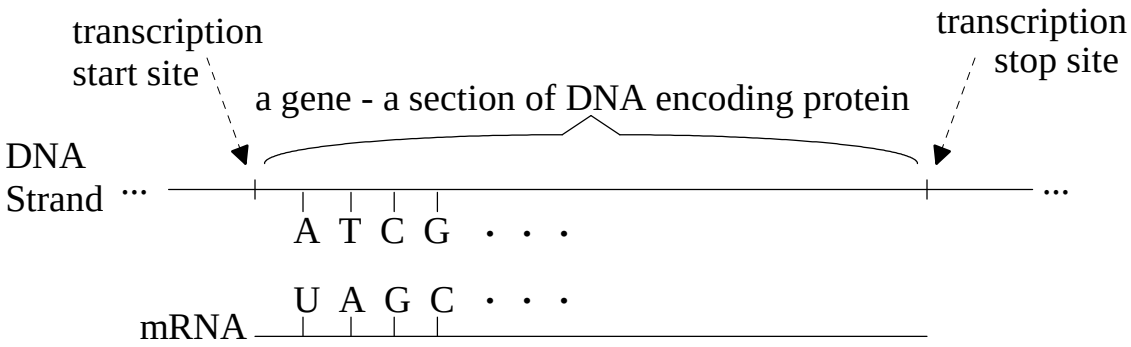
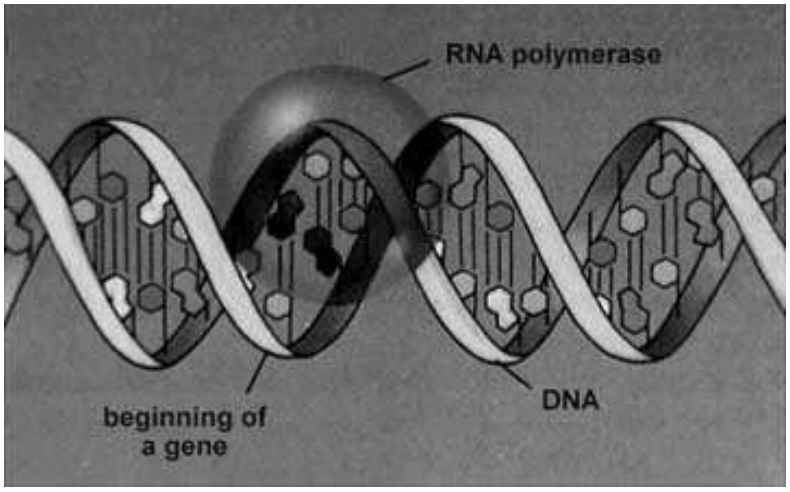
Central Dogma of Molecular Biology

Transcription Process



Following images from
http://www.emc.maricopa.edu/faculty/farabee/BIOBK/BioBookPROTSYn.html



Prokaryotic RNA polymerase scans DNA for a **promoter sequence** that consists of a specific set of about 13:

- 1 nt serves as start site
- 6 nts that are 10 nts 5' to the start site
- 6 nts that are 35 nts 5' to the start site

How often would we expect a promoter sequence to occur by random chance?
(Note: Prokaryotic genomes are only a few million nucleotides in length)

Prokaryotes vs. Eukaryotes

Organisms are classified into two types:

- **Prokaryotes**: lack a true membrane-bound nucleus and organelles (single-celled, includes bacteria)
- **Eukaryotes**: contain a membrane-bound nucleus and organelles (plants, animals, fungi,…) *Note: Not all single celled organisms are prokaryotes!*

Eukaryotic Genes

Eukaryotic genes can be pieced together

- Exons: coding regions
- Introns: non-coding regions

Splicing: mRNA processing removes introns, splices exons together

Processed mRNA can be translated into a protein sequence

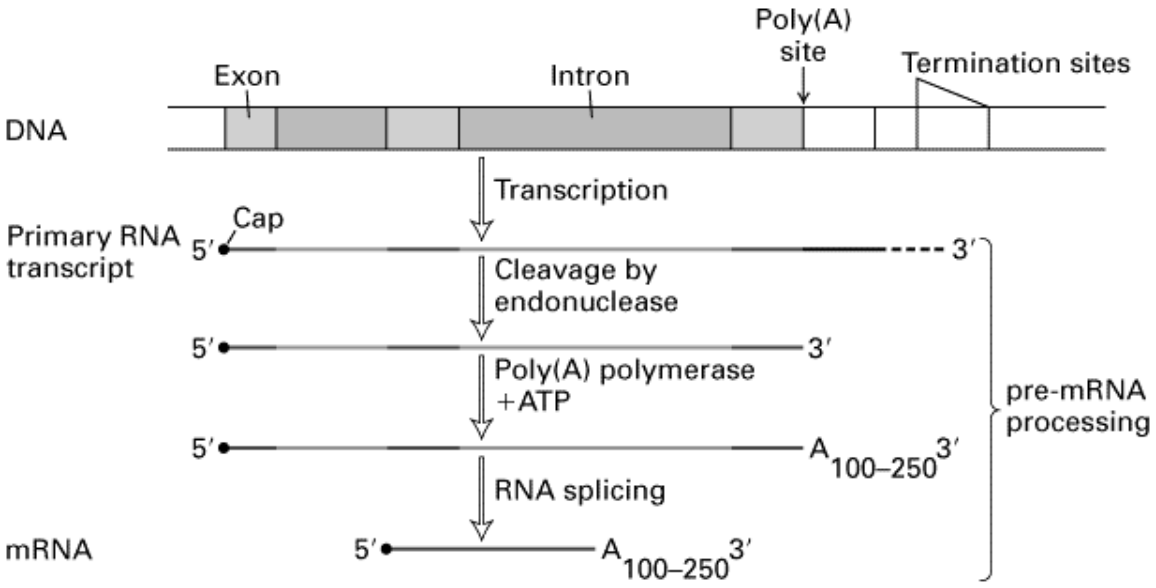
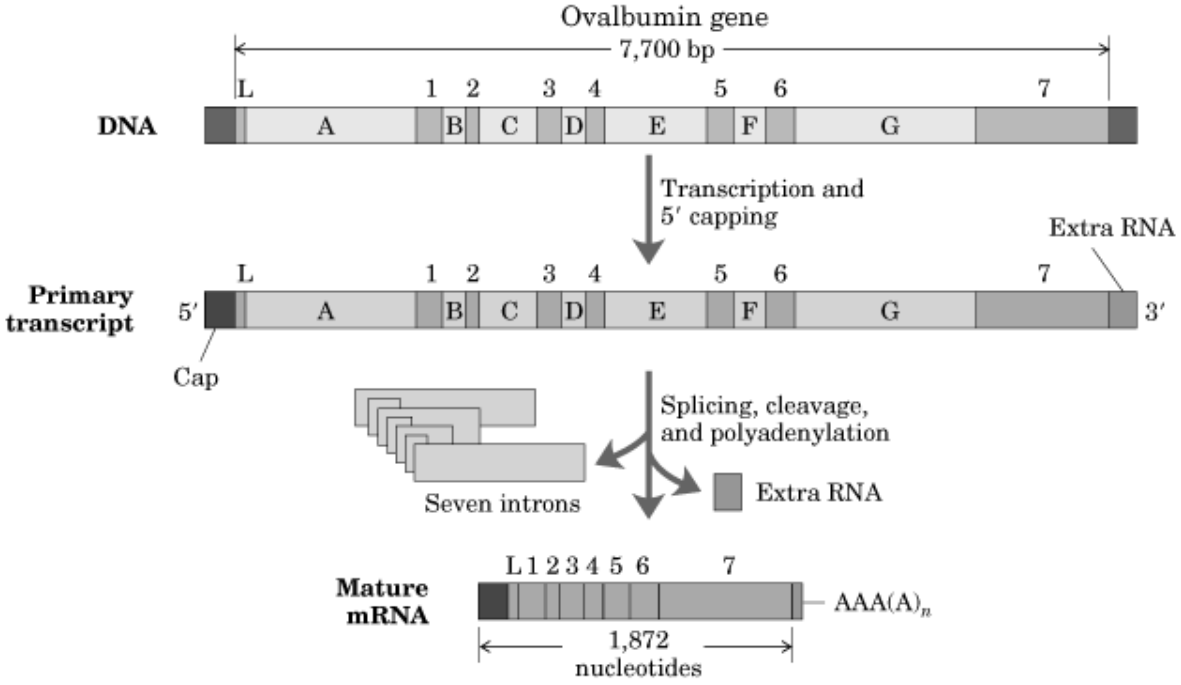


Image source: http://departments.oxy.edu/biology/Stillman/bi221/111300/processing_of_hnrnas.htm

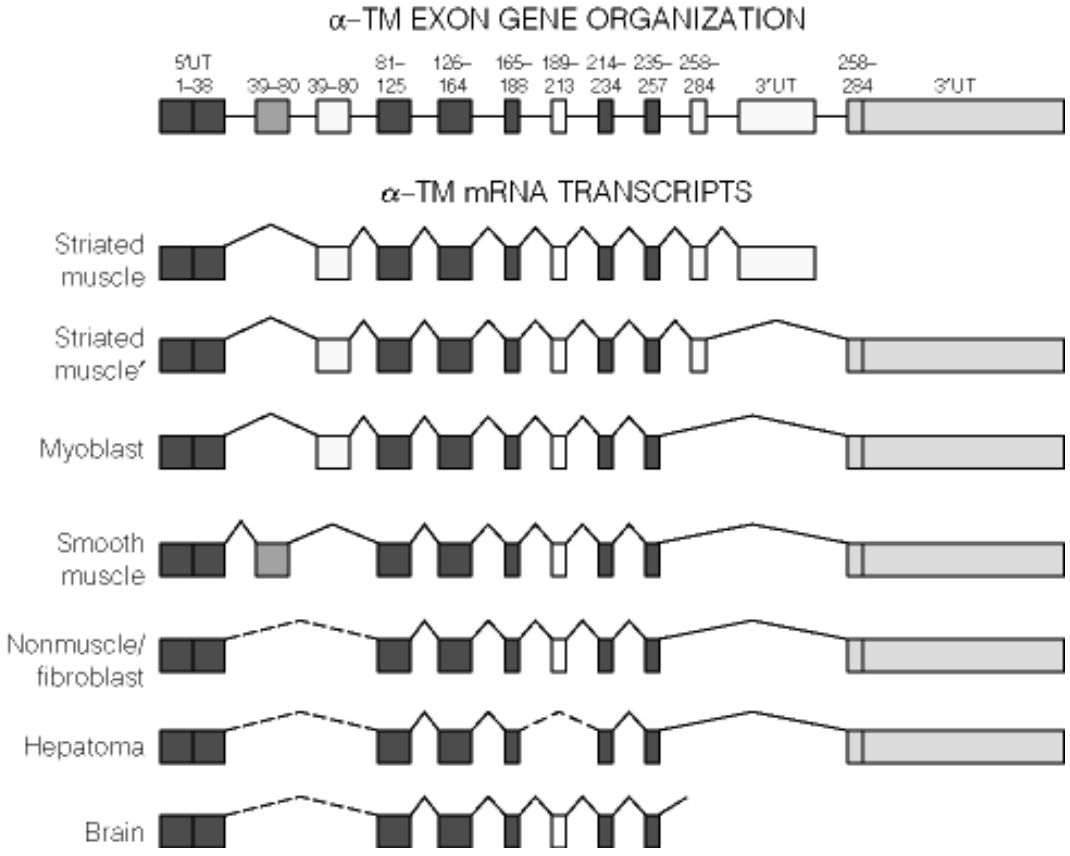
Typically, introns in DNA begin with GT and end with AG (the "GT - AG rule")

About 6 additional nucleotides at the 5' and 3' ends of the introns are also scrutinized. Scrutinization can vary depending on cell type, e.g., tissue type

Splicing in the ovalbumin gene: Introns labled by letters, exons by numbers.



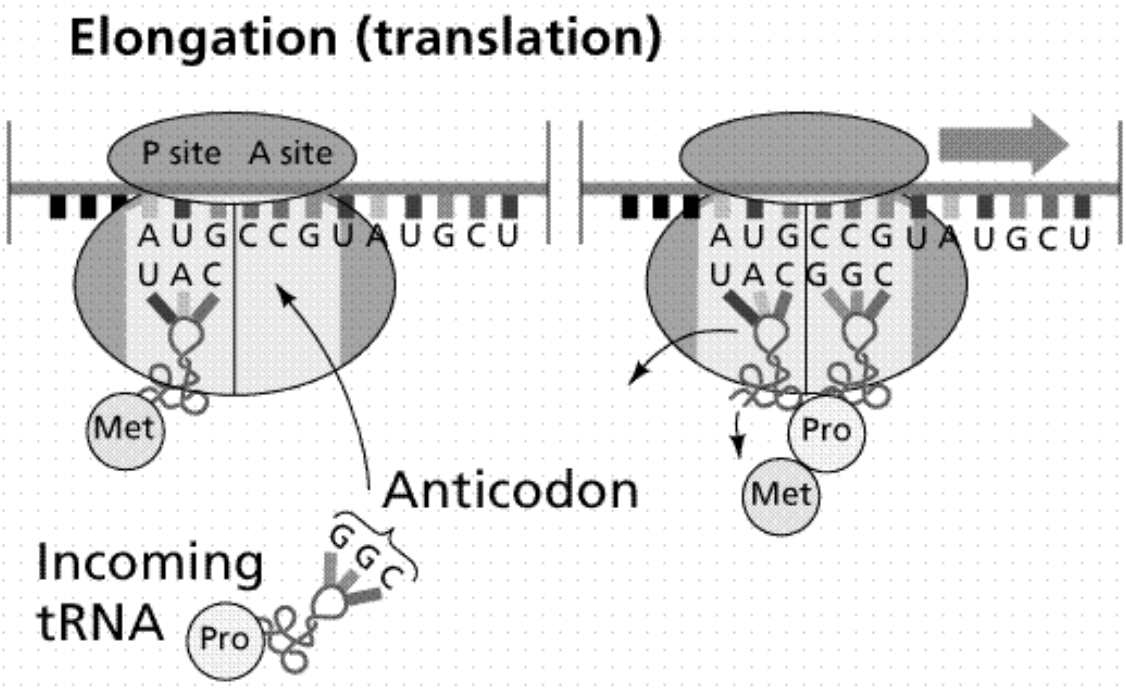
Alternative splicing example: tissue specific gene expression of α-tropomyosin



Translation Process

DNA —(Transcription Process; RNA Polymerase)→ mRNA —(Translation Process; Ribosomes)→ Protein

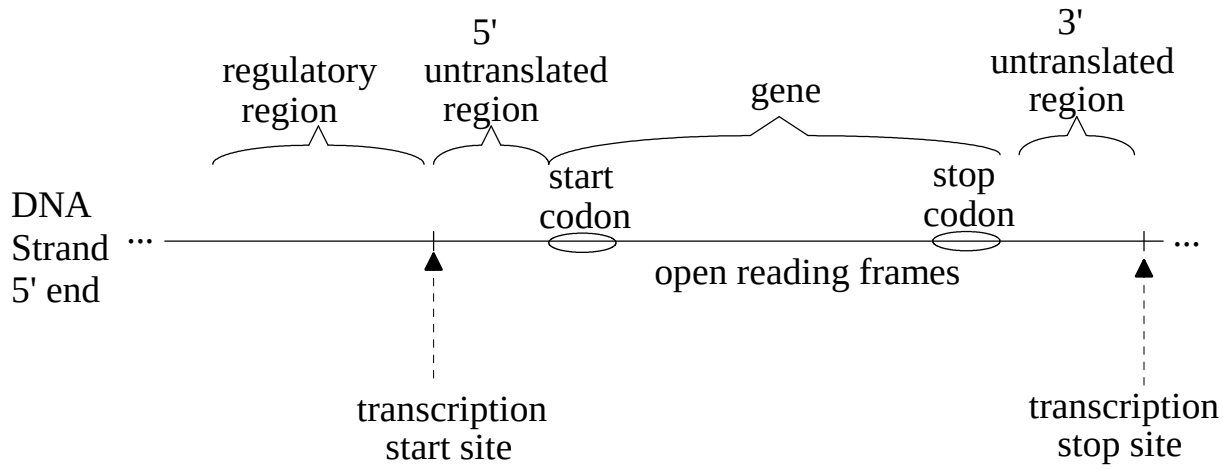
Translation Process: mRNA is "read" by ribosomes to produce a protein with the help of tRNAs



Three types of RNA:

- mRNA- messenger RNA. The template for protein synthesis
- tRNA- transfer RNA. The "adapter" molecule that converts nucleic acid sequence to protein sequence. tRNA contains a **anticodon** which is a base sequence which is complementary to a codon.
- rRNA- ribosomal RNA. The structural and sometimes catalytic molecule of the ribosome.

Regulation of Protein Production



Regulatory region contains promoters which are specific DNA sites where regulatory proteins called transcription factors can bind and regulate gene expression.

A transcription factor might bind to the promoter to affect the ability of the RNA polymerase to perform its task of transcription.

Translation regulation is possible too, e.g., regulatory factor binds to mRNA and affects the ability of the ribosome to perform its task of translation.

In eukaryotes, multiple regulatory regions are possible that might be far from an exon in either the 5’ or 3’ direction.

Patterns of Substituion within Genes

Molecular evolution is the study of genetic material (nucleic acids) and gene products and the molecular processes that describe its alteration over evolutionary time.

Knowledge gained about these processes has permitted researchers to reconstruct evolutionary histories of genes and organisms by comparison of homologous sequences (molecular phylogenetics).

Mutation Rate, $r = K/(2T)$, where K is the number of substitutions two sequences have undergone since they last shared a common ancestor, and T is a divergence time.

Mutations categorization:
1) deleterious - mutations that are disadvantageous to living cells/organism
2) those that are advantageous to the living cell/organism
3) those that are effectively neutral to the organism

Mutations that diminish an organism's ability to survive are typically removed from the gene pool by the process of natural selection.

Changes to a gene's nucleotide sequence that impact the corresponding protein's catalytic or structural properties are especially subject to natural selection.

Functionally constrained is the term used to indicate that a portion of a gene is especially import.

Would you expect changes to the functionally constrained regions of a gene to have a higher or lower observed mutation rate than changes to regions of the gene that have no effect on the amino acid sequence or expression levels of the protein?

Table 3.1 Average pairwise divergence among different regions of the human, mouse, rabbit, and cow beta-like globin genes.

Region	**Length of Region (bp) in Human**	**Average Pairwise Number of Changes**	**Standard Deviation**	**Substitution Rate (substitutions/site/10^9 years)**
Noncoding, overall	913	67.9	14.1	3.33
Coding, overall	441	69.2	16.7	1.58
5’ Flanking sequence	300	96.0	19.6	3.39
5’ Untranslated sequence	50	9.0	3.0	1.86
Intron 1	131	41.8	8.1	3.48
3’ Untranslated sequence	132	33.0	11.5	3.00
3’ Flanking sequence	300	76.3	14.3	3.60

Summary of Trend:

Mutation rate in introns and flanking regions > Mutation rate in other regions that are transcribed but not translated > Mutation rate of coding region

Synonymous vs. Nonsynonymous Substitutions

	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Glycine (G, GLY)
•Alanine (A, ALA)
•Valine (V, VAL)
•Leucine (L, LEU)
•Isoleucine (I, ILE)
•Phenylalanine (F, PHE)
•Proline (P, PRO)
•Serine (S, SER)
•Threonine (T, THR)
•Cysteine (C, CYS)
•Methionine (M, MET)
•Tryptophan (W, TRP)
•Tyrosine (T, TYR)
•Asparagine (N, ASN)
•Glutamine (Q, GLN)
•Aspartic acid (D, ASP)
•Glutamic Acid (E, GLU)
•Lysine (K, LYS)
•Arginine (R, ARG)
•Histidine (H, HIS)
•START: AUG
•STOP: UAA, UAG, UGA

Synonymous substitutions - changes in the nucleotide coding sequence that do not change the amino acid sequence, e.g, CUU, CUG, UUG, etc. all encode for Leucine

Nonsynonymous substitutions - changes in the nucleotide coding sequence that change the amino acid sequence, e.g., UUG encodes for Leucine, but UUU encodes for Phenylalanine.

Condon sites can be categorized by the likelyhood of a nonsynonymous substitution occurs:

- nondegenerate sites - condon positions where mutations always results in amino acid substitutions, e.g., GUU (Val), GCU (Ala), GAU (Asp), GGU (Gly)
- twofold degenerate sites - condon positions where two different nucleotides result in the translation of the same amino acid, but the other two nucleotides code for a different amino acid, e.g., GAU and GAC encode for aspartic acid, but GAA and GAG encode for glutamic acid
- fourfold degenerate sites - condon positions where all 4 nucleotides result in the translation to the same amino acid, e.g., GUU, GUC, GUA, GUG all encode for glycine

Nonsynonymous substitutions have about 3 times as many opportunities than synonymous substitutions for occuring, but which would natural selection favor?

Table 3.2 Divergence between different kinds of sites within the coding sequence of the human and rabbit beta-like globin genes.

Region	**Number of Sites (bp)**	**Number of Changes**	**Substitution Rate (substitutions/site /10^9 years)**
Nondegenerate	302	17	0.56
Twofold degenerate	60	10	1.67
Fourfold degenerate	85	20	2.35

Of the 47 substitutions, 27 are synonymous and 20 are nonsynonymous

Indels and Pseudogenes

Strong bias against indels (insertion and deletions) because of their tendency to alter the reading frame used by ribosomes.

Indels are roughly 10 times less likely to occur than simple exchanges for one nucleotide for another.

Duplication of an entire gene allows one copy to provide the necessary function of the original and the other copy to accumulate substitutions in a way that is free of selective constraint.

The copy might evolve into a new gene with a new function or it might evolve into a *pseudogene* that make it nonfunctional and transcriptionally inactive

Mammalian genomes have many pseudogenes that tend to accumulate substitutions at a very fast rate ~ 4 substitutions per site per 100 million years.

Substitutions vs. Mutations

Mutations are changes in nucleotide sequences that occur due to mistakes in DNA replication or repair processes

Substitutions are mutations that have passed through the filter of selection on at least some level

Substitution rates can be observed by the data, but mutation rates are very difficult to reliably estimate since natural selection can be so subtle and pervasive.

Comparisons between substitution and mutation rates give the best indication as to how functionally constrained a sequence actually is, so good mutation rates are important.

Synonymous and pseudogene substitution rates (K_s), are considered to be fairly reflective of actual mutation rates.

Nonsynonymous substitution rates (K_a) are not because they are subject to natural selection.

Fixation

Populations of an organism contains a substantial amount of genetic variation, e.g., humans differ from each other at an average of 1 base pair out of every 200.

Different versions of any given gene within a species of organism are called **alleles**.

New alleles arise from mutations occurring to an existing allele within a single member of a population.

If the mutation makes the organism less likely to survive and reproduce, then what will happen to the mutation?

If the mutation makes the organism more likely to survive and reproduce, then what will happen to the mutation?

What accounts for the relatively high levels of allele variations seen within naturally occurring populations of organisms?

If we perform comparative sequence analyses between genes within a species, portions of a gene that vary correspond to regions that are functionally constrained or unconstrained?