LESSON 4

RNA. Genetic code. Transcription. Translation.

DNA is essentially a storage molecule. It contains all of the instructions a cell needs to sustain itself. These instructions are found within ***genes***.

Gene *is defined as a segment of DNA that is used to make a functional product, either an RNA molecule or a polypeptide.*

In order to be implemented, the instructions contained within genes must be expressed, or copied into a form that can be used by cells to produce the proteins needed to support life. The instructions stored within DNA are read and processed by a cell in two steps: transcription and translation.

Structural genes encode the amino acid sequence of a polypeptide. When a structural gene is transcribed, the first product is an RNA molecule known as ***messenger RNA*** **(m-RNA).** During polypeptide synthesis — a process called ***translation*** — the sequence of nucleotides within the m-RNA determines the sequence of amino acids in a polypeptide. One or more polypeptides then assemble into a functional protein. The synthesis of functional proteins ultimately determines an organism's traits. This model is called the ***central dogma of genetics*** (the central dogma of molecular biology). The flow of genetic information occurs from DNA to m-RNA to polypeptide (fig.1).

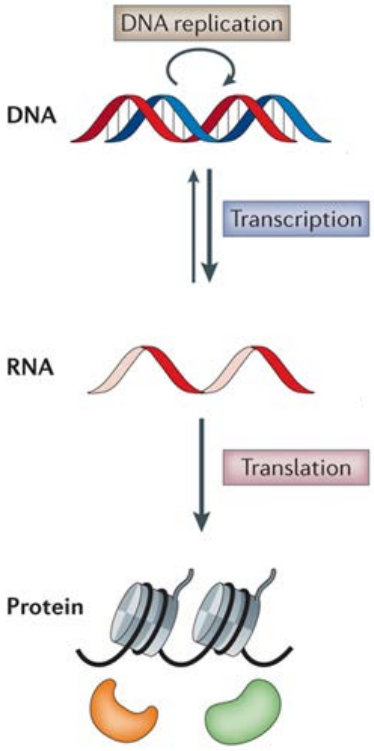

Fig.1. The central dogma of genetics[1]

During transcription, a portion of the cell's DNA serves as a template for creation of an **RNA** molecule.

RNA

Ribonucleic acid or RNA is a single-stranded nucleic acid. It is a polymer with ***ribonucleotides*** or as monomers.

The polynucleotide chain is formed by joining of ribonucleotides, with the help of 3' – 5' *phosphodiester bonds* in the same fashion as in case of DNA. But RNA is more stable than DNA.

RNA is made up of three primary components:

- ✓ a nitrogen-containing (nitrogenous) base;
- ✓ a carbon-based (pentose) sugar molecule called ribose;
- ✓ a phosphate group attached to the sugar molecule.

The difference between DNA and RNA chemical composition is RNA has ***ribose sugar***, has hydroxyl group (- OH) at the 2' position and N-base thymine is replaced by ***uracil***.

Nitrogen bases are of two types: *purines*: ***Adenine*** (**A**) and ***Guanine*** (**G**); *pyrimidines*: ***Cytosine*** (**C**) and ***Uracil*** (**U**).

Ribonucleic acid (RNA) can be classified into three types:

- ✓ m-RNA – messenger RNA
- ✓ rRNA – ribosomal RNA
- ✓ t-RNA – transfer RNA

Messenger RNA (m-RNA)

Messenger RNA (m-RNA) is a large family of RNA molecules that convey genetic information from DNA to the ribosome, where they specify the amino acid sequence of the protein products of gene expression. Following transcription of primary transcript m-RNA (known as ***pre-m-RNA***) by RNA polymerase, processed, ***mature*** **m-RNA** is translated into

[1] Nature Reviews Genetics **15**,293–306(2014)

a polymer of amino acids. Pre-m-RNA is composed of ***exons*** and ***introns,*** whereas and mature m-RNA contains only exons. Mature RNA produced during maturation (processing), which is followed by excision of introns from exons crosslinking. 5% of total cellular RNA is m-RNA.

Ribosomal RNA (r-RNA)

In combination with r-RNA ribosomal proteins forms the ***ribosome***. In eukaryotic cells the synthesis of r-RNA occurs in the nucleolus. 80 – 85% of total cellular RNA is r-RNA.

t-RNA is a family of nearly 60 small sized ribonucleic acids. 10 – 15% of total cellular RNA is t-RNA. t-RNAs are small molecules with about 74 – 95 ribonucleotides. t-RNAs are made up of a single stranded polynucleotide chain. A cell makes many different t-RNAs, all t-RNAs share common structural features. As originally proposed by Robert W. Holley in 1965, the secondary structure of t-RNAs exhibits a ***cloverleaf model***.

Clover Leaf Model

According to this model single polynucleotide chain is folded upon itself to form 4 or 5 arms, because of this folding 3’ and 5’ ends of this t-RNA polynucleotide chain come near to each other. Arm consists of a stem and a loop. Arms in t-RNA are:

- ✓ Acceptor arm
- ✓ DHU or D arm
- ✓ Anticodon arm
- ✓ TψC arm
- ✓ Variable arm

In the stem, complementary bases are joined together by hydrogen bonds. This maintains the structure of t-RNA. There is no base pairing in the loops.

Acceptor arm is formed by complementary base pairing between nucleotides from the 5’ and 3’ ends of the t-RNA (7 base pairs are present). It provides the point of attachment for amino acids.

DHU or D arm consists of stem and loop with unusual pyrimidine nucleotide dihydrouracil. There are 15 – 18 nucleotides in this loop, with 3 – 4 base pairs in the stem and 7 – 11 unpaired nucleotides in the loop. The loop of this arm is called as DHU loop or Loop I or D loop.***Anticodon arm*** contains stem and loop. Stem consists of 5 base pairs and loop (called as anticodon loop or loop II) contains 7 unpaired nucleotides. Out of these 7 unpaired nucleotides the middle three form anticodon. ***Anticodon*** recognizes and ***codon*** of m-RNA and binds to it.

TψC arm is named for the presence of sequence TψC (thymine – pseudouridine (ψ) – cytosine), where pseudouridine is unusual base. This arm also consists of stem and loop. This loop contains a ribosome recognition site. Clover Leaf Model of t-RNA is shown at picture below (fig. 2).

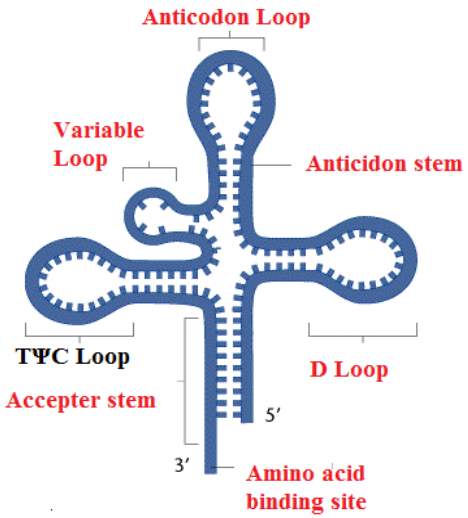

Fig. 2. Clover Leaf Model of t-RNA[2]

To function correctly, each type of t-RNA must have the appropriate amino acid attached to its 3′ end. The t-RNA with its attached amino acid is called a ***charged t-RNA*** or an ***aminoacyl-t-RNA.***

Transcription

Transcription is the process of synthesis of molecule of m-RNA from a DNA template. In the transcription process, only one strand is actively used as a template it is

[2] Klug & Cummings 1997

known as the sense strand ***or template strand***; the another complementary DNA strand which is not used, is called the ***nonsense*** or ***antisense strand.***

An enzyme known as ***RNA polymerase*** catalyzes transcription. In Eukaryotes, there are three different RNA polymerase, designated as RNA polymerase I, II, and III. Each of the three RNA polymerases transcribes different categories of genes:

RNA polymerase I transcribes all of the genes that encode ribosomal RNA;

RNA polymerase II plays a major role in cellular transcription because it is responsible for the synthesis of all m-RNA;

RNA polymerase III transcribes all t-RNA genes.

In Prokaryotes, a single RNA polymerase makes all types of RNA.

RNA polymerases are large enzymes that work together with a number of other specialized cell proteins. These cell proteins, called ***transcription factors***, help determine which DNA sequences should be transcribed and precisely when the transcription process should occur.

Transcription occurs in three stages: ***initiation; elongation,*** or synthesis of the RNA transcript; and ***termination***.

Transcription does not need a primer to start; it requires a promoter like sequence.

Initiation

Transcription begins at a specific site called as ***promoter*** (In Eukaryotes, a ***TATA box***, a nucleotide sequence like TATAAAA…, is typically present in the promoter.) The TATA box aids in the recognition of the promoter. Promoter itself is not transcribed. First transcription factor locate and attach to the TATA box. Then RNA ***polymerase II*** binds to the promoter. This complete assembly of transcription factors and RNA polymerase along with promoter is called as ***transcription initiation complex***. RNA polymerase II begins to unwind the DNA helix and transcribing the DNA sequence. Once the complex has been opened, Initiation starts and the first *phosphodiester bond* is formed. This is the end of step initiation. In Eukaryotes transcription may start at a variety of different locations.

Promoter clearance (***Abortive Initiation***) occurs between initiation and elongation. After synthesize of first bond, RNA polymerase must clear the promoter. During this time

RNA transcript is released and truncated transcripts produced. This is called abortive initiation. When the transcript reaches length of approximately 23 nucleotides it no longer slips and elongation can occur.

Elongation

RNA polymerase II begins moving down the DNA template strand in the 3' to 5' direction, and as it does so, it strings together complementary nucleotides.

Transcription can involve multiple RNA polymerases on a single DNA template and multiple rounds of transcription, so a single gene may be transcribed thousands of times.

Termination

RNA chain is terminated when the enzyme encounters a ***terminator*** or a sequence of nucleotides that signals the end of transcription. The enzyme RNA polymerase transcribe the terminator sequence and then continues for about 10-15 nucleotides before the pre-m-RNA strand is released. Then the pre-m-RNA is released and DNA helix reforms.

In Eukaryotes, structural genes include regulatory elements. There are two categories of regulatory elements: activating sequences, known as ***enhancers***, are needed to stimulate transcription and sequences, known as ***silencers*** that inhibit transcription. Before the pre-m-RNA strand or m-RNA leaves the nucleus, it must undergo RNA processing.

RNA processing

In Eukaryotes, the transcription of structural genes produces a long transcript known as ***pre-m-RNA*** (***precursor m-RNA***) which contains coding sequences – ***exons***, and sequences that are not translated into protein – ***introns***. Subsequently, as pre-m-RNA matures, the introns are removed and the exons are connected, or spliced, together. The pre-m-RNA is modified into m-RNA. This process is called ***RNA splicing***.

It may be an advantage for a species to have genes that contain introns. One benefit is a phenomenon called ***alternative splicing.*** When a pre-m-RNA has multiple introns, variation may occur in the pattern of splicing, so the resulting m-RNAs contain alternative combinations of exons. The biological advantage of alternative splicing is that two or more different proteins can be derived from a single gene. This allows an organism to carry fewer genes in its genome.

In Eukaryotes, in addition to splicing, pre-m-RNAs both ends of m-RNA are protected by two different ways:

5' end – the 5'end is capped with a *methylated guanosine* during the early stages of transcription. The process is known as ***capping***. This cap protects the m-RNA from degradation in cytoplasm and helps in recognition by the ribosome.

3' end - a sequence of adenine nucleotides called a ***poly-A tail*** is added to the 3' end of the m-RNA molecule. The process is called as ***polyadenylation***. This sequence signals to the cell that the m-RNA molecule is ready to leave the nucleus and enter the cytoplasm. The steps of transcription are shown in picture below (fig.3).

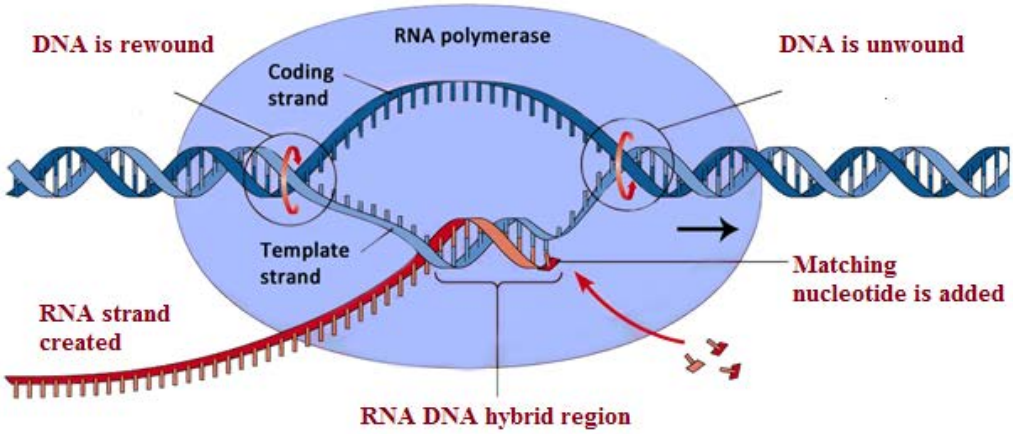

Fig.3. Transcription[3]

Once an m-RNA molecule is complete, that molecule can go on to play a key role in the process known as ***translation***. During translation, the information that is contained within the mRNA is used to direct the creation of a protein molecule. In order for this to occur, however, the m-RNA itself must be read by a special, protein-synthesizing structure within the cell known as a ***ribosome***.

Ribosomes are a complex of proteins and specialized RNA molecules called ribosomal RNA (r-RNA). Ribosomes are composed of two subunits, one large (**LSU**) and one small (**SSU**). They only bind together during protein synthesis. They have three binding sites (fig.4):

- ✓ ***A site*** (***Aminoacyl site***) - the binding site for t-RNA, which holds the ***aminoacyl-t-RNA complex;***
- ✓ ***P site*** (***Peptidyl site***), which binds to the t-RNA attached to the growing polypeptide chain;

[3] Copyright © Pearson Education, Inc., 2006

✓ ***E*** (***exit***) site, which are occupied by tRNA molecules.

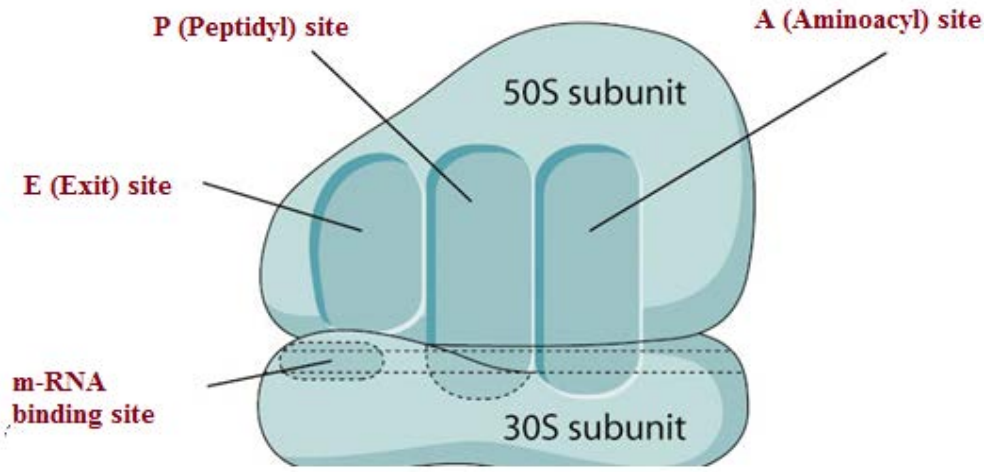

Fig.4. Ribosom binding sites[4]

Genetic code

The ***genetic code*** is the set of "rules" that helps to translate genetic information from a nucleotide chain to a sequence of amino acids.

The main features of genetic code:

1. The code consists of *triplet*s known as ***codons***[5]. The three-letter nature of codons means that the four nucleotides found in mRNA — A, U, G, and C — can produce a total of 64 different combinations.
2. The code is ***degenerate***, because there are only 20 different amino acids but 64 possible codons thus in many cases, more than one codon specifies the same amino acid (fig.5). It minimizes the harmful effects that incorrectly placed nucleotides can have on protein synthesis.
3. Despite its degeneracy, the code is ***unambiguous***, because each codon specifies only one amino acid.

---

[4] Copyright © Pearson Education, Inc., 2010

[5] The idea of codons was first proposed by Francis Crick and his colleagues in 1961. During that same year, Marshall Nirenberg and Heinrich Matthaei began deciphering the genetic code, and they determined that the codon UUU specifically represented the amino acid phenylalanine. Following this discovery, Nirenberg, Philip Leder, and Har Gobind Khorana eventually identified the rest of the genetic code and fully described which codons corresponded to which amino acids.

4. The codons are ***nonoverlapping*** (for example, in the mRNA sequence 5 GAAGUUGAA 3 the first three nucleotides (GAA) form one codon; nucleotides 4 through 6 (GUU) form the second; and nucleotides 7 through 9 (GAA), the third. Each nucleotide is part of only one codon).
5. The code includes three ***stop codons*** (***nonsense codon)***: UAA, UAG, and UGA. These codons do not encode an amino acid and thus terminate translation.

The genetic code is largely ***universal*** but some exceptions are known to occur. For example, *Selenocysteine* (Sec) and *pyrrolysine* (Pyl) are sometimes called the 21st and 22nd amino acids in polypeptides. Selenocysteine is found in several enzymes involved in oxidation-reduction reactions in bacteria, archaea, and eukaryotes. Pyrrolysine is found in a few enzymes of methane-producing archaea. Selenocysteine and pyrrolysine are encoded by UGA and UAG codons, respectively, which normally function as stop codons.

1 st position	2nd position				3rd position
	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

Fig. 5. Genetic code[6]

Reading the genetic code

[6] Ala: Alanine; Arg: Arginine; Asn: Asparagine; Asp: Aspartic acid; Cys: Cysteine; Gln: Glutamine; Glu: Glutamic acid; Gly: Glycine; His: Histidine; Ile: Isoleucine; Leu: Leucine; Lys: Lysine; Met: Methionine; Phe: Phenylalanine; Pro: Proline; Ser: Serine; Thr: Threonine; Trp: Tryptophan; Tyr: Tyrosine; Val: Valine

Redundancy in the genetic code means that most amino acids are specified by more than one mRNA codon. For example, the amino acid phenylalanine (Phe) is specified by the codons UUU and UUC, and the amino acid leucine (Leu) is specified by the codons CUU, CUC, CUA, and CUG. Methionine is specified by the codon AUG, which is also known as the start codon. Consequently, methionine is the first amino acid to dock in the ribosome during the synthesis of proteins. Tryptophan is unique because it is the only amino acid specified by a single codon. The remaining 19 amino acids are specified by between two and six codons each. The codons **UAA, UAG**, and **UGA** are the stop codons that signal the termination of translation.

Translation

Translation can be broken into three distinct phases: ***initiation***, ***elongation***, and ***termination***. All three phases of translation involve the ribosome, which directs the translation process. Multiple ribosomes can translate a single mRNA molecule at the same time, but all of these ribosomes must begin at the first codon and move along the mRNA strand one codon at a time until reaching the stop codon. This group of ribosomes, also known as a ***polysome***, allows for the simultaneous production of multiple strings of amino acids, called ***polypeptides***, from one mRNA.

Initiation

At the start of the initiation phase of translation small subunit of ribosome attaches to the mRNA strand and finds the beginning of the genetic message, called the ***start codon***. This codon is almost always **AUG**, which corresponds to the amino acid *methionine*. But an RNA contains more than one AUG. Prokaryotes use a recognition sequence that is near the AUG initiation codon - ***the Shine-Dalgarno sequence***. Usually this sequence is **AGGAGG**. It indicates that the initiation triplet 5'-AUG-3' is a few nucleotides downstream. A charged tRNA-met (with a 3'-UAC-5' anticodon) binds to the initiation triplet in the small subunit (fig.6).

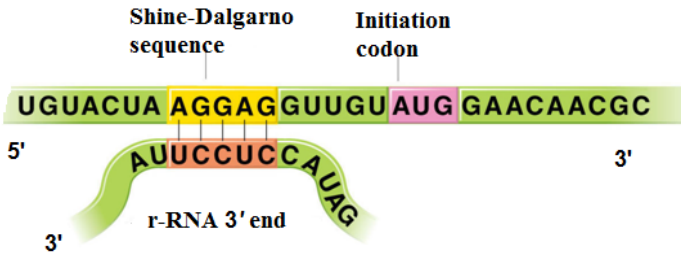

Fig.16. Shine-Dalgarno sequence[7]

[7] © Parson Education, Inc. 2010

In eukaryotes, the ribosome seems to scan from the 5' cap, looking for the first AUG follows ***Kozak's rules***. In many, but not all, cases, the ribosome uses the first AUG codon that it encounters as a start codon. The consensus sequence for initiation in vertebrates (also called ***Kozak box***) is: **ACCATGG,** more general it is: **(GCC)RCCATGG** where R is a purine (A or G).

Then the large subunit joins to form the ribosomal monosome. At this point, the initiation phase of translation is complete.

Elongation

During the elongation stage of translation, amino acids are added, one at a time, to the polypeptide chain[8].

To begin elongation, a charged t-RNA brings a new amino acid to the ribosome so it can be attached to the end of the growing polypeptide chain. A charged t-RNA carrying a single amino acid binds to the A site. This binding occurs because the ***anticodon*** in the t-RNA is complementary to the codon in the m-RNA. The next step of elongation is the ***peptidyl transfer reaction*** — the polypeptide is removed from the – t-RNA in the P site and transferred to the amino acid at the A site. This transfer is accompanied by the formation of a ***peptide bond*** between the amino acid at the A site and the polypeptide chain, lengthening the chain by one amino acid. The peptidyl transfer reaction is catalyzed by a ***peptidyl transferase***, which is composed of several proteins and r-RNA. The discharged t-RNAs are released from the E site. The ribosome shifts to the next codon. This process is repeated along the entire length of the mRNA, thereby elongating the polypeptide chain that is emerging from the top of the ribosome (fig.17).

[8] Even though this process is rather complex, it occurs at a remarkable rate. Under normal cellular conditions, a polypeptide chain can elongate at a rate of 15 to 20 amino acids per second in bacteria and 2 to 6 amino acids per second in eukaryotes.

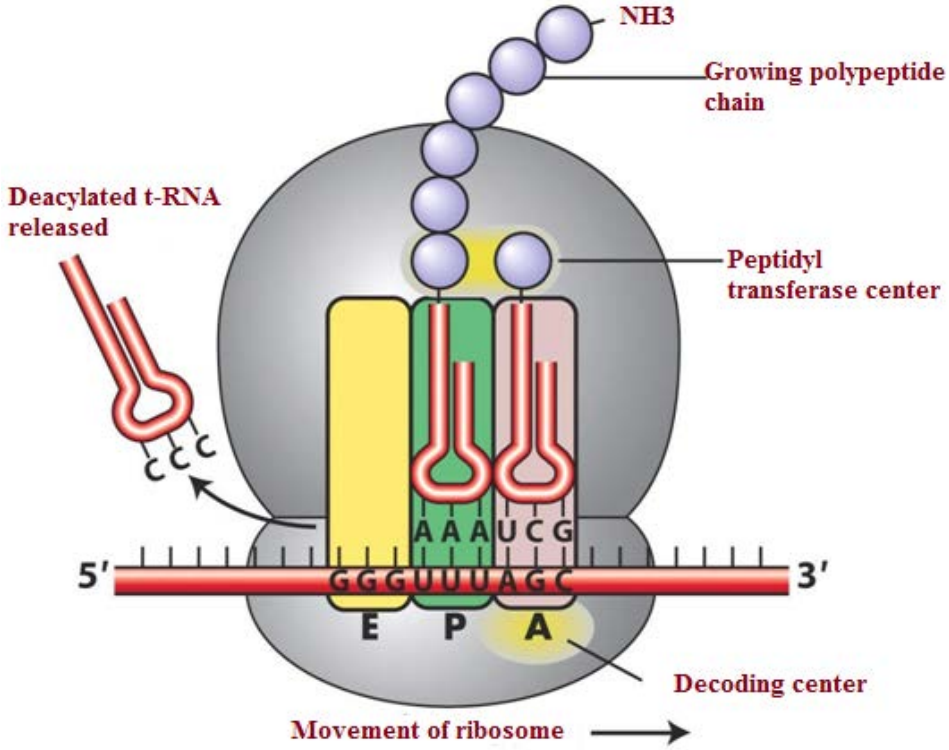

Fig.17. Schematic model of Elongation

Termination

Eventually, after elongation has proceeded for some time, the ribosome comes to a ***stop codon***, which signals the end of the genetic message. As a result, the ribosome detaches from the mRNA and releases the amino acid chain. This marks the final phase of translation, which is called termination

Post-Translational Modification

For many proteins, translation is only the first step in their life cycle. Moderate to extensive post-translational modification is sometimes required before a protein is complete. The peptide chain must be converted from a long chain of amino acids into a three dimensional structure. The protein does this action by making numerous connections between different amino acids. A special group of proteins, called ***chaperones***, assist with this process. This family of proteins can bind to proteins that are not folded correctly. They then provide a thermodynamically favorable environment for the protein to reach its correctly folded state. These modifications can alter the three dimensional structure of a protein, regulate its ability to bind to different proteins, and alter its activity.

Regulation of gene expression

Not all nucleotide sequences in a strand of DNA code for the production of proteins. Rather, some of these noncoding sequences serve as binding sites for the various protein molecules required to start or regulate the transcription process. For example, a group of nucleotides known as a ***promoter sequence*** lies near the beginning of most genes and provides a binding site for RNA polymerase to begin transcription. Similarly, other noncoding sequences near the promoter sequence function as protein binding sites that can either induce or block transcription. This basic system affects gene expression in both prokaryotes and eukaryotes, albeit in different ways.

Regulation of gene expression in prokaryotes

In Prokaryotes, system that regulates the gene expression is known as ***operon***. One especially well-known operon is the ***lac operon*** found in E. coli bacteria. It produces some enzymes to break down lactose (lactose is a sugar molecule that bacteria often use as a source of energy).

The model of the *lac operon* was proposed by Jacob and Monod for which they were awarded the Nobel Prize. According to this model an operon is a cluster of bacterial genes along with an adjacent promoter that controls the transcription of those genes. It consists of three structural genes, a promoter, a terminator, a regulator, and an operator (fig.17).

The *three structural genes* are: lacZ, lacY, and lacA. They implement lactose metabolism.

The *operator* is a short region of DNA that lies partially within the promoter and that interacts with a regulatory protein that controls the transcription of the operon.

The *promoter* is a binding site for RNA polymerase to begin transcription.

The *regulatory gene* produces an m-RNA that produces a *repressor* protein, which can bind to the operator of the *lac operon*.

In the absence of lactose, the repressor binds to the operator and keeps RNA polymerase from transcribing the *lac* genes. The effect of the repressor on the *lac* genes is referred to as *negative regulation* (fig.17).

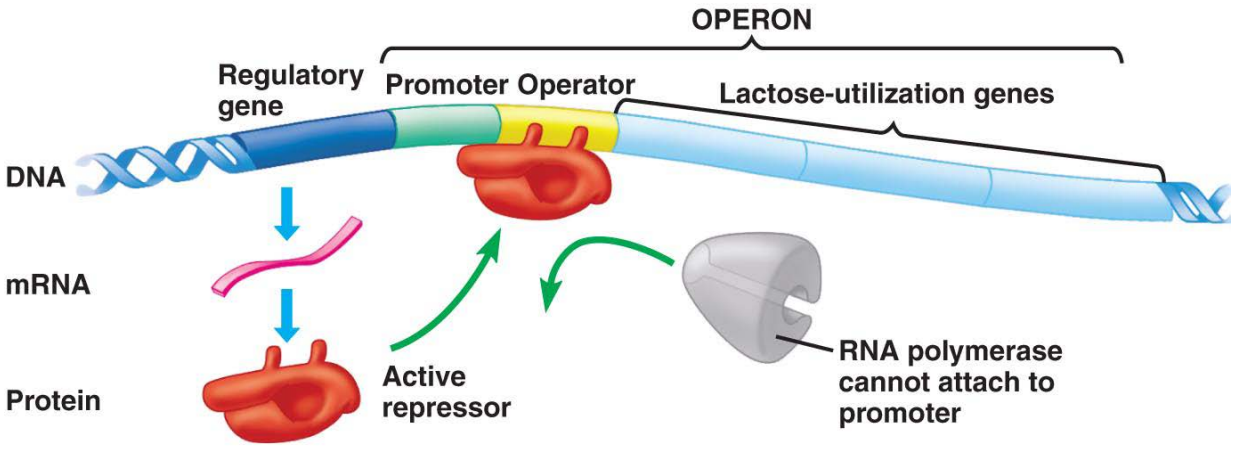

Fig.17. Negative regulation (lactose is absent)[9]

When lactose is present, the *lac* genes are expressed because lactose binds to the repressor protein and keeps it from binding to the *lac* operator. In this case lactose acts as an inducer and the effect is referred to as *positive regulation* (fig.18).

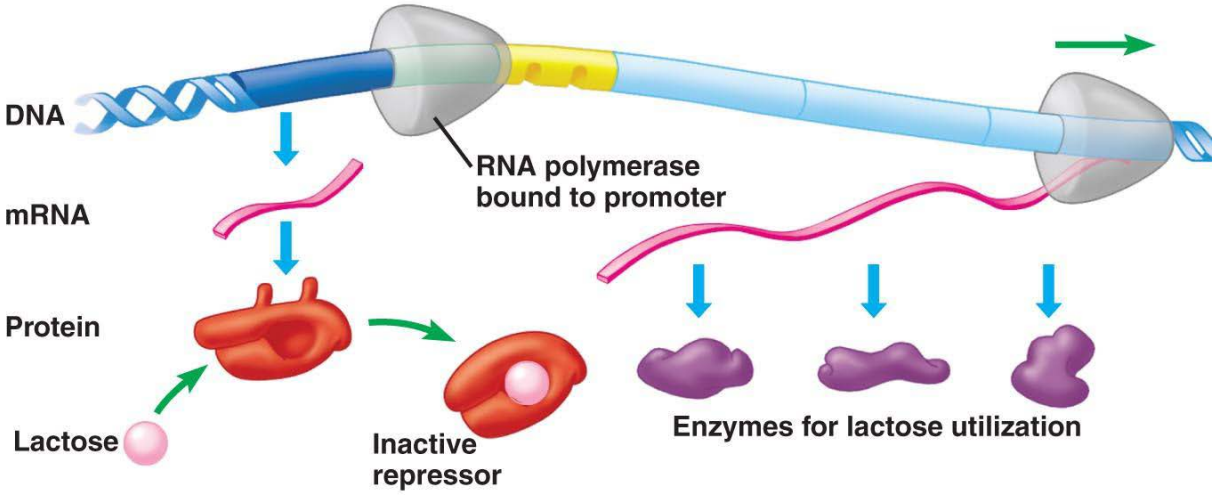

Fig.18. Positive regulation (lactose is present and it acts as an inducer)[10]

When the enzymes encoded by the *lac operon* are produced, they break down lactose and lactose, eventually releasing the repressor to stop additional synthesis of m-RNA. Messenger RNA breaks down after a relatively short amount of time.

[9] Copyright © Parson Education, Inc. 2009, publishing by Benjamin Cummings

[10] Copyright © Parson Education, Inc. 2009, publishing by Benjamin Cummings