.2 was constructed as explained in the previous slides; in other words, by taking the log of the ratio of the observed frequency over the expected frequency.

Copyright © 2006 A. Malcolm Campbell ©2018 Sami Khuri



Using Log-Odds Tables (II)

Table MM2.2 Position weight matrix.


A	-0.84	-2.77	1.69	-5.18	1.70	1.30	1.76	1.03	0.51	-0.96	-0.39	-0.41	-0.41	-0.68	-0.50
C	0.76	-0.90	-99.00	-3.10	-99.00	-99.00	-4.80	-5.42	-0.96	0.66	0.78	0.57	0.46	0.32	0.24
G	0.83	-2.25	-5.42	-5.42	-4.10	-99.00	-3.06	-0.96	0.88	0.81	0.58	0.58	0.58	0.70	0.71
T	-1.81	1.50	-1.64	1.78	-1.86	0.15	-4.14	0.15	-1.72	-1.18	-1.81	-1.07	-0.84	-0.54	-0.62

Table MM2.3 PWM score of the 15 bp sequence ACATATATAAGCTGG.

	A	C	A	T	A	T	A	T	A	A	G	C	T	G	G
A	-0.84	-2.77	1.69	-5.18	1.70	1.30	1.76	1.03	0.51	-0.96	-0.39	-0.41	-0.41	-0.68	-0.50
C	0.76	-0.90	-99.00	-3.10	-99.00	-99.00	-4.80	-5.42	-0.96	0.66	0.78	0.57	0.46	0.32	0.24
G	0.83	-2.25	-5.42	-5.42	-4.10	-99.00	-3.06	-0.96	0.88	0.81	0.58	0.58	0.58	0.70	0.71
T	-1.81	1.50	-1.64	1.78	-1.86	0.15	-4.14	0.15	-1.72	-1.18	-1.81	-1.07	-0.84	-0.54	-0.62

To see if a sequence of length 15 is a TATA box, we simply add the corresponding values from the PWM and see if we get a value above some threshold.
 In the example above, we add the 15 highlighted numbers to get 6.78.

Copyright © 2006 A. Malcolm Campbell ©2018 Sami Khuri



Consensus Sequence and PWM

- All current methods for representing DNA motifs involve either consensus sequences or probabilistic models (such as PWM) of the motif.
- Consensus sequences do not adequately represent the variability seen in promoters or transcription factor binding sites.
- Both consensus sequences and PWM models assume positional independence. Neither method can accommodate correlations between positions.
- Probabilities calculated from PWM models can be highly misleading.

©2018 Sami Khuri