

Video Tutorial 2.1: The importance of reading frames

Nucleotide sequence is translated to amino acid sequence using a three-letter code. The nucleotide sequence in DNA and RNA, and the amino acid sequence in proteins are all linear. If you have a nucleotide sequence and the genetic code table, you can translate the sequence yourself.

As you'll already know, the DNA in a cell consists of a pair of molecules that run antiparallel to one another. Here, displayed in this sequence browsing software is a fragment of DNA, which contains the white gene of the fruit fly *Drosophila melanogaster*. We have obtained this gene from GenBank, a genetic sequence database, and uploaded it in a plain sequence format known as FASTA.

The DNA strand represented here and running left to right on the top of the number line, is the coding strand for the gene we will be looking at, and is oriented 5' to 3'. You are looking at the actual nucleotide sequence, which reads beginning GTT.

You will notice that the strand at the bottom is complementary to the coding strand. All the As are paired with Ts and the Cs with Gs. This complementary strand is the template strand, so named because it can be used to generate another copy of the coding strand by complementary base pairing.

The bottom or template strand is running 5' to 3' in the opposite direction, which is why we say that the strands are antiparallel to one another.

Now let us take the coding strand from the very start. You can now see the sequence of amino acids encoded by this DNA. The first three nucleotides are GTT. In the genetic code, this translates to Valine, represented as V in the single-letter shorthand for amino acids. When I highlight codons in this browser, I can see the translation above. So you can see that the second 'TCG' codon translates to a Serine or S and the third 'TGA' is a STOP codon. Notice that we are reading the amino acid translation on the top line of letters here.

Now that is the translation that we got beginning with the first nucleotide. Let's see what happens if we begin translating from the second nucleotide instead. Our second nucleotide is T and the first codon beginning with that nucleotide is therefore TTT, encoding phenylalanine (F).

As I go on translating, what you can see is that translation beginning at the second nucleotide generates a completely different sequence of amino acids, which we are reading on the second line of letters above the nucleotide sequence. This is a second possible reading frame for the nucleotide sequence.

If I begin translating at the third nucleotide, I will generate this third reading frame.

What happens if I begin on the fourth nucleotide? I skip the first codon but otherwise I'm effectively reading our first reading frame again.

Thus there are three possible reading frames for the coding strand.

I am going to scroll all along and take us all the way to the other end of the sequence to translate the complementary template strand in the 5' to 3' direction.

To translate 5' to 3', I have to read this strand backwards. Our first codon beginning at the first nucleotide is therefore CTA or Leucine and that is followed by a TAG or stop codon. Here is the second reading frame. Reading the first codon TAT from the second letter from the end and followed by AGG encoding R or Arginine.

Starting at the third nucleotide from the end, I read a frame that begins with ATA for isoleucine followed by GGT for glycine (G).

As you can see, the three amino acid sequences below the nucleotide sequence represent three possible reading frames for the bottom strand read in the 5' to 3' direction.

For any double-stranded DNA molecule, there are six possible reading frames.

In other words, the two strands of DNA you see here have six possible reading frames. There are six very different hypothetical ways of translating this gene into protein.

This video was narrated by Iruka N. Okeke and recorded and edited by Katherine Heston and Charles Woodard at Haverford College, PA, USA. We are grateful to Rachel Hoang and Philip Meneely for helpful suggestions.

You can get more information about the white gene from Flybase at <http://flybase.org/reports/FBgn0003996.html>

This video used unannotated (FASTA) records for the white gene (Genbank accession number NC_004354) retrieved from the NCBI Genbank database <https://www.ncbi.nlm.nih.gov/gene/31271> on 3 June 2015.

The sequences were viewed in Artemis, a genome browser available from <http://www.sanger.ac.uk/science/tools/artemis>. We thank Pathogen Informatics, The Wellcome Trust Sanger Institute for permission to use Artemis for this instructional video.