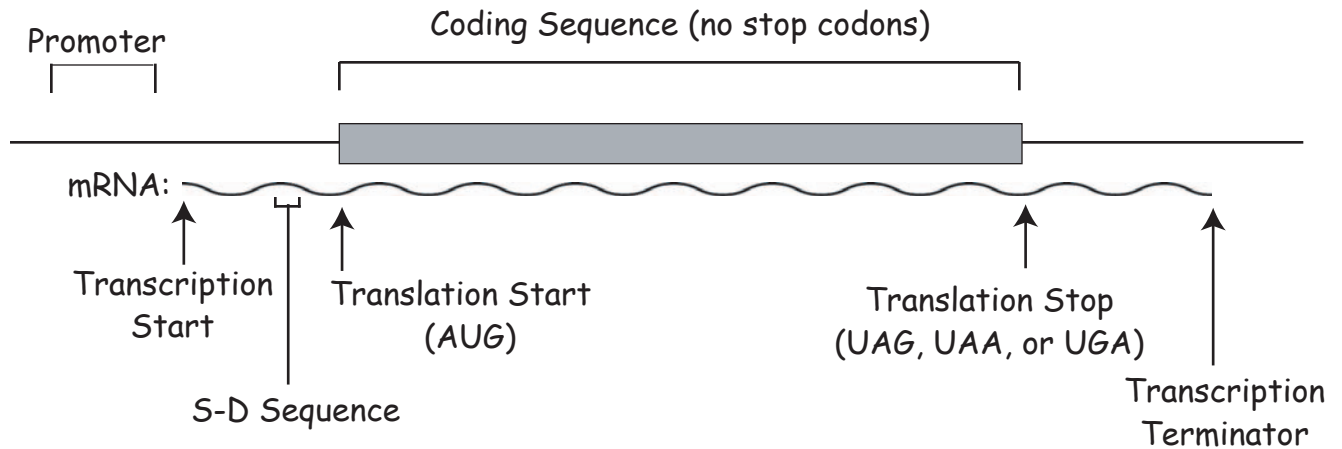


Lecture 11

Analysis of Gene Sequences

Anatomy of a bacterial gene:



Sequence Element

Function

Promoter	To target RNA polymerase to DNA and to start transcription of a mRNA copy of the gene sequence.
Transcription terminator	To instruct RNA polymerase to stop transcription.
Shine-Dalgarno sequence	S-D sequence in mRNA will load ribosomes to begin translation. Translation almost always begins at an AUG codon in the mRNA (an ATG in the DNA becomes an AUG in the mRNA copy). Synthesis of the protein thus begins with a methionine.
Coding Sequence	Once translation starts, the coding sequence is translated by the ribosome along with tRNAs which read three bases at a time in linear sequence. Amino acids will be incorporated into the growing polypeptide chain according to the genetic code.
Translation Stop	When one of the three stop codons [UAG (amber), UAA (ochre), or UGA] is encountered during translation, the polypeptide will be released from the ribosome.

Example: A gene coding sequence that is 1,200 nucleotide base pairs in length (including the ATG but not including the stop codon) will specify the sequence of a protein $1200/3 = 400$ amino acids long. Since the average molecular weight of an amino acid is 110 da, this gene encodes a protein of about 44 kd — the size of an average protein.

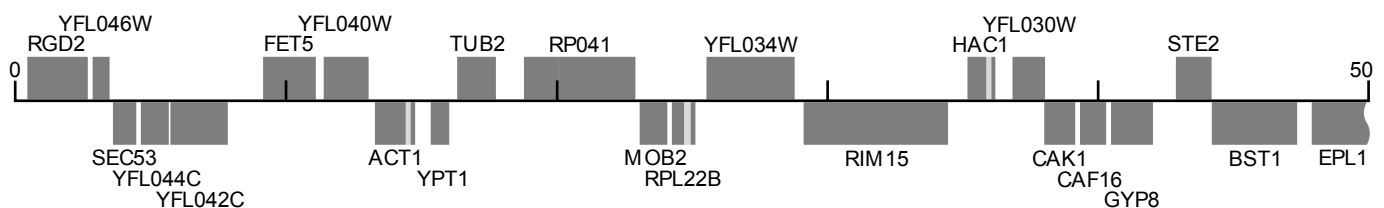
The Genetic Code

1st position (5' end) ↓	2nd position				3rd position (3' end) ↓
	U	C	A	G	
U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G
C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G

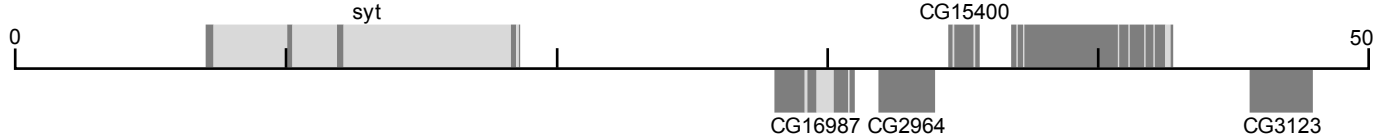
Classically, genes are identified by their function. That is the existence of the gene is recognized because of mutations in the gene that give an observable phenotypic change. Historically, many genes have been discovered because of their effects on phenotype. Now, in the era of genomic sequencing, many genes of no known function can be detected by looking for patterns in DNA sequences. The simplest method which works for bacterial and phage genes (but not for most eukaryotic genes as we will see later) is to look for stretches of sequence that lack stop codons. These are known as "open reading frames" or **ORFs**. This works because a random sequence should contain an average of one stop codon in every 20 codons. Thus, the probability of a random occurrence of even a short open reading frame of say 100 codons without a stop codon is very small $(61/64)^{100} = 8.2 \times 10^{-3}$

Identifying genes in DNA sequences from higher organisms is usually more difficult than in bacteria. This is because in humans, for example, gene coding sequences are separated by long sequences that do not code for proteins. Moreover, genes of higher eukaryotes are interrupted by **introns**, which are sequences that are spliced out of the RNA before translation. The presence of introns breaks up the open reading frames into short segments making them much harder to distinguish from non-coding sequences. The maps below show 50 kbp segments of DNA from yeast, *Drosophila*, and humans. The dark grey boxes represent coding sequences and the light grey boxes represent introns. The boxes above the line are transcribed to the right and the boxes below the line are transcribed to the left. Names have been assigned to each of the identified genes. Although the yeast genes are much like those of bacteria (few introns and packed closely together), the *Drosophila* and human genes are spread apart and interrupted by many introns. Sophisticated computer algorithms were used to identify these dispersed gene sequences.

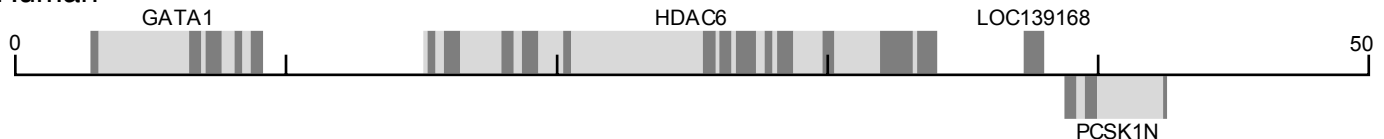
Saccharomyces cerevisiae



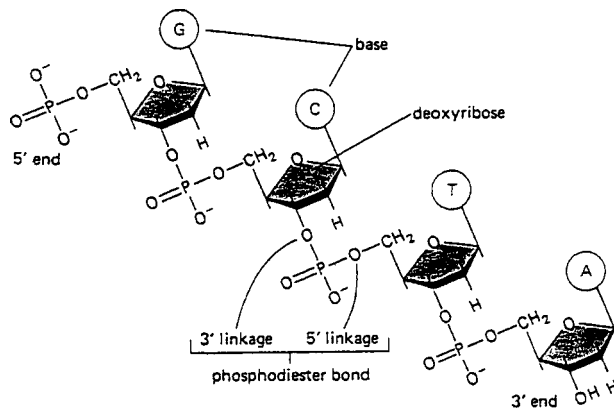
Drosophila melanogaster



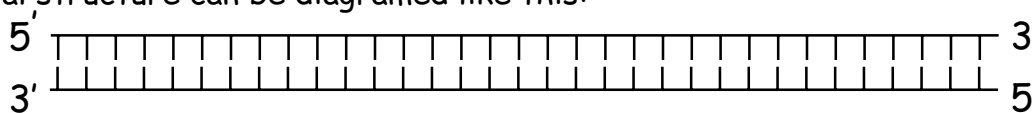
Human



To see how gene sequences are actually obtained, we will first need to consider some fundamentals of the chemical structure of DNA. Each strand of DNA is directional. The different ends are usually called the 5' and 3' ends; referring to different positions on the ribose sugar ring where the linking phosphate residues attach.



In a double stranded DNA molecule the two strands run anti-parallel to one another and the general structure can be diagrammed like this:



• Note about representation of DNA sequences.

- 1) Single strands are always represented in direction of synthesis - 5' to 3'
- 2) For double stranded DNA, usually one strand is represented in the 5' to 3' direction. For a gene, the strand represented would correspond to the sequence of the mRNA.

DNA polymerases are the key players in the methods that we will be considering. The general reaction carried out by DNA polymerase is to synthesize a copy of a DNA template starting with the chemical precursors (nucleotides) dATP, dGTP, dCTP, and dTTP (dNTPs). All DNA polymerases have two fundamental properties in common.

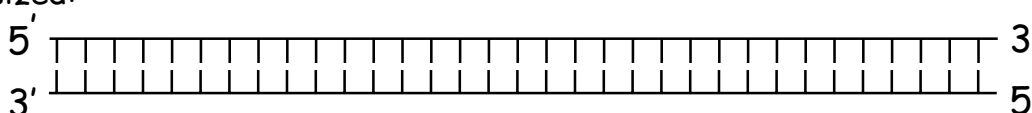
- (1) New DNA is synthesized only by elongation of an existing strand at its 3' end.
- (2) Synthesis requires nucleotide precursors, a free 3' OH end, and a template strand.

A general substrate for DNA polymerase looks like this:



Note that the template strand can be as short as 1 base or as long as several thousand bases.

After addition of DNA polymerase and nucleotide precursors this product will be readily synthesized:



DNA Sequencing

Consider a segment of DNA that is about 1000 base pairs long that we wish to sequence.

(1) The two DNA strands are separated. Heating to 100°C to melt the base pairing hydrogen bonds that hold the strands together does this.

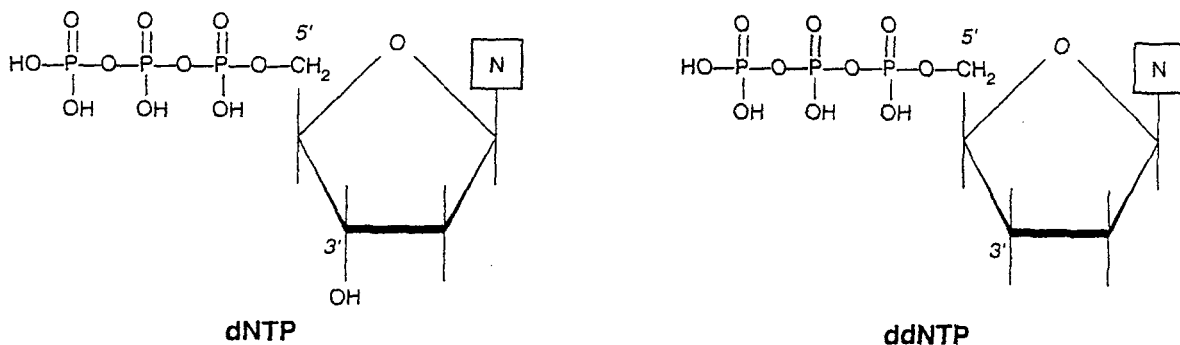
(2) A short oligonucleotide (ca. 18 bases) designed to be complimentary to the end of one of the strands is allowed to anneal to the single stranded DNA. The resulting DNA hybrid looks much like the general polymerase substrate shown previously.

(3) DNA polymerase is added along with the four nucleotide precursors (dATP, dGTP, dCTP, and dTTP). The mixture is then divided into four separate reactions and to each reaction a small quantity different dideoxy nucleotide precursor is added. Dideoxy nucleotide precursors are abbreviated ddATP, ddGTP, ddCTP, and ddTTP.

(4) The polymerase reactions are allowed to proceed and, using one of a variety of methods, radiolabel is incorporated into the newly synthesized DNA.

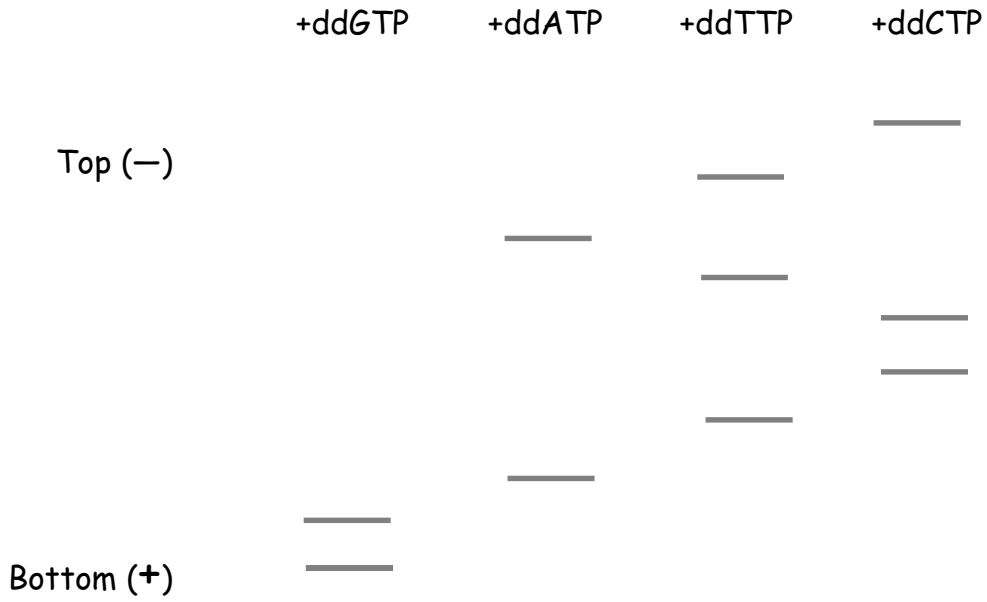
(5) After the DNA polymerase reactions are complete, the samples are melted and run on a gel system that allows DNA strands of different lengths to be resolved. The DNA sequence can be read from the gel by noting the positions of the radiolabeled fragments.

The crucial element of the sequencing reactions is the added dideoxynucleotides. These molecules are identical to the normal nucleotide precursors in all respects except that they lack a hydroxyl group at their 3' position (3' OH).



Thus dideoxynucleotides can be incorporated into DNA, but once a dideoxynucleotide has been incorporated further elongation stops because the resulting DNA will no longer have a free 3' OH end. Each of the four reactions contains one of the dideoxynucleotides added at about 1% the concentration of the normal nucleotide precursors. Thus, for example, in the reaction with added ddATP about 1% of the elongated chains will terminate at the position of each A in the sequence. Once all of the elongating chains have been terminated there will be a population of labeled chains that have terminated at the position of each A in the sequence.

A part of the final gel will look like this:



(Note that larger molecules migrate more slowly to the cathode on these gels)

The deduced DNA sequence obtained from this gel is: 5' GGATCCTATC 3'

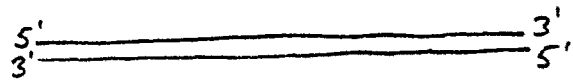
Polymerase Chain Reaction

Now let's consider how to obtain DNA segments that are suitable for sequencing. At first, DNA sequences were obtained from cloned DNA segments (we will discuss some methods to clone new genes in a subsequent lecture). Presently the entire DNA sequence for *E. coli*, as well as a variety of other bacterial species, has been determined. If we want to find the sequence of a new mutant allele of a known gene we need an easy way to obtain a quantity of this DNA from a culture of bacterial cells. The best way to do this is to use a method known as PCR or polymerase chain reaction that was developed by Kary Mullis in the mid-1980's. The steps in a PCR reaction are as follows.

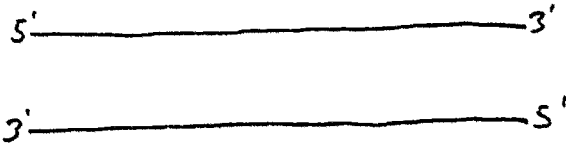
- (1) A crude preparation of chromosomal DNA is extracted from the bacterial strain of interest.
- (2) Two short oligo nucleotide primers (each about 18 bases long) are added to the DNA. The primers are designed from the known genomic sequence to be complementary to opposite strands of DNA and to flank the chromosomal segment of interest.
- (3) The double stranded DNA is melted by heating to 100°C and then the mixture is cooled to allow the primers to anneal to the template DNA.
- (4) DNA polymerase and the four nucleotide precursors are added and the reaction is incubated at 37°C for a period of time to allow a copy of the segment to be synthesized.
- (5) Steps 3 and 4 are repeated multiple times. To avoid the inconvenience of having to add new DNA polymerase in each cycle a special DNA polymerase that can withstand heating to 100°C is used.

The idea is that in each cycle of melting, annealing and DNA synthesis the amount of the DNA segment is doubled. This gives an exponential increase in the amount of the specific DNA as the cycles proceed. After 10 cycles the DNA is amplified 10^3 fold and after 20 cycles the DNA will be amplified 10^6 fold. Usually amplification is continued until all of the nucleotide precursors are incorporated into synthesized DNA.

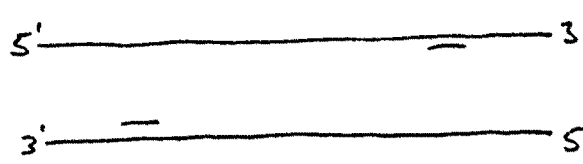
A PCR Reaction



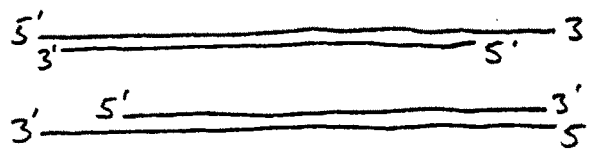
Separate DNA strands



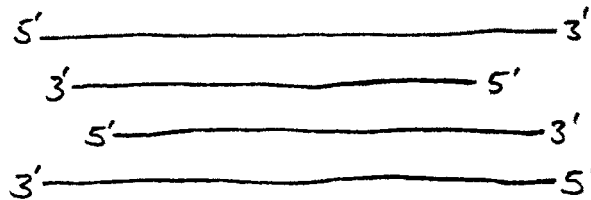
Anneal primers



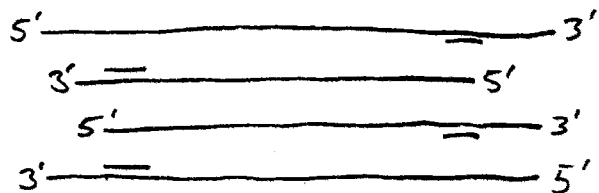
DNA synthesis



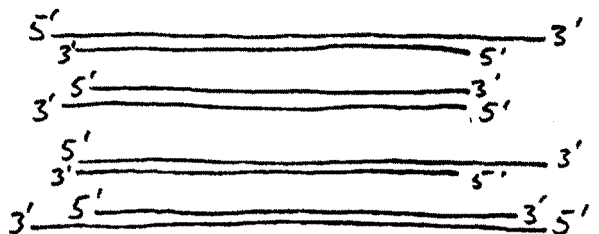
Separate DNA strands



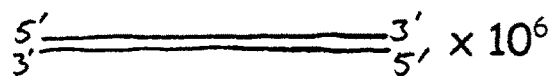
Anneal primers



DNA synthesis



Repeat cycle 20 times



Lecture 12

Gene Mutations

Let's say that we are investigating the LacZ gene, which encodes the lactose hydrolyzing enzyme β -galactosidase. There is a useful compound known as X-gal that can be hydrolyzed by β -galactosidase to release a dark blue pigment. When X-gal is added to the growth medium in petri plates, Lac⁺ E. coli colonies turn blue whereas Lac⁻ colonies with mutations in the LacZ gene are white. By screening many colonies on such plates it is possible to isolate a collection of E. coli mutants with alterations in the LacZ gene. PCR amplification of the LacZ gene from each mutant followed by DNA sequencing allows the base changes that cause the LacZ⁻ phenotype to be determined. A very large number of different LacZ mutations can be found but they can be categorized into three general types.

Mutation Type	Description
Missense	A base change that converts one codon into another. Many missense mutations are silent because the encoded amino acid remains the same or the amino acid substitution is sufficiently subtle so as not to compromise activity of the enzyme. Missense mutations that have a marked effect often lie in the active site or grossly disrupt protein folding.
Nonsense	A base change that converts a codon within the coding sequence into a stop codon. Note that there is only a limited set of sense codons that can be converted to a stop codon by a single base change. Nonsense mutations lead to a truncated protein product. Nonsense mutations that lie early in the gene sequence will completely inactivate the gene. Sometimes nonsense mutations that lie late in the gene sequence will not disrupt gene function.
Frameshift	The addition or deletion of a base or bases such that the coding sequence is shifted out of register. Note that addition or deletion of a multiple of three bases does not cause a frameshift. After the frameshift mutation is encountered, missense codons will be read up to the first stop codon. Like nonsense mutations, frameshift mutations usually lead to complete inactivation of the gene.

Although many different kinds of mutations occur spontaneously, the frequency with which mutations occur can be increased as much as 10^3 fold by treatment of cells with a mutagen. Here are some general categories of mutagens

Type of Mutagen	Mechanism	Examples	Type of Mutations
Base Analog	Analog is incorporated into DNA and can pair with more than one base	5-bromouracil 2-aminopurine	$A \cdot T \rightarrow G \cdot C$, $G \cdot C \rightarrow A \cdot T$ $A \cdot T \rightarrow G \cdot C$
Base modifying agent	Chemical or photo damage to DNA can be repaired, but repair itself is error prone	Hydroxylamine EMS UV	$G \cdot C \rightarrow A \cdot T$ $G \cdot C \rightarrow A \cdot T$, $C \cdot G$, or $T \cdot A$ All changes
Intercalating agent	Polycyclic compounds can fit between bases and cause mis-copying by polymerase to add or delete bases	Acridine Proflavine ICR-191	Frameshifts (+ or -) " "

Suppressor mutations

A powerful mode of genetic analysis is to investigate the types of mutations that can reverse the phenotypic effects of a starting mutation. Starting with a $LacZ^-$ mutant of *E. coli* (white on X-gal), we could mutagenize and then plate out a large number of colonies on X-gal plates to screen for rare colonies that were blue on X-gal. These **revertants** could have either been mutated such that the starting mutation was reversed or they could have acquired a new mutation that somehow compensates for the starting mutation.

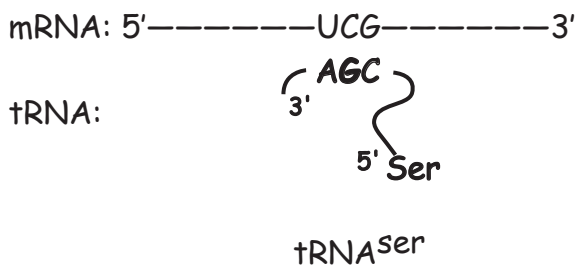
The possibilities are:

- 1) back mutation - true wild type
- 2) intragenic suppressor - compensating mutation in same gene
- 3) extragenic suppressor - compensating mutation in different gene

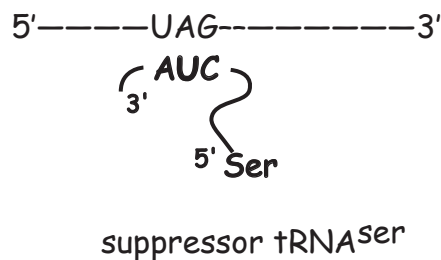
Nonsense suppressors.

An important class of extragenic suppressor mutations can suppress nonsense mutations by changing the ability of the cells to read a nonsense codon as a sense codon. Such extragenic revertants were originally isolated by selecting for reversion of amber (UAG) mutations in two different genes. Since simultaneous back mutations at two different sites is highly improbable the most frequent mechanism for suppression is a single mutation in the gene for a tRNA that changes the codon recognition portion of the tRNA. For example, one of several possible nonsense suppressors occurs in the gene for a serine tRNA (tRNA^{ser}). One of six tRNA^{ser} normally contains the anticodon sequence CGA which recognizes the serine codon UCG (by convention sequences are given in the 5' to 3' direction). A mutation that changes the anticodon to CUA allows the mutant tRNA^{ser} to recognize a UAG codon and insert tryptophan when a UAG codon appears in a coding sequence.

Recognition of UCG (serine codon)
by wild type tRNA^{ser}



Recognition of UAG(stop codon)
by amber suppressor mutant tRNA^{ser}



The presence of an amber suppressing mutation is usually designated Su⁺ whereas a wild-type (nonsuppressing) strain would be designated Su⁻.

Example: LacZ^{am} designates an amber (nonsense) mutation in the LacZ gene. Whereas a LacZ^{am} E. coli strain is Lac⁻ and therefore is white on X-gal plates, a strain with both a LacZ^{am} mutation and an amber suppressor (Su⁺) mutation in a tRNA gene might be phenotypically Lac⁺ because of extragenic suppression of the LacZ^{am} mutation.

The combined use of amber mutations and an amber suppressor produces a conditional mutation, ie a mutation that is expressed under some circumstances but not under others. Conditional mutants are especially useful for studying mutations in essential genes. Another kind of conditional mutation is a temperature sensitive mutation for which the mutant trait is exhibited at high temperature but not at low temperature.