

Core Course X

Unit 4: Central Dogma & Genetic Code

Central Dogma of Molecular Biology

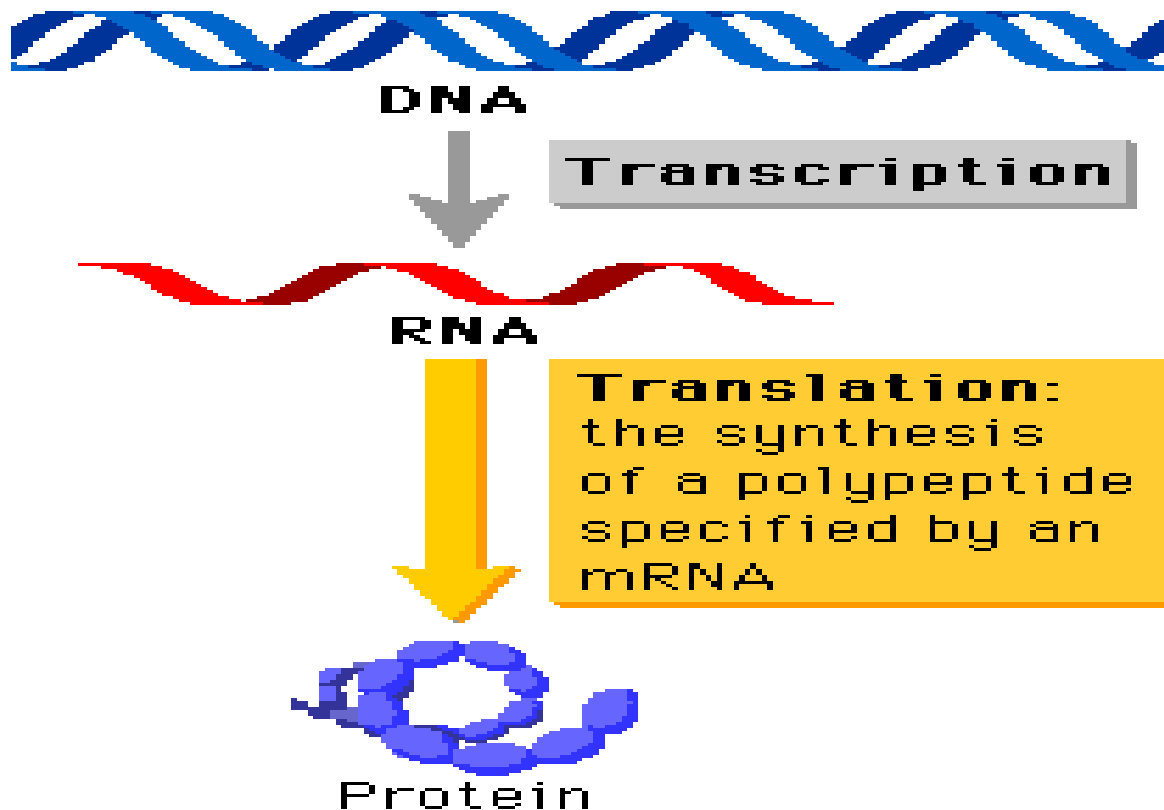
“The central dogma deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to either protein or nucleic acid.”



Francis Crick, 1958

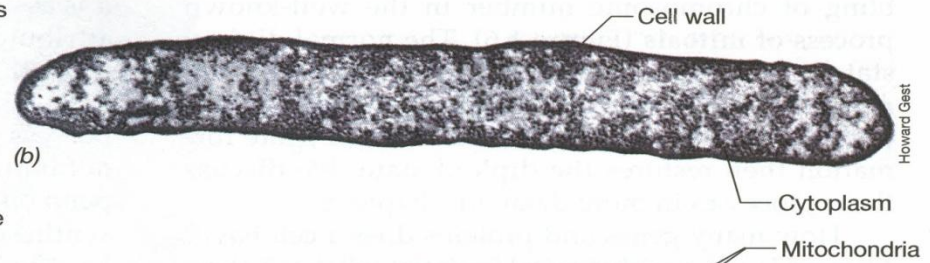
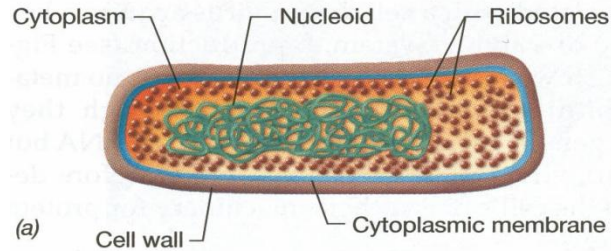
... in other words

The Central Dogma

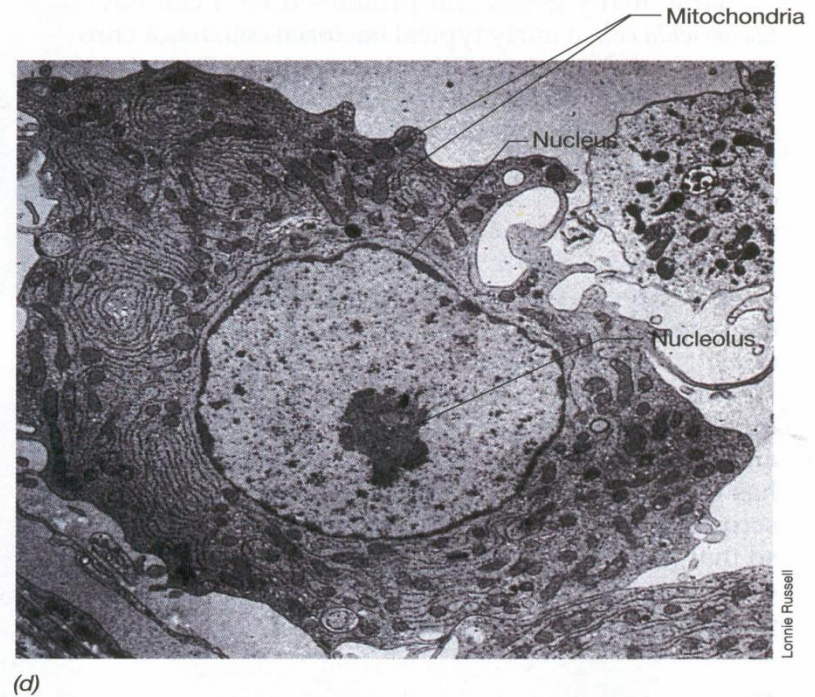
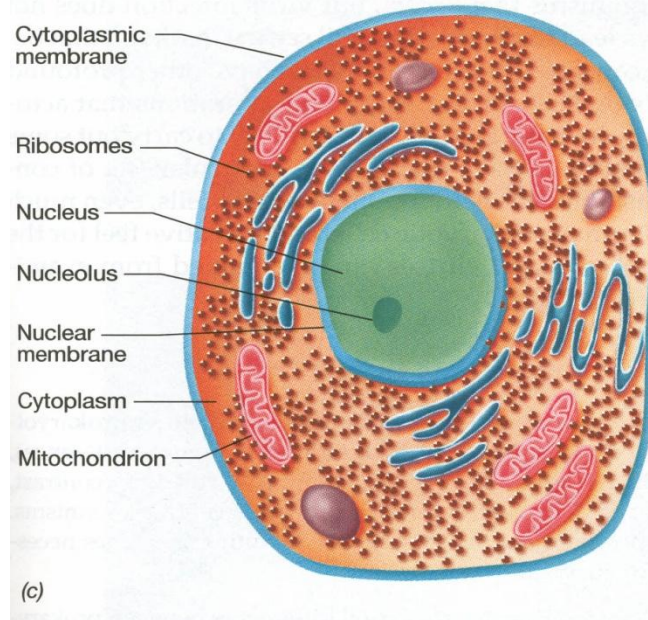


The Basics: Cell Organization

Prokaryotes



Eukaryotes



Central Dogmas of Molecular Biology

- 1) The concept of genes is historically defined on the basis of genetic inheritance of a phenotype. (Mendelian Inheritance)
- 2) The DNA an organism encodes the genetic information. It is made up of a double stranded helix composed of ribose sugars.

Adenine(A), Cytosine (C), Guanine (G) and Thymine (T).

[note that only 4 values need be encoded ACGT.. Which can be done using 2 bits.. But to allow redundant letter combinations (like N means any 4 nucleotides), one usually resorts to a 4 bit alphabet.]

Central Dogmas of Molecular Biology

3) Each side of the double helix faces it's complementary base.

$A \leftrightarrow T$, and $G \leftrightarrow C$.

4) Biochemical process that read off the DNA always read it from the 5'' side towards the 3' side. (replication and transcription).

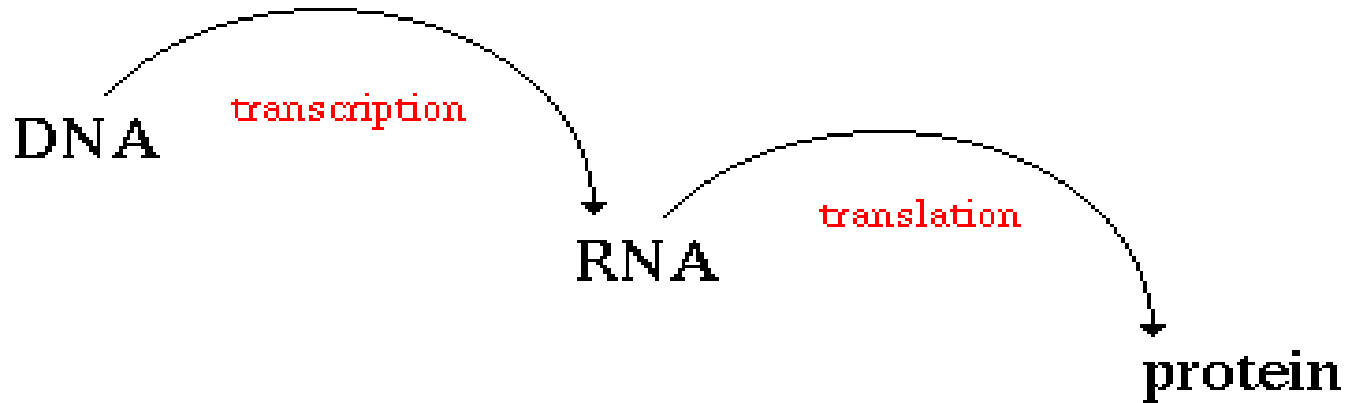
5) A gene can be located on either the 'plus strand' or the minus strand. But rule 4) imposes the orientation of reading .. And rule 3 (complementarity) tells us to complement each base E.g.

If the sequence on the + strand is ACGTGATCGATGCTA, the – strand must be read off by reading the complement of this sequence going 'backwards'

e.g. TAGCATCGATCACGT

Central Dogmas of Molecular Biology

6) DNA information is copied over to mRNA that acts as a template to produce proteins.

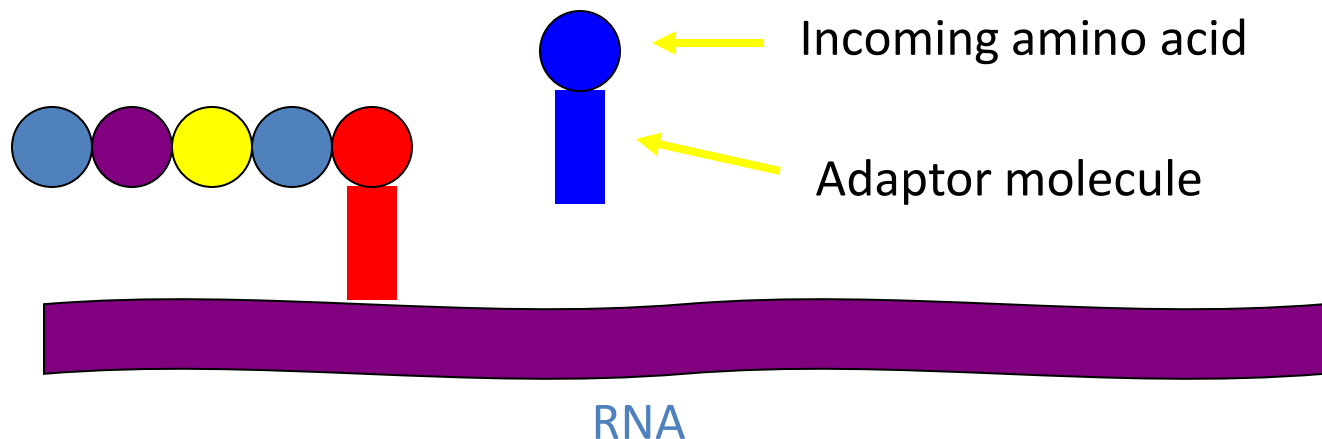


We often concentrate on protein coding genes, because proteins are the building blocks of cells and the majority of bio-active molecules. (but let's not forget the various RNA genes)

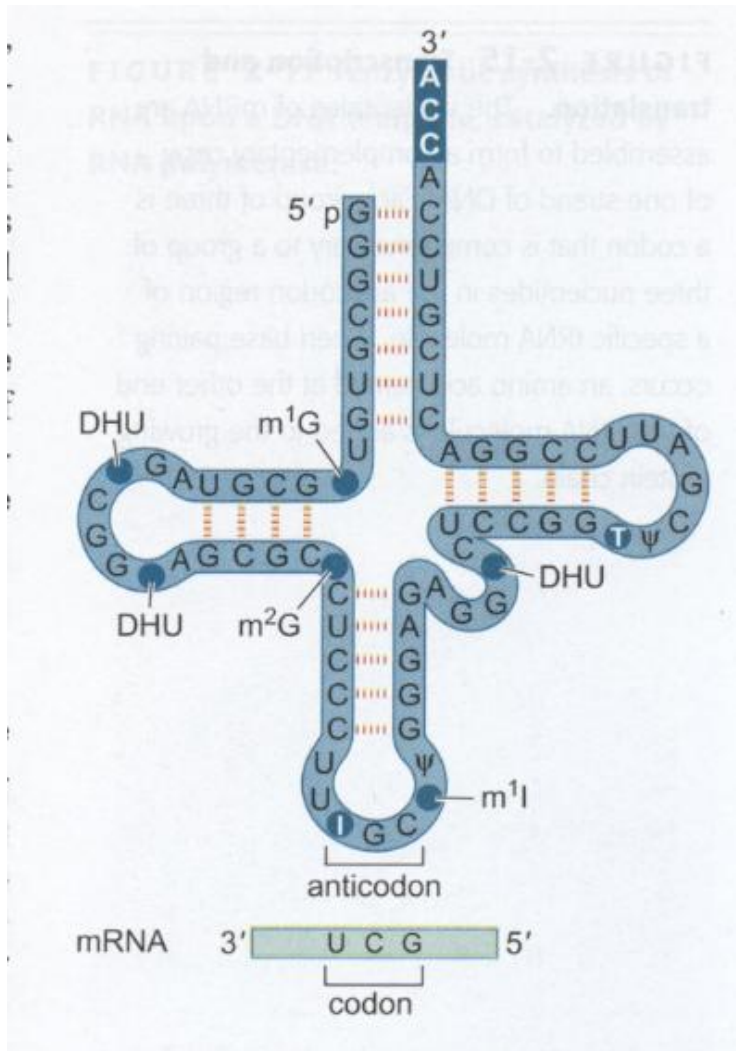
Crick's adaptor hypothesis

- Can folded RNA act directly as the template for protein synthesis?
- Seems unlikely:
 - the nucleosides chemically want to react with water soluble groups
 - but many amino acids are polar
 - no clear way to discriminate chemically similar amino acids

Crick proposes that an adaptor molecule must fit between RNA and the incoming amino acids, but its nature is unknown



Crick's adaptors (tRNAs) are themselves RNA molecules



- Self-folding by complementary base pairs gives a structure with several functional domains
- Account for ~10% of cellular RNA abundance
- Typically includes several modified, non-standard bases.

Discovery of mRNA

mRNA carries the information from DNA to the ribosomes in a cell where it acts as a template for protein synthesis.

In 1953, Hershey reported that shortly after infection with phage, bacteria produced a form of RNA at high level and this RNA was also broken down rapidly. However, there was skepticism that this could have been a result of pathological infection.

Discovery of mRNA

First clear indication by Elliot Volkin and Lazarus Astrachan in 1956

Infected *E. coli* with T2 bacteriophages and put them into a medium containing ^{32}P sources.

Finding the missing link between the genetic code and proteins. Protein synthesis of *E. coli* was halted and phage proteins were synthesized.

There was very little change in total RNA content except the formation of very unstable ribonucleic acids for which the base composition matched those of viral DNA.

Discovery of mRNA

In 1961, Spiegelman clearly demonstrated that the base sequence of mRNA is complementary to the DNA template through the formation of RNA-DNA hybrid

Using ^{32}P label on the single stranded mRNA synthesized in *E. coli* after the infection by T2 bacteriophage and T2 DNA labeled with ^3H .

The samples were analyzed by density-gradient centrifugation and it was found that no other DNA molecule was able to hybridize with the T2 mRNA.

Why does Nature bothers with the mRNA?

Why would the cell want to have an intermediate between DNA and the proteins it encodes?

- Gene information can be amplified by having many copies of an RNA made from one copy of DNA.
- Regulation of gene expression can be effected by having specific controls at each element of the pathway between DNA and proteins. The more elements there are in the pathway, the more opportunities there are to control it in different circumstances.
- In Eukaryotes, the DNA can then stay pristine and protected, away from the caustic chemistry of the cytoplasm.

Zamecnick and Hoagland discover aminoacyl synthetases

Enzymes that added an adaptor (that we now know to be tRNA) to amino acids *prior* to their incorporation in proteins

It turns out these tRNAs are Crick's proposed adaptors



Mahlon Hoagland

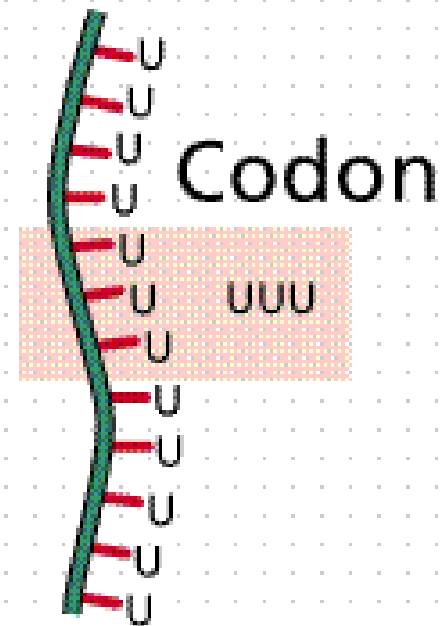


Paul Zamecnik

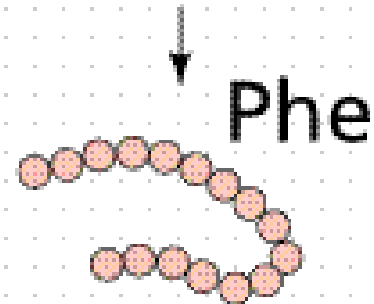
The Genetic Code:

- The correlation between **nucleotide** sequence (DNA or RNA) and **amino acid** sequence (protein)
- **Codons:** combination of 3 mRNA nucleotides that code for a specific amino acid

Poly U



Translation



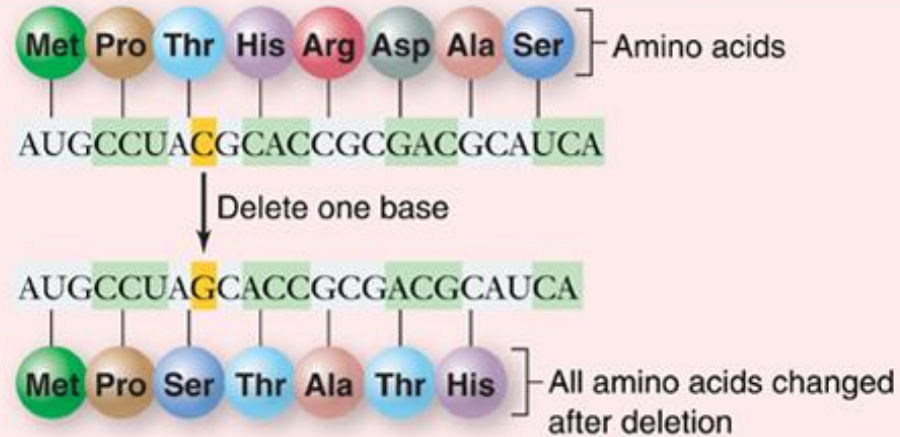
Polyphenylalanine

Deciphering Genetic Code

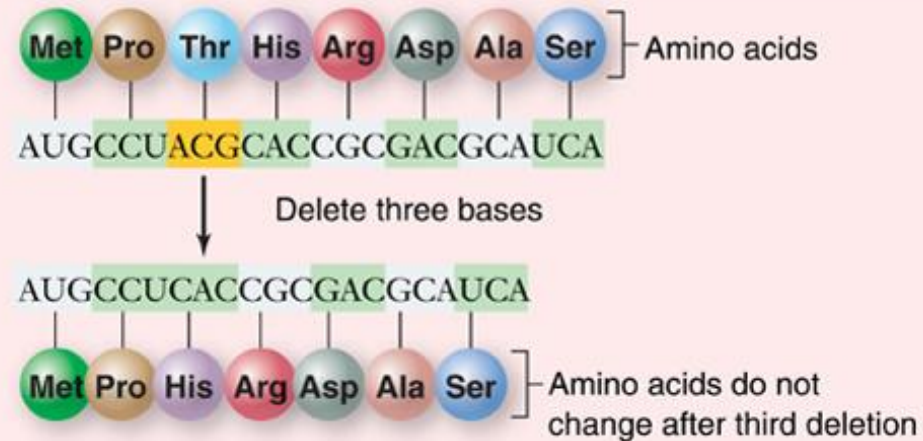
Deciphering the genetic code required determining how 4 nucleotides (A, T, G, C) could encode more than 20 amino acids.

Francis Crick and Sydney Brenner determined that the DNA is read in sets of 3 nucleotides for each amino acid.

One Base Deleted



Three Bases Deleted



The Genetic Code

codon: set of 3 nucleotides that specifies a particular amino acid

reading frame: the series of nucleotides read in sets of 3 (codon)

- only 1 reading frame is correct for encoding the correct sequence of amino acids

The Genetic Code

Marshall Nirenberg identified the codons that specify each amino acid.

RNA molecules of only 1 nucleotide and of specific 3-base sequences were used to determine the amino acid encoded by each codon.

The amino acids encoded by all 64 possible codons were determined.

Khorana's synthetic RNA approach to cracking the genetic code

Example RNA with two repeating units

RNAs with two repeating units:

(**UCUCUCU** → **UCU CUC UCU**) produced a polypeptide consisting of alternating Serine and **CUC** codes for Leucine

RNAs with three repeating units:

(**UACUACUA** → **UAC UAC UAC**, or **ACU ACU ACU**, or **CUA CUA CUA**) produced three different strings of amino acids

RNAs with four repeating units including **UAG**, **UAA**, or **UGA**, produced only dipeptides and tripeptides thus revealing that **UAG**, **UAA** and **UGA** are **stop codons**.

Types of codons:

- 64 codons code for 20 amino acids
 - Thus more than one codon codes for an AA
- **Start codon:** (AUG) starts the process of translation
- **Stop codons:** (UAA, UAG, UGA) ends the process of translation

Nature of the genetic code

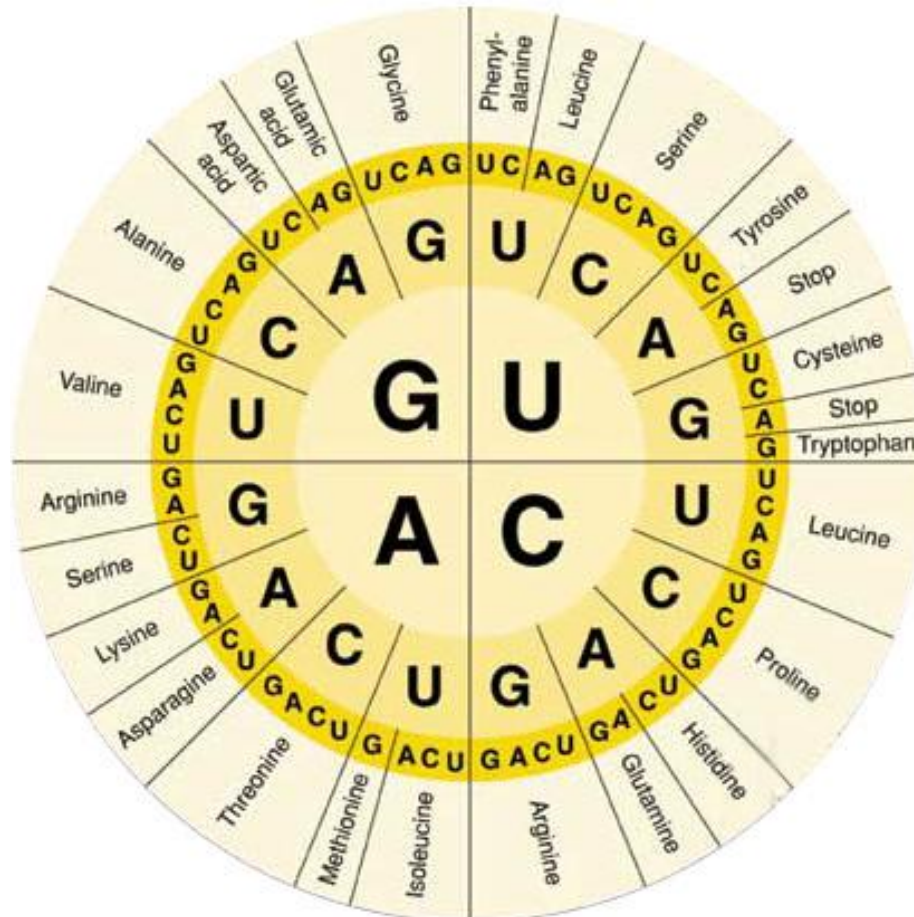
- Obvious early on most likely a triplet code in order to code 20 amino acids:
 - 4 x 4 nucleotides can specify $4^2 = 16$ amino acids
 - 4 x 4 x 4 nucleotides can specify $4^3 = 64$ amino acids
- Code must be redundant
- Not overlapping – Sydney Brenner's thought experiment
- Marshall Nirenberg and Heinrich Matthaei showed that a homopolymer (UUUUUU.... *etc.*) produced a poly-phenylalanine protein

Second letter

First letter

	U	C	A	G	
U	UUU UUC Phenylalanine	UCU UCC UCA UCG Serine	UAU UAC Tyrosine	UGU UGC Cysteine	U
	UUA UUG Leucine		UAA UAG Stop codon Stop codon	UGA Stop codon	A
				UGG Tryptophan	G
C	CUU CUC CUA CUG Leucine	CCU CCC CCA CCG Proline	CAU CAC Histidine	CGU CGC CGA CGG Arginine	U
			CAA CAG Glutamine		C
					A
A	AUU AUC AUA Isoleucine	ACU ACC ACA ACG Threonine	AAU AAC Asparagine	AGU AGC Serine	U
	AUG Methionine; initiation codon		AAA AAG Lysine	AGA AGG Arginine	C
					A
G	GUU GUC GUA GUG Valine	GCU GCC GCA GCG Alanine	GAU GAC Aspartic acid	GGU GGC GGA GGG Glycine	U
			GAA GAG Glutamic acid		C
					A
					G

Circular Genetic Code



Genetic Code

How does an mRNA specify amino acid sequence? The answer lies in the genetic code. It would be impossible for each amino acid to be specified by one nucleotide, because there are only 4 nucleotides and 20 amino acids. Similarly, two nucleotide combinations could only specify 16 amino acids. The final conclusion is that each amino acid is specified by a particular combination of three nucleotides, called a **codon**:

Each 3 nucleotide code for one amino acid.

- The first codon is the start codon, and usually coincides with the Amino Acid Methionine. (M which has codon code 'ATG')
- The last codon is the stop codon and does NOT code for an amino acid. It is sometimes represented by '*' to indicate the 'STOP' codon.
- A coding region (abbreviation CDS) starts at the START codon and ends at the STOP codon.

Codon table

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

Third Nucleotide

Abbreviation	Amino acid
Ala A	Alanine
Arg R	Arginine
Asn N	Asparagine
Asp D	Aspartic acid
Cys C	Cysteine
Gln Q	Glutamine
Glu E	Glutamic acid
Gly G	Glycine
His H	Histidine
Ile I	Isoleucine
Leu L	Leucine
Lys K	Lysine
Met M	Methionine
Phe F	Phenylalanine
Pro P	Proline
Ser S	Serine
Thr T	Threonine
Trp W	Tryptophan
Tyr Y	Tyrosine
Val V	Valine

Note the **degeneracy** of the genetic code. Each amino acid might have up to six codons that specify it.

- Different organisms have different frequencies of codon usage.
- A handful of species vary from the codon association described above, and use different codons for different amino acids.

Start
AUG

The Genetic Code

- Code is triplet, 64 codons (4*4*4 bases)
- Start and Stop codons
- Codon synonyms for 18 codons
- Code is universal (exc. mitochondria)

How do tRNAs recognize to which codon to bring an amino acid? The tRNA has an **anticodon** on its mRNA-binding end that is complementary to the codon on the mRNA. Each tRNA only binds the appropriate amino acid for its anticodon.

Second Genetic Code

- An imprecise term that sometimes refers to the nature of the amino acid residues of a protein which determine its secondary and tertiary structure, and sometimes to the features of tRNA molecule that make it recognizable by one amino acid synthase but not by others.
- During 1988-89, this problem of interaction of specific tRNAs with corresponding aminoacyl tRNA synthetase enzymes (aaRSs) for aminoacylation of tRNAs (transfer of amino acid to tRNA) was described as the second genetic code.

- Although we generally say that individual codons are meant for specific amino acids, we should recognize the fact that in reality, codons do not specify amino acids, instead they specify tRNA molecules, which in turn specify amino acids with the help of enzymes called 'amino acyl RNA synthetases' (aaRSs).
- X-ray crystallographic structures of several 'tRNA-synthetase complexes' representing intermediates in the formation of 'amino-acyl tRNAs' have been studied.
- the most important of these is a complex between 'glutamy I-tRNA synthetase' with tRNA^{Gln} (tRNA for glutamine) and ATP.

- It was shown that there are multiple points of contact all along the inner side of the L-shaped structure of tRNA, which help in recognition of tRNA by synthetase enzyme and in the formation of linkage between tRNA and amino acid .

Second Half of the Genetic code

- The 'second Half of the genetic code' refers to the of a 'protein folding' problem or to the rules, which govern the formation of a three dimensional structure from the primary structure (amino acid sequence) of a protien.
- This problem also received attention of molecular biologists during 1980s and 1990s, when the phrase 'the second half of Genetic code' was popular.
- The problem continues to be important and is now more often described as 'protein folding' problem.

Importants of Second genetic code

- **Second genetic code led to the burgeoning field of genetic engineering in which the genetic material of cells is altered to tailor make proteins, such as those used as medicines to treat diabetics and heart diseases.**
- **The newly deciphered code plays a key role in a later step in protein synthesis.**
- **The newly revealed code helps explain a crucial aspect of protein synthesis : the attraction of building blocks, amino acids , to genetic materials inside the cell.**

Deciphering or Cracking of Genetic Code:

The most important feature of the genetic code is that it is a triplet codon. Three consecutive nucleotides of a single strand of DNA contain the information for coding a specific amino acid. It is known as a triplet codon. Translation takes place in such a way that these nucleotide triplets are read in a successive non-overlapping fashion. The information is first transcribed into messenger RNA, which has a sequence of bases complementary to DNA from which it is copied.

DNA has four types of bases C, T, G, A while RNA has four complementary bases G, A, C, U. The four base language of DNA is translated into language of 20 amino acids.

Deciphering or cracking of genetic code is the outcome of research of various scientists like Marshal Nirenberg, Steve Ochoa, Hargobind Khorana, Francis Crick, Mattaei and many others.

They discovered that the order in which the nucleotides are arranged – mRNA would determine the sequence of amino acids in polypeptides. Nirenberg and Mattaei gave the first experimental proof for the triplet codon. They used artificial mRNA raide of only uracil nucleotides (Poly U) in a cell free system. It resulted in the synthesis of polypeptide chain made up of only one kind of amino acid, phenylalanine. It was concluded that codon for phenylalanine was uridylic acid basis (uracil), UUU.

Similarly poly C(CCC) codon represented amino acid proline and poly A(AAA) codon represents am: no acid chain of lysine.

Later Hargobind Khorana confirmed the genetic code to be triplet codon. Using synthetic mRNA have alternating polynucleotides in a cell free system, discovered the chain of alternating amino acid using alternating uracil (U) and guanine (G) triplets which showed the following results.

GUG	UGU	GUG	UGU	GUG
Valine	Cystine	Valine	Cystine	Valine

Similarly, alternating ACA and CAC triplets produced a chain of following amino acids.

ACA	CAC	ACA	CAC	ACA
Threonine	Histidine	Threonine	Histidine	Threonine

This also confirmed that each codon is a triplet. The cell free protein-synthesizing system was an extract of E. coli without walls. It contained ribosomes, tRNA, tRNA synthetase enzymes, ATP and radioactive amino acids. Use of artificial trinucleotide templates resulted in determination of base composition of all genetic codons.

Khorana Technique:

H. G. Khorana and associates synthesized DNA of known sequences in vitro conditions. From this DNA, the mRNA is transcribed. These mRNA molecules, which have known sequences of nucleotides are then directed to synthesize polypeptides.

These polypeptides are then sequentially degraded to know their amino acid sequences. The nucleotide sequence is then compared with amino acid sequence. This nucleotide sequence specifies codons of amino acids. This technique led to the determination of codons for all amino acids.

DNA Genetic Codon is Triplet:

If a genetic codon consisted of two consecutive bases, the number of codons would be $4^2 = 16$. Since the number of amino acids is 20, this is insufficient. Therefore, three is the minimum number of bases needed to code for 20 amino acids $4^3 = 64$. George Gamow in 1964 pointed out that the code would contain at least three consecutive bases.

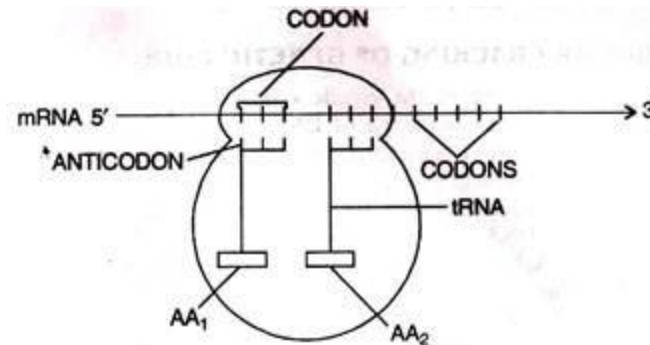


Fig. 11.1. Triplet codons on mRNA dictate the arrangement of amino acids of a polypeptide chain.

The length of the coding portion of a gene called reading frame depends upon the length of the message to be translated. For example, a sequence of 600 nucleotides will code for a polypeptide having a chain of 200 amino acids. Therefore, the length of mRNA depends upon the length of polypeptide it codes for.

Universal Genetic Code:

64 Types of codons					
First Base 5'-end	Second Base				Third Base 3'-end
	U	C	A	G	
U	UUU } Phe	UCU }	UAU } Tyr	UGU } Cys	U
	UUC }	UCC } Ser	UAC }	UGC }	C
	UUA } Leu	UCA }	UAA Stop	UGA Stop	A
	UUG }	UCG }	UAG Stop	UGG Tryp	G
C	CUU }	CCU }	CAU } His	CGU }	U
	CUC } Leu	CCC } Pro	CAC }	CGC } Arg	C
	CUA }	CCA }	CAA } Glu	CGA }	A
	CUG }	CCG }	CAG }	CGG }	G
A	AUU }	ACU }	AAU } Asp N	AGU } Ser	U
	AUC } Ileu	ACC } Thr	AAC }	AGC }	C
	AUA }	ACA }	AAA } Lys	AGA } Arg	A
	AUG Met	ACG }	AAG }	AGG }	G
G	GUU }	GCU }	GAU } Asp	GGU }	U
	GUC } Val	GCC } Ala	GAC }	GGC } Gly	C
	GUA }	GCA }	GAA } Glu	GGA }	A
	GUG }	GCG }	GAG }	GGG }	G

Phe-Phenylalanine
Cys-Cystaine
Glu-Glutamic acid
Thr-Threonine
Gly-Glycine.

Leu-Leueine
Tryp-Tryptophan
Arg-Arginine
Val-Valine

Ser-Serine
Pro-Proline
Ileu-Isoleucine
Ala-Alanine

Tyr-Tyrosine
His-Histidine
Met-Methionine
Asp-Aspartic acid

Properties of Genetic Codons:

As the genetic codon is read on mRNA, it is described in terms of four kinds of bases A, G, C, U. The sequence of template strand of DNA is read in the direction of $5' \rightarrow 3'$.

1. Genetic codon is a triplet codon:

The consecutive three nucleotides of the coding strand of DNA code for one amino acid.

2. Redundancy of the code:

Out of 64 codons, 61 codons represent amino acids, the remaining three are stop codons. As there are only 20 amino acid, coded by 61 codons, several codons specify the same amino acids. In this way the codons are synonyms.

This phenomenon is called redundancy of the code or degenerate code. Except for methionine and tryptophan all other codons are multiple codons. Each of the three amino acids — leucine, serine and arginine is represented by six different codons.

Number of codons coding for different amino acids:

Amino acids		Number of codons
1.	Methionine, Tryptophan	1
2.	Histidine, Glutamine, Lysine, Aspartic acid, Phenylalanine, tyrosine, asparagine, glutamic acid, cystine.	2
3.	Isoleucine	3
4.	Proline, Valine, Theonine, glycine, Alanine	4
5.	Leucine, serine, arginine	6

Wobble Hypothesis:

This was put forward by Francis Crick in 1965. According to this, hydrogen bonding between the codon of mRNA and anticodon of tRNA, there is a strict base pairing rule only for first two bases of the codon, while the base pairing involving the third base of codon appears to be less important. This is known as wobble hypothesis.

The first two bases of each codon are primary determinants of specificity. The third base pairing is not very stable and wobbles. For example, CUU, CUG, CUC, CUA codons, which differ only at the third base represent the same amino acid leucine. The first two bases of the codon form strong base pairs with the corresponding bases of the anticodon but the third base forms weak hydrogen bond.

At third position even unusual base pairing which does not conform to Watson and Crick base pairing rule can occur. Like adenine, cytosine and Uracil from A-I, C-I and U-I base pairs respectively at the third position, where I is inosine base.

Several codons meant for the same amino acids are recognized by the same tRNA. In this way a minimum of 32 tRNAs are required to translate 61 codons.

3. The codons are non-overlapping:

The same base cannot be a part of the two consecutive codons. They lie adjacent to each other.

4. The codons are comma less:

The three bases on DNA code for one amino acid and next three bases will code for next amino acid and so on. There is no gap or pause between the consecutive triplets.

5. Start codons:

The codons which initiate protein synthesis is called the start codon. The first amino acid of the polypeptide chain is always methionine coded by AUG codon. Therefore AUG is the start codon. Rarely the first amino acid is valine coded by GUG. In prokaryotes AUG codon codes for a modified amino acid, formyl methionine (f-Met).

6. Stop codons:

Protein synthesis stops before UAA codon, UAG codon and UGA codon. This indicates that these three codons are stop codons. They terminate the protein synthesis. The completed polypeptide chain is released. Release factors (RF) enter the 'A' site of the ribosome and trigger hydrolysis of the peptidyl-tRNA occupying the 'P' site resulting in the release of newly synthesized protein.

Stop codons are also called non-sense codons. No tRNA can bind to these codons. UAG is called amber codon, UAA is called ochre codon and UGA is called opal codon.

7. Genetic code is universal:

A particular codon codes for the same amino acid in all organisms from prokaryotes to plants and animals including viruses.

The universality of genetic code provides strong evidence that life on the earth started from a common ancestor. When living forms appeared on the earth, the genetic code was established. It has not changed since then, throughout the evolution of living forms and has been preserved throughout the biological evolution.

8. Co-linearity:

The sequence of codons on mRNA and the sequence of corresponding amino acids in polypeptide chain are co-linear.

**9. Translation of mRNA occurs in $5' \rightarrow 3'$ direction.
Codons on mRNA and anticodons on tRNA are written as follows:**

	1	2	3	
5'	A	C	G	3' codon on mRNA
3'	U	G	C	5' Anticodon on tRNA
	3	2	1	

First base of the codon pairs with the third base of the anticodon. Codons are written in $5' 3'$ but anticodon sequence is written with a backward arrow.

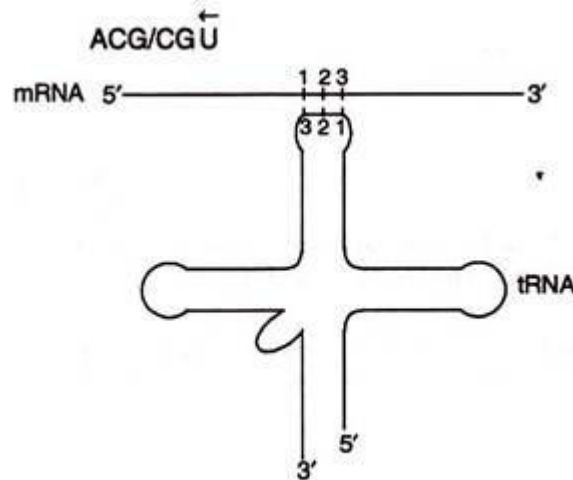


Fig. 11.2. Codon-anticodon base pairing