Cell Biology and Genetics

Gene Expression and Regulation

Manoj K. Dhar

Department of Biotechnology, University of Jammu, Jammu-180006,
J & K, INDIA

manojkdhar@rediffmail.com

Contents:

Introduction

Sequencing of the entire genome of model organisms has clearly brought to light that DNA is composed of coding regions (genes) and non-coding (repetitive) regions. The proportion of these types of DNA varies in different organisms. However, in majority of the cases the repetitive DNA constitutes about 70 to 90%. Consequently, therefore, the number of genes also varies. One very important realization is that not all genes in a particular tissue are active and the genes active in one tissue may not necessarily be active in other tissues. In other words the expression of genes is tissue specific. This tissue specific expression results in the production of different types of proteins and therefore, a distinct phenotype. As pointed out earlier, while genes are the coded messages, these are transcribed into messenger RNAs, which are in turn translated into proteins, which are responsible for a particular phenotype.

The nucleic acid sequences can be considered as coded messages, written using only four letters or bases – Adenine (A), Thymine or Uracil (T or U), Cytosine (C) and Guanine (G). The genetic code has since been proved to be triplet, i.e. three nucleotides constitute a code for an amino acid. Therefore, if a stretch of DNA is read from one end to other, we are likely to decode this information into a series of amino acids called proteins. However, one must understand every stretch of DNA cannot be decoded in this fashion; one which can be is known as a gene. The actual number of genes is very less as compared to long stretches of non-coding DNA.

The central dogma of Molecular biology (Fig. 1) is that the information usually flows during, i) **DNA replication**, from DNA to DNA ultimately resulting in its transfer from one generation to another and, ii) **Gene Expression**, from DNA to proteins via RNA ultimately determining the phenotypic expression. Gene expression can be further divided into two steps: i) **transcription** involves transfer of information from DNA to RNA and, ii) **translation** involving transfer of information from RNA to protein. The exception to the central dogma is the transfer of information from RNA to DNA as in the case of reverse transcription, also called as Teminism.

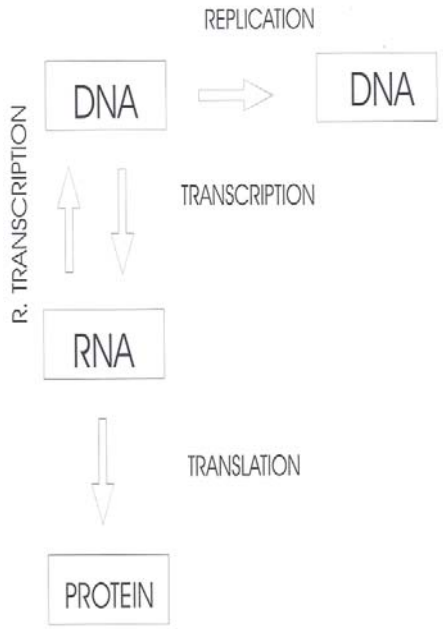

Fig. 1: CENTRAL DOGMA OF MOLECULAR BIOLOGY

Transcription

In order to fully appreciate the process of transcription, we must understand the various components involved in this process. One strand of DNA is used as a template to synthesize the RNA strand. The complementary base pairs are A-U and G-C such that if A is present in DNA, the RNA strand will have U at that position. Thus RNA strand is the complementary sequence of the DNA strand (template strand) and is in a way like the other strand of DNA which was not used for transcription (coding strand) except that DNA has T while RNA has U (Fig. 2). The RNA synthesized in this fashion is called messenger RNA (mRNA) because it is this RNA, which gets translated to form the protein. There are three more types of RNA molecules, which play active role in gene expression. Transfer RNAs (tRNAs) are small, about 75 nucleotides long and act as adaptors between amino acids and codons in the mRNA, during the translation. Ribosomal RNA (rRNAs) are the important structural component of the ribosome, where the mRNA is translated into series of amino acids. The third type of molecule is small nuclear RNA (snRNA), which is the integral part of spliceosome which excises introns from genes.

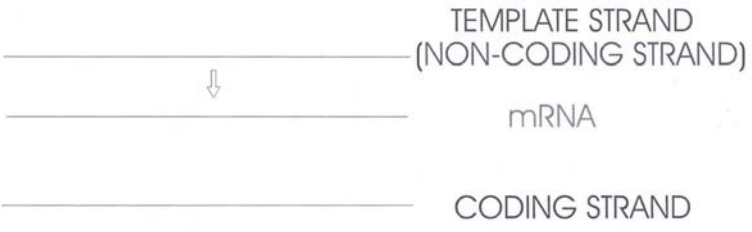

Fig. 2: Coding and non-coding strands

Transcription takes place from 5' – 3' direction with addition of ribonucleotides to the free 3' – hydroxyl group. The chemical reaction whereby nucelophilic attack by 3' –OH on phosphorus of ribonucleotide releases pyrophospahate is catalyzed by enzyme RNA polymerase.

RNA polymerase

RNA polymerase is the universal enzyme, which catalyzes synthesis of RNA on DNA template. The enzyme isolated from prokaryotes and eukaryotes is similar in many respects especially the structure of the catalytic part. The enzyme is generally made up of several subunits. While bacteria have only one type of RNA polymerase, eukaryotes have three types; Pol I, Pol II and Pol III. Out of these, Pol II is responsible for transcription of most of the protein coding genes. Pol I transcribes the large rRNA precursor gene and Pol III transcribes tRNA and other RNA genes.

The prokaryotic polymerase consists of five subunits namely two α, and one each of β, β' and ω subunits. The enzyme has a cleft in the centre, acting as the active site.

Promoter

For transcription to begin, the first step is binding of RNA polymerase to the DNA template. The initial binding site at 5' site, upstream of the gene, is called promoter. The promoter is a DNA sequence that regulates efficiency of initiation of transcription. Sequencing of promoters of several genes in bacteria has led to identification of two consensus sequences. Consensus sequences are defined as similar sequences found in different genes of the same organism or genes of related organisms. The consensus sequences in bacterial promoters are: TATAAT and TTGACA. The former sequence is located 10 nucleotides before (upstream) the start of transcription initiation site. This sequence is called the – 10 (minus 10) region or Pribnow box. The latter sequence is located 35 nucleotides upstream and is called the – 35 region (Fig. 3) Those promoters whose sequences match the consensus sequences are stronger than the ones with less matching sequences. Strength of the promoter gets reflected in the number of RNA transcripts formed in given time. It can therefore, be inferred that genes showing higher expression will have strong promoters and those showing low expression will have weak promoter.

RNA polymerase can initiate transcription at any point. However, in order to initiate the transcription at promoter site only, an initiation factor called sigma (σ) binds with the RNA polymerase (core enzyme), which is now called **holoenzyme**. The σ factors are of various types; the one predominant in prokaryotes is a 70 kilodalton protein called σ^{70}.

Initiation

The transcription initiates by binding of the holoenzyme with the promoter sequences while DNA is still in double helical form. This structure is called a closed complex. In the next step the DNA helix opens between –11 and +3 positions, a process called 'melting'. One of the strands functions as the template and the other as non-template. The transition of the enzyme from the closed complex to open complex where it is intimately associated with template strand is called isomerization (Fig. 4). RNA polymerase does not require any primer for initiation of synthesis. RNA polymerase brings together ribonucleides as per the template sequence and joins them. Subsequently the enzyme moves further and opens up the DNA strand ahead. However, several initial attempts to synthesize RNA abort. Several short (about 10 bases) RNA molecules are released. Once the RNA polymerase overcomes these initial hiccups and goes beyond ten bases, it forms a stable complex involving itself with DNA and RNA. This complex is called ternary complex.

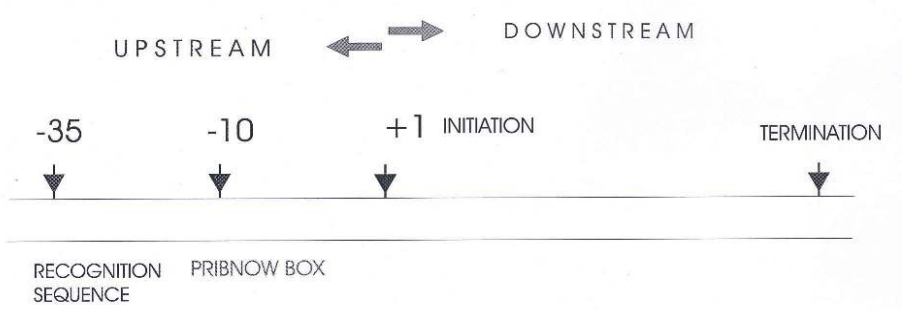

Fig. 3: Organization of a gene in prokaryotes

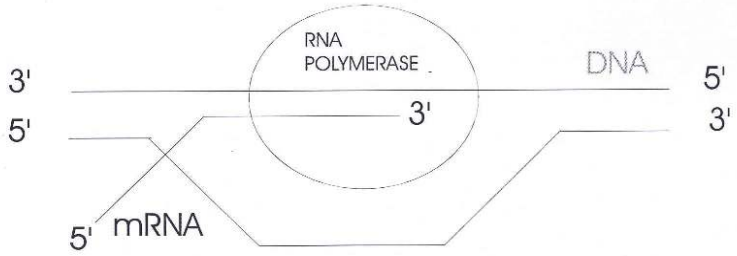

Fig. 4: Initiation of transcription

Elongation

Once the RNA polymerase starts polymerization, the σ subunit gets released from the enzyme. RNA polymerase can initiate transcription without σ, however, it will bind at random positions in the DNA sequence and will not remain restricted to the promoter. The elongation of chain takes place by addition of ribonucleotides by the polymerase at average rate of 50 nucleotides per second at 37º C.

Termination

As the polymerase reaches a specific nucleotide in the template DNA, the synthesis continues. Once the RNA synthesis has passed beyond the initial ten bases, it has greater chances of completion. The synthesis stops as the enzyme encounters the termination signal. At this stage the ternary complex dissociates with the release of RNA molecule. The termination is of two types; rho-independent and rho-dependent. In the first type termination takes place without the involvement of other factors and in the second type termination requires an additional protein rho. The rho independent terminators consist of two inverted repeats followed by several AT base pairs, such that when polymerase transcribes this

stretch it results into RNA sequence, which can base pair with itself. This stem loop structure is called hairpin and is believed to terminate transcription by disrupting the elongation complex.

On the other hand rho dependent terminators need the activity of a ring shaped enzyme called rho. The rho consists of six identical subunits and uses energy derived from ATP hydrolysis to release freshly synthesized RNA molecule. The net result of transcription is an RNA molecule whose sequence is complementary to the DNA template.

Transcription in Eukaryotes

Transcription in eukaryotes is much different than in prokaryotes. One of the major differences is that RNA polymerases exist in three forms and each of these transcribes different genes. Pol I transcribes ribosomal RNA genes and is located in the nucleolus. Pol III is responsible for transcription of 5S ribosomal RNA genes and tRNA genes. It is located in nucleoplasm. Pol II is the main enzyme, which transcribes all messenger RNAs in eukaryotes.

In eukaryotes the RNA polymerase does not interact directly with the promoter. The promoter is generally 40 nucleotides long and may extend from upstream to downstream of transcription initiation site. There are four elements in the promoter which are important for transcription. These are; BRE (TFIIB recognition element), TATA box, Initiator (Inr) and DPE (Downstream promoter element). First General Transcription factor (GTF) TFIID is formed between TATA-binding protein (TBP) and at least 12 TBP-associated factors or TAFs. Both these DNA binding proteins, TBP and TAFs are important for recognition of promoter. In the next step, TFIIA joins the complex, followed by TFIIB. Another transcription factor TFIIF first associates with RNA polymerase II and then these two together join the transcription initiation complex. TFIIF consists of two subunits, one of which is responsible for unwinding of DNA for initiating transcription. This is followed by joining of TFIIE to the initiation complex, downstream of transcription start point. Position of other two factors TFIIH and TFIIJ is not known.

Promoters of eukaryotic genes have some sequences, which are most important for transcription. One of such sequences is called TATA or Goldberg-Hogness Box. The consensus sequence is composed of ~ 7 nucleotides, TATAAAA located at – 30 position. Its function seems to be similar to that of Pribnow box in prokaryotes. Another sequence which has also been observed in promoters is called CAAT box. The consensus sequence, GGCCAATCT is located at – 80 position.

Going by the complexity of eukaryotic system, many other sequences have been observed to influence the efficiency of promoters in eukaryotes. One such type is Enhancers. These are *cis*-elements (*cis* means adjacent or on the same strand) and can be located upstream or downstream.

Genetic code

How the four nucleotides, Adenine, Guanine, Cytosine and thymine/uracil encode different amino acids which constitute a protein was a matter of enquiry. The problem was however, solved during 1960s. The main features of code are discussed briefly as follows and are given in Fig. 5.

i) Code is triplet:

The total number of amino acids for which codes are needed is twenty. Therefore, if one base is used as a code, only four codes are possible. With two base code the number of codons would be sixteen, which is again insufficient. If three nucleotides per codon are considered, the possible codons would be 64, more than what is required.

THE GENETIC CODE

First (5') Letter	Second Letter: U	Second Letter: C	Second Letter: A	Second Letter: G	Third (3') Letter
U	UUU } Phe UUC UUA } Leu UUG	UCU } Ser UCC UCA } Ser UCG	UAU } Tyr UAC UAA[1] UAG[2]	UGU } Cys UGC UGA[3] Trp UGG	U C A G
C	CUU } Leu CUC CUA } Leu CUG	CCU } Pro CCC CCA } Pro CCG	CAU } His CAC CAA } Gln CAG	CGU } Arg CGC CGA } Arg CGG	U C A G
A	AUU } Ileu AUC AUA AUG[4] Met	ACU } Thr ACC ACA } Thr ACG	AAU } Asn AAC AAA } Lys AAG	AGU } Ser AGC AGA } Arg AGG	U C A G
G	GUU } Val GUC } Val GUA } Val GUG	GCU } Ala GCC GCA } Ala GCG	GAU } Asp GAC GAA } Glu GAG	GGU } Gly GGC GGA } Gly GGG	U C A G

[1]UAA- Ochre (terminator), [2]UAG- Amber (terminator)
[3]UGA- Opal (terminator), [4]AUG- Initiation codom

Fig. 5

In 1961 Francis Crick and his co-workers were the first to report evidence in support of triplet code. They used a mutagenic agent, Proflavin which causes single base-pair additions or deletions, in inducing mutations in rII locus of bacteriophage T4. The mutants were unable to grow in the cells of E. coli strain K12. However, surprisingly, they were able to isolate revertants in proflavin induced mutants. They reasoned that insertion or deletion of one nucleotide will result in shifting of the reading frame i. e. the sequence of bases beyond the mutant site will get changed. Such mutations are called frameshift mutations. The altered sequence will lead to altered protein. Further experimentation by these scientists using additional insertion or deletion of one or two nucleotides revealed that original frame was restored when two additional insertions or deletions were introduced, strongly supporting the triplet nature of the code. Further support came from the experiments on *in vitro* translation. a) Using specific trinucleotides, specific tRNA charged with particular amino acid would bind to the ribosome. b) Using chemically synthesized mRNA containing repeating dinucleotide sequence resulted in synthesis of protein with alternating amino acids. The final confirmation came from comparison of DNA and amino acid sequences of polypeptide products.

Several scientists namely, M. Nirenberg, S. Ochoa, H. G. Khorana, P. Leder and their co-workers determined the meaning of all the 64 triplet codons by combining results of the experiments described above. Three scientists, Nirenberg and Khorana were awarded 1968 Nobel prize in Physiology and Medicine for this work.

ii) Commaless code

The code does not have internal punctuation i. e. it is commaless. However, the initiation codon starts the translation and termination codon signals the completion of translation. The codon AUG is used for initiation in prokaryotes and eukaryotes. This codon, in prokaryotes, is recognized by an initiator $tRNA_f^{Met}$ (transfer RNA charged with methionine whose one end is formylated to prevent extensions on 5' side). Eukaryotes have $tRNA^{Met}$ instead of $_f^{Met}$. If AUG codon is present in the internal sequence, it is recognized by $tRNA^{Met}$ (not formylated). Three codons UAG, UAA and UGA sometimes called *amber*, *ochre* and *opal* terminators, specify the chain termination. These codons do not specify any amino acid, but are recognized by release factors. Release factors are the proteins that terminate the protein synthesis.

iii) Degenerate code

Since there are 20 amino acids and 64 codons, therefore, most of the amino acids are specified by more than one codon. Occurrence of more than one codon is called degeneracy of the code. While three amino acids, serine, arginine and leucine have the maximum, six different codons. Others have 2, 3 or 4 codons. Interestingly, in those cases where more than one codon specifies the amino acid, the difference is only with regard to third base or 3' base. This degeneracy will help in minimizing the effect of mutations. A change in third base to any of the four bases in GUU will still code for valine. This only proves that degeneracy is not random but ordered.

iv) Universal code

The genetic code is nearly universal. Except mitochondria, the code means the same in all species. In mitochondria, UGA specifies Tryptophan and not chain termination, AUA is the codon for methionine not isoleucine and AGA and AGG represent chain termination codon.

v) Wobble

Crick while observing the nature of degeneracy at 3' base proposed that pairing between tRNA anticodon and mRNA at third base is not as stringent. He called it a wobble and proposed that there must be at least two tRNAs for each amino acid which has since been proved true.

Polyadenylation

The process of addition of a poly A tail to the 3' end of an mRNA molecule is termed as polyadenylation (Fig. 6). This is the first step in production of mature messenger RNA for translation. In eukaryotic organisms, polyadenylation is the mechanism by which most messenger RNA molecules are terminated at their 3' ends. The polyA tail protects the mRNA molecule from exonucleases and is important for, transcription termination, export of the mRNA from the nucleus and translation.

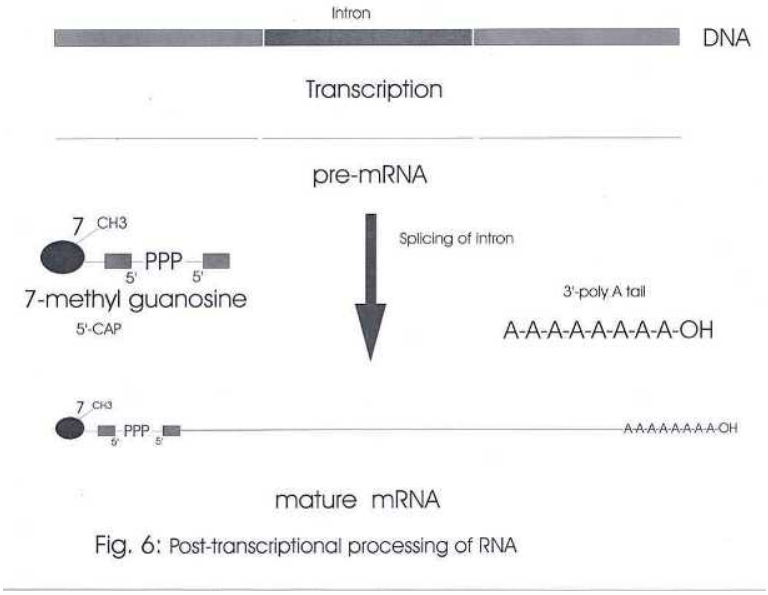

Fig. 6: Post-transcriptional processing of RNA

Polyadenylation occurs during and immediately after transcription of DNA into RNA in the nucleus. After transcription termination, the mRNA chain is cleaved through the action of an endonuclease complex associated with RNA polymerase. The cleavage site is characterized by the presence of the sequence AAUAAA. After the mRNA has been cleaved, 50 to 250 adenosine residues are added to the free 3' end at the cleavage site. This reaction is catalyzed by enzyme polyadenylate polymerase.

Capping

In eukaryotes, at the 5' end of the mRNA is a specially altered dinucleotide structure called the cap (Fig. 6). The process of capping is very important for producing mature mRNA. Capping ensures the messenger RNA's stability while it undergoes translation and is a highly regulated process which occurs in the nucleus.

The 5' cap consists of a guanosine nucleotide connected to the mRNA via an unusual 5'- 5' triphosphate linkage. Further modifications include the methylation of the guanosine, and the possible methylation of the 2' hydroxy groups of the first three ribose sugars of the 5' end of the mRNA.

In the first step, one of the terminal phosphate groups of the first nucleotide is removed by a phosphatase, leaving two terminal phosphates. Two phosphate groups from GTP are added to the terminal phosphates by a guanylyl transferase resulting in 5' to 5' triphosophate linkage. This is followed by methylation of guanosine catalyzed by methyl transferase. Other methyltransferases are optionally used to carry out methylation of 5' proximal nucleotoides.

RNA Splicing

One of the major differences in the organization of genes in pro- and eukaryotes is that the genes are split in eukaryotes. The genes are composed of coding sequences called exons and non-coding sequences called introns. For a particular gene, both introns and exons are transcribed together to form a single mRNA. However, after transcription, the introns are removed from the preliminary mRNA and the exons are joined together giving rise to mature mRNA. The process of removal of introns is called splicing. Splicing occurs by a series of biochemical reactions, which are catalyzed by proteins, RNA, or both.

RNA splicing can occur in several ways. The type of splicing depends on the structure of the spliced intron and the catalysts required for splicing to occur.

A) Spliceosomal

This mechanism of splicing takes place in those introns present within the protein-coding genes. The intron must have a 3' splice site, 5' splice site, and branch site for splicing. Splicing is catalyzed by the spliceosome, which is a large RNA-protein complex composed of five small nuclear RNAs and about 40 different proteins. The five snRNAs are rich in uridine residues and are called U1, U2, U4, U5 and U6 (Fig. 7). These snRNAs are about 100 nucleotides long and do not exist as free RNA molecules. Together with proteins, these five snRNAs constitute the four small nuclear ribonucleoproteins (snRNPs, pronounced "snurps"). While U1, U2 and U5 exist as independent snRNPs, U4 and U6 combine into one snRNP. The first step in splicing involves base pairing between 5’ end of the intron to a homologous sequence in U1 snRNA. This is followed by addition of other snRNPs forming a complex called spliceosome. Two trans-esterification take place; one involving interaction of 2’-OH group of Adenine present within intron called branch point. The 5’ end is

joined to Adenine and a loop structure is formed called the lariat. Subsequently, after second reaction, the exons are joined and the lariat containing the intron is released.

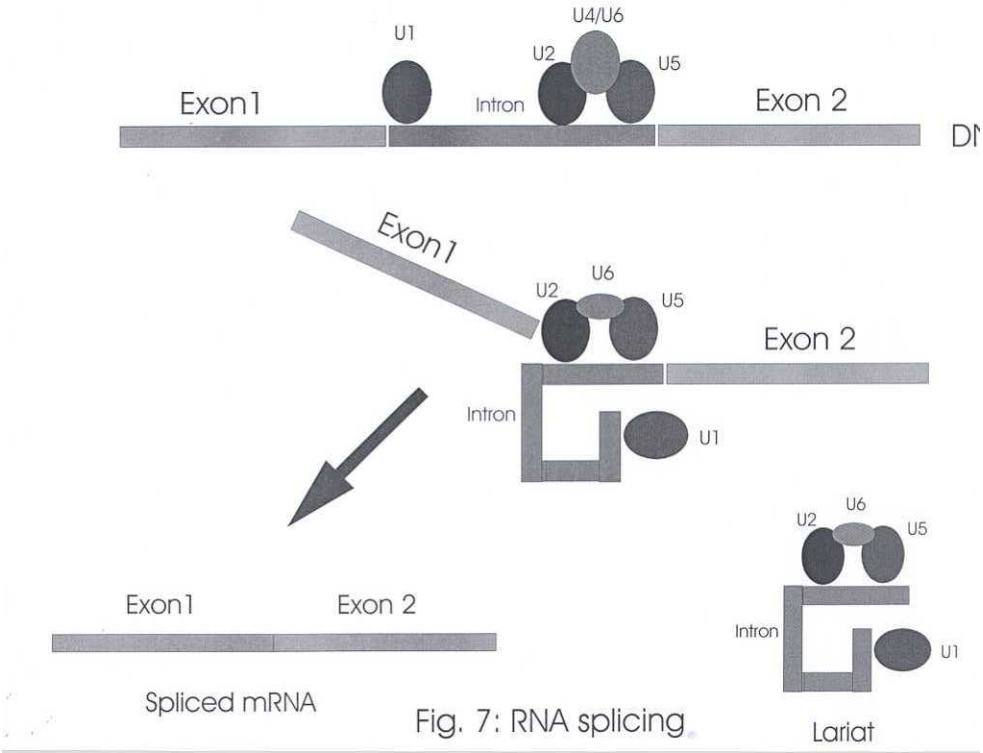

Fig. 7: RNA splicing

b) Self-splicing

Self-splicing occurs for rare introns that form a ribozyme which actually performs the functions of the spliceosome, except that it is composed of only RNA. There are three kinds of self-splicing introns, group I, II, and III. Group II and III introns perform splicing similar to the spliceosome without requiring any protein. This similarity suggests that group II and III introns may be evolutionarily related to the spliceosome. Self splicing involves two nucleophilic (also called trans-esterification) reactions. In the first reaction 3'-OH group of guanosine is transferred to nucleotide adjacent to 5' end of the intron. In the second reaction, the 3'-OH group reacts with phosphate on other intron. The intron is excised out and the two exons join.

c) tRNA splicing

This type of splicing is another rare form of splicing that usually occurs in tRNA. The splicing reaction involves a different biochemistry than the spliceomsomal and self-splicing pathways. Ribonucleases cleave the RNA and ligases join the exons together. This form of splicing does also not require any RNA components for catalysis.

Translation

The information stored in mRNA is to be translated into protein. The proteins constitute the molecules, which put into action the information encoded in the gene. While in prokaryotes the mRNA is translated while it is still being synthesized (transcribed). Therefore, transcription and translation in prokaryotes is coupled. However, in eukaryotes the two processes are not coupled. The mRNA is transcribed in the nucleus and transported to the cytoplasm where it gets attached to the ribosome and is translated into protein.

The requirements of the translation process are enormous. These include; i) ribosomes, ii) tRNA molecules, iii) enzymes involved in amino acid activation, iv) proteins involved in initiation, elongation and termination of polypeptide chain. Ribosomes constitute the major portion of the protein synthesizing machinery of the cell. By an estimate molecules involved in protein synthesis constitute about one-third of the total dry mass of most of the cells. Therefore, protein synthesis can be considered as the most vital life process.

Ribosomes are composed of protein and RNA in about equal proportion. Each ribosome has two subunits, one large and one small. The prokaryotic and eukaryotic ribosomes differ in size. The size is expressed as Svedberg units or S. The prokaryotic ribosome is 70S, small subunit 30S and large subunit 50S. The 30S subunit is composed of 21 ribosomal proteins and 16S rRNA while 50S subunit is made up of 31 ribosomal proteins, 5S rRNA and 23S rRNA. The eukaryotic ribosome is 80S; two subunits 40S and 60S. The 40S subunit is made of 33 ribosomal proteins and 18S rRNA. The 60S subunit is composed of 49 ribosomal proteins, 5S, 5.8S and 28S rRNA.

The part of the mRNA that is translated into protein is called an open reading frame or ORF. Another class of RNA, which is very important for translation, is transfer RNA (tRNA). This molecule acts as an adaptor between mRNA and amino acids. It is a small (4S) molecule 70-90 nucleotides long and has a triplet anticodon which is complementary to mRNA sequences (8a). During translation, tRNA picks the amino acid on one hand and also base pairs with the complementary sequence in mRNA thereby playing a crucial role. There is at least one aminoacyl-tRNA synthetase and tRNA for each of the 20 amino acids.

The attachment of amino acids to tRNA, the process called activation, is catalyzed by the enzyme aminoacyl-tRNA synthetase, in two steps. In the first step amino acid is activated by the enzyme using energy derived from ATP. In the next step the amino acid~AMP reacts with the specific tRNA resulting in amino acid~tRNA molecule. The latter molecules recognize the codon on the mRNA and get attached at the appropriate place.

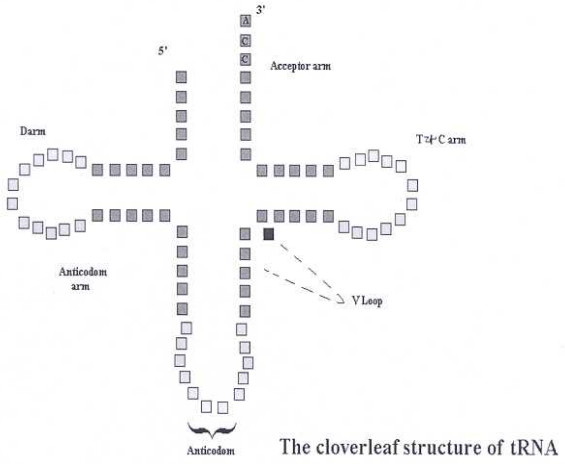

Fig. 8a

The first step in translation involves interaction of 30S ribosomal subunit, mRNA, initiation factors, IF-1, IF-2 and IF-3, and an initiator tRNA (Fig. 8b). The initiator tRNA is called tRNAfMet which binds to the initiation codon, AUG in mRNA. This tRNA carries the amino acid methionine whose amino group is blocked with a formyl (CHO) group. Since it is possible that the mRNA can have an internal AUG codon, therefore, the tRNA which responds to that codon is designated tRNAMet. The blocking of amino group on first methionine prevents addition of amino acids on this side while translation is going on. Although all polypeptides will begin with methionine, in some cases it is subsequently removed.

Initially, two separate complexes are formed; one between IF-2 and methionyl tRNAfMet and another involving 30S ribosomal subunit, mRNA and IF-3. In prokaryotes formation of the latter complex is dependent on base pairing between 16S rRNA (component of 30S subunit) and a sequence near 5' end of mRNA. Majority of prokaryotes have the sequence as AGGAGG located at –7 position. After the discoverers, this sequence is called Shine-Delgarno sequence. The two complexes combine to give rise to complete 30S complex, with the involvement of IF-1 and GTP, the IF-3 gets released. At his stage the large ribosomal subunit (50S) also joins resulting in complete 70S ribosome, while initiation factors IF-1 and IF-2 are released. Addition of 50S subunit creates three binding sites for tRNA molecules.

These are; A or amioacyl site, P or peptidyl site and E or exit site. The fresh aminoacyl-tRNA binds to A site while the tRNA carrying growing polypeptide chain binds to P site.

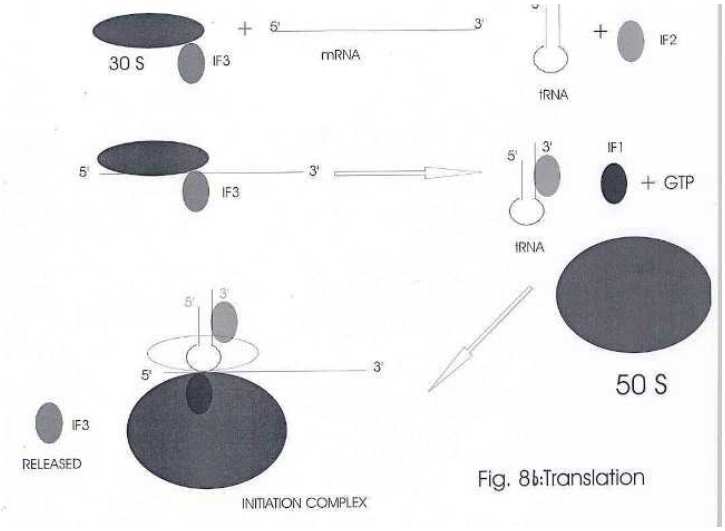

Fig. 8b:Translation

At the beginning, therefore, the aminoacyl-tRNAfMet aligned with AUG codon of the mRNA is located in P site (Fig. 9). The second codon being in register with A site will base pair with the anticodon of fresh aminoacyl-tRNA, thereby leading to chain elongation. For this step to occur, elongation factor Tu and energy from a molecule of GTP is required. After successful binding of aminoacyl tRNA, EF-Tu-GDP is released from the ribosome. This inactive form is again made active by elongation factor Ts using the energy of GTP molecule.

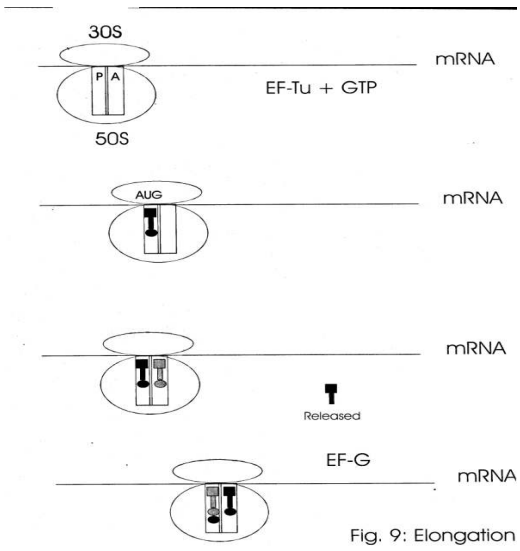

Fig. 9: Elongation

In the next step a peptide bond is formed between the amino group of the amino acid attached with tRNA in the A site with carboxyl group of the terminal amino acid attached with tRNA in the P site. This way the growing polypeptide chain gets detached from the P site tRNA and attached with the A site tRNA. The 23S rRNA molecule of 50S subunit of the ribosome has an inbuilt enzymatic activity called peptidyl transferase which brings about this important reaction. The tRNA in P site becomes devoid of polypeptide chain moves to E site wherefrom it is released. The ribosome moves three steps towards the 3' side so that a new codon comes into register with A site. Consequently, the tRNA alongwith the growing chain translocates into P site. This translocation requires elongation factor G (EF-G) and GTP. The previous step is now repeated and as the ribosome moves over the mRNA molecule, the polypeptide chain keeps on growing.

As soon as any one of the chain termination codons (UAA, UAG or UGA) comes into register with A site, it is recognized by the some proteins called release factors (RFs). While in prokaryotes each codon

is recognized by different release factors, in eukaryotes single release factor recognizes all of them. The polypeptide chain is released from the tRNA molecule located in P site. The ribosome gets dissociated into individual subunits (Fig. 10).

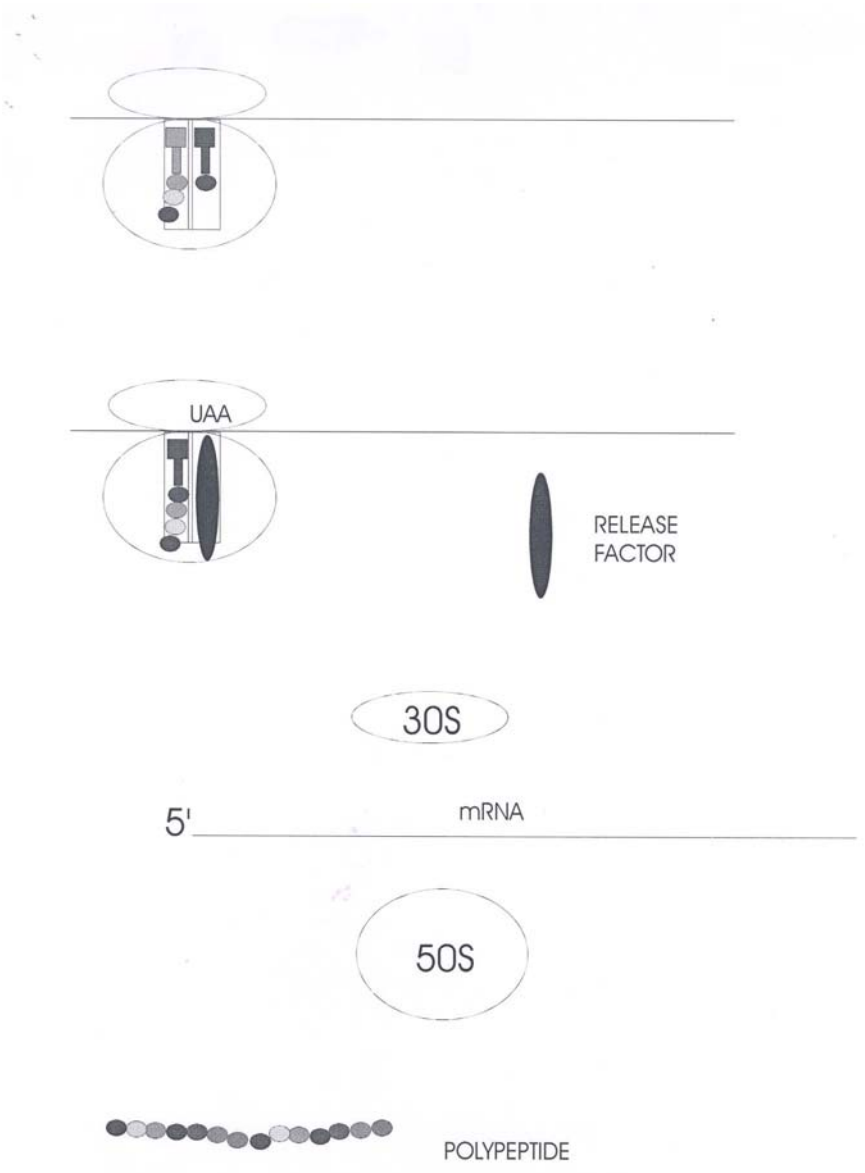

Fig. 10: Chain termination

The initial process of translation is complex in eukaryotes. Besides involvement of several initiation factors, there are four major differences.

i) Presence of 7-methylguanosine 'cap' at 5' end of the mRNA, which is essential for proper translation

ii) amino group of methionine in the initiator tRNA is not formylated

iii) translation begins at AUG codon very near the 5' end and not at Shine-Delgarno sequence

iv) a short recognition sequence exists around AUG codon, 5'-ACCAUGG-3' called KOZAK sequence which helps in the initial binding of mRNA to ribosome, like Shine-Delgarno sequence

REGULATION OF GENE EXPRESSION

Living organisms are exposed to various environmental factors, some of which are not suitable for the very survival of the organism. However, more often than not, the organisms have developed various evolutionary mechanisms, which help them to overcome the vagaries of environment and adapt to the changes. Prokaryotes are much more vulnerable to such changes than the eukaryotes. Therefore, prokaryotes have developed specific mechanisms to regulate the gene expression. It is no surprise that prokaryotes, especially *E. coli* has been studied very extensively to understand the process of gene regulation.

E. coli is the common colon bacillus which is exposed to variety of growth conditions during its life time. If the bacterium starts producing all sorts of enzymes, which can help in metabolizing various metabolites, it would be sheer wastage. Therefore, in order to conserve the energy resources, the bacterium synthesizes some of the enzymes only when needed. Such a system when genes get activated in response to an external stimulus is called inducible. The substrate, which activates the genes, is called the inducer. On the other hand there are some genes whose products are needed continuously such as those coding for ribosomal RNA and proteins, are called constitutive or house keeping genes. In some cases the presence of a particular molecule inhibits gene expression. Such a system is called repressible and the substrate is called repressor. Whether inducible or repressible, the regulation is under negative or positive control. To illustrate various modes of regulation we will take examples from *E. coli*.

Inducible system

Operon model

During 1961 two scientists, namely Francois Jacob and Jacques Monod proposed a model for gene regulation in prokaryotes. In prokaryotes a group of genes with related functions, regulated and expressed together as a unit, is called an operon. The operon consists of three structural genes *lacZ*, *lacY* and *lacA* and the DNA sequence adjacent to these genes called operator (Fig. 11). The structural genes code for enzymes involved in lactose utilization. The first gene, *lacZ*, 3510 bp, codes for β-galactosidase, the enzyme which breaks lactose into glucose and galactose. This reaction creates glucose which can be used by the bacterium as energy source. The second gene, *lacY*, 780 bp, encodes the enzyme permease which facilitates the entry of lactose molecules into the cells. *LacA*, 825 bp, is the third enzyme of the group which codes for enzyme transacetylase, whose function is still not clear. However, some evidence of its involvement in removing toxic by-products of lactose digestion has been collected. An interesting aspect of these genes is that all of these are transcribed as a single unit, resulting in polycistronic RNA. This way these genes are coordinately regulated and it is also ensured that products of these genes are produced simultaneously.

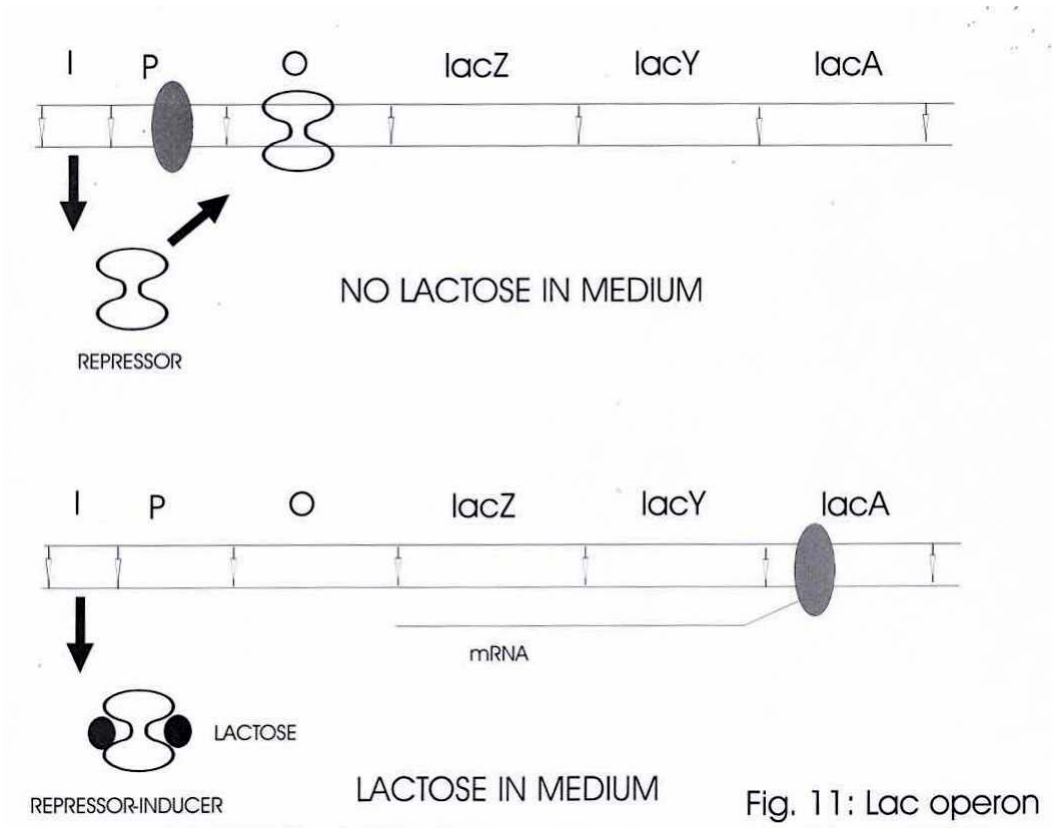

Fig. 11: Lac operon

RNA polymerase binds to the operator and initiates transcription of structural genes. However, another gene designated as *I* acts as a regulator. It produces a repressor molecule, 360 amino acids long, which, in the absence of lactose in the medium, binds to the operator region and does not allow RNA polymerase to bind. The active form of the repressor is a tetramer composed of four copies of the product of *I* gene. If the medium contains lactose, these sugar molecules bind reversibly with repressor molecule causing conformational change in the repressor and thereby making it incapable of binding the operator. As a result, RNA polymerase binds normally to the operator and the transcription of structural genes initiates. Since in this case binding of the repressor to the operator hinders transcription, this type of gene regulation is said to be under negative control. The original model of Jacob and Monod had only one operator, now called O_1. However, recently two additional operators (O_2 and O_3) have also been identified. It has been shown that besides O_1, either O_2 or O_3.

One of the intriguing questions is that when the bacterium suddenly moves to a medium containing lactose, how do initial molecules of lactose enter the cell for induction, since at that time β-galactoside permease is not available. The answer to this question lies in the fact that few molecules of these enzymes are always produced even in the uninduced state, so as to maintain the background level. This level of enzymes is essential to convert the first few molecules of lactose, which enter the cell, into allolactose. Allolactose is the actual inducer of lac operon as it binds with the repressor molecule leading to its release from the operator.

Initially, lacI gene, promoter and operator were identified genetically by isolating series of mutants showing altered expression of operon genes. Mutants synthesizing constitutively the enzymes for lactose utilization (in absence of lactose) were found to have mutations either in *I* gene or operator. Such mutations were designated I^- and O^c. In addition, partial diploids, called merozygotes were constructed, using fertility factors. The bacterial cells having F' factor carrying lac operon genes, in addition to normal genome, represent partial diploids for lac operon. The wild type alleles of structural genes are designated Z^+, Y^+ and A^+, mutant alleles as Z^-, Y^- and A^-. Therefore, wild type, monoploid individuals will have the genotype $I^+ P^+ O^+ Z^+ Y^+ A^+$ and will be inducible for utilization of lactose. The merozygotes with genotype $I^+ P^+ O^+ Z^+ Y^+ A^+ / I^+ P^+ O^+ Z^- Y^- A^-$ or $I^+ P^+ O^+ Z^- Y^- A^- / I^+ P^+ O^+ Z^+ Y^+ A^+$ are also inducible. The wild type alleles of structural genes and the regulator are dominant to respective mutant alleles thereby suggesting that the latter do not code for functional products. Many such combinations of dominant and mutant alleles were generated in merozygotes. The genotype $I^+ P^+$

$O^+ Z^+ Y^+ A^+ / I^- P^+ O^+ Z^+ Y^+ A^+$ is inducible, as expected, but also suggestive that the repressor is a diffusible product which is synthesized by I^+ on one DNA fragment and controls the transcription of structural genes on other DNA fragment. Therefore, the lac repressor can control the expression in *cis* as well as *trans* configuration. Similarly it was shown that the operator acts only in *cis* configuration , thereby indicating that operator does not code for any product, but is a site where repressor binds. The mutations in the promoter region do not affect the inducibility of the operon, but change the level of expression of structural genes.

Catabolite repression

Under optimal conditions of growth, the bacteria has ample quantity of glucose in the medium to meet the energy needs. Therefore, the lac operon remains switched off. However, in the situations where glucose as well as lactose is present, which of the two will the bacterium prefer. It has been established beyond doubt that the bacterium will prefer glucose utilization. This is achieved by the activity of another protein called **c**atabolite **a**ctivator **p**rotein or **CAP**. The CAP is involved in repression of the lac operon in the presence of glucose, the process called catabolite repression or glucose effect (Fig. 12). CAP in its dimeric form, binds a small effector molecule called cAMP (cyclic Adenosine –3',5'-monophosphate), therefore, sometimes CAP is also called crp (cyclic AMP receptor protein). The lac promoter has two distinct sites; one for the binding of RNA polymerase and another for CAP-cAMP complex. For normal induction of lac operon, binding of CAP-cAMP complex in its designated site in the promoter is a must. This type of control where binding of a molecule leads to induction is called positive control as opposite to binding of repressor leading to switching of the operon.

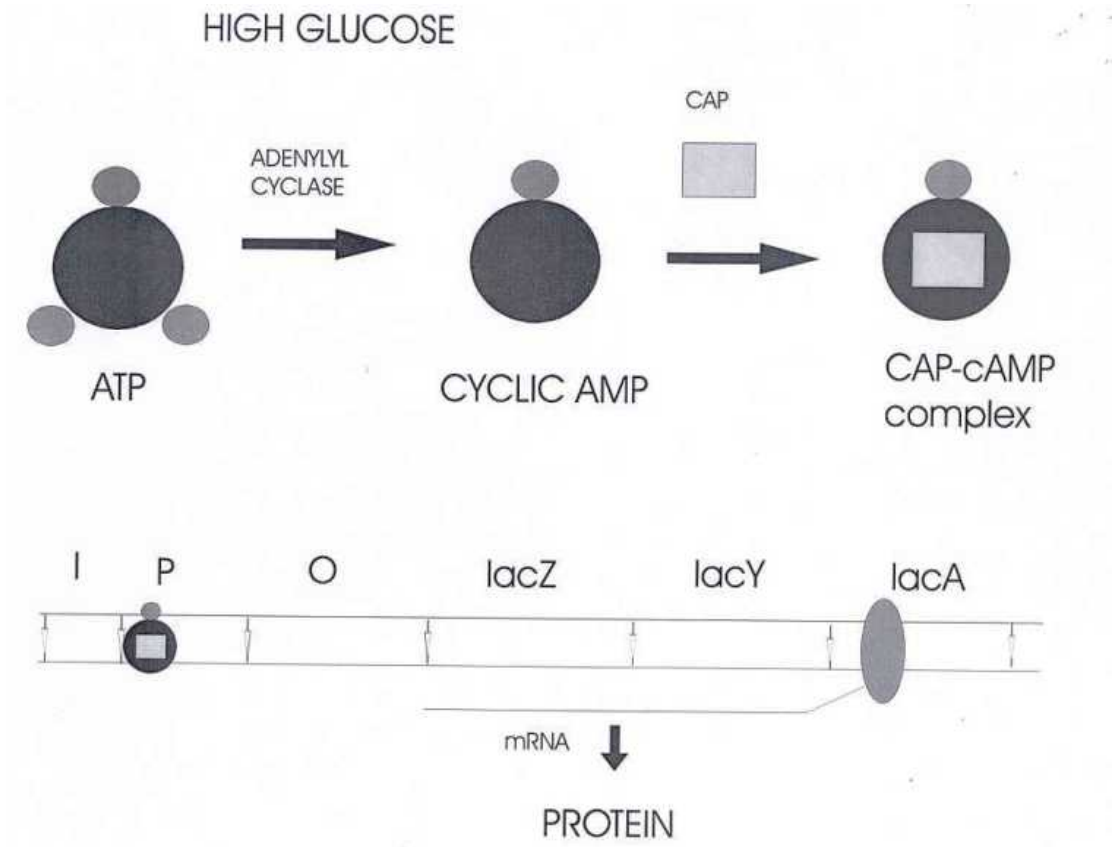

Fig. 12: Catabolite repression

Interestingly, CAP alone cannot bind to the promoter. Therefore, camp acts as an effector molecule. The intracellular concentration of cAMP is dependent on glucose; high concentration leading to decline in concentration of cAMP. Cyclic AMP is synthesized from ATP by the activity of the enzyme adenylcyclase. Glucose prevents the activation of this enzyme, thereby leading to decrease in intracellular concentration of cAMP, not sufficient enough to bind with CAP, which in turn cannot bind to the promoter alone. Since no CAP-cAMP complex is bound at promoter, therefore, no induction of lac operon takes place.

Repression and attenuation

In E. coli, other operons exist whose regulation is different than the lac operon. Best example to demonstrate this is tryptophan (*trp*) operon. The *trp* operon consists of five structural genes, which code for the enzymes responsible for biosynthesis of the amino acid tryptophan from chorismic acid. Besides, there are regulatory sequences upstream of structural genes. The five structural genes are designated *trpE* (1560 bp), *trpD* (1590), *trpC* (1353), *trpB* (1191) and *trpA* (804 bp). These genes are activated only when there is no tryptophan in the medium. The details of the reactions and the products have been shown in Fig. 13.

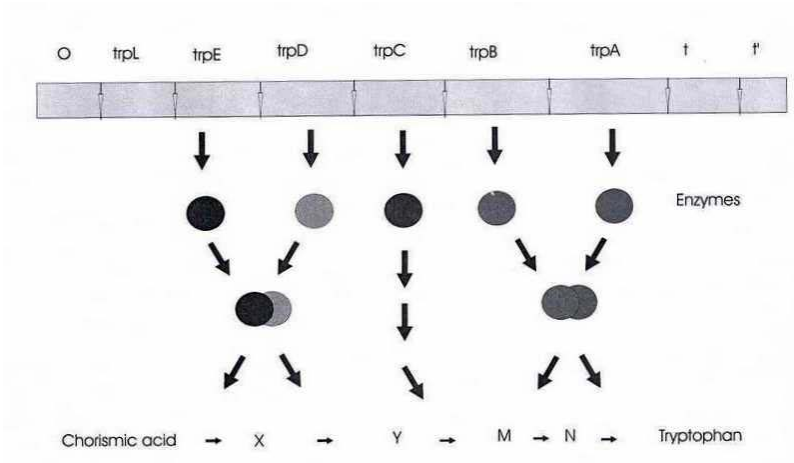

Fig. 13: Tryptophan operon in E. coli

This operon is an example of negative repressible operon. The regulator gene designated *trpR,* codes for a repressor and is not closely linked to the operon. The operator locus is located within the promoter. It is followed by *trpL* region about 162 bp long which encodes a mRNA leader sequence. The structural genes are followed at the end by two transcription termination sequences.

The trp operon is activated only in the absence of tryptophan. RNA polymerase binds to the promoter region and the structural genes are transcribed. Tryptophan acts as a corepressor because, when it is present in the medium, it binds to the repressor molecule and binds to the operator, thereby blocking the transcription. In the absence of tryptophan in the medium, the rate of transcription is about 70 times more than what it is in presence of tryptophan in the medium.

There is another level of regulation of trptophan synthesis in E. coli. trpL encodes a leader sequence (Fig. 14), which plays a very important role in regulation. This second level of regulation is independent of repression mechanism described above and is called attenuation. The DNA sequence within the trpL region responsible for this process is called attenuator. Attenuation is a mechanism of control of termination of transcription near the end of the mRNA leader sequence. However, for attenuation presence of tRNA charged with tryptophan amino acid is necessary. Analysis of trpL sequence in detail reveals that it has between 110 and 141 bp regions a typical sequence consisting of GC rich palindromic sequence followed by several AT base pairs. This region is called the attenuator sequence. This sequence has similarity with the transcription termination sequence found at the end of several bacterial operons. Transcription of attenuator region results in the formation of RNA, which as

per the sequence can form a hairpin structure. The hairpin structure causes conformational change in RNA polymerase leading to termination of transcription.

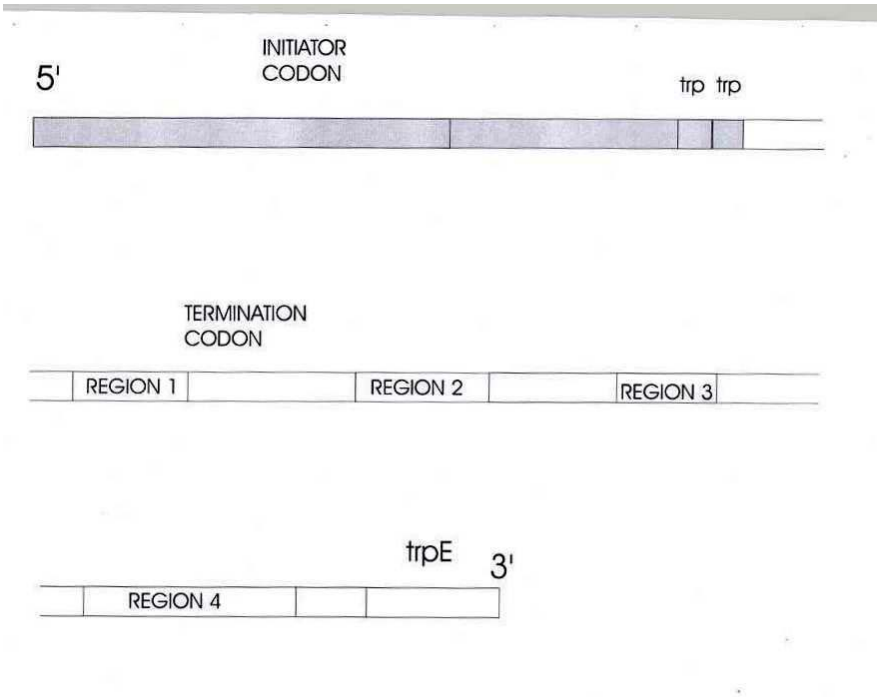

Fig. 14: trpL region of tryptophan operon

In addition to attenuator sequence, there are two more regions in the leader sequence which can form alternate hairpin structures. In total four regions exist; between, nucleotides 60-68, 75-83, 110-121 and 126-134 (Fig. 15). Out of these the first two and last two regions can pair to form loops. However, sometimes a loop is formed between second and third region, leaving first and fourth regions free. Pairing between third and fourth regions leads to transcription termination, described above. If the second region pairs with third region then it cannot form the termination loop.However, formation of a particular loop will depend upon the presence or absence of tryptophan.

The leader sequence has two adjacent UGG codons, the codes for tryptophan. Since in bacteria, the transcription and translation is coupled, therefore as soon a the initial part of the leader sequence is transcribed, the ribosome gets attached to it for initiating the translation. If the tryptophan is missing or is present in low concentration, thereby leading to shortage of tRNA charged with tryptophan, the ribosome gets stuck up at UGG codons. At this stage a loop is formed between second and third region, therefore, the transcription continues beyond attenuator to structural genes of *trp* operon. When tryptophan is present in the medium, the ribosome does not stop at UGG codons and therefore, disrupts the pairing between second and third region. However, termination loop is formed between third and fourth region (attenuator) leading to termination of transcription. This way the trancription of *trp* operon is regulated to a great extent by the joint efforts of repression and attenuation.

In addition to regulation of gene expression at transcriptional level, there is ample evidence that it is also happening at translational level.

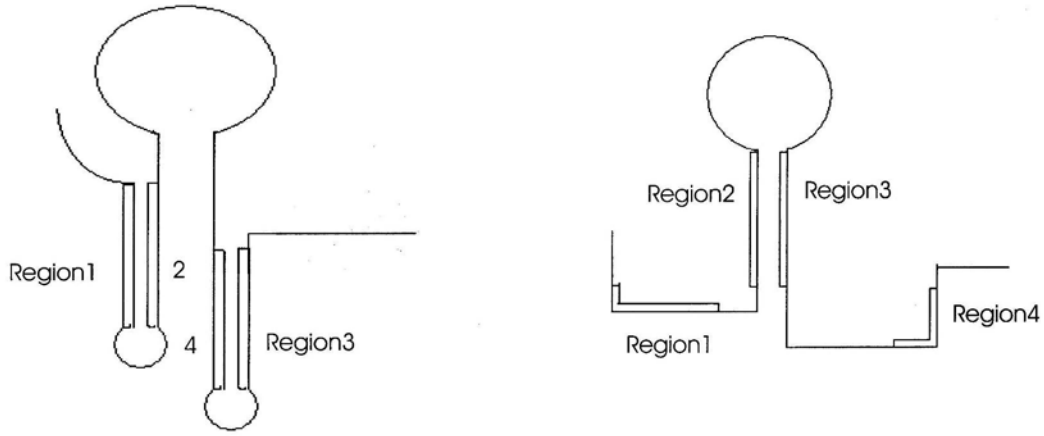

Fig. 15 : Secondary structures of trpL transcript

Gene regulation in eukaryotes

Eukaryotic genomes are highly complex. The DNA is associated with histones to form chromatin. The chromatin is organized in the form of chromosomes. The chromosomes are found inside the nucleus, which is a membrane bound structure. The genetic information is therefore, physically separated from the protein synthesizing machinery. The transcription and translation is not coupled. The DNA sequence gets transcribed into mRNA, which is transported out of the nucleus, where ribosome gets attached and translation takes place. Interestingly, the mRNA is processed in the nucleus before it is transported to the cytoplasm.

The number of genes also varies from one organism to another. The situation gets further complex if one takes into account the multicellular nature of most of the eukaryotes. Not all genes are expressed in all tissues. Depending upon the placement of the cell in the body, a specific set of genes is switched on and that too in a cascade fashion. Cascade would mean that the product of one gene activates another gene and its product activates third gene and so on. The expression of genes is either spatial (related to space) or temporal (related to time). In case of spatial expression, a gene expressed in leaves of a plant may not express in flowers. For example, the tubulin genes in *Arabidopsis*. Tubulin polypeptides are either α-type or β-type. While α-tubulin gene is expressed only in pollen grains and not in leaves, β-tubulin gene is expressed only in roots. In temporal expression, different genes are expressed at different times during the development. The example which clearly demonstrate this type of expression is hemoglobin, which is made up of two α and two β globin polypeptide chains. These genes are expressed at different times during development.

Transcriptional control

Despite these complexities, the basic process of transfer of information from DNA to RNA is the same. As expected , the transcriptional control of gene expression is complicated. Unlike prokaryotes, being located in the nucleus, genes cannot get easily activated in response to external signals. The signals are received at the cell memebrane, and through a complex mechanism of transfer called signal transduction, the signals reach the gene. The signal molecules can be those activated by the environment or the hormones. The various aspects of transcriptional control are discussed below:

a) Enhancer elements

The eukaryotic genes have some specific DNA sequences either upstream or downstream, which play very important role in gene regulation. These *cis* elements are called enhancers. These elements increase the efficiency of promoter function. Enhancers interact with regulatory proteins and

transcription factors. These are responsible for spatial and temporal expression of genes. It is speculated that enhancers mediate the effect by altering the chromatin configuration or by bringing the promoters, enhancers and transcription factors in close contact so as to increase the binding of polymerase. In certain respects, the enhancer elements resemble the operator sequence of prokaryotes.

b) Environmental stimulus

Environmental stresses constitute important factors in inducing the expression of genes. In response to heat stress, a set of genes is transcribed to produce Heat Shock Proteins (HSPs). These proteins help the cells to tolerate the stress by stabilizing the cellular machinery. In fruit fly, when the temperature increases to 33° C, heat shock transcription factor induces transcription of the genes to form HSP70 (70 kilodaltons is the molecular weight of the protein). These genes exist in five to six copies. The transcription factor binds to a sequence located upstream of the gene called heat shock response element.

c) Hormonal stimulus

Hormones are the signal molecules which are synthesized in one part of the body and circulated to other parts, to cause the effect. The hormones go to specific target cells and initiate a series of events leading to control of gene expression. In animals and human beings the hormones are of two main types; steroid and peptide hormones. The steroid hormones are small, lipid soluble molecules, which can pass the cell membrane very easily. Inside the cell these interact with cytoplasmic or nuclear proteins called receptors. The hormone-receptor complex interacts with the DNA and regulates the expression of the particular gene or set of genes. The examples are, estrogen, progesterone, (female hormones) and testosterone (male hormone).

On the contrary, peptide hormones are large molecules composed of amino acids. Since these are not able to pass through the cell membrane their effect is indirect. At the cell surface the signal is perceived by the membrane bound receptor, which transmits it to the interior of the cell. Binding of the hormone to the receptor causes change in its conformation, which in turn leads to change in the intracellular proteins. The change in one leads to change in another protein and finally the signal is transferred to the nucleus whereby the transcription of specific genes is regulated. The important example of this category is Insulin.

There are specific DNA sequences, which act as initial receivers of the signal and are responsible for activation of transcription. These sequences are known as hormone response elements. Specific proteins bind to these sequences and act like transcription factors.

Post-transcriptional control

In eukaryotes, expression of majority of genes is regulated at transcriptional level. However, recently various mechanism of gene regulation have been demonstrated to work at post-transcriptional level. In some of these cases, small, non-coding RNA molecules play an important role. These RNA molecules base pair with target regions in the mRNA, thereby preventing the expression. The phenomenon whereby the small RNA molecules regulate the expression at post-transcriptional level, is called RNA interference or in short RNAi.

Two types of RNA molecules are involved in RNA interference; short interfering RNAs (siRNAs) or microRNAs (miRNAs). These molecules are 21 to 2 nucleotides long and derived from double stranded (ds) RNA molecules. A special class of endonucleases, which are specific for dsRNAs, chop (dice) the large molecule into small pieces and are therefore called Dicer enzymes. Subsequently, the siRNA or miRNA get associated with proteins, where one of the strands is eliminated preferentially. The single stranded RNA complexed with protein is called RNA-induced silencing complex (RISC) and is now ready to base pair with a complementary region in the mRNA thereby inhibiting its translation.

If the pairing between the target sequence in mRNA and RNA in RISC is perfect, the RISC cleaves the mRNA in the middle leading to its degradation. It may subsequently get attached with another mRNA and cause its cleavage. This process can get repeated. Such RNA molecules in RISC which cleave the mRNA are called siRNA. In alternate situation if the pairing is not perfect, mRNA does not get cleaved

but translation is inhibited, such RNA molecules are called miRNA. The targets of RISC are located in 3' untranslated regions of mRNA molecules.

Post translational modifications

The chemical modification of a protein after it has been translated is known as post-translational modification. It is one of the later steps in protein biosynthesis for many proteins. The polypeptide is composed of several different amino acids. The post-translational modification of amino acids includes, i) attachment of functional groups like acetate, phosphate, various lipids and carbohydrates, ii) changing the chemical nature of an amino acid or by making structural changes, like the formation of disulfide bridges. Also, enzymes may remove amino acids from the amino end of the protein, or cut the peptide chain in the middle. For instance, the peptide hormone insulin is cut twice after disulfide bonds are formed, and a propeptide is removed from the middle of the chain; the resulting protein consists of two polypeptide chains connected by disulfide bonds.

Other modifications, like phosphorylation, are part of common mechanisms for controlling the behavior of a protein. The common types of post-translational modifications are given below.

	Modification	**Meaning**
1.	Acetylation	Addition of an acetyl group
2.	Alkylation	Addition of an alkyl group (e.g. methyl, ethyl)
3.	Methylation	Addition of a methyl group
4.	Glycosylation	Addition of a glycosyl group resulting in a glycoprotein
5.	Phosphorylation	Addition of a phosphate group
6.	Sulfation	Addition of a sulfate group

Protein folding

Immediately after synthesis, the protein exists as a chain of unbranched amino acids which is called the primary structure. However, all proteins undergo coiling and folding thereby assuming a specific three-dimensional shape in order to perform their functions. Therefore, protein folding can be defined as the process by which a protein structure assumes its functional shape or conformation.

On the contrary, protein denaturation is defined as the process whereby a native protein is loses its functional conformation, and becomes an amorphous, and non-functional amino acid chain. Denatured proteins may lose their solubility, and precipitate, becoming insoluble solids. In some cases, denaturation is reversible, and proteins may refold. In many other cases, however, denaturation is irreversible.

The folding of amino acid sequence into a particular shape is determined to a great extent by the sequence itself. However, importance of other factors like nature of the solvent, temperature, salt concentration and molecular chaperones, which play a significant role in protein folding.

The first step in the folding of the protein is the transition from primary to secondary structure. The secondary structure is organization into α – helix and β pleated sheets. In the next step due to covalent bonding in the form of disulfide bridges formed between two cysteine residues**,** tertiary structure of the protein is produced. The quaternary structure involves the grouping of two polypeptide chains.

Acknowledgements

The author is thankful to Dr. Sanjana Kaul and Mr. Khalid Zafar for going through the manuscript very critically and also to Ms. Pooja Sharma and Mrinal for help in preparing the illustrations.

Further reading

i) Snustad, DP and Simmons, MJ (2003) Principles of Genetics. John Wiley and Sons, Inc. USA.
ii) Watson, JD et al. (2004) Molecular Biology of the Gene. Pearson Education (Singapore) Pte. Ltd.
iii) Klug, WS and Cummings, MR (2003) Concept of Genetics. Pearson Education (Singapore) Pte. Ltd.
iv) Hartl, DL and Jones, EW (2006) Essential Genetics: A Genomics Perspective. Jones and Bartlett Publishers, USA