

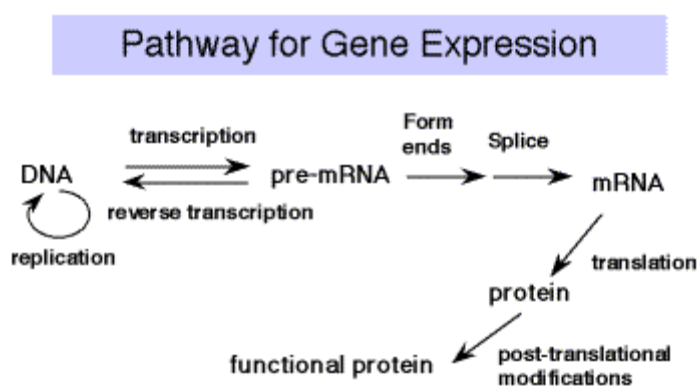
# GENETIC CODE

## TIG

### Introduction:

Experiments testing the effects of frameshift mutations showed that the deletion or addition of 1 or 2 nucleotides caused a loss of function, whereas deletion or addition of 3 nucleotides allowed retention of considerable function. This demonstrated that the coding unit is 3 nucleotides. **The nucleotide triplet that encodes an amino acid is called a codon.** Each group of three nucleotides encodes one amino acid. Since there are 64 combinations of 4 nucleotides taken three at a time and only 20 amino acids, the code is degenerate (more than one codon per amino acid, in most cases). The adaptor molecule for translation is tRNA. A charged tRNA has an amino acid at one end, and at the other end it has an anticodon for matching a codon in the mRNA; ie. it "speaks the language" of nucleic acids at one end and the "language" of proteins at the other end. The machinery for synthesizing proteins under the direction of template mRNA is the ribosome.

Figure: tRNAs serve as an adaptor for translating from nucleic acid to protein



### A. Size of a codon: 3 nucleotides

1. Three is the minimum number of nucleotides per codon needed to encode 20 amino acids.

a. 20 amino acids are encoded by combinations of 4 nucleotides

b. If a codon were two nucleotides, the set of all combinations could encode only

$4 \times 4 = 16$  amino acids.

c. With three nucleotides, the set of all combinations can encode

$4 \times 4 \times 4 = 64$  amino acids (i.e. 64 different combinations of four nucleotides taken three at a time).

2. Results of combinations of frameshift mutations show that the code is in triplets. Length-altering mutations that add or delete one or two nucleotides have severe defective phenotype (they change the reading frame, so the entire amino acid sequence after the mutation is altered.). But those that add or delete three nucleotides have little or no effect. In the latter case, the reading frame is maintained, with an insertion or deletion of an amino acid at one site. Combinations of three different single nucleotide deletions (or insertions), each of which has a loss-of-function phenotype individually, can restore substantial function to a gene. The wild-type reading frame is restored after the 3rd deletion (or insertion).

### B. Experiments to decipher the code:

1. Several different cell-free systems have been developed that catalyze protein synthesis. This ability to carry out translation in vitro was one of the technical advances needed to allow investigators to determine the genetic code.

a. Mammalian (rabbit) reticulocytes: ribosomes actively making lots of globin.

b. Wheat germ extracts

c. Bacterial extracts

2. The ability to synthesize random polynucleotides was another key development to allow the experiments to decipher the code. S. Ochoa isolated the enzyme polynucleotide phosphorylase, and showed that it was capable of linking nucleoside diphosphates (NDPs) into polymers of NMPs (RNA) in a reversible reaction.



The physiological function of polynucleotide phosphorylase is to catalyze the reverse reaction, which is used in RNA degradation. However, in a cell-free system, the forward reaction is very useful for making random RNA polymers.

3. Homopolymers program synthesis of specific homo-polypeptides (Nirenberg and Matthei, 1961).

a. If you provide only UDP as a substrate for polynucleotide phosphorylase, the product will be a homopolymer poly(U).

b. Addition of poly(U) to an in vitro translation system (e.g. E. coli lysates), results in a newly synthesized polypeptide which is a polymer of polyphenylalanine.

c. Thus UUU encodes Phe.

d. Likewise, poly(A) programmed synthesis of poly-Lys; AAA encodes Lys.

Poly(C) programmed synthesis of poly-Pro; CCC encodes Pro.

Poly(G) programmed synthesis of poly-Gly; GGG encodes Gly.

4. Use of mixed co-polymers

a. If two NDPs are mixed in a known ratio, polynucleotide phosphorylase will make a mixed co-polymer in which nucleotide is incorporated at a frequency proportional to its presence in the original mixture.

b. For example, consider a 5:1 mixture of A:C. The enzyme will use ADP 5/6 of the time, and CDP 1/6 of the time. An example of a possible product is:

AACAAAACAACAAAAAAACAAAAACAAAC...

Table 3.4.1. Frequency of triplets in a poly(AC) (5:1) random copolymer

Composition	Number	Probability	Relative frequency
3 A	1	0.578	1.0
2 A, 1 C	3	3 x 0.116	3 x 0.20
1 A, 2 C	3	3 x 0.023	3 x 0.04
3 C	1	0.005	0.01

c. So the frequency that AAA will occur in the co-polymer is

$$(5/6)(5/6)(5/6) = 0.578.$$

This will be the most frequently occurring codon, and can be normalized to 1.0 ( $0.578/0.578 = 1.0$ )

d. The frequency that a codon with 2 A's and 1 C will occur is

$$(5/6)(5/6)(1/6) = 0.116.$$

There are three ways to have 2 A's and 1 C, i.e. AAC, ACA and CAA. So the frequency of occurrence of all the A2C codons is  $3 \times 0.116$ .

Normalizing to AAA having a relative frequency of 1.0, the frequency of A2C codons is  $3 \times (0.116/0.578) = 3 \times 0.2$ .

e. Similar logic shows that the expected frequency of AC2 codons is  $3 \times 0.04$ , and the expected

Table 3.4.2. Amino acid incorporation with poly(AC) (5:1) as a template

Radioactive amino acid	Precipitable cpm		Observed incorporation		Theoretical incorporation
	- template	+ template			
Lysine	60	4615	100.0	100	
Threonine	44	1250	26.5	24	
Asparagine	47	1146	24.2	20	
Glutamine	39	1117	23.7	20	
Proline	14	342	7.2	4.8	
Histidine	282	576	6.5	4	

frequency of CCC is 0.01.

These data are from Speyer et al. (1963) Cold Spring Harbor Symposium in Quantitative Biology, 28:559. The theoretical incorporation is the expected value given the genetic code as it was subsequently determined.

f. When this mixture of mixed copolymers is used to program in vitro translation, Lys is incorporated most frequently, which can be expressed as 100. This confirms that AAA encodes Lys.

g. Relative to Lys incorporation as 100, Thr, Asn, and Gln are incorporated with values of 24 to 26, very close to the expectation for amino acids encoded by one of the A2C codons. However, these data do not show which of the A2C codons encodes each specific amino acid. We now know that ACA encodes Thr, AAC encodes Asn, and CAA encodes Gln.

h. Pro and His are incorporated with values of 6 and 7, which is close to the expected 4 for amino acids encoded by AC2 codons. E.g. CCA encodes Pro, CAC encodes His. ACC encodes Thr, but this incorporation is overshadowed by the "26.5" units of incorporation at ACA. Or, more accurately, "26.5" @ 20 (ACA) + 4 (ACC) for Thr.

##### 5. Defined trinucleotide codons stimulate binding of aminoacyl-tRNAs to ribosomes

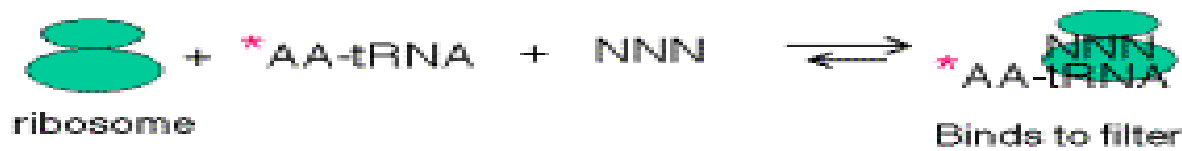
a. At high concentrations of  $Mg^{2+}$  cations, the normal initiation mechanism, requiring f-Met-tRNA<sup>f</sup>, can be overridden, and defined trinucleotides can be used to direct binding of particular, labeled aminoacyl-tRNAs to ribosomes.

b. E.g. If ribosomes are mixed with UUU and radiolabeled Phe-tRNA<sup>phe</sup>, under these conditions, a ternary complex will be formed that will stick to nitrocellulose ("Millipore assay" named after the manufacturer of the nitrocellulose).

c. One can then test all possible combinations of triplet nucleotides.

Figure : Data from Nirenberg and Leder (1964) Science 145:1399.

Defined trinucleotides stimulate binding of particular aminoacyl-tRNAs to ribosomes



Which trinucleotide will allow binding of a particular AA-tRNA to ribosomes?

AA-tRNA	pmoles AA-tRNA bound with:			
	no NNN	UUU	AAA	CCC
Phe-tRNA	0.34	1.56	0.20	0.30
Lys-tRNA	0.80	0.56	6.13	0.60
Pro-tRNA	0.24	0.20	0.18	0.73

6. Repeating sequence synthetic polynucleotides (Khorana)

a. Alternating copolymers: e.g. (UC)<sub>n</sub> programs the incorporation of Ser and Leu. So UCU and CUC encode Ser and Leu, but cannot tell which is which. But in combination with other data, e.g. the random mixed copolymers in section 4 above, one can make some definitive determinations. Such subsequent work showed that UCU encodes Ser and CUC encodes Leu.

b. poly(AUG) programs incorporation of poly-Met and poly-Asp at high Mg concentrations. AUG encodes Met, UGA is a stop, so GUA must encode Asp.

C. The genetic code

By compiling observations from experiments such as those outlined in the previous section, the coding capacity of each group of 3 nucleotides was determined. This is referred to as the genetic code. It is summarized in the table below. This tells us how the cell translates from the "language" of nucleic acids (polymers of nucleotides) to that of proteins (polymers of amino acids).

Table : **The Genetic Code**

**Position in Codon**

1st	2nd								3rd
	U	C	A	G					
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys					U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys					C
	UUA Leu	UCA Ser	UAA Term	UGA Term					A
	UUG Leu	UCG Ser	UAG Term	UGG Trp					G
C	CUU Leu	CCU Pro	CAU His	CGU Arg					U
	CUC Leu	CCC Pro	CAC His	CGC Arg					C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg					A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg					G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser					U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser					C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg					A

	AUG*	Met	ACG	Thr	AAG	Lys	AGG	Arg	G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
	GUG*	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

\* Sometimes used as initiator codons.

2. Of the total of 64 codons, 61 encode amino acids and 3 specify termination of translation.

### 3. Degeneracy:

The degeneracy of the genetic code refers to the fact that most amino acids are specified by more than one codon. The exceptions are methionine (AUG) and tryptophan (UGG). The degeneracy is found primarily the third position. Consequently, single nucleotide substitutions at the third position may not lead to a change in the amino acid encoded. These are called silent or synonymous nucleotide substitutions and do not alter the encoded protein. This is discussed in more detail below.

The pattern of degeneracy allows one to organize the codons into "families" and "pairs". In 9 groups of codons, the nucleotides at the first two positions are sufficient to specify a unique amino acid, and any nucleotide (abbreviated N) at the third position encodes that same amino acid. These comprise 9 codon "families". An example is ACN encoding threonine.

There are 13 codon "pairs", in which the nucleotides at the first two positions are sufficient to specify two amino acids. A purine (R) nucleotide at the third position specifies one amino acid, whereas a pyrimidine (Y) nucleotide at the third position specifies the other amino acid.

These examples add to more than 20 (the number of amino acids) because leucine (encoded by UUR and CUN), serine (encoded by UCN and AGY) and arginine (encoded by CGN and AGR) are encoded by both a codon family and a codon pair. The UAR codons specifying termination of translation were counted as a codon pair. The three codons encoding isoleucine (AUU, AUC and AUA) are half-way between a codon family and a codon pair.

4. Chemically similar amino acids often have similar codons.

5. The major codon specifying initiation of translation is AUG

Bacteria can also use GUG or UUG, and very rarely AUU and possibly CUG. Using data from the 4288 genes identified by the complete genome sequence of E. coli, the following frequency of use of codons in initiation was determined:

AUG is used for 3542 genes.

GUG is used for 612 genes.

UUG is used for 130 genes.

AUU is used for 1 gene.

CUG may be used for 1 gene.



Regardless of which codon is used for initiation, the first amino acid incorporated during translation is f-Met in bacteria.

#### 6. Three codons specify termination of translation: UAA, UAG, UGA.

Of these three codons, UAA is used most frequently in *E. coli*, followed by UGA. UAG is used much less frequently.

UAA is used for 2705 genes.

UGA is used for 1257 genes.

UAG is used for 326 genes.

#### 7. The genetic code is almost universal

In the rare exceptions to this rule, the differences from the genetic code are fairly small. For example, one exception is RNA from mitochondrial DNA, where both UGG and UGA encode Trp.

#### D. Differential codon usage

1. Various species have different patterns of codon usage: E.g. one may use 5' UUA to encode Leu 90% of the time (determined by nucleotide sequences of many genes). It may never use CUR, and the combination of UUG plus CUY may account for 10% of the codons.

2. tRNA abundance correlates with codon usage in natural mRNAs: In this example, the tRNA<sup>Leu</sup> with 3' AAU at the anticodon will be the most abundant.

3. The pattern of codon usage may be a predictor of the level of expression of the gene: In general, more highly expressed genes tend to use codons that are frequently used in genes in the rest of the genome. This has been quantitated as a "codon adaptation index". Thus in analyzing complete genomes, a previously unknown gene whose codon usage profile matches the preferred codon usage for the organism would score high on the codon adaptation index, and one would propose that it is a highly expressed gene. Likewise, one with a low score on the index may encode a low abundance protein.

The observation of a gene with a pattern of codon usage that differs substantially from that of the rest of the genome indicates that this gene may have entered the genome by horizontal transfer from a different species.

4. The preferred codon usage is a useful consideration in "reverse genetics": If you know even a partial amino acid sequence for a protein and want to isolate the gene for it, the family of mRNA sequences that can encode this amino acid sequence can be determined easily. Because of the degeneracy in the code, this family of sequences can be very large. Since one will likely use these sequences as hybridization probes or as PCR primers, the larger the family of possible sequences is, the more likely that one can get hybridization to a target sequence that differs from the desired one. Thus one wants to limit the number of possible sequences, and by referring to a table of codon preferences (assuming they are known for the organism of interest), then one can use the preferred codons rather than all possible codons. This limits the number of sequences that one needs to make as hybridization probes or primers.

#### E. Wobble in the anticodon

This flexibility at the "wobble" position allows some tRNAs to pair with two or three codons, thereby reducing the number of tRNAs required for translation. The following "wobble" rules mean that the 61 codons (for 20 amino acids) can be read by as few as 31 anticodons (or 31 tRNAs).

In addition to the usual base pairs, one can have G-U pairs and I in the anticodon 1st position can pair with U, C or A (wobble rules).

5' base of the anticodon = 3' base of the codon =

first position in the tRNA third position in the mRNA

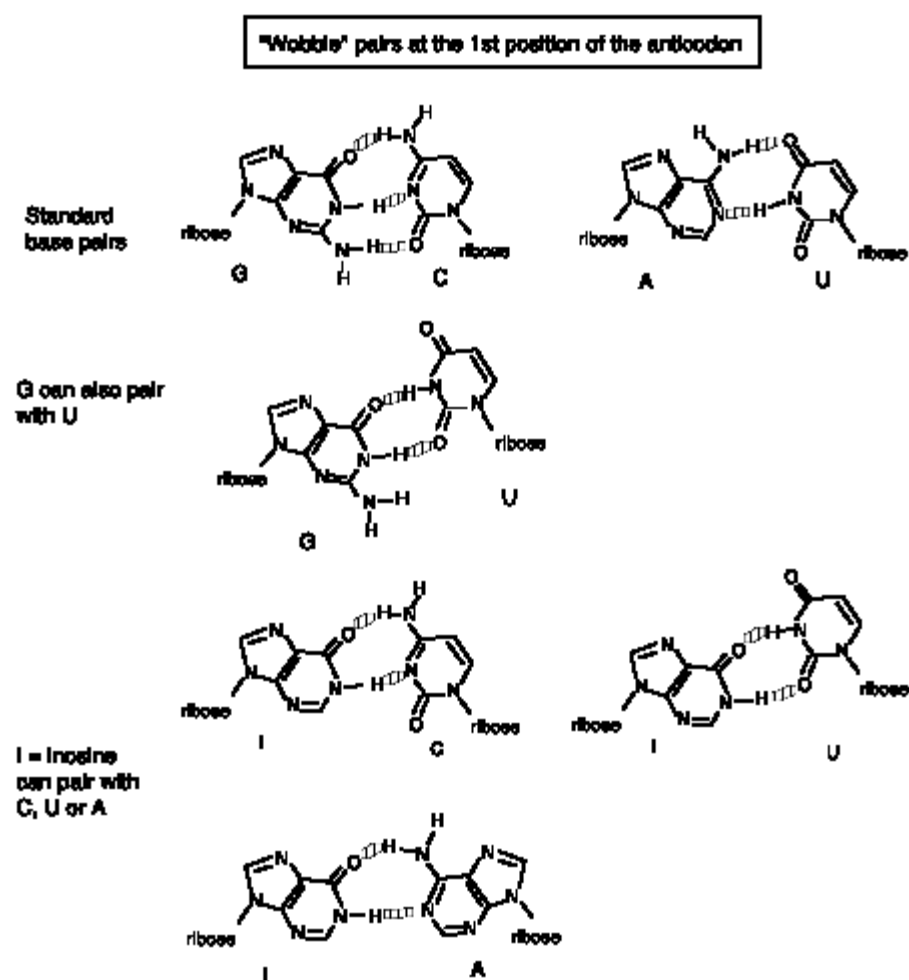
C G

A U

U A or G

G C or U

I U, C or A



## F. Types of mutations

### Base substitutions

There are two types of base substitutions:

**Transitions:** A purine substitutes for a purine or a pyrimidine substitutes for another pyrimidine. The same class of nucleotide remains. Examples are A substituting for G or C substituting for T.

**Transversions:** A purine substitutes for a pyrimidine or a pyrimidine substitutes for a purine. A different class of nucleotide is placed into the DNA, and the helix will be distorted (especially with a purine-purine base pair). Examples are A substituting for T or C, or C substituting for A or G.

Over evolutionary time, the rate of accumulation of transitions exceeds the rate of accumulation of transversions.

### Effect of mutations on the mRNA:

Missense mutations cause the replacement of an amino acid. Depending on the particular replacement, it may or may not have a detectable phenotypic consequence. Some replacements, e.g. a valine for an leucine in a position that is important for maintaining an  $\alpha$ -helix, may not cause a detectable change in the structure or function of the protein. Other replacements, such as valine for

a glutamate at a site that causes hemoglobin to polymerize in the deoxygenated state, cause significant pathology (sickle cell anemia in this example).

**Nonsense mutations cause premature termination of translation.** They occur when a substitution, insertion or deletion generates a stop codon in the mRNA within the region that encodes the polypeptide in the wild-type mRNA. They almost always have serious phenotypic consequences.

Frameshift mutations are insertions or deletions that change the reading frame of the mRNA. They almost always have serious phenotypic consequences.

Not all base substitutions alter the encoded amino acids

The base substitution may lead to an alteration in the encoded polypeptide sequence, in which case the substitution is called nonsynonymous or nonsilent.

If the base substitution occurs in a degenerate site in the codon, so that the encoded amino acid is not altered, it is called a synonymous or silent substitution.

Example:

ACU → AAU is a nonsynonymous substitution that results in Thr → Asn

while,

ACU → ACC is a synonymous substitution that results in Thr → Thr

Examination of the patterns of degeneracy in the genetic code shows that nonsynonymous substitutions occur mostly in the first and second positions of the codon, whereas synonymous substitutions occur mostly in the third position. However, there are several exceptions to this rule.

In general, the rate of fixation of synonymous substitutions in a population is significantly greater than the rate of fixation of nonsynonymous substitutions. This is one of the strongest supporting arguments in favor of model of neutral evolution, or evolutionary drift, as a principle cause of the substitutions seen in natural populations.

Reference:

Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)