

The size of introns can vary, ranging from 70 bp to more than 30,000 bp. The average human gene is 30,000 bp long, but the dystrophin gene is over 2 million bp. In addition to the complexity of real genes, the mammalian genome also contains sections of approximately 225 bp

per kb that look like genes but are not transcribed. These 225 bp gene look-alikes are called **pseudogenes**, because a mutation has rendered them nonfunctional. So finding a real gene amidst all this complexity is a very difficult task.

Math Minute 2.2 How Do You Find Motifs?

To locate genes in newly sequenced eukaryotic genomes, we can use the fact that the sequence upstream of a gene contains certain motifs (nucleotide sequence patterns of functional significance). However, we do not know in advance what the sequence patterns are. To complicate matters further, the patterns can vary significantly from gene to gene and organism to organism. Every transcription factor that regulates gene expression has a preferred motif to which it binds—but how can we find these binding sites if we do not know what the transcription factors are, much less which genes they regulate? To get started, bench scientists had to find some binding sites through painstaking molecular biology methods. Once a few binding sites had been identified, a computational analysis using a position weight matrix (PWM) helped find more.

The TATA box is a motif that helps RNA polymerase find the transcription start site in many eukaryotic genes. A similar motif appears in prokaryotic genes as well. We will use the TATA box to explore the PWM method. Although the consecutive letters TATA are the heart of this motif, these four letters also occur randomly in many places that are not immediately upstream of a gene. By discovering some true TATA boxes, investigators began to characterize the sequence pattern that distinguishes true TATA boxes from the random TATA-like sequences. Table MM2.1 summarizes the TATA box sequences for 389 different eukaryotic genes (note that the sum of every column is 389).


The four letters TATA for which the motif is named appear in positions 2–5, but even these four letters are not always TATA. For example, position 2 contains a T only about 80% of the time. Table MM2.1 also shows that position 6 is either A or T, position 7 is almost always A, position 8 is usually A or T, and positions 1 and 10–15 have a slight tendency to be G or C.

Suppose we are examining 15 bp in a potential promoter of a newly sequenced genome. How can we encapsulate all the information in Table MM2.1 to help decide if these 15 bp form a real TATA box? The basic idea is to compare the probability of these 15 letters occurring in a TATA box to the overall probability of these 15 letters occurring anywhere in the genome. For example, if the genome-wide average GC content is 44%, the probability of an A in position 1 is 0.28 (0.56×0.5 ; probability of not-GC times the probability of A if not GC). However, if these 15 bp form a TATA box, the probability of seeing an A in position 1 is given by the relative frequency of A in position 1, $61/392 \approx 0.1556$.

The TATA probability (e.g., 0.1556) divided by the overall probability in this species (e.g., 0.28) indicates whether a letter is more or less likely to occur at a given position in a TATA box than it is to occur overall. To get the final PWM score, we must multiply the ratios across all 15 positions, but it is easier to work with the logarithm of the ratio, called the **log odds**, and add across all 15 positions. Traditionally, log base 2 is used, because it is easy to detect doublings in the probability.

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

LINKS 
 JASPAR
MATH MINUTE
 pwm.xls

If the letter is more likely to occur at a given position in a TATA box than it is to occur overall, the ratio of probabilities will be bigger than 1.0 and the log odds will be positive. In contrast, if the letter is less likely to occur at that position of a TATA box than it is to occur overall, the ratio of probabilities will be smaller than 1.0 and the log odds will be negative. In our example of determining the probability of seeing an A in position 1 of a real TATA box, the log odds is $\log_2(0.1556 / 0.28) = -0.84$, which means that an A is not very likely to be in the first position of a 15 bp TATA box. We compute the log odds for each of the other three letters that could be in position 1, and then repeat this process for all 15 positions. The log odds scores for the TATA box motif are given in a position weight matrix in Table MM2.2. Note that we use a large negative number (-99) whenever the log odds is undefined (i.e., ratio of probabilities is 0).

To measure the likelihood that a 15 bp query sequence is a TATA box, we sum the log odds scores for the 15 letters in the sequence. For each position, we use the row that matches the letter in that position. For example, suppose our sequence is ACATATATAAGCTGG. The log odds scores to be added are highlighted in Table MM2.3. The sum of all 15 highlighted scores (6.78) is the total PWM score of this sequence. By considering every 15 bp sequence in the genome using a sliding window (see Math Minute 2.1), we can identify sequences that are most likely to be TATA boxes (i.e., with the highest scores). By repeating this scoring process with many different PWMs, representing many known binding-site motifs, we can begin to deduce the location of genes in our newly sequenced genome.

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

MATH MINUTE DISCOVERY QUESTIONS

1. Go to [JASPAR](#) and select “Browse profiles by class”. Scroll down to the TATA box and click on the “View” button in this row. Verify that the values in Table MM2.1 are displayed in this window. Explain how the sequence logo represents the information in Table MM2.1.
2. Return to the JASPAR “Browse profiles by class” page. Find transcription factors with ID numbers MA0040, MA0041, and MA0047. By looking at the sequence logos, explain which of these three transcription factors in rat is most likely to bind to DNA containing the motif TGTTTA.
3. Use the spreadsheet [pwm.xls](#) to compute the total TATA PWM score of the following three sequences, and determine which one is most likely to be a true TATA box: ATATATATAGGCTGG, CTATATATATGCTGG, CTATAAATAGCCGG.
4. Use the spreadsheet [pwm.xls](#) to compute the total TATA PWM score of CCGCCTATTTATAG. Explain why the score is so high, even though this sequence does not look like a true TATA box. (Hint: How is this sequence related to one of the sequences in Math Minute Discovery Question 3?)