

Genes: DNA to Protein

a) Introduction

A gene is a segment of DNA that encodes a protein. The nucleotides of the gene (the G's, A's, T's, and C's) are the instructions that tell the cell which amino acids must be linked together to make one particular protein. Human beings have around 25,000 genes. In contrast, bacteria have only around 3,000.

When a cell is using a gene to make a protein we say the cell is “expressing” the gene. Chapter 15 in your textbook describes all the steps of gene expression for both eukaryotic cells and prokaryotic cells. A brief review of the steps in eukaryotic cells is given below:

- 1) The enzyme RNA Polymerase II (RNA Pol II) binds to the gene at a region known as the promoter.
- 2) RNA Pol II makes an RNA molecule (called the “pre-mRNA” or the “primary transcript”) that is complementary to one strand of the gene. The 5' end is the beginning of the pre-mRNA and the 3' end is the end of the pre-mRNA. This step is called transcription.
- 3) When the whole gene has been transcribed, the pre-mRNA detaches from the gene.
- 4) A 5' cap is added to the pre-mRNA.
- 5) A poly-adenosine tail is added to the 3' end of the pre-mRNA.
- 6) The introns (segments of nucleotides within the pre-mRNA that do not encode any amino acids) are removed. After this step, the RNA is called the mature mRNA.
- 7) The mature mRNA exits the nucleus and enters the cytoplasm.
- 8) A ribosome organelle binds to the 5' end of the mRNA
- 9) Using the codons of the mRNA as instructions, the ribosome assembles a protein from the 20 different types of amino acids. This step is called translation.

b) Sequences in the gene and the mRNA

Many of the nine steps described above rely on specific sequences of nucleotides in the gene or the mRNA. For example, the promoter has a specific sequence that allows RNA POL II to begin transcription of the gene. Transcription ends when a specific sequence near the end of the gene is transcribed into the mRNA.

The important nucleotide sequences are described below. Biologists use the term 5' to mean toward the beginning of the gene or mRNA, and 3' to mean toward the end of the gene or mRNA.

Promoter sequences: The promoters of almost all eukaryotic genes have sequences that mark where the RNA Pol II enzyme should attach to the DNA. Although there are several different sequences that allow RNA Pol II to attach to the DNA, one of the most common is the “CAT box”. It has the sequence:

5' GGCCAATCT 3'
3' CCGGTTAGA 5'

The CAT box is usually located within 40 nucleotides of another promoter sequence called the “TATA box”. The TATA box is important because it tells RNA Pol II where to begin transcription and what strand of the gene to transcribe. The TATA box sequence is:

5' TATAAAA 3'
3' ATATTTT 5'

The ATATTTT strand is called the **template** strand and the TATAAAA strand is called the **coding** strand (or the non-template strand). The enzyme RNA Polymerase II does the following:

- (1) It binds to the template strand then it moves along the template strand in the 5' direction of the template strand.
- (2) Starting about 30 nucleotides from TATA box, it begins transcription (it begins making the pre-mRNA).
- (3) It makes the pre-mRNA complementary to the template strand (which means that the pre-mRNA will have the same sequence as the coding

strand, except that it has U's instead of T's). And the pre-mRNA is made starting at the pre-mRNA's 5' end and finishing at the pre-mRNA's 3' end.

The Poly-A signal: A sequence in the pre-mRNA, called the “polyadenylation signal” (or “Poly-A signal”) is NOT the 3' end of the pre-mRNA, but it does help determine where the 3' end will be. The Poly-A signal has the sequence:

5' AAUAAA 3'

The RNA polymerase enzyme transcribes the poly-A signal and continues transcribing for dozens of nucleotides beyond the poly-A signal. After transcription, enzymes in the nucleus cut the pre-mRNA about 20 nucleotides 3' from the Poly-A signal. After the pre-mRNA is cut, another nuclear enzyme will add 50 – 250 adenosine (A) nucleotides to the 3' end of the pre-mRNA. This is called the “Poly-A tail”.

Removal of introns: The intron sequences must be removed from the pre-mRNA before it leaves the nucleus. Enzymes called snRNPs (small nuclear Ribonucleoproteins) bind to sequences at the exon/intron junctions. The spliceosome (all the snRNPs together) brings the two exons on either side of the intron together, which causes the intron to form a loop. The spliceosome ligates (joins together) the two exons and cuts off the looped intron.

The sequences that mark the beginning and end of the intron are shown below. In this diagram, the exon sequences are shown in **boxed and bold** and the intron sequences are not boxed and shown in plain text. “N” means any nucleotide and “.....” means any number of nucleotides.

5' NNAGGUAAGU.....CCUCUUCUCUCUNCAG**NNNNN 3'**

There can be some variation in the above sequences. For example, note that the “.....” within the intron represents any number of nucleotides. This means introns can be of any length. After these dots, there is a region of twelve C and U nucleotides. The C's and U's in this region can be in any order and any mixture of C and U. The sequences at the beginning of the intron and the end of the intron can also vary somewhat, but generally they are the same as the underlined sequences shown in the figure above.

Start of translation codon: After the mRNA has left the nucleus, a ribosome will bind to its 5' end and begin moving toward the 3' end. The first nucleotides of the mRNA are not translated into protein. These nucleotides are called the 5' untranslated region (5' UTR).

How does the ribosome know where to start translation? The ribosome knows because the first AUG that the ribosome encounters is always the start of translation. This codon is for the amino acid methionine (see the genetic code table on page 7), so all proteins are initially made with methionine as their first amino acid (However, often the first few amino acids of the protein are removed after translation, so not every protein in the cell ends up with methionine as its first amino acid).

Note that the first AUG codon not only tells the ribosome where to begin translation, it also sets the “reading frame” for the ribosome. The reading frame means the pattern that the ribosome uses to group the mRNA’s nucleotides into codons (groups of three nucleotides).

As an example, consider the following mRNA sequence:

5' NNNNNAUGCUCUUUGAC 3'

The N’s represent the 5' UTR. The first codon is AUG (methionine), the second is CUC (Leucine), the third is UUU (Phenylalanine), and the fourth is GAC (Aspartic acid). If, somehow, the ribosome made a mistake and skipped the first A of the AUG codon, this would change all the codons of the entire mRNA. The first codon would be UGC (Cysteine), the second would be UCU (Serine), the third would be UUG (Leucine). In both cases, the nucleotide sequence in the mRNA is the same, but the difference is the reading frame (how the ribosome groups the nucleotides).

End of translation codon: When the ribosome encounters a codon with the sequence UGA, UAG, or UAA, it stops translation. These three codons are called “stop codons”. They don’t encode any amino acid. When the ribosome encounters a stop codon it releases the mRNA.

mRNAs have nucleotides 3' of the stop codon, but because they are 3' of the stop codon they are never translated. This region of the mRNA is called

the 3' untranslated region (3' UTR). The Poly-A tail is attached to the end of the 3' UTR.

c) Instructions for today's exercise

Your lab group will be given a gene sequence (double stranded DNA). You will construct a protein from the gene by performing all 9 steps outlined on page 1 of this handout.

1) Obtain the sequence for gene #1 from pages 8-9 of this handout. Using scissors or a cutting board, cut out each part of the DNA sequence, trim away its right margins, and tape the parts together in the correct order. Note that each part is numbered to help you assemble them correctly.

2) Tape the entire gene onto your bench top.

3) Locate the promoter and transcribe the gene: Begin by scanning the ends of the gene for promoter sequences. Remember that the promoter is may be at the left or the right end of the gene.

When the promoter sequences are located, tape a blank roll of paper next to the gene and “transcribe the pre-mRNA” by writing the complementary RNA bases to the template strand. Refer to the section of this handout on promoter sequences to confirm where you should begin the transcript and which DNA strand you should use as the template.

Refer to the section on the poly-A signal to learn where you should end transcription.

When done with this step, have your instructor inspect your pre-mRNA before starting the next step.

4) After the pre-mRNA is transcribed, add a 5' cap by taping a square of paper with “GTP” written on it to the 5' end of the transcript.

5) Locate the poly-A signal and clip the pre-mRNA. Read the section of this handout on the poly-A signal to see exactly where you should cut the pre-mRNA.

6) Add a Poly-A tail. Read the section of this handout on the poly-A signal to see exactly where you should add the tail. The tail that you add on should be a piece of the roll paper with 50 A's written on it.

7) Identify all exon/intron boundaries. Read the section of this handout on the exon/intron boundaries help you find them.

8) Remove the introns. Cut out the introns with scissors and tape the exons together. Remember that genes usually have more than one intron.

9) After the introns are removed, the mRNA moves from the nucleus to the cytoplasm, where a ribosome will translate it. The "ribosome" student should begin by scanning the mRNA for the start codon. Once this codon is located, they should "translate" the mRNA by writing the corresponding amino acid on the roll paper under each codon. Stop translation when the "ribosome" reaches a stop codon.

Show your instructor your protein and enter the data about the protein in data table 1 on page 11.

After you have finished with gene #1, save your mRNA and protein (you will need them to answer some review questions). Repeat the procedure with gene #2.

		Second letter				
		U	C	A	G	
First letter	U	UUU } Phe	UCU } Ser	UAU } Tyr	UGU } Cys	Third letter
		UUC } Phe	UCC } Ser	UAC } Tyr	UGC } Cys	
		UUA } Leu	UCA } Ser	UAA Stop	UGA Stop	
		UUG } Leu	UCG } Ser	UAG Stop	UGG Trp	
	C	CUU } Leu	CCU } Pro	CAU } His	CGU } Arg	
		CUC } Leu	CCC } Pro	CAC } His	CGC } Arg	
		CUA } Leu	CCA } Pro	CAA } Gln	CGA } Arg	
		CUG } Leu	CCG } Pro	CAG } Gln	CGG } Arg	
	A	AUU } Ile	ACU } Thr	AAU } Asn	AGU } Ser	
		AUC } Ile	ACC } Thr	AAC } Asn	AGC } Ser	
		AUA } Ile	ACA } Thr	AAA } Lys	AGA } Arg	
		AUG Met	ACG } Thr	AAG } Lys	AGG } Arg	
	G	GUU } Val	GCU } Ala	GAU } Asp	GGU } Gly	
		GUC } Val	GCC } Ala	GAC } Asp	GGC } Gly	
		GUA } Val	GCA } Ala	GAA } Glu	GGA } Gly	
		GUG } Val	GCG } Ala	GAG } Glu	GGG } Gly	

Gene #1

5' GCTATCCATATTTTTGGTTTGGCACCCAGTG
3' CGATAGGTATAAAAACCAAACCGTGGTCAC

2GCCAATCTGACTTACGTGTCAGTACGTATAAT
CGGTTAGACTGAATGCACAGTCATGCATATTA

3GTGACCGCAGTAGCTATAAAAGTATGTCCCTG
CACTGGCGTCATCGATATTTTCATACAGGGAC

4TACGTAGACAGTAAGACTTTTTGTCGTCGTTG
ATGCATC TGTCATTCTGAAAAACAGCAGCAAC

5TGCAGTATCGTAGGGTATCGTGACTAGATGGC
ACGTCATAGCATCCCATAGCACTGATCTACCG

6CGATGAGGTAAGTGTCGAGAATGCTTCGATAT
GCTACTCCATTCACAGCTCTTACGAAGCTATA

7GCTTTAGTCTGATTGTAGCTAGTTGCGCGTATC
CGAAATCAGACTAACATCGATCAACGCGCATAG

8TTCTTTTTTCCTTCAGGTTGTGGCCGCACTCCC
AAGAAAAAGGAAGTCCAACACCGGCGTGAGGG

9CCAGGTAAGTGTACGTCATGCTTCGCTCAA
GGTCCATTCACATGCAGTACGAAGCGAGTT

10ATGATGCTAGCTGGCCTATGCTTTTTTCTCCC
TACTACGATCGACCGGATACGAAAAAGAGGG

11TCTGCAGAGAATACTTTAAAGTCGATTCGCT
AGACGTCTCTTATGAAATTTTCAGCTAAGCGA

12AGAGCAAAGTACACAGTGATTTAGCATGAC
TCTCGTTTCATGTGTCACTAAATCGTACTG

13GTGATCCAGTAGATCGTGAAATAAACTAGTA
CACTAGGTCATCTAGCACTTTATTTGATCAT

14GCGATAGCAACCCGTGCATGATTGGCAGTAT
CGCTATCGTTGGGCACGTACTAACCGTCATA

15ACAGATGATTGTGAAACGATACAGTAG 3'
TGTCTACTAACACTTTGCTATGTCATC 5'

Gene #2

5' GGC ACTACAGCTACTGTCACACACTAGG
3' CCGTGATGTCGATGACAGTGTGTGATCC

2CACGTTTTATTGGTCATAAACTGGGTCACCAC
GTGCAAATAACCAGTATTTGACCCAGTGGTG

3 TGTGTCACCGACACACTTCAGGCTGGGGA
ACACAGTGGCTGTGTGAAGTCCGACCCCT

4GAGGAGGGGGCTACGTCACACTGTCACTTCT
CTCCTCCCCCGATGCAGTGTGACAGTGAAGA

5GTCAA ACTACTTACCTCTTCTTTGGGCATCT
CAGTTTGATGAATGGAGAAGAAACCCGTAGA

6TCAGTCTACTGCTACCGCTACTGGCACTCAT
AGTCAGATGACGATGGCGATGACCGTGAGTA

7TTTATAGGTCACTAGCACTGTGTCACGCACC
AAATATCCAGTGATCGTGACACAGTGCGTGG

8GTCTACTGCATCGTCACGTCTACTTGTGAGA
CAGATGACGTAGCAGTGCAGATGAACACTCT

9TTGGCCCCCACGGGGTCATCG 3'
AACCGGGGGTGCCCCAGTAGC 5'

e) Review questions (You may need to consult the textbook for some questions)

1) How many genes do human's have? _____ Bacteria? _____

2) Write the nine basic steps from gene to protein, as outlined in this handout.

3) What are the functions of the 5' cap and the poly-A tail?

4) Two promoter sequences are the CAT box and the TATA box. These two sequences have different roles in starting transcription. Contrast their two roles.

5) An intron is always located between two exons. One exon is upstream of the intron and the other exon is downstream of the intron. Review the sequences that are found at exon/intron junctions (see page 3) then answer this question:

Which of the following contain sequences necessary for splicing out the intron?

The upstream exon The intron The downstream exon

6) If a gene contains 10 exons, how many introns does it have? _____

7) If a gene contains 7 introns, how many exons does it have? _____

8) Define "reading frame" as it was defined in this handout and explain why the reading frame is important in translation.

9) Inspect the mRNA and the protein that you made from gene #1. The 5th codon of the mRNA is GUU (which encodes Valine amino acid). If the gene had been mutated (changed) so that an extra A had been added to the end of this codon (in other words, the GUU became GUUA), what would be the amino acid sequence of the protein from the mutated gene? Does the protein from the mutated gene differ greatly in sequence from the protein from the original gene?

10) Assume that five hydrophobic amino acids in a row are sufficient to form a transmembrane domain. Recall that a transmembrane domain is a hydrophobic region of a protein that anchors the protein to the cell membrane. In other words, having a transmembrane domain is what causes a protein to be a membrane protein.

Does gene #1 encode a membrane protein? Justify your answer, using the exact amino acid sequence of the protein as part of your answer.

f) Review question answers

1) Humans have about 25,000 genes. A typical bacteria has between 3000 – 5000 genes.

2)

(1) The enzyme RNA Polymerase II binds to the gene's promoter.

(2) RNA Pol II makes the pre-mRNA molecule. This step is called transcription.

(3) When the whole gene has been transcribed, the pre-mRNA detaches from the gene.

(4) A 5' cap is added to the pre-mRNA.

(5) A poly-adenosine tail is added to the 3' end of the pre-mRNA.

(6) The introns are removed from the pre-mRNA. After this step, the RNA is called the mature mRNA.

(7) The mature mRNA exits the nucleus and enters the cytoplasm.

(8) A ribosome organelle binds to the 5' end of the mRNA

(9) The ribosome assembles a protein from amino acids. This step is called translation.

3) The 5' cap and the poly-A tail both have three functions:

1) They help the mRNA exit the nucleus

2) They protect the mRNA from being degraded

3) They help the ribosome bind the mRNA

4) The CAT box is where RNA Polymerase II enzyme (the enzyme that makes mRNAs) binds to the gene. The TATA box tells the RNA Polymerase II enzyme where to begin transcription and which of the two DNA strands to transcribe.

5) The upstream exon and the intron contain sequences that are necessary for splicing out the intron.

6) 9

7) 8

8) The reading frame means the pattern by which the ribosome groups the mRNA's nucleotides into codons. If a ribosome is not translating an mRNA in the correct reading frame, the ribosome will group the mRNA's nucleotides into incorrect codons (incorrect groups of three nucleotides). And when the codons are not correct, the protein made by the ribosome will contain incorrect amino acids, which will lead to the protein not functioning.

9) The original (unmutated) mRNA's coding region (the region from the start codon and the stop codon) is this:

AUG GCC GAU GAG GUU GUG GCC GCA CUC CCC CAG AGA AUA
CUU UAA

Note that this sequence has been arranged into codons (groups of three consecutive nucleotides) that the ribosome will use to translate the mRNA into a protein.

The above mRNA is translated by the ribosome into the correct protein sequence of:

Met-Ala-Asp-Glu-Val-Val-Ala-Ala-Leu-Pro-Gln-Arg-Ile-Leu

If the gene had been mutated such that an extra A had been added to the end of the codon 5 (GUU), the mutated mRNA's coding region would be this:

AUG GCC GAU GAG GUU AGU GGC CGC ACU CCC CCA GAG AAU
ACU UAA

This would be translated by the ribosome into the protein sequence of:

Met-Ala-Asp-Glu-Val-Ser-Gly-Arg-Thr-Pro-Pro-Glu-Asn-Thr-Leu

Notice that the amino acid sequence of this protein is the same as the original protein until the point of the mutation (amino acid 5), and after that point the amino acid sequence is completely different than the original protein. Because of the large change in the protein's sequence, the protein from the mutated gene would probably not function at all.

It may seem confusing that adding only a single nucleotide to the gene (an extra A nucleotide, in this example) can cause such a large change in many of the amino acids of the protein.

The explanation is that adding (or removing) even one nucleotide from a gene can change the ribosome's reading frame of the mRNA. In other words, after the location of the extra A in the mRNA, the ribosome groups (reads) the mRNA into the wrong groupings of three nucleotides. In general, anytime a mutation adds a nucleotide to a gene or removes a nucleotide from a gene, all of the amino acids downstream of the mutation will be changed because the mutation causes a frame shift when the ribosome translates the mRNA.

10) Gene #1 does encode a membrane protein. We know this because the protein's amino acid sequence ...

Met-Ala-Asp-Glu-Val-Val-Ala-Ala-Leu-Pro-Gln-Arg-Ile-Leu

...contains 6 hydrophobic amino acids in a row: Val-Val-Ala-Ala-Leu-Pro