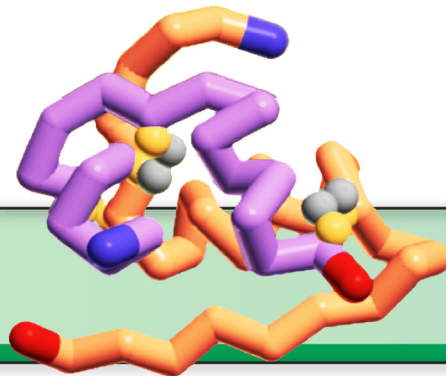




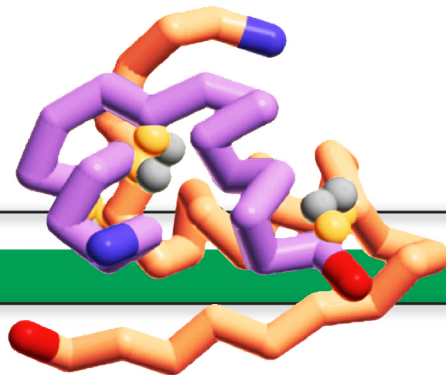
Exploring Further Determining a Consensus Sequence Activity

A **consensus sequence** is determined by aligning many nucleotide (or protein) sequences that share a common function, then determining the most commonly expressed nucleotide (or amino acid) at each position. Often conserved sequences reflect a common function or binding domain. In this exercise you will identify important nucleotides in the region for the initiation of translation.

Below is a table that lists the nucleotide sequence surrounding the AUG start codon (highlighted) for ten human genes. For each column, tally the frequency of each nucleotide (A, G, C and U) in the table below.

Position	-8	-7	-6	-5	-4	-3	-2	-1	1	2	3	4	5	6	7	8	9	10	11
Beta globin	C	A	G	A	C	A	C	C	A	U	G	G	U	G	C	A	C	C	U
Insulin	C	U	U	C	U	G	C	C	A	U	G	G	C	C	C	U	G	U	G
Amylase	A	A	A	G	C	A	A	A	A	U	G	A	A	G	C	U	C	U	U
Catalase	C	G	C	A	C	G	C	U	A	U	G	G	C	U	G	A	C	A	G
DNA polymerase alpha 1	U	C	G	G	G	A	C	C	A	U	G	G	C	A	C	C	U	G	U
Myosin	G	C	A	C	A	G	C	C	A	U	G	G	G	A	G	A	U	U	C
Alpha globin	A	A	C	C	C	A	C	C	A	U	G	G	U	G	C	U	G	U	C
Tubulin	G	A	C	U	C	A	C	C	A	U	G	C	G	U	G	A	A	U	G
Actin	U	A	G	A	C	A	C	A	A	U	G	U	G	C	G	A	C	G	A
tRNA synthetase	U	U	C	U	C	A	A	G	A	U	G	G	C	G	C	U	G	C	A

A																			
G																			
C																			
U																			
Position	-8	-7	-6	-5	-4	-3	-2	-1	1	2	3	4	5	6	7	8	9	10	11



Determining a Consensus Sequence Activity (continued)

Because there are four nucleotides that are possible at each position in the sequence, if the distribution of these nucleotides is totally random, you would expect the frequency of each nucleotide to be $\frac{1}{4}$, or 25%. This is a small sample, so we don't expect a perfect match to this distribution. Look for any position at which one of the nucleotides is present very frequently – let's say 7-10 times. Write the letter representing these consensus nucleotides in the table below. If no nucleotide occurs at least 7 times in a column, leave the box empty.

Consensus																			
Position	-8	-7	-6	-5	-4	-3	-2	-1	1	2	3	4	5	6	7	8	9	10	11

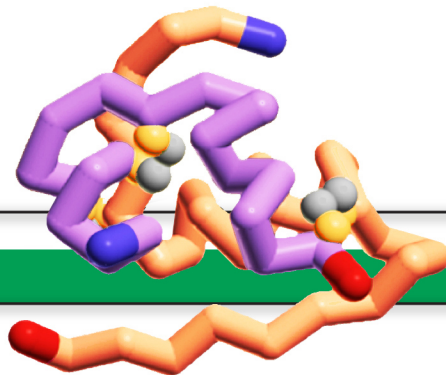
In 1986, Marilyn Kozak examined thousands of human genes to determine the consensus sequence surrounding the initiation of translation site. The sequence is called the Kozak sequence in recognition of her work. In addition to lining up the genes as you did above, Dr. Kozak made changes in the nucleotide sequence in the region of the consensus. When the changes were more similar to the consensus, the protein was made more often; when the changes were more different than the consensus, the protein was made less often. The gene she used for these experiments was the insulin gene. We now know that there are slight variations in the consensus sequence between eukaryotic species.

Frequently a consensus sequence is written like this:



At a given position, the size of each nucleotide reflects its frequency. The most frequently occurring nucleotide appears on top. Compare your consensus with the Kozak sequence. How well do they match?

Note that the actual frequency of each nucleotide differs between your consensus and the Kozak sequence. Why do you think there is such a difference?



Determining a Consensus Sequence Activity (continued)

Dr. Kozak discovered that not all positions within the consensus were of equal importance. She classified sequences as strong or weak, depending on how readily proteins were made. The sequences are identified below:

Classification	-3	-2	-1	1	2	3	4
Strong	A			A	U	G	G
Adequate	A			A	U	G	
Adequate				A	U	G	G
Weak				A	U		

Look back at the sequences you compared. Place a letter beside each protein name to indicate whether it is strong (S), adequate (A) or weak (W).

We now know that genes containing a weak Kozak sequence can still be translated, but additional factors are necessary for the ribosome to bind to these sequences.

Sequencing of all of the DNA within the human genome (Human Genome Project) was completed in 2003. Analysis of the information contained within the genome is part of a new field of study called genomics. The activity you just completed is one of many types of activities done by genome scientists (and the computers they use!).