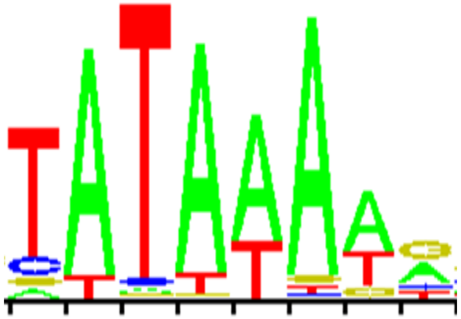
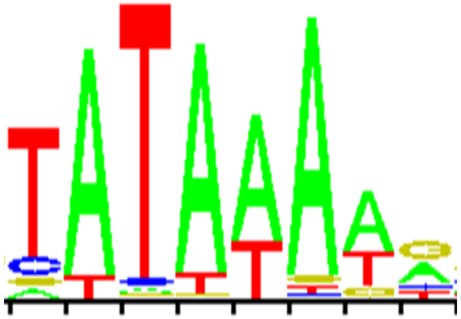
Bioinformatics

Nine
Motifs and Logos



Wendy Lee
Department of Computer Science
San José State University

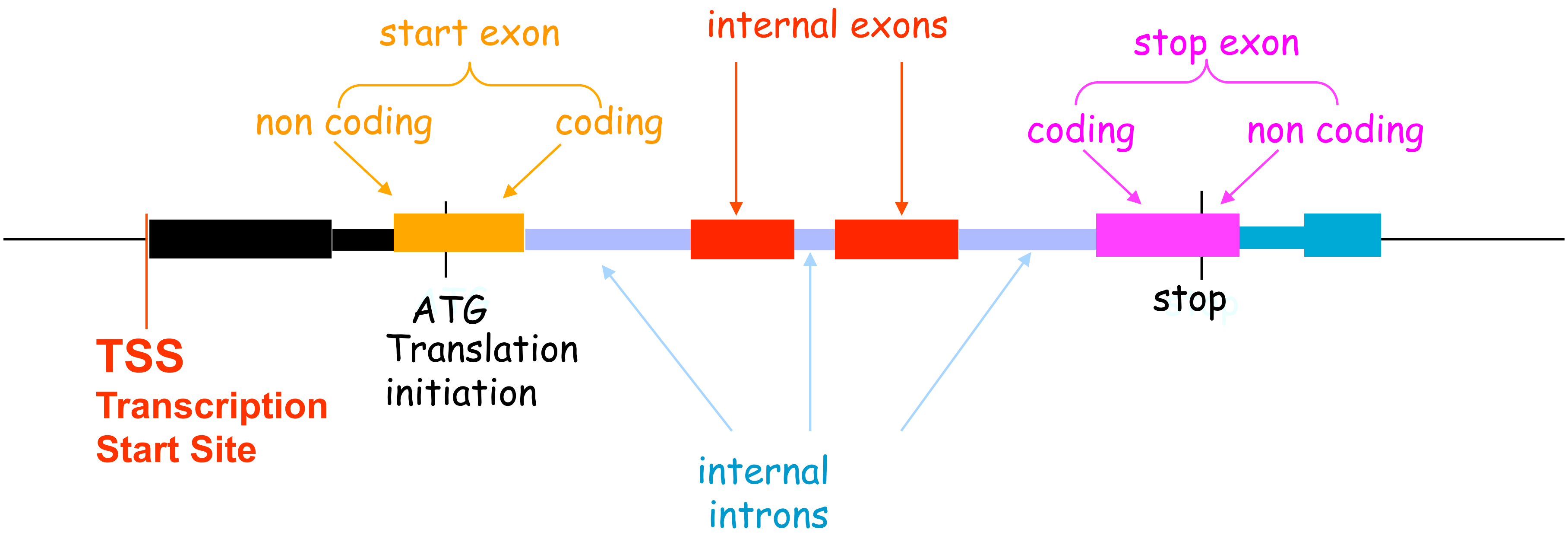
Biology/CS/SE 123A
Fall 2014



Importance and Abundance of Motifs

- DNA **motifs** are nucleotide sequence patterns of functional significance.
- **Examples**:
  - The **TATA box** is a motif that helps RNA polymerase find the transcription start site (TSS) in many eukaryotic genes.
  - The **CAT box** is another highly conserved region used for the initiation of transcription.

Getting the CDS



start exon

non coding

coding

internal exons

stop exon

coding

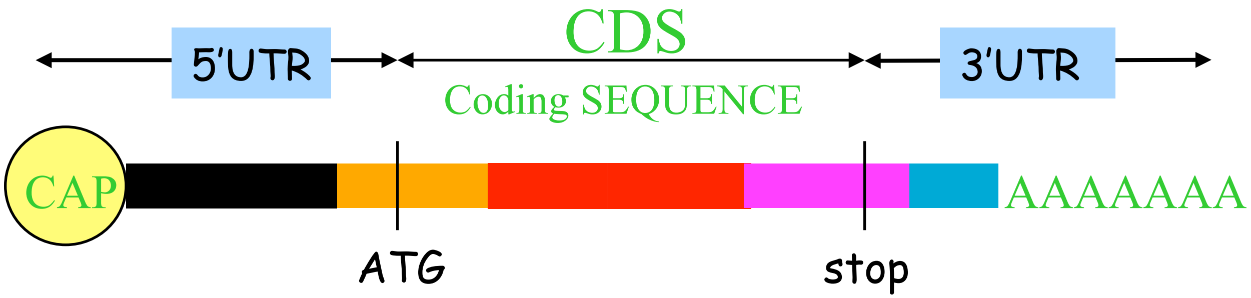
non coding

ATG
Translation
initiation

stop

TSS
Transcription
Start Site

internal
introns



5'UTR

CDS
Coding SEQUENCE

3'UTR

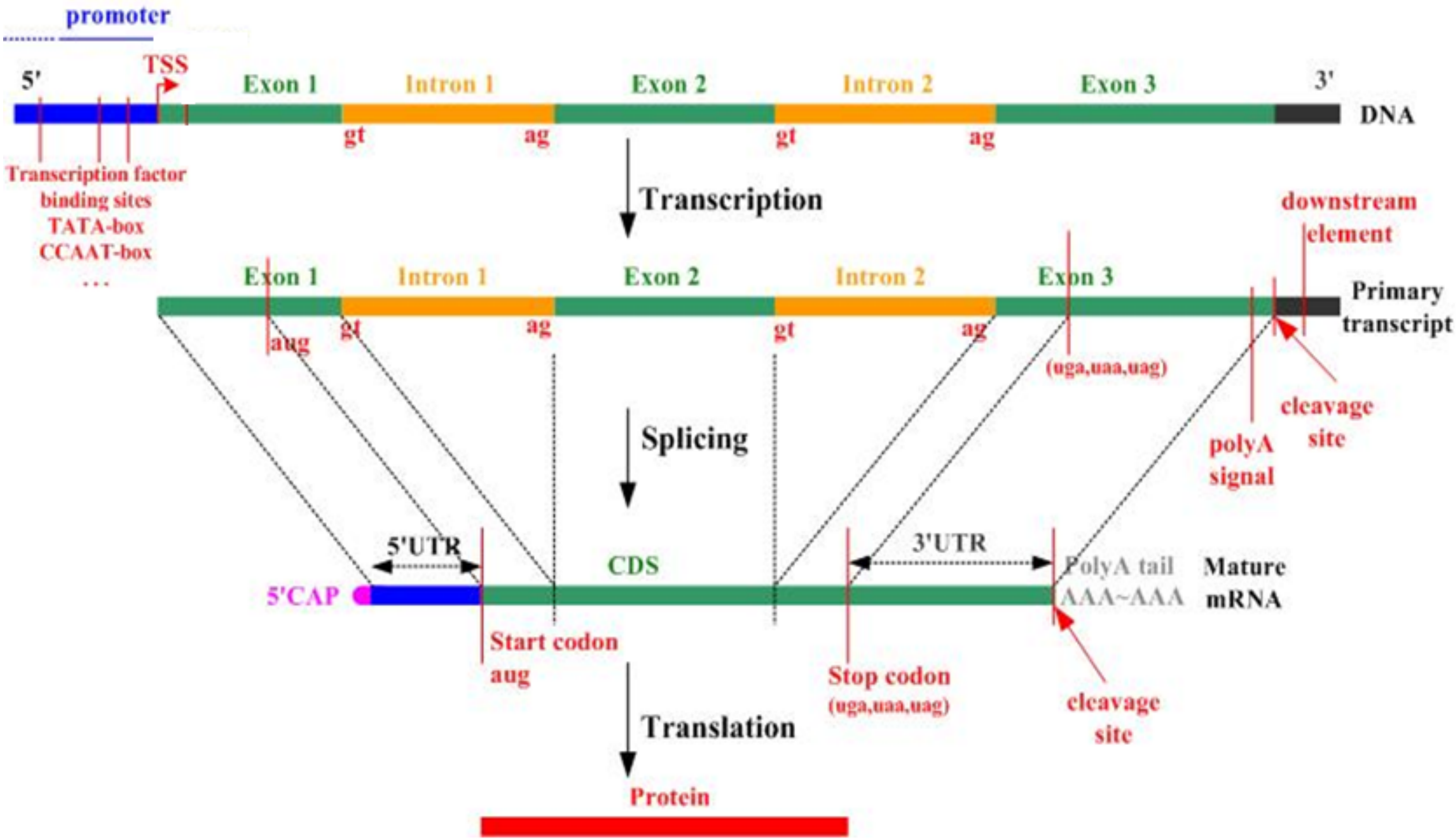
CAP

AAAAAAA

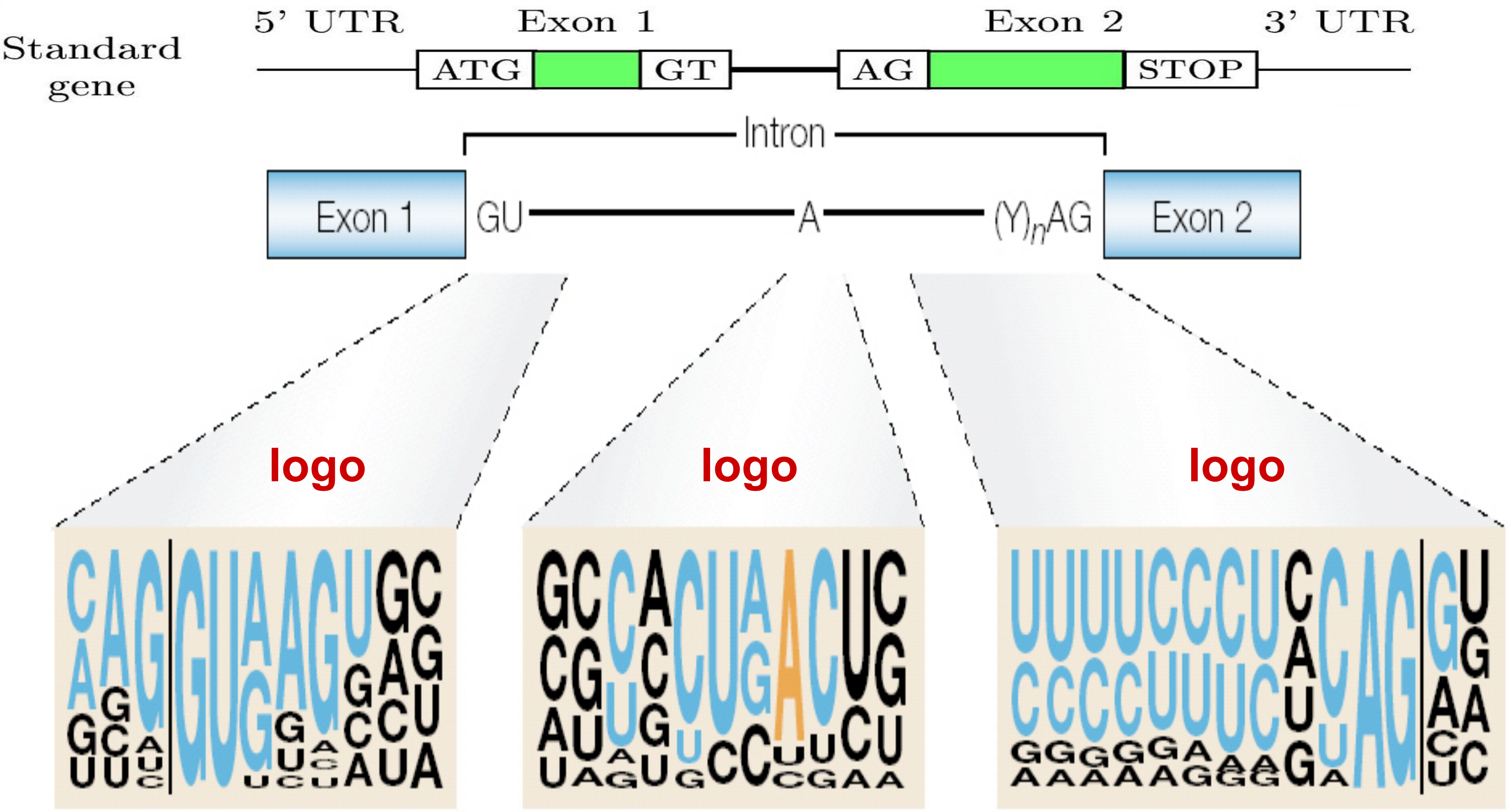
ATG

stop

From DNA to Protein



Anatomy of an Intron



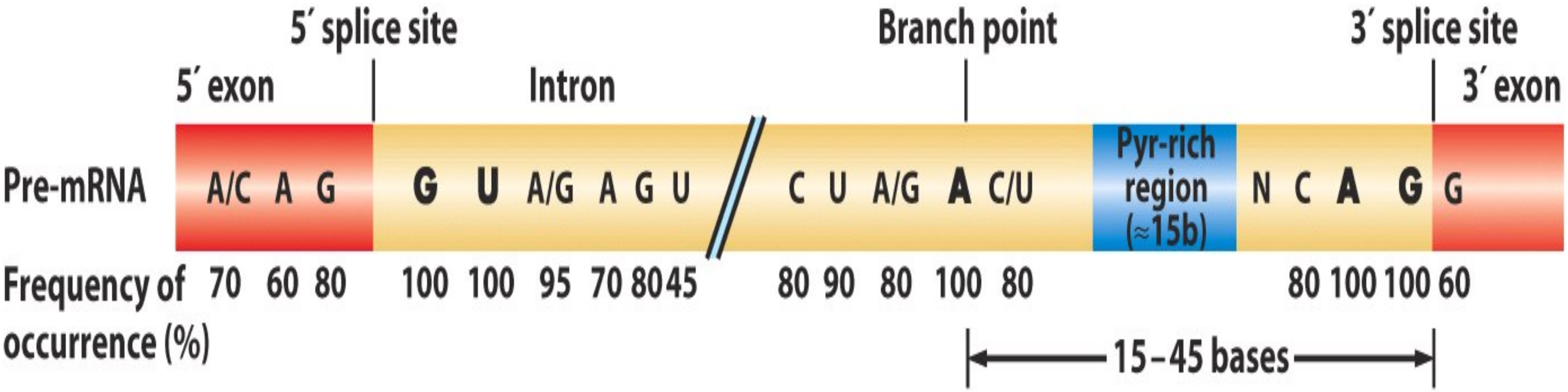
5′ splice site

Branch site

3′ splice site

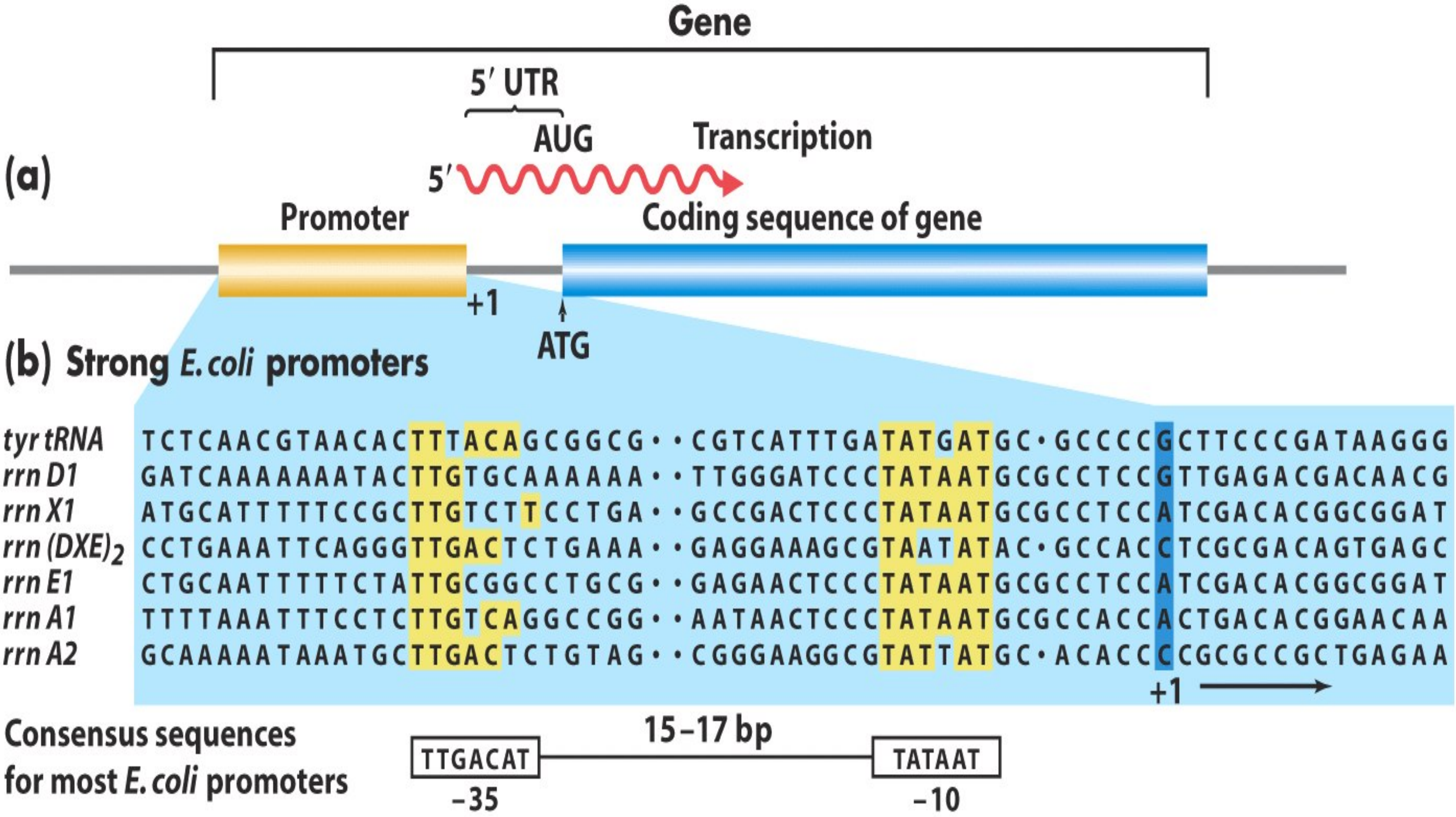
©2014 Wendy Lee

Conserved Sequences in Introns



The conserved nucleotides in the transcript are recognized by small nuclear ribonucleoprotein particles (snRNPs), which are complexes of protein and small nuclear RNA. A functional splicing unit is composed of a team of snRNPs called a spliceosome.

E.Coli Promoter Sequences



tyr tRNA	TCTCAACGTAACACTTTACAGCGGCG••CGTCATTTGATATGATGC•GCCCCGCTTCCCGATAAGGG
rrn D1	GATCAAAAAAATACTTGTGCAAAAAA••TTGGGATCCCTATAATGCGCCTCCGTTGAGACGACAACG
rrn X1	ATGCATTTTTCCGCTTGTCTTCCTGA••GCCGACTCCCTATAATGCGCCTCCATCGACACGGCGGAT
rrn (DXE)$_2$	CCTGAAATTCAGGGTTGACTCTGAAA••GAGGAAAGCGTAATATAC•GCCACCTCGCGACAGTGAGC
rrn E1	CTGCAATTTTTCTATTGCGGCCTGCG••GAGAACTCCCTATAATGCGCCTCCATCGACACGGCGGAT
rrn A1	TTTTAAATTTCCTCTTGTCAGGCCGG••AATAACTCCCTATAATGCGCCACCACTGACACGGAACAA
rrn A2	GCAAAAATAAATGCTTGACTCTGTAG••CGGGAAGGCGTATTATGC•ACACCCCGCGCCGCTGAGAA

Sequence Motifs

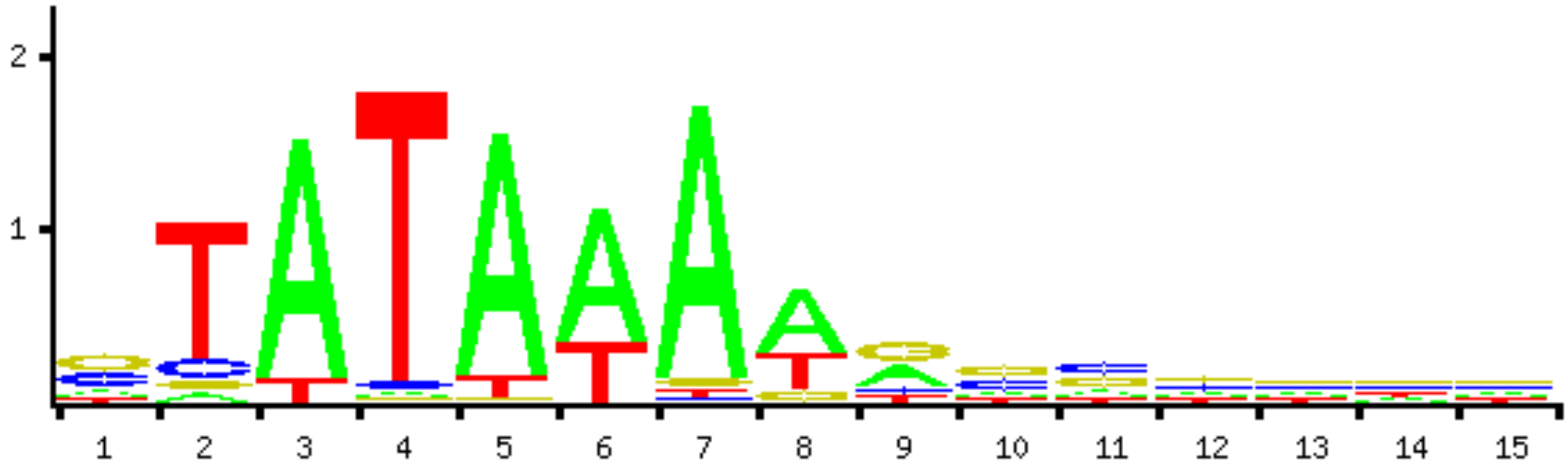


Table MM2.1 Nucleotide frequencies in 389 known TATA boxes.

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

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.

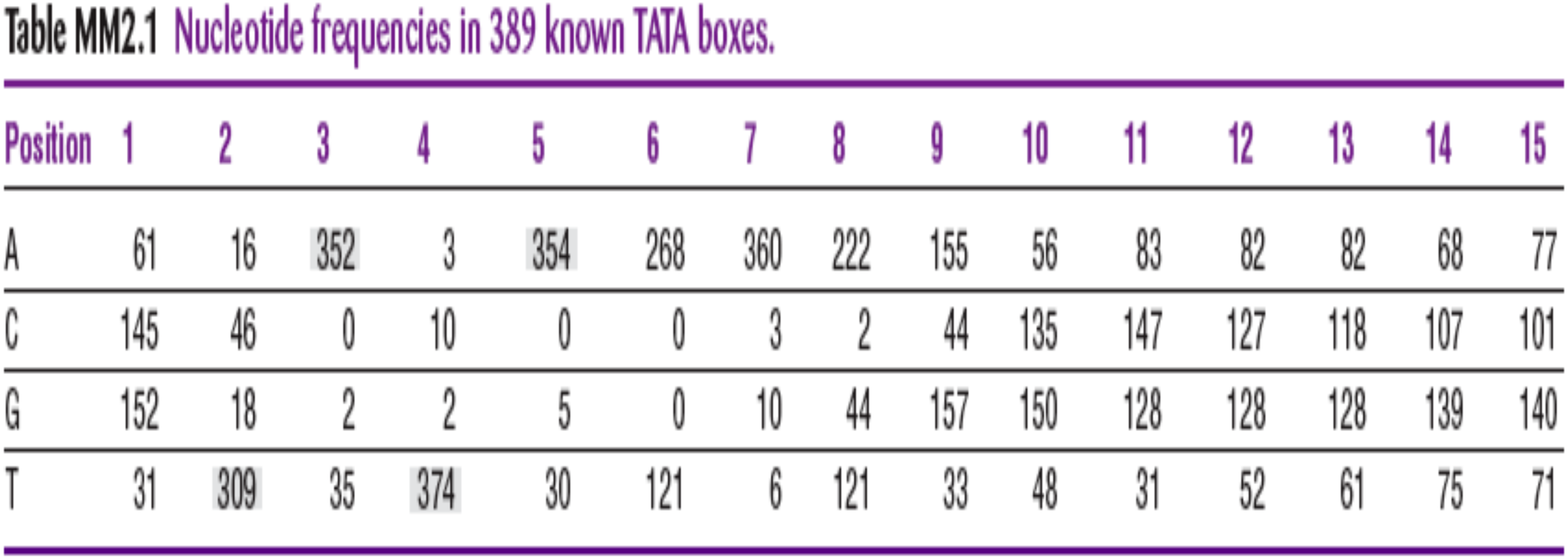


Table MM2.1 Nucleotide frequencies in 389 known TATA boxes.

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

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.

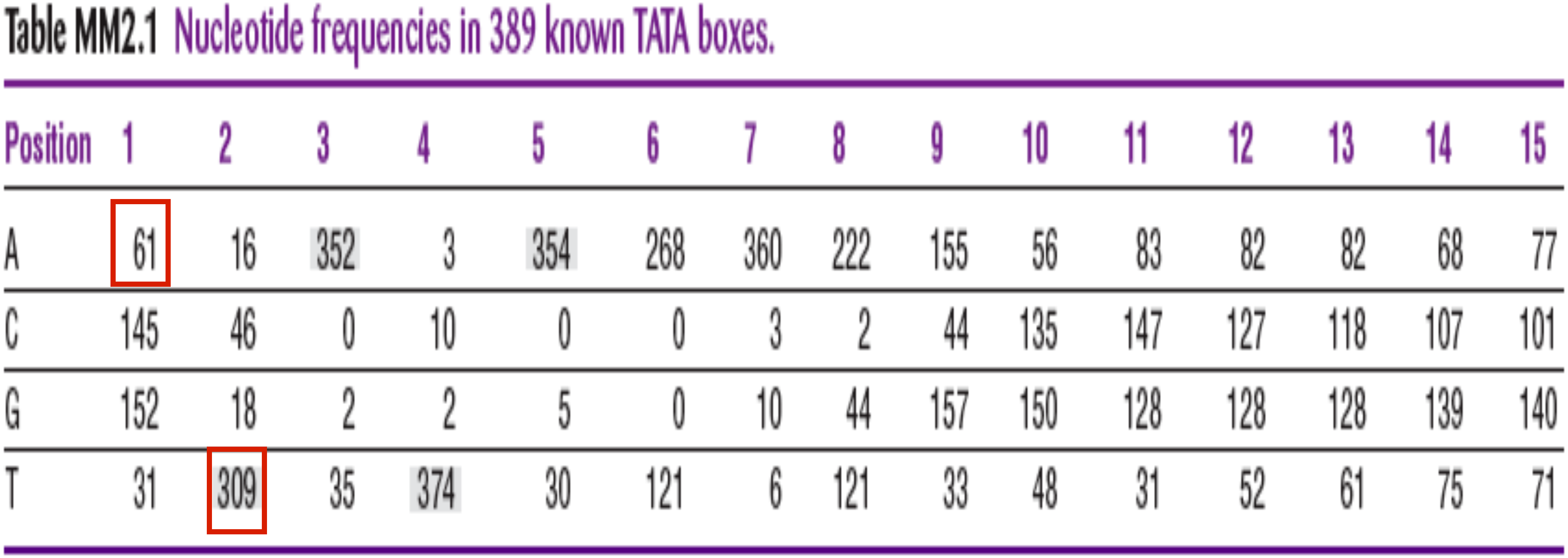


Table MM2.1 Nucleotide frequencies in 389 known TATA boxes.

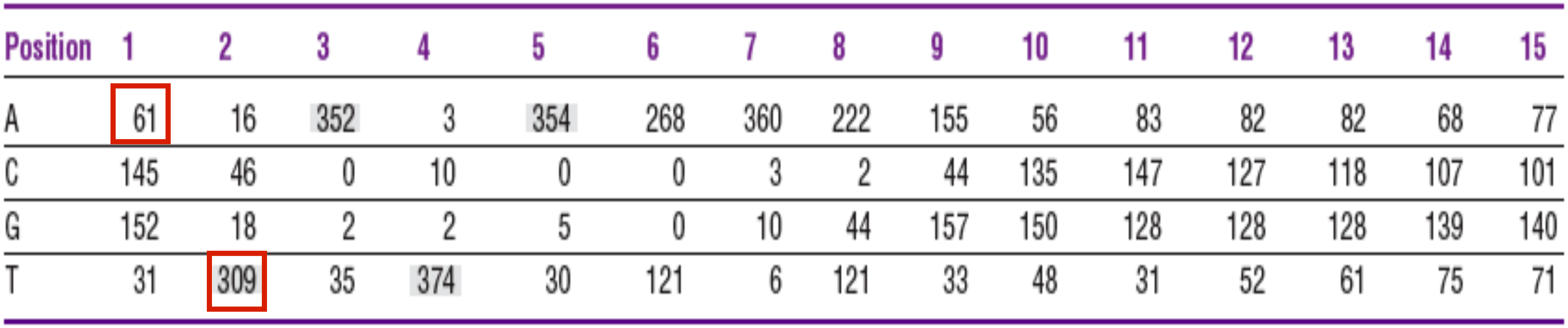
Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	61	16	352	3	354	268	360	222	155	56	83	82	82	68	77
C	145	46	0	10	0	0	3	2	44	135	147	127	118	107	101
G	152	18	2	2	5	0	10	44	157	150	128	128	128	139	140
T	31	309	35	374	30	121	6	121	33	48	31	52	61	75	71

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

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

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**.

Using Log-Odds Tables (I)

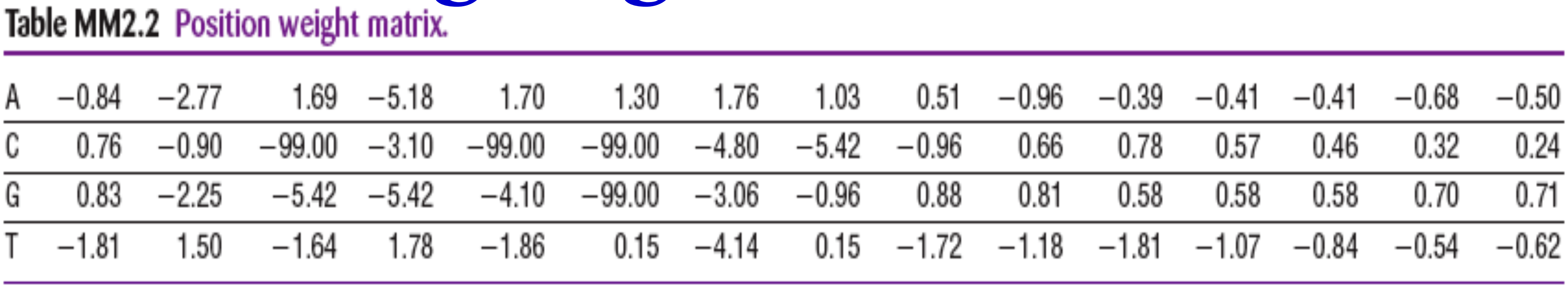


Table MM2.2 Position weight matrix.

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

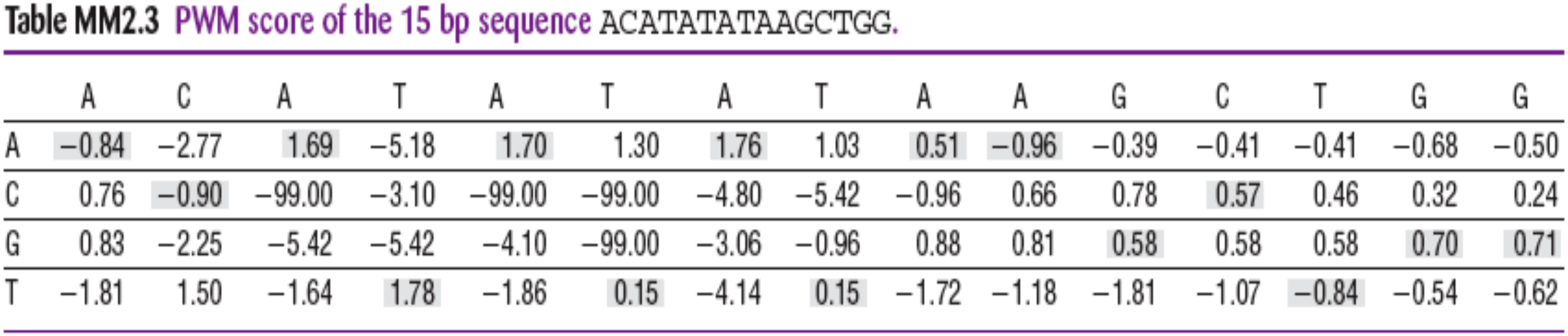


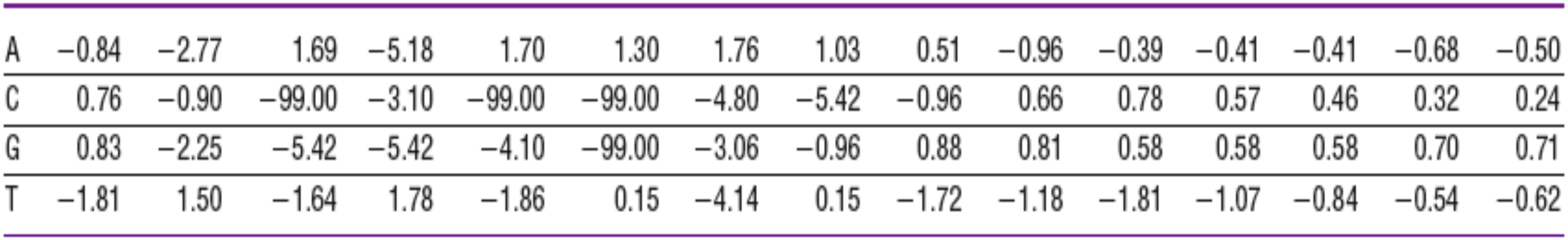
Table MM2.3 PWM score of the 15 bp sequence ACATATATAAGCTGG.

	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.

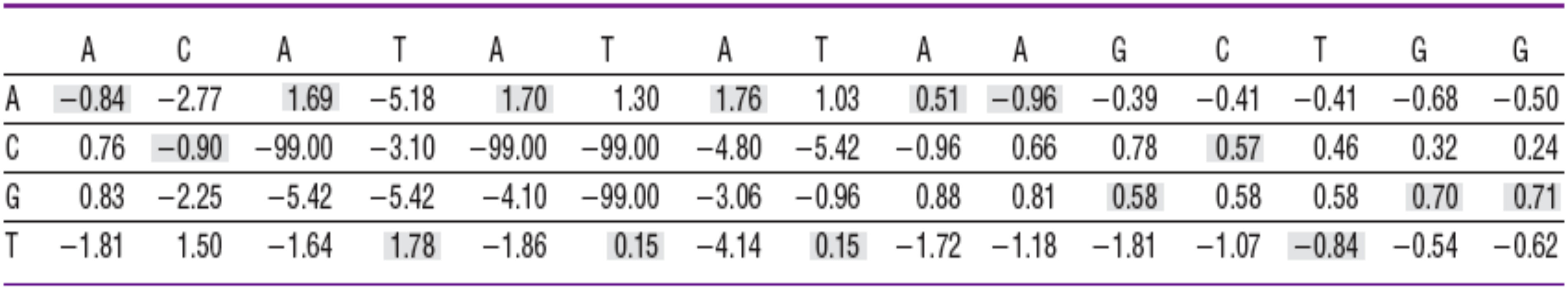
Using Log-Odds Tables (II)

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 threshhold.
In the example above, we add the 15 highlighted numbers to get 6.78.

Designing Logos

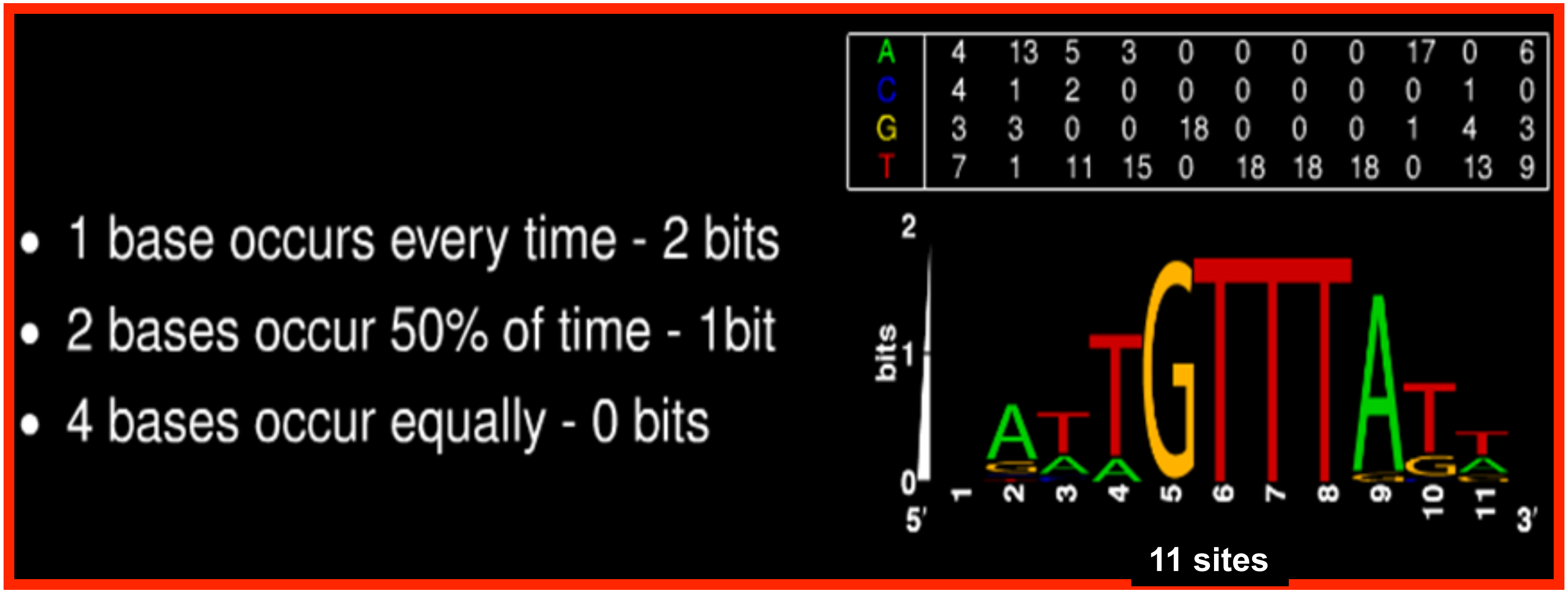
- A **logo** is a visual representation of a set of aligned sequences that indicates the positional preferences as given by **information theory**.
- A **logo** gives a visual representation of the motif.
- The size of the character in the stack of characters is proportional to the character’s frequency in that position.
- The total height of each column is proportional to its **information** content.
- **Information theory** quantifies the amount of information

Logos with Bases

- Define:

$$I_j = log_2(4) - H_j = 2 + \sum f_{x,j} \, log_2 \, (f_{x,j})$$

where $f_{x.j}$ is the frequency of character x at position j.



- 1 base occurs every time - 2 bits
- 2 bases occur 50% of time - 1bit
- 4 bases occur equally - 0 bits

A	4	13	5	3	0	0	0	0	17	0	6
C	4	1	2	0	0	0	0	0	0	1	0
G	3	3	0	0	18	0	0	0	1	4	3
T	7	1	11	15	0	18	18	18	0	13	9

11 sites