Interpreting a Sequence Logo

When initiating translation, ribosomes bind to an mRNA at a ribosome binding site upstream of the AUG start codon. Because mRNAs from different genes all bind to a ribosome, the genes encoding these mRNAs are likely to have a similar base sequence where the ribosomes bind. Therefore, candidate ribosome binding sites on mRNA can be identified by comparing DNA sequences (and thus the mRNA sequences) of several genes in a species, searching the region upstream of the start codon for shared (conserved) base sequences.

The DNA sequences of 149 genes from the *E. coli* genome were aligned with the aim to identify similar base sequences as potential ribosome binding sites. Rather than presenting the data as a series of 149 sequences aligned in a column (a sequence alignment), the researchers used a sequence logo.

The potential ribosome binding regions from 10 *E. coli* genes are shown in the sequence alignment in Figure 1. The sequence logo derived from the aligned sequences is shown in Figure 2. Note that the DNA shown is the nontemplate (coding) strand, which is how DNA sequences are typically presented.

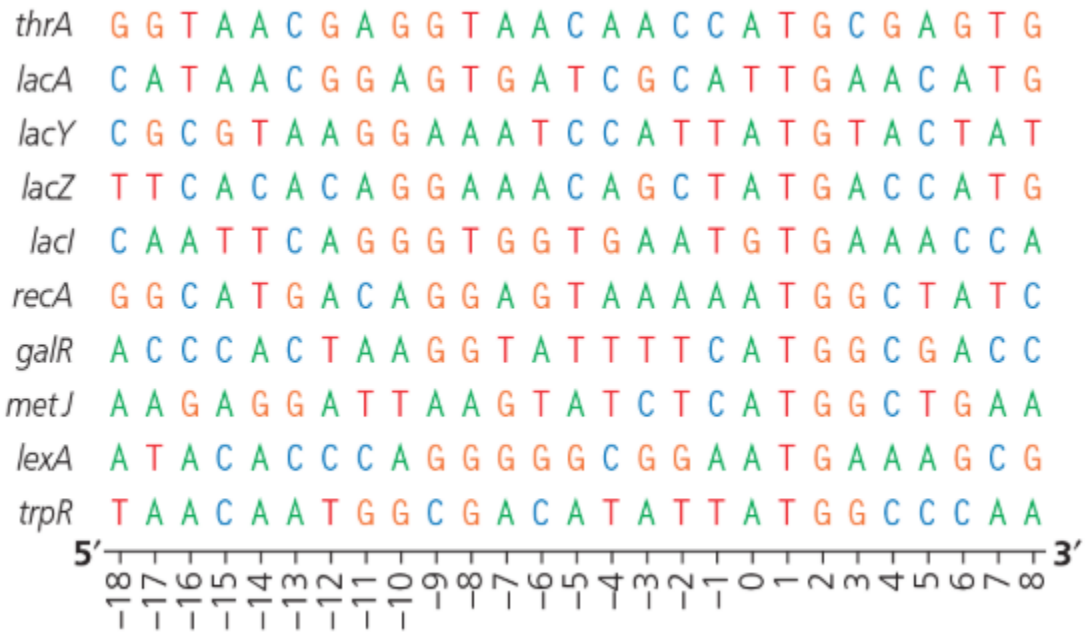

Figure 1 Sequence alignment for 10 *E. coli* genes.

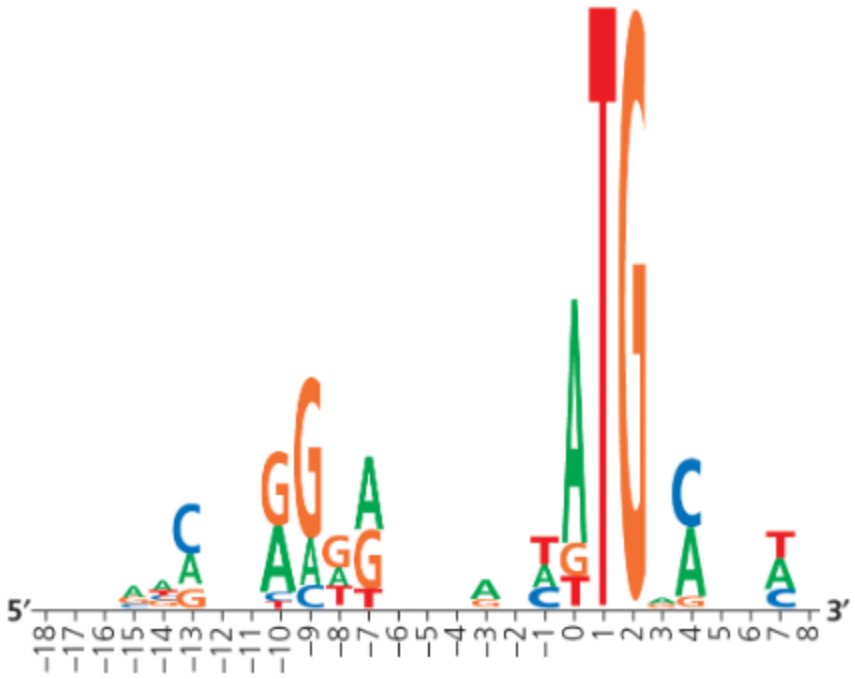

Figure 2 Sequence logo derived from sequence alignment

1) In the sequence logo, the horizontal axis shows the primary sequence of the DNA by nucleotide position. Letters for each base are stacked on top of each other according to their relative frequency at that position among the aligned sequences, with the most common base as the largest letter at the top of the stack. The height of each letter represents the relative frequency of that base *at that position*.
   a) Describe the frequency of bases at positions 0, 1 and 2 of the sequence alignment.
   b) Explain the significance of the frequency of bases at positions 0, 1 and 2 of the sequence alignment.
   c) Explain the significance of the height of a stack of letters in the sequence logo.
2) The height of a stack of letters in a logo allows us to predict what base will be in that position if a new sequence is added to the logo.
   a) Looking at the sequence logo, identify the two positions that have the most predictable bases.
   b) State the bases you predict would be at those positions in a newly sequenced gene.
   c) Predict the probability of the new sequence having T at position -14.
3) In the actual experiment, the researchers used 149 sequences to build their sequence logo, which is shown in Figure 3. Identify the three positions in the sequence logo in Figure 3 that have the most predictable bases and identify those bases.

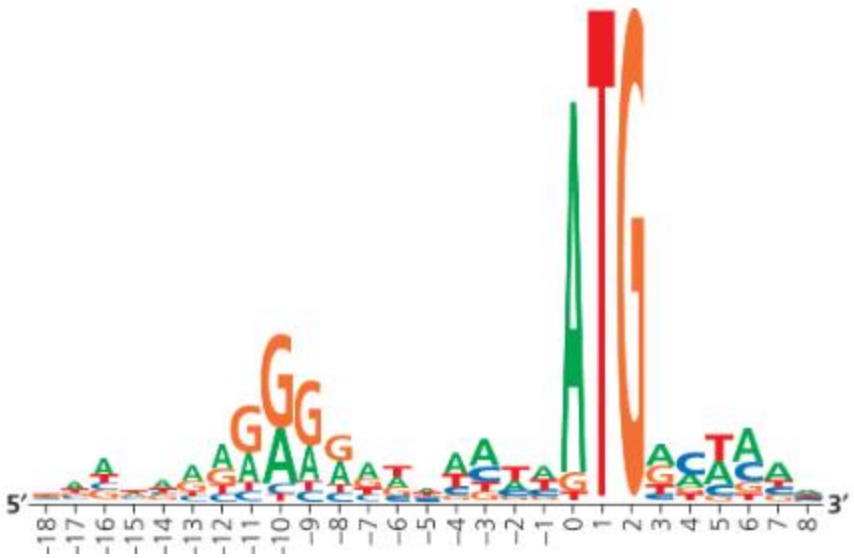

Figure 3 Sequence logo from 149 *E. coli* genes

4) A consensus sequence identifies the base occurring most often at each position in the set of sequences.
   a) Write the consensus sequence of this (the nontemplate) strand. In any position where the base can't be determined, use a dash.
   b) The consensus sequence does not show the relative frequencies of the bases, making it impossible to predict which base will be at each position in a newly-discovered sequence. Propose a value for a consensus sequence.
5) Based on the logo, identify the five adjacent base positions in the 5′ UTR region that are most likely to be involved in ribosome binding? Justify your response.