MOLECULAR BIOLOGY

Regulation of Gene Expression

Abhay Anand Shukla, Misti Jain and Shyam S. Chauhan
Department of Biochemistry
All India Institute of Medical Sciences
Ansari Nagar
New Delhi – 110 029

22-Feb-2008

CONTENTS

Keywords

Gene expression regulation; Operons; Post-transcriptional regulation; mRNA; Gene amplification; Transcription; Translation; RNA splicing; Post translational modification (PTM)

Introduction

Gene expression is the process by which a gene's DNA sequence is used as a template for the synthesis of effector molecules. These effectors which may be proteins (synthesized from mRNA), tRNAs or rRNAs are in turn responsible for the implementation of genetic blue print. Cells of different organs of multi-cellular organisms contain same genome but exhibit differences in expression profile of these effector molecules. Similarly gene expression profile of unicellular organism varies depending upon their environmental conditions and stage of development. On the other hand expression of some genes, whose products are essential for the cells/organism (house keeping genes), is unaltered by the stage of development or environmental conditions. Such genes are said to be constitutively expressed. The genes whose expression level alters under different conditions and stages of development are termed as regulatable. The gene regulatory mechanisms operate at many levels which include: through alterations in the DNA (chromatin structure); transcription(RNA synthesis by RNA polymerase using DNA as a template), RNA processing and stability; translation of mRNA into protein (including initiation and rate of translation) as well as control of protein trafficking to proper cellular compartment and post translational modifications (protein activation by folding, cleavage, modification, assembly of polypeptides). It has been found that both eukaryotes and prokaryotes chemically exert gene control in two general ways: affecting molecules that interact with DNA, RNA and/or the polypeptide chains, or controlling the synthesis of an enzyme or the activity of an enzyme in the cell. Gene controls can be positive – inducing gene activity, or negative – repressing gene activity. Whereas, in prokaryotes, control of transcriptional initiation is the major point of regulation, in eukaryotes the regulation of gene expression is controlled at many different points.

Prokaryotes

Both in prokaryotes and eukaryotes, RNA polymerase initiates gene transcription after binding to the promoter region in a sequence specific manner. This binding is aided by protein factors. The prokaryotes are considered to be in a non-restrictive ground state. This state is defined as the inherent activity of promoters *in vivo* in the absence of specific regulatory sequences. This state is independent of the strength/quality of individual promoters, which varies considerably as a function of the DNA sequence. Experiments have shown that *in vivo*, an isolated prokaryotic promoter region is sufficient to initiate transcription at a rate comparable to that achieved *in vitro* confirming the fact that there is no inherent restriction on the ability of prokaryotic RNA polymerase to gain access to the DNA template and initiate transcription. It is becoming increasingly clear that supercoiled state of the chromosomes is a part of the cellular regulatory repertoire.

Supercoiling affects the gene expression by changing the shape of the DNA. Negative super coils not only facilitate promoter melting, a prerequisite for transcription initiation, but can also have also opposite effects of repressing it, by an as yet unknown mechanism that might be related to DNA twist. Micro array studies on *Escherichia coli* genome have revealed that expression of 7% of the genes is rapidly and significantly affected by the loss of chromosomal super coiling. As DNA super coiling affects functionally diverse genes it has been hypothesized to function as a second messenger in relaying the environmental signals to various circuits in the cells.

Operons and regulation of gene expression

An *E. coli* contains as few as 10 copies to as many as 10^5 copies of proteins. This differential expression of proteins is the result of regulation of gene expression. It is a cost effective method used by cells to save valuable resources so that no energy is wasted in unnecessary mRNA and protein synthesis. In prokaryotes the gene activity is primarily regulated at the level of transcription. F. Jacob and J. Monod (1961) in their historic paper for the first time proposed operon model for regulation of gene expression in prokaryotes. According to this model genes encoding functionally related proteins (structural genes) are present in a cluster. All these structural genes are transcribed from the same promoter into a single transcript encoding multiple proteins. A single transcript having a separate open reading frame for each protein encoded by it also known as polycistronic mRNAs allows coordinate expression of structural genes. Thus a gene cluster is regulated together by the common regulatory elements namely promoter, operator and regulatory gene. The structural genes along with the regulatory elements constitute an operon. They exhibit both positive and negative regulation. In negative regulation, a repressor molecule binds to the operator locus of an operon with very high affinity and terminates transcription of the structural genes. Where as, in positive regulation, an activator interacts with the promoter region and facilitates binding of RNA polymerase to initiate transcription of the downstream genes.

Lactose (Lac) operon

Fig. 1 depicts the structural organization of *E. coli*. lac operon as proposed by Jacob and Monod. This bacterium normally uses glucose as the source of energy during normal conditions. However, when lactose is the sole source of energy in the medium, expression of the structural genes encoding the following proteins is induced:

(a) β-galactosidase: this enzyme hydrolyzes lactose into, glucose and galactose. The gene encoding the enzyme is denoted by *lacZ*.

(b) Lactose permease: is an integral membrane protein required for the transport of lactose into the bacterial cell. The membrane is otherwise essentially impermeable to lactose. The gene encoding the enzyme is denoted by *lacY*.

(c) Thiogalactoside transacetylase: This enzyme is not required for lactose metabolism and its physiological function is uncertain. *In vitro,* it catalyses the transfer of an acetyl group from acetyl CoA to the C-6 hydroxyl group of a thiogalactoside. The gene encoding the enzyme is denoted by *lac A*.

Lac Z, Y and A are located adjacent to each other in the *E. coli* genome. Their protein products enable cells to take up and metabolize β-galactosides, such as lactose. The elements that control the transcription of these genes are the promoter region, the operator region and the regulatory gene (*lacI*) encoding a repressor protein. The *lac I* located near the structural genes is constitutively transcribed from its promoter. The resulting transcript is translated into a 38 Kda repressor subunit. Four such units assemble into a functionally active tetramer. This tetramer binds to the operator locus with high affinity. The 27 bp operator region located between the promoter and the structural genes includes nucleotides starting from 5 base pairs upstream to the

21 base pairs downstream of the transcription initiation site of *lacZ*. Thus binding of repressor to the operator stalls RNA polymerase and therefore blocks the transcription of structural genes *lac Z, Y* and *A*. Therefore, when lactose is not the only source of energy for *E. coli* the lac operon is repressed i.e. *lac Z, Y and A* are not expressed and hence very few molecules of Z, Y and A are present in the bacterial cells. (Fig. 1)

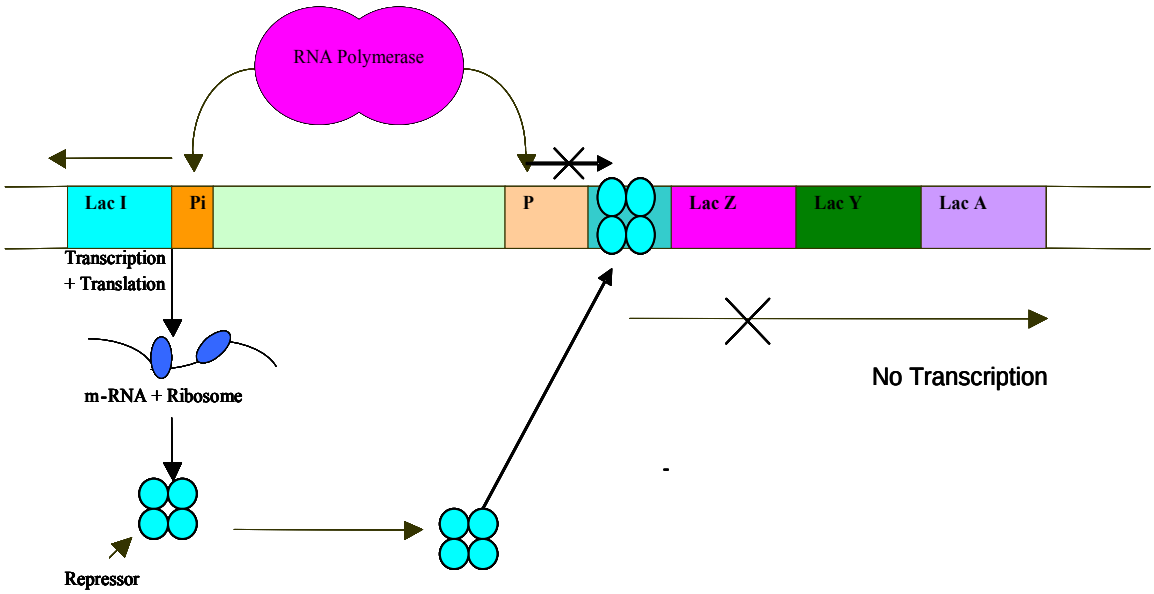


Fig. 1: Regulation of Lactose operon in absence of Lactose

The repressor has two binding sites, one for binding to the operator region and the other for binding to the inducer. When glucose, glycerol etc. are consumed and lactose remains the only and abundant source of energy some of it is converted to allo-lactose. It is the physiologic inducer of lac operon. Lac repressor molecules both those attached to operator locus and those present in cytosol binds to the inducer with high affinity. This binding brings about a conformational change in repressor and reduces its affinity for the operator and dissociation thereby relieving the repression (Fig. 2). Binding of two molecules of inducer to the repressor causes allosteric transition that is sufficient to release repression. Studies on synthetic inducers have shown that some of the compounds act as inducers without being metabolized and are termed as 'gratuitous inducers'. Isopropylthiogalactoside (IPTG) is one such very efficient inducer of the *lac* operon and therefore used for prokaryotic expression of recombinant proteins using lac expression system.

RNA polymerase binding to the majority of prokaryotic promoters is controlled by sequence elements that are approximately 35 bases and 10 bases, respectively, upstream of the site of transcriptional initiation and therefore referred as the '-35' (TTGACA) and '-10'(TATAAT) promoter elements. These elements are recognized and contacted by RNA polymerase for

binding. However, for RNA polymerase to bind to lac promoter also requires another protein termed as catabolite gene activator protein (CAP /CRP). This protein is synthesized as an inactive allosteric precursor which binds cAMP to form a CAP-cAMP complex. The dimer of CAP-cAMP complex in turn binds to a 14bp region (CAP binding motif) upstream to the lac promoter and exposes a binding site for RNA polymerase (called the acidic patch). This facilitates RNA polymerase to bind CAP and the *lac* promoter and initiate transcription of the structural genes. However, only CAP-cAMP complex not free CAP can bind its cognate motif. Thus availability of cAMP is vital for the binding of RNA polymerase and transcription/expression of the lac operon and therefore it exerts positive regulation. The cAMP levels in the cell increase with decrease in the levels of glucose. Under normal conditions when both glucose and lactose are present in the growth media, lac operon is repressed because lac I is bound to the operator due to the absence of inducer (allolactose) and cAMP levels are low so cAMP-CAP complex is not formed. When glucose is exhausted by the bacterial cell cAMP levels increase allowing the formation of cAMP-CRP complex which facilitates binding of RNA polymerase to the promoter and lac I dissociates from the operator due to the presence of inducer. In this way as summarized in Table 1, lac operon is amenable to both positive and negative regulation.

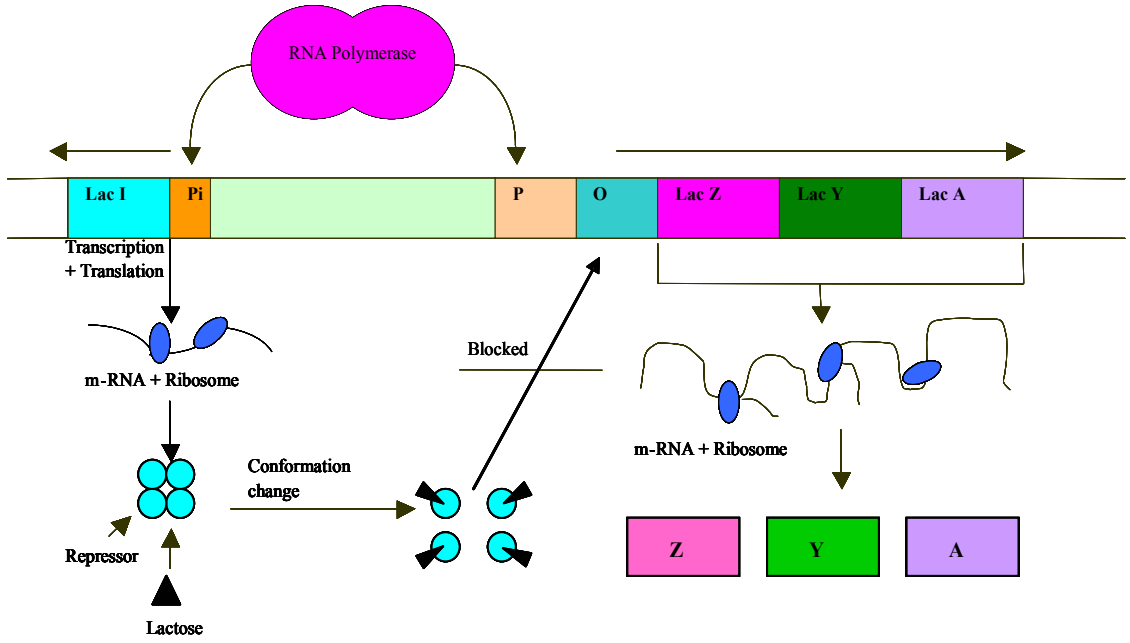


Fig. 2: Regulation of Lactose operon in presence of Lactose

In the case lac operon, lac I functions as a regulator protein for negative regulation and cAMP-CAP complex serves as a regulatory protein for positive regulation. The regulation of operon by CAP in conjunction with cAMP is also called catabolite repression. The phenomenon of catabolite repression has been observed in many operons involved in sugar and amino acid metabolism such as arabinose (ara) operon; maltose (mal) operon; and, histidine utilization (hut) operon.

Table1

	Regulator protein	
Type of control	Bound to DNA	Not bound to DNA
Negative control	Operon switched off	Operon switched on
Positive control	Operon switched on	Operon switched off

Tryptophan operon

Adequate supply of amino acid is essential of protein synthesis in free living unicellular organism for adapting to the changes in environment. *E. coli* is able synthesize all 20 amino acids from their precursors. The enzymes involved in amino acid biosynthesis are organized in operons. This kind of operons called biosynthetic operons are not regulated by cAMP-CAP complex but by the amino acid synthesized. Thus tryptophan operon (*trp* operon) is an example of a biosynthetic operon whose expression is regulated by tryptophan itself. The operon consists of 5 structural genes namely *trp* E, *trp*D, *trp*C, *trp* B and *trp* A. They encode the following three enzymes required to convert chorismic acid into tryptophan:

1) *Anthranilate synthetase*

   component I - encoded by a gene denoted by *trpE*
   component II - encoded by a gene denoted by *trpD*

2) *N-(5'-phosphoribosyl)-anthranilateisomerase/Indole-3-glycerolphosphate synthase*

   encoded by a gene denoted by *trpC*

3) *Tryptophan synthase*
   β-subunit - encoded by a gene denoted by *trpB*
   α-subunit - encoded by a gene denoted by *trpA*

As shown in Fig. 3 a promoter and an operator region precede the structural genes. Beyond the transcription initiation site is a transcribed leader region (*trpL*) of 162bp containing a regulated site of transcription termination termed as the attenuator.

The full-length polycistronic *trp* mRNA encoding the five-trp polypeptides is about 6800 nucleotides in length. Rho-dependent transcription termination occurs in the region following the last gene, *trpA*. The trp operon is controlled by two separate mechanisms: a) repression and b) attenuation.

Repression

The tryptophan repressor (TrpR) is a dimer of 107 amino acids encoded by *trpR* gene. It is encoded by a gene located away from the *trp* operon on the *E. coli* chromosome. TrpR actually regulates the expression of two other operons in addition to the *trp* operon. TrpR cannot bind to the operator by itself but requires a co-repressor to do so. In case of trp operon tryptophan serves as a co-repressor. Each sub unit of TrpR dimer contains a tryptophan binding site. As long as there is insufficient tryptophan (co-repressor) in the cell, repressor cannot bind the operator hence, structural genes are transcribed. However, as the level of tryptophan builds up, the *Trp* repressor forms a complex with tryptophan. This repressor-co-repressor complex binds tightly to the operator and blocks further transcription of the operon and, as a result, the synthesis of the three enzymes declines. The tryptophan repressor dimer contains three domains: a central core (formed by the amino terminus of both the subunits), and two flexible DNA-reading heads (formed by the carboxyl terminus half) the central core plays the part of a spacer between the reading heads. When the repressor is not bound to tryptophan (apo-repressor) the reading heads are quite close to each other and hence cannot fit into the successive major grooves of the DNA. However, when the tryptophan binds to the repressor it acts as wedge to separate the two heads and thus the complex now snugly fits into the major grooves. Tryptophan also serves to mould the side chains of the protein to make stabilizing contacts with the DNA. It itself is linked to the DNA by hydrogen bonds between its indole ring and the phosphate backbone of the DNA. The repression mechanism has been schematically shown in Fig. 3.

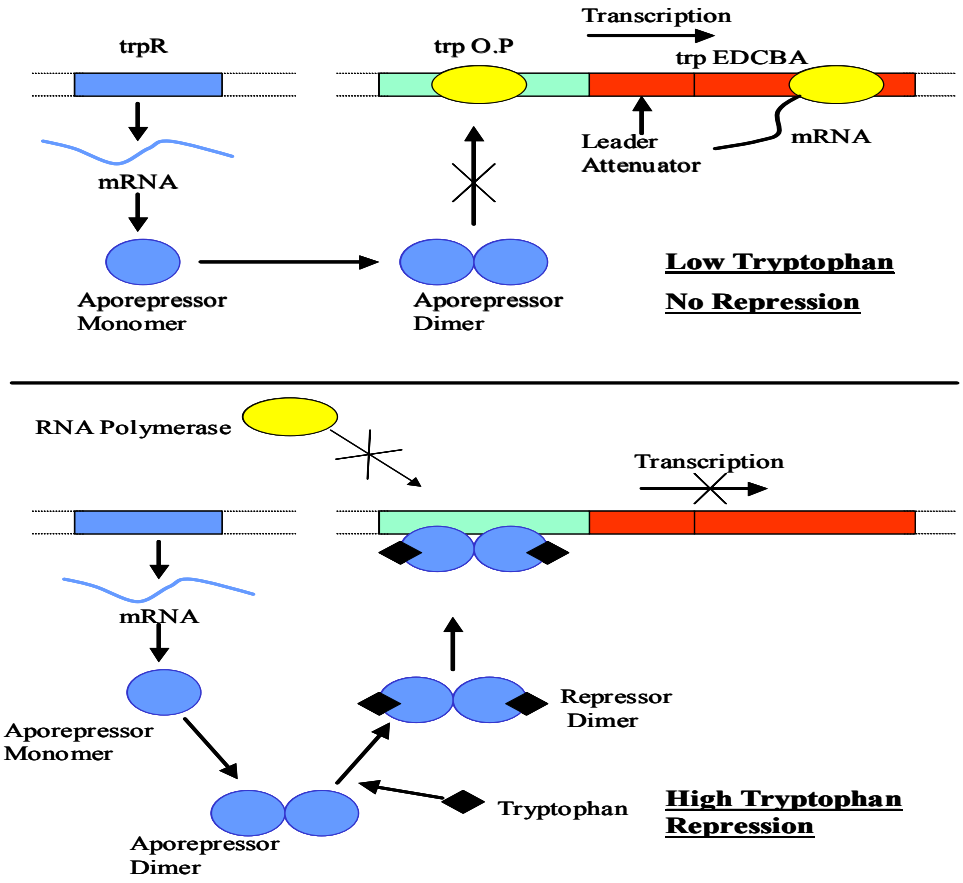


Fig. 3: Regulation of *trp* operon by tryptophan

Attenuation

Charles Yanofsky and his colleagues were among the first ones to discover the regulation of trp operon by attenuation. They observed that mutants with deletions between operator and the first structural gene in the trp operon showed increased production of *trp* mRNA. This led to the discovery of a leader region of the *trp* mRNA that encodes a 14-residue peptide with adjacent tryptophan residues at positions 11 and 12. It was observed that the non-mutants produced only about 130 nucleotide long transcript when sufficient level of tryptophan was available in the cell. But when its levels were low a 7000 nucleotide *trp* mRNA was produced. This led to the conclusion that an intrinsic terminator, the attenuator, controls the transcription of trp operon even after its initiation. The attenuator contains a GC-rich region followed by an AT-rich region and exhibits a two-fold symmetry. The transcription and translation of the tryptophan genes are tightly coupled. The attenuation mechanism is based on the principle that the ribosome location on the leader segment of the transcript, while RNA polymerase ahead of it is synthesizing the segment, determines the secondary structure of the transcript. This in turn directs the transcribing RNA polymerase to terminate or to continue transcription.

When the tryptophan is plentiful, initiation of the transcription is blocked by the tryptophan-repressor complex, which tightly binds to the operator and prevents the RNA polymerase from binding to the promoter. However, a decrease in cellular concentration of tryptophan favors its dissociation from the tryptophan-repressor complex. As a result of which more and more free repressor is generated which cannot bind the operator and therefore, transcription is initiated at a faster rate. Still some of the RNA polymerases are unable to transcribe the entire *trp* mRNA and get dissociated prematurely. The proportion of RNA polymerases transcribing the entire mRNA increases as the levels of tryptophan decrease.

As shown in Fig. 4 mRNA encoding the leader sequence can acquire alternative base-paired structures. The entire leader sequence can be divided into 4 regions. These regions are short stretches containing purines and pyrimidines, which complement each other to form hairpin like structures. Hence, region 2 can base pair with region 1 and region 3. Similarly, region 3 can base pair with region 2 as well as region 4.

When tryptophan is present, ribosomes are able to synthesize the leader peptide. The *tryptophanyl-tRNA* is easily available and hence they progress through the tandem tryptophan codons (UGG-UGG) and continue to translate region 2 of the leader peptide. As a result, region 2 is unavailable for base pairing with region 3. Region 3, thus, base pairs with region 4 generating the terminator hairpin (Fig. 4). Under these conditions, the polymerase terminates at the attenuator.

When tryptophan is scarce, the levels of *tryptophanyl-tRNA* are also low. Hence, the ribosomes stall at the tandem trptophan codons. So, region 1 is sequestered within the ribosome and cannot base pair with the region 2. Regions 2 and 3 become base paired before region 4 has been transcribed (as shown in the Fig. 4). In the absence of the terminator hairpin, RNA polymerase continues transcription past the attenuator.

Thus availability of tryptophan plays a crucial part in the attenuation mechanism. RNA polymerase remains paused after transcribing 90 bases along the leader sequence till the ribosomes translate the leader peptide. As the translation proceeds it determines the fate of the secondary structure at the attenuator site and in turn determines the fate of the RNA polymerase too.

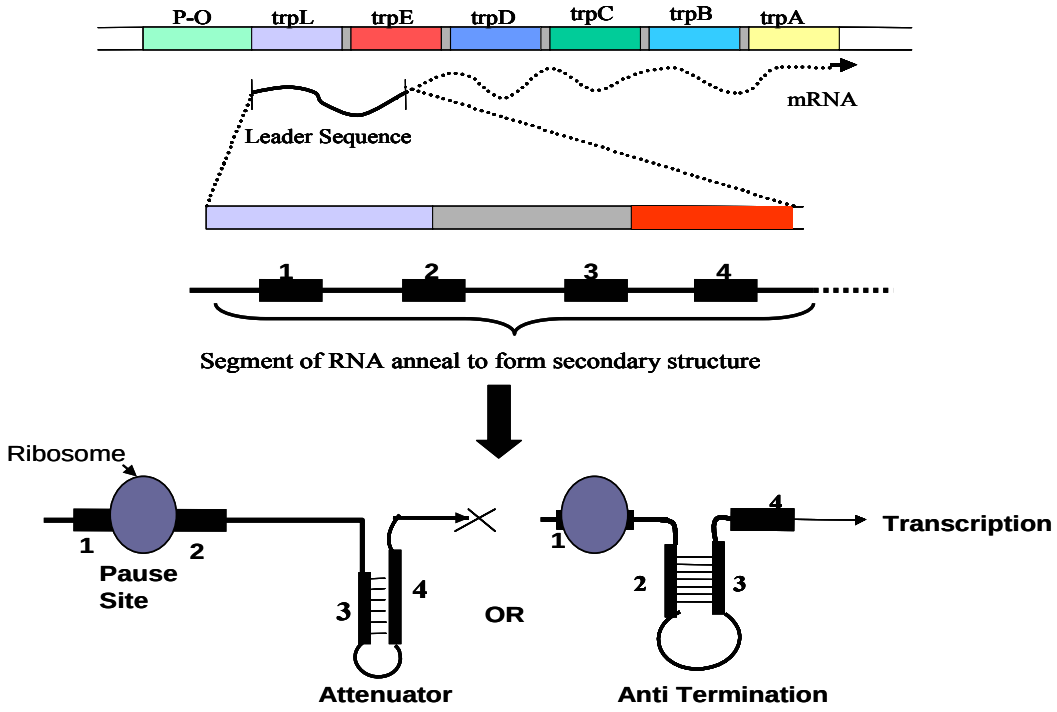


Fig. 4: Regulation of *trp* operon by attenuation

Similarly threonine, phenylalanine and histidine operons are also regulated by attenuation**.**

Post-transcriptional regulation in prokaryotes

It has been observed that unequal amount of proteins are synthesized from the different cistrons of a poly cistronic mRNA. For example the ratios of number of copies of β-galactosidae, permease and transacetylase made from lac operon mRNA are, 100:50:10. These differences can only be explained by the posttranscriptional regulation of this operon.

Translational

The ribosomes binds independently to each cistron of the poly cistronic mRNA by means of a purine rich tract (Shine-Delgarno sequence) upstream to the translation initiation codon. The complementarity of the SD sequence to 3'end of 16S ribosomal RNA determines the efficiency of ribosomal binding and therefore its translational efficiency. In this way the observed difference in level of three enzymes encoded by lac operon is due to the difference in the efficiency of their cistrons in binding to the ribosomes.

mRNA stability

The degradation of mRNA normally starts from its 3'end and proceeds towards its 5'end. The cistron of transacetylase is present in the most 3'end of the lac operon mRNA and β - galactosidase at its 5'end. Therefore, at given point in time there are more copies of β-galactosidase mRNA compared to that permease and transacetylase. Thus amount of protein made from them is dependent on the relative abundance of their mRNA.

Regulatory small RNAs

It is rapidly becoming apparent that there exists an entire family of gene expression regulators that have remained undetected by the conventional transcription-based approaches such as site directed mutagenesis or mutant hunts or study of the genome sequence. These are small regulatory RNAs that control the translation and degradation of many mRNAs. The vast diversity of RNA-dependent regulatory mechanisms has been linked to the dynamic properties and versatility of the RNA structure. These small regulatory RNAs are a part of a larger family known as "non-coding RNAs". These are not translated into proteins. They may be approximately 50 to 400 nucleotides in length. The importance of small regulatory RNAs was first appreciated in the elegant studies of plasmid-encoded antisense RNAs. The number of such noncoding regulatory RNAs has increased over the last decade, and their role not only in stimulating and inhibiting gene expression but also as key effectors of adaptive responses, such as environmental cue recognition, stress response, and virulence control has been firmly established.

By the early 1980s the bacterial phages, transposons and plasmids were shown to control their copy number, life cycle using these small antisense RNAs. Till now, in addition to phages, even bacterial genomes have been shown to harbor many loci that encode these sRNAs. One of these RNAs is *micF* RNA, a *trans*-encoded RNA, which participates in the repression of outer membrane protein F (OmpF) expression when *E. coli* cells are grown under high temperature or other environmental stress conditions. *micF* RNA binds to and regulates the stability of *ompF* mRNA in response to various enviromental stimuli. Another example of antisense RNA regulating gene expression is dicF, also found in *E. coli*, involved in the regulation of cell division. In the past few years, many small non-coding RNAs have been identified as crucial regulatory elements in bacterial stress responses and in bacterial virulence.

The small regulatory RNAs exert their effect by various mechanisms. A single sRNA is capable of regulating multiple genes. The most common one is the antisense mechanism, which involves base pairing of sRNA to target RNAs in order to affect post-transcriptional regulation of gene expression. Generally this physical binding leads to translational inhibition and/or facilitated mRNA decay; however, in some cases these antisense RNA have also been shown to promote the conversion of a translational inert mRNA conformation to an active conformation. This mechanism of gene regulation has been seen in plasmid replication and conjugation, control of bacterial genes such as *ompF* and *crp* gene of *E. coli*; temporal control of bacteriophage development to name a few. A second mechanism of control involves sequestration of regulatory proteins. The possible mechanism of RNAs acting as temperature sensors has been attributed to changes in the conformation of these RNAs resulting in regulation of gene expression.

Regulation of gene expression in Eukaryotes

The multicellular eukaryotes consist of hundreds of different cell types, each differentiated to perform a specialized function. Each cell type differentiates by activating a different subset of genes and hence, the need for a higher degree of gene regulation. A major difference between the prokaryotes and eukaryotes is the presence of "nucleus" in the eukaryotes. Hence, the two major events of gene expression, transcription and translation are spatially and temporally separated in eukaryotes. This is not the case in prokaryotes. The pathway of gene expression starts from transcription through translation till a properly folded functional protein is produced. In eukaryotes, regulation can occur at any point in this pathway; i.e. at the levels of chromatin structure, transcription, RNA processing, mRNA stability (longevity), and translation.

Eukaryotic chromosome

The eukaryotic genome is enormously large as compared to prokaryotes but the cell size is not proportionately that large. Therefore, in eukaryotes but not in prokaryotes the genome is organized into highly compact nucleoprotein structures called chromosomes. Thus the substrate for the essential biological processes of transcription, replication, recombination, DNA repair, and cell division is not naked DNA; rather, they are protein-DNA complexes known as chromatin. The physical structure of the DNA, as it exists compacted into chromatin, can affect the ability of transcriptional regulatory proteins (transcription factors) and RNA polymerases to find access to specific genes and to initiate transcription from them. Thus the presence of histones the methylation status of DNA affect accessibility of the chromatin to RNA polymerases and transcription factors to a great extent.

Histones and gene expression

DNA in chromatin is closely associated with a number of highly conserved proteins known as histones that fold the DNA in a hierarchical series of stages, ultimately yielding a 10,000-fold linear compaction preparatory to cell division. Histones can be separated from DNA by treating chromosome with salt or dilute acid. The initial or lowest level of chromatin organization consists of the local wrapping of a short stretch of DNA, 147 bp in length, in the form of 3/4 turns of a flat superhelix around an octameric histone protein core. This ocatmeric core is composed of two molecules each of histones H2A, H2B, H3, and H4. The complex of histone octamer and 147 bp of DNA wrapped around it is known as the "nucleosome core particle". This local packing motif is repeated at intervals, millions of times along the entire DNA length, with short variable-length stretches of "linker" DNA between consecutive core particles. In most cases *in vivo*, each nucleosome core particle is associated with one additional molecule termed as "linker histone," H1. Particles containing the complete core particle plus histone H1 and the linker DNA at each end are called "nucleosomes". Nucleosomes are traditionally considered to be the fundamental units of chromatin structure.

Binding of histone octamer to DNA is primarily at the sugar phosphate backbone over short stretches where the minor groove (and hence the DNA backbone) faces in towards the octamer surface. A striking feature of histones is their high content of positively charged side chain as about one in four residues is either lysine or arginine. Contacts between the histones and DNA include extensive salt bridges and hydrogen bonds to the phosphate groups contributed by both main-chain and side-chain groups; extensive non polar contacts with the DNA sugar;

electrostatic interactions of the positively charged N-termini of alpha helices with DNA phosphates; and a smaller number of base-specific contacts, including non polar contact of the 5-methyl group of thymidine in the major groove. Each of the core histones has a 10–40 amino acid long highly positively charged N-terminal region. Histones H2A and H3 have shorter but analogous domains at their C-termini as well. These domains are referred to as "tails" because of being highly extended and mobile. They are the sites of numerous posttranslational modifications known to be essential in chromatin function.

All core histones are acetylated. Histone acetylation appears to be increased in a domain containing active genes, and acetylated chromatin is more sensitive to the action of DNAse I and to micrococcal nuclease. It also occurs during S phase, when histones are being incorporated into nucleosomes. The absence of acetyl group may be a prerequisite for a more condensed, inactive structure. Acetylation occurs in both replication and transcription processes. It allows the loosening of nucleosome core thus facilitating the binding of transcription factors required for recruitment of RNA polymerase. Thus acetylation of histones promotes gene expression and their deacetylation blocks it.

Methylation and gene expression

DNA methylation is another factor that controls transcription. One common feature among eukaryotic organisms is the presence of methyl ($-CH_3$) groups attached to DNA. Although they are not defined sequences, it has been suggested that DNA methylation is involved in controlling gene expression. Specifically, the methyl group appears to be preferentially associated with cytosine bases, and the modified base is called 5'-methylcytosine (5'-mC). Physically, the 5'-mC group occupies an open space inside the double-stranded DNA molecule. This observation has suggested that this group could act as a specific signal to other molecules that act to regulate gene expression. This is one of the several regulatory events that influence the activity of promoter. There are two general mechanisms by which DNA methylation inhibits gene expression: first, modification of cytosine bases can inhibit the association of some DNA binding factors with their cognate DNA recognition sequences and second, proteins that recognize methyl-CpG can elicit the repressive potential of methylated DNA. Methyl-CpG-binding proteins (MBPs) use transcriptional co-repressor molecules to silence transcription and to modify surrounding chromatin, providing a link between DNA methylation and chromatin remodeling and modification.

CpG islands are regions in the DNA containing abundant repeats of dinucleotide sequence CpG. They are usually present in 5' region of most of the genes and are connected with the effect of methylation on gene expression. These islands have an average 60% GC content, compared with the 40% average in bulk DNA. These regions can stretch upto 1-2 kb. There are near about 45,000 such islands in the human genome. The nucleosome at the islands is known to have low content of histone H1 and the cytosine residue of island is in unmethylated state. Methylation of a CpG island present in the promoter of a gene usually prevents its expression. However, methylation of a silencer region can upregulate the expression of a gene.

Gene amplification

Under certain conditions many copies of a gene are produced in a cell. This process, which is termed as gene amplification occurs due to repeated rounds of replication of a specific gene. The

extra copies of the gene thus produced may either be present in the cell as free extra chromosomal molecules (episome, double minute) or as tandem array of sequences within the chromosome. Gene amplification makes multiple copies of a gene available for transcription leading to a net increase in its level of expression. Amplification of genes takes place under both physiological and pathological conditions. Ribosomal RNA genes get amplified in amphibian oocytes. Similarly cancer cells utilize gene amplification as tool to evade their killing by cytotoxic drugs. Methotrexate, a commonly used chemotherapeutic agent specifically inhibits dihydrofolate reductase (DHFR). The later is involved the synthesis of dTTP and therefore required for DNA replication. Thus methotrexrate kills rapidly proliferating tumor cells by blocking DNA replication. In most of the methotrexate resistant cancers DHFR gene is amplified leading to its increased expression and thereby making the drug ineffective. Similarly amplification p-glycoprotein (p-gp) gene confers multi-drug resistance phenotype to cancer cells. P-gp is a 170 kDa membrane transporter capable of pumping out a variety of cytotoxic drugs. Over expression of this transporter in tumor cells make them drug resistant by keeping the intracellular concentration of the chemotherapeutic agent(s) below their cytotoxic levels.

Transcription in Eukaryotes

Transcription in eukaryotes is more evolved in comparison to the prokaryotes. This process occurs in the membrane bound nucleus and the translation occurs in the cytoplasm. Thus, whereas in prokaryotes the processes of transcription and translation are tightly coupled, this is not the case in eukaryotes. The spatial and temporal separation of the two processes enables the eukaryotes to regulate gene expression. This regulation allows for selective transport of RNAs across the nuclear membrane for protein synthesis.

Prokaryotes possess a single RNA polymerase for the synthesis of RNA. However, the eukaryotes make use of three types of RNA polymerases differing in specificity, localization and sensitivity to inhibitors. RNA polymerase I is responsible for the synthesis of ribosomal RNAs where as t-RNA, 5S-rRNA, small nuclear RNA U6 are synthesized by RNA polymerase III. All protein-coding genes are transcribed by RNA polymerase II. In this chapter we will restrict our discussion to the regulation of protein coding gene expression.

Unlike prokaryotes RNA pol II does not initiate transcription on naked DNA instead it is a complex and highly regulated process that requires the assembly of many transcription factors along with the enzyme into a pre-initiation complex (PIC) at the core promoter. The **core** promoter may be defined as the minimal DNA region that is capable of directing low levels of basal transcription by RNAP II *in vitro*. Promoter elements of the genes transcribed by RNAP II usually contain a TATA box and INR (initiator) element. While the first nucleotide of INR (PyPyAN(T/A)PyPy ;Py-pyrimidine; N-any nucleotide), represents the transcription initiation site , the TATA- box ,an AT rich sequence, is located approximately 30 bases upstream of transcription start site. As shown in Fig. 5 RNAP II requires a number of protein factors called general transcription factors (TF-II A, B, D, H, F etc) to bind the promoter region and initiate transcription. Many of these factors have multiple subunits e.g. TF-IID consists of a TATA binding proteins (TBP) and several TBP associated factors (TAF). As the name suggests TBP binds TATA box and cause DNA to bend which facilitates the binding of RNAP II to the TATA box and INR. As depicted in figure- this is followed by binding of other general transcription factors which stabilize this transcription initiation complex. The promoters of a large number

genes transcribed by RNAP II lack TATA box (house keeping genes). The INR binding protein may be involved in formation of ignition complex on such promoters. Recently it has been demonstrated that in Drosophila that downstream promoter element (DPE) functions cooperatively with INR for the binding of TFII D in the transcription of core promoter in the absence of a TATA box. The DPE element (PuG(A/T)CGTG) is centered around position +30 with respect to the INR.

In addition to the above mentioned general transcription factors required for the formation of initiation complex many other specific transcription factors (STF) are involved in this process. STF bind to the motifs upstream to the TATA box called upstream promoter elements such as the GC box, CAAT box etc. Binding of STF to these promoter elements enhance the stability as well as the activity of the initiation complex thereby increasing the rate of transcription.

Fig. 5: General features of Pol II Promoter

Some of these STF may also bind to a region located upstream or down stream to the promoter and increase the rate of transcription (enhancer region). Similarly binding of these factors to another region may affect the rate of transcription negatively (silencer region). In order to influence the rate of transcription the STFs bound the enhancers/silencers (regulatory) interact with the initiation complex with the help of other protein factors called adaptors. In this way factor bound to the regulatory region comes close to the initiation complex by DNA looping.

Thus the factors involved in the Pol II mediated transcription can be classified into three major categories:

1. **General transcription factors:** They are required for transcription of all gene transcribed by Pol. II These general transcription factors (GTFs) bind to the promoter elements (TATA box, Initiator element etc) and facilitate the formation of pre-initiation complex (PIC).
2. **Sequence specific DNA-binding transcription regulators (i.e. activators and repressors):** they bind to proximal promoter elements as well as distal regulatory sequences (enhancers and silencers). Initiation of specific transcription is mediated by the recognition of core promoter by GTFs followed by the recruitment of Pol II and co-operative assembly of all the proteins into PIC at the promoter.
3. **Transcription co-factors/co-regulators (co-activators and co-repressors):** They interact with the regulators and mediate their effect on the basal transcription machinery. Their effect may be mediated through direct protein interaction as well as indirectly by modification of the chromatin structure.

The entire process of transcription in eukaryotes has been schematically depicted below in Fig. 6.

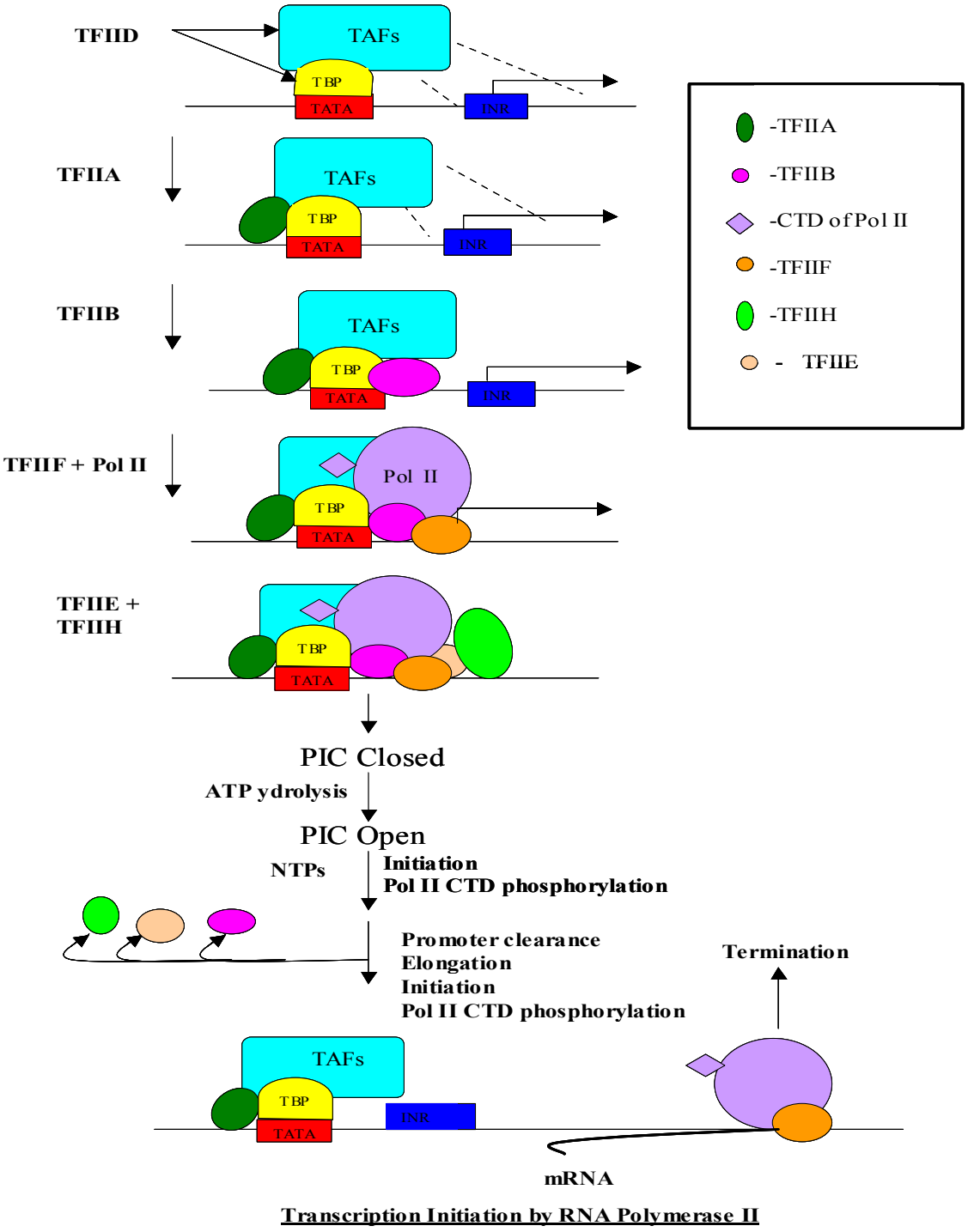


Fig. 6: Schematic representation of Transcription in eukaryotes

Polyadenylation and capping of mRNA

Eukaryotic mRNAs (all 3 classes) undergo significant post-transcriptional processing such as 5'capping and polyadenylation. The 5' end of all eukaryotic mRNAs is capped with a unique 5' ---→ 5' linkage to a 7-methylguanosine residue. The capped end of mRNA (Fig.7) is thus protected from exonucleases thereby providing stability to the mRNA and more importantly is recognized by specific proteins of the translational machinery thereby increasing its translational efficiency.

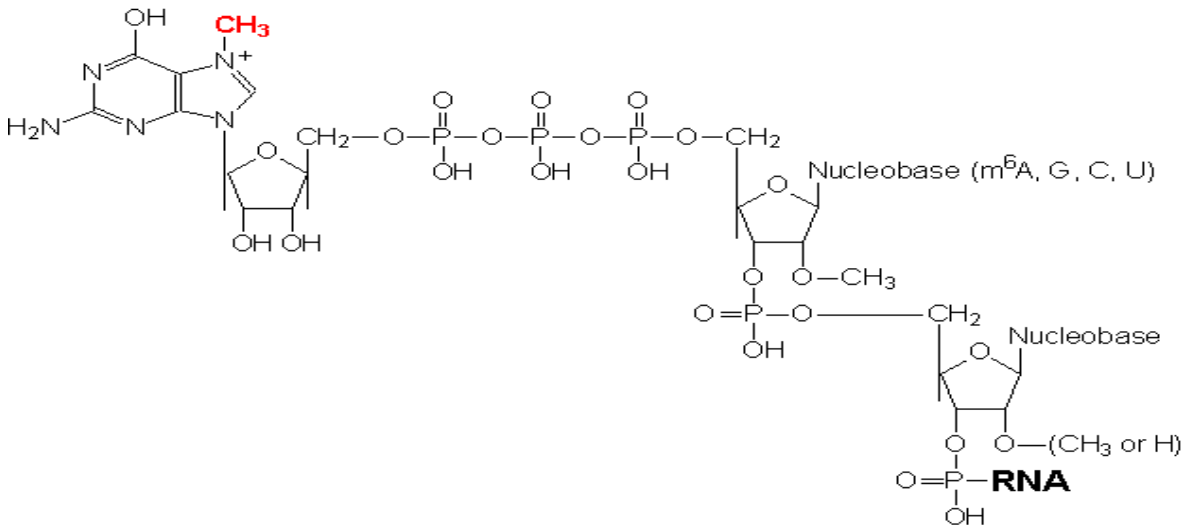


Fig. 7: 5' Capped mRNA

Similarly a poly A tail is added to a majority of mRNAs at their 3' ends. For this purpose a specific sequence, AAUAAA, at the 3' end of primary transcript is recognized by the polyadenylate polymerase enzyme whose endonuclease activity cleaves the primary transcript approximately 11 – 30 bases 3' of this sequence element. A stretch of 20 – 250 Adenine residues is then added by the polyadenylate polymerase activity of the above mentioned enzyme. (Fig. 8)

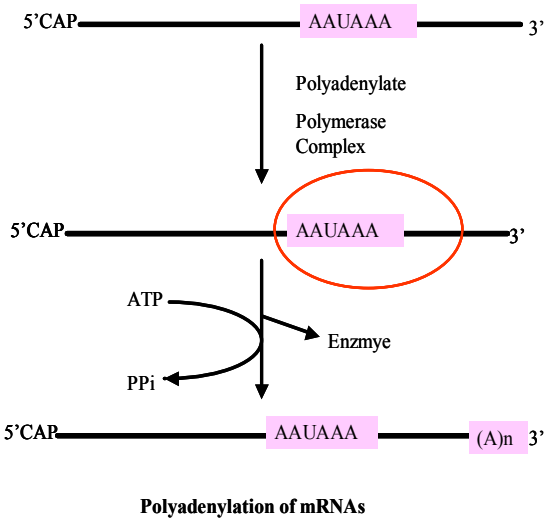


Fig. 8

RNA splicing

Unlike most prokaryotes, the coding sequences of eukaryotic genes are interrupted by the non-coding sequences. Therefore, it necessary to remove some stretches of nucleotides from the primary transcript before converting it into translatable mRNA (mature mRNA). The sequences which are retained in the mature mRNA are termed as exons and the ones which are removed from the primary transcripts are called introns. Thus introns are intervening sequences between exons. The gene for the β chain of hemoglobin is interrupted within its amino acid coding sequence by a long intervening sequence of 550 base pairs and a short one of 120 base pairs. Thus, the β-globin gene has three exons and two introns (Fig. 9).

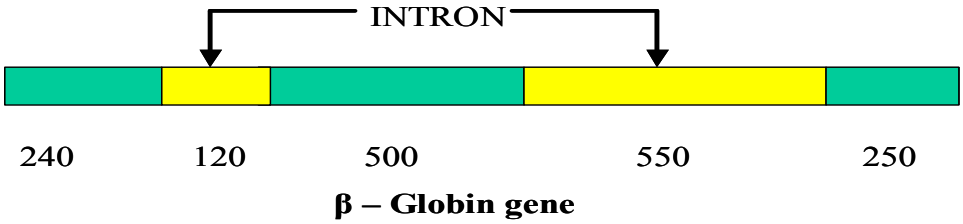


Fig. 9

Analysis of the human genome using the exon and intron distribution profiles has revealed that the 26,564 annotated genes in the human genome contain 233,785 exons and 207,344 introns. Thus on an average, there are 8.8 exons and 7.8 introns per gene. About 80% of the exons on each chromosome are <200 bp in length. It has been seen that <0.01% of the introns are <20 bp in length and <10% of introns are more than 11,000 bp in length. These results suggest constraints on the splicing machinery to splice out very long or very short introns and provide insight to optimal intron length selection. Interestingly, the total length of introns and intergenic DNA on each chromosome is significantly correlated to chromosome size.

The sequences encoded by the intronic DNA must be removed from the primary transcript prior to the RNAs being biologically active. The process of removal introns from the primary transcript and ligating the exons together to make a mature mRNA is known as RNA splicing. RNA splicing has been a subject of intensive investigation and this process has been elucidated in considerable details.

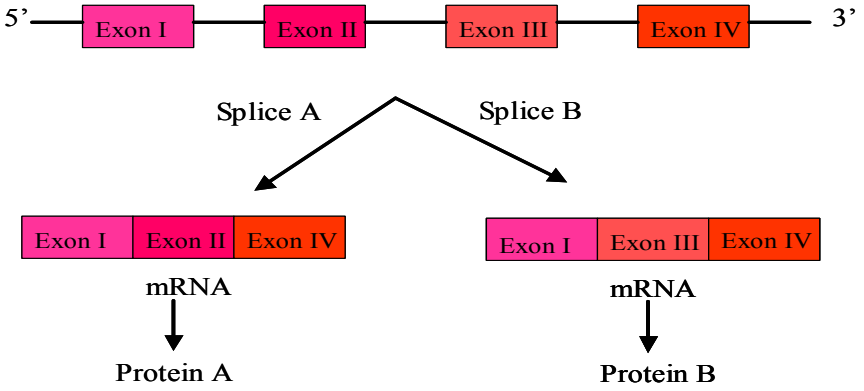
Alternative and Aberrant Splicing

The presence of introns in eukaryotic genes would appear to be an extreme waste of cellular energy when considering the number of nucleotides incorporated into the primary transcript only to be removed later as well as the energy utilized in the synthesis of the splicing machinery. However, the presence of introns can protect the genetic makeup of an organism from genetic damage by outside influences such as chemicals or radiation. An additional important function of introns is to allow alternative splicing to occur, thereby, increasing the genetic diversity of the genome without increasing the overall number of genes. By altering the pattern of exons, from a single primary transcript, that are spliced together different proteins can arise from the processed

mRNA from a single gene. Alternative splicing can occur either at specific developmental stages or in different cell types.

This process of alternative splicing (Fig. 10) has been identified to occur in the primary transcripts from at least 40 different genes. Depending upon the site of transcription, the calcitonin gene yields an RNA that synthesizes calcitonin (thyroid) or calcitonin-gene related peptide (CGRP, brain). Even more complex is the alternative splicing that occurs in the α-tropomyosin transcript. At least 8 different alternatively spliced α-tropomyosin mRNAs have been identified.

Abnormalities in the splicing process can lead to various disease conditions. Many defects in the β–globin genes are known to exist leading to β–thalassemias. Some of these defects are caused by mutations in the primary sequence of the gene required for intron recognition and, therefore, result in abnormal processing of the β-globin primary transcript.



ALTERNATE SPLICING

Fig. 10

Gene regulation at the level of translation

The regulation of gene expression in eukaryotes is not only restricted to transcription but also extends to translation. A brief overview of the process is given below followed by its control over gene regulation. Translation process can be divided into three phases namely initiation, elongation and termination.

The initiation step is the rate-limiting step of the entire process and is assisted by more than 25 polypeptides. It involves the positioning of an elongation-competent 80S ribosome at the initiation codon (AUG). Initially, the mRNA associates with the 40S ribosomal subunit, initiator methionyl tRNA, eIF2 (eukaryotic initiation factor 2) and GTP to form the 43S pre-initiation complex (Fig.11). A group of initiation factors termed as "eIF4F" help recruit the 43S complex to the 5' end of the mRNA. More specifically, the various factors of this group are:

(i) eIF4E: this physically binds to the 'cap structure';
(ii) eIF4A: it has helicase activity and thus it unwinds the secondary structures of RNA that may be present in the 5'UTR to facilitate binding of the 43S complex;

(iii) eIF4G: it functions as a Scaffold protein as it interacts with eIF4E, eIF4A and eIF3 as well as PABP. This complex then scans the mRNA till it reaches the first AUG codon (translation initiation codon). Recognition of this codon triggers eIF5 to hydrolyze GTP, as a result the initiation factors dissociate from the complex and lastly the 60S subunit joins the complex to create a fully functional translation apparatus. The 3'end of the mRNA is bound to the poly-A binding protein, which in turn is associated with eIF4G. The interaction of PAB and eIF4G synergistically stimulates translation, although the mechanism of how this occurs is not clear.

The elongation phase of translation also involves a set of elongation factors that facilitate in the production of amino acid chain. It starts with the binding of the aminoacyl-tRNA to the aminoacyl site. A peptide bond forms between the amino group of the incoming aminoacyl-tRNA and the carboxyl group of the methionine carried by the initiator tRNA. The entire process is powered by the hydrolysis of the GTP.

The termination phase occurs when a stop codon is read by the release factor and causes the disassociation of the synthesized polypeptide chain from the ribosome.

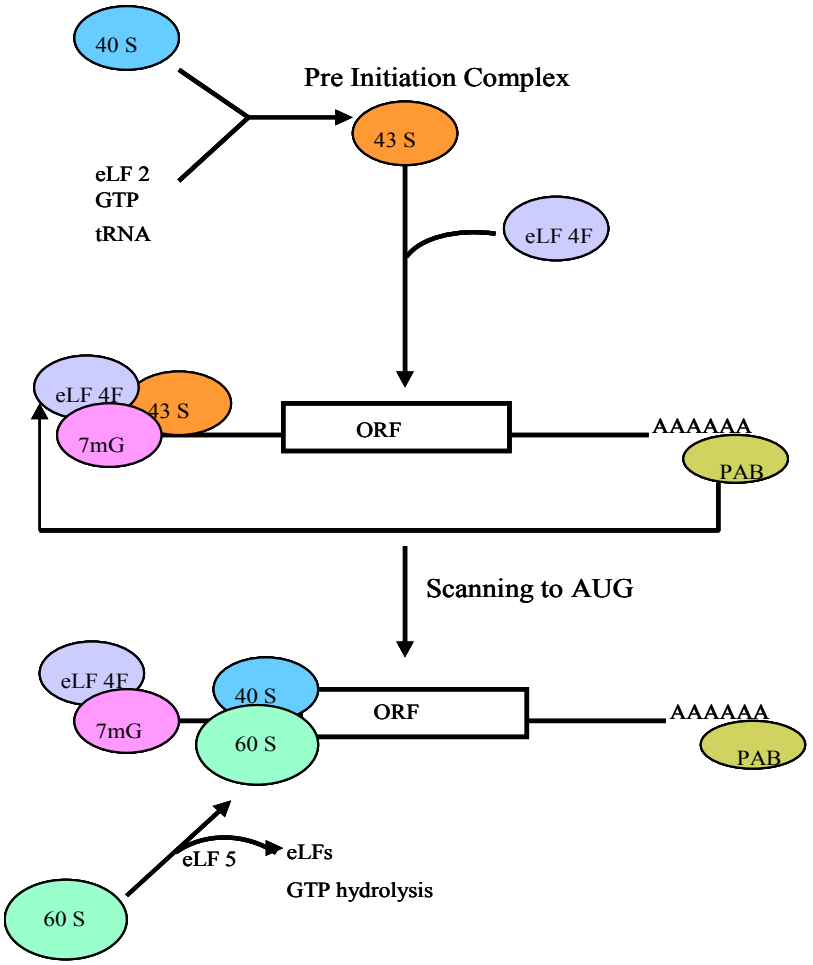


Fig. 11: Schematic representation of Translational in eukaryotes

Translational control of gene expression

Translation represents the last step of gene expression through which the genomic information is finally converted into proteome. It has been widely demonstrated that cells also use 'translation' as a mechanism to fine tune the gene expression for e.g. during embryogenesis as well as in stem-cell proliferation, sex determination, neurogenesis and erythropoiesis.

Two general mechanisms of translational control of gene expression have been widely studied:

(a) **Global control** (mainly occurs by the modification of translation-initiation factors)
This type of control is mediated by the changes in the phosphorylation state of initiation factors or the protein regulators that interact with them. e.g. phosphorylation of a subunit of eIF2B inhibits the GTP hydrolysis required for translation initiation as soon as the initiation codon is recognized. Another example is that of the translation factor eIF4E. The availability of this factor also regulates the general translation rates. Hypophosphorylation of the proteins (4E binding proteins) inhibits its association with eIF4G in turn inhibits the association of the 43S complex with the mRNA and consequently, in translational repression.

Similarly, caspase 3 is capable of cleaving eIF4G and PABP and this mechanism of interference with the translation of cellular mRNAs has been observed in viral proteases.

(b) **mRNA specific control** (mediated by regulatory protein complexes that recognize particular elements that are usually present in the 5' and / or 3' untranslated regions:
(UTRs) of the target mRNA) e.g. translation of specific mRNAs at defined locations such as in anterior / posterior pole of oocyte, neuronal synapse etc. The structural features and sequences that influence the translation rate are:

- **i)** The '**Cap structure' and the poly-A tail**: they are strong promoters of translation initiation.
- **ii)** The **Internal ribosomal entry sites (IRESs)**: These are located in the 5'UTR of some mRNAs. They mediate cap-independent translation initiation by recruiting the ribosome directly to an internal site in the mRNA.
- **iii)** The **upstream open reading frames (uORFs)**: They normally reduce translation from the main ORF.
- **iv)** Secondary and tertiary structures of the RNA (hairpins and pseudoknots): They commonly block translation. **Psuedoknots** are homopolymeric stretches of usually 25–200 adenine nucleotides that are present at the 3' end of most eukaryotic mRNAs. e.g. presence of a stem loop motif termed as 'an iron responsive element' in the 5' UTR of the ferritin mRNAs. Similarly, during oocyte maturation and early development a protein known as the cytoplasmic-polyadenylation element-binding protein (CPEB) binds to a uridine rich sequence — the cytoplasmic polyadenylation element (CPE) — located in the 3'UTR of target mRNAs and promotes both silencing of the mRNA before oocyte maturation as well as subsequent cytoplasmic polyadenylation and translational activation (Mendez and Richter, 2001). The mechanisms of translation regulation are in general inhibitory in nature thus hinting to the fact that most (not all) of the mRNAs are translationally active by default.

(c) Micro RNAs

Researchers have in the past decade discovered that regulation of translation is controlled not only by protein factors but also by small RNAs. These RNAs are usually 22 nucleotides long and are known as micro RNAs (miRNAs). So far, a number of these RNAs have been discovered and they have been described to regulate a broad spectrum of biological processes ranging from cell metabolism to its differentiation, its growth as well as apoptosis.

They hybridize by incomplete base pairing to several sites in the 3' UTR of target mRNAs. Thus, they facilitate in the repression of translation and have no role in the degradation of target RNAs. However, the mechanism of tranlstional repression is largely unknown till now.

The miRNAs are quite similar to another species of regulating RNAs namely '**silencing RNAs (siRNAs)**'. They are double stranded in nature and are capable of degrading the target RNAs.

Posttranslational modification

A eukaryotic protein is often destined for further processing and modification after translation. Processing involves proteolytic cleavage and splicing of the polypeptide. Similarly, post translational modification (PTM) extends the range of functions of the protein by a) attaching to it other biochemical functional groups such as acetate, phosphate, various lipids and carbohydrates, b) by changing the chemical nature of an amino acid (e.g. citrullination) or c) by making structural changes, like the formation of disulfide bridges, to name a few. These changes are important for their biological activities. Therefore, modulation of the activity of a protein after their synthesis may serves as a control point in the regulation of their biological functions and hence their expression

Processing of proteins

a) Proteolytic cleavage

Most eukaryotic proteins undergo cleavage after translation. The simplest form of cleavage is the removal of the translation initiation amino acid methionine. Many proteins are synthesized as inactive precursors (or proproteins) that are activated under proper physiological conditions by limited proteolysis for e.g. pancreatic enzymes and enzymes involved in clotting. Another example of this is the production of Corticotropin, Lipotropin, Melanocyte – Stimulating hormone, Endorphin and Corticotropin like intermediate lobe peptides. These all come from the same polypeptide, but due to the cleavage at specific points they produce smaller polypeptides, that form the hormones.

The proteins that are membrane bound or are destined for secretion contain an N- terminal signal sequence or signal peptide composed of usually 13-36 predominantly hydrophobic residues. It is recognized by a multi-protein complex called signal recognition particle (SRP) in the

endoplasmic reticulum. The removal of the signal peptide is catalyzed by signal peptidase. Proteins that contain a signal peptide are called preproproteins to distinguish them from proproteins.

b) Splicing involves the addition of other proteins or peptides for e.g.

Isgylation is the covalent linkage of a polypeptide to the ISG15 protein (Interferon-Stimulated Gene 15).

Sumoylation is the covalent linkage to of a polypeptide to the SUMO protein (Small Ubiquitin-related MOdifier). SUMO proteins are a family of small proteins (around 100 amino acids in length) that are covalently attached to and detached from other proteins in cells and hence modify their function for e.g. protein stability, nuclear-cytosolic transport, transcriptional regulation. Sumo modification of hNinein leads to its movement from the centrosome to the nucleus. In most cases Sumo attachment to transcriptional regulators correlates with inhibition of transcription.

Ubiquitination is the covalent linkage of a polypeptide to one or more ubiquitin monomers. Mono-ubiquitination has been associated with targeting of membrane proteins to the lysosome. Poly-ubiquitination, the process in which a chain of at least four ubiquitin peptides are attached to a lysine on a substrate protein, most commonly results in the degradation of the substrate protein via the proteasome.

Post translation modifications (PTMs) involve:

I) Addition of functional groups

Acetylation : the addition of an acetyl group, usually at the N-terminus of the protein for example, histones and tubulins. Histones are acetylated and deacetylated on lysine residues in the N-terminal tail. These reactions are catalyzed by enzymes with "histone acetyltransferase" (HAt) or "histone deacetylase" (HDAc) activity. The source of the acetyl group in histone acetylation is Acetyl-Coenzyme A is the source of the acteyl group. Tubulin acetylation and deacetylation system has been best studied in Chlamydomonas. Tubulin acetyltransferase located in the axoneme of the organism acetylates a specific lysine residue in the α-tubulin subunit of the assembled microtubule. A separate deacetylase is present in the cytosol of the organism to reverse the acetylation.

Alkylation: it involves the addition of an alkyl group (e.g. methyl, ethyl). Addition of methyl group (methylation) usually occurs at the lysine or the arginine residues.

Biotinylation : it is the acylation of conserved lysine residues with a biotin moiety

Glutamylation: it is the covalent linkage of glutamic acid residues to proteins such as tubulin.

Glycylation: it is the covalent linkage of one to more than 40 glycine residues to proteins for e.g. as seen in the C terminal tail of tubulin.

Glycosylation: the addition of a glycosyl group to asparagines / hydroxylysine / serine or threonine, resulting in a glycoprotein

Isoprenylation: the addition of an isoprenoid group (e.g. farnesol and geranylgeraniol) lipoylation, attachment of a lipoate functionality

Phosphopantetheinylation: the addition of a 4'-phosphopantetheinyl moiety from coenzyme A, as in fatty acid, polyketide, non-ribosomal peptide and leucine biosynthesis

Phosphorylation: the addition of a phosphate group, usually to serine, tyrosine, threonine or histidine

Sulfation: the addition of a sulfate group to a tyrosine.

Myristoylation: in this case a myristoyl group (derived from myristic acid) is covalently attached via an amide bond to the alpha-amino group of an N-terminal glycine residue of a nascent polypeptide. The reaction is catalysed by the enzyme N-myristoyltransferase. This process plays a vital role in membrane targeting and signal transduction in plant responses to environmental stress.

II) Change in the chemical nature of amino acids

Citrullation: also known as **deimination** is the conversion of arginine to citrulline.

Deamidation: it is the conversion of glutamine to glutamic acid or asparagine to aspartic acid

III) Change in structure

Formation of **disulfide bridges** (the covalent linkage of two cysteine amino acids) for e.g. insulin. It is secreted from the pancreas it has a prepeptide. Following cleavage of the 24 amino acid signal peptide the protein folds into proinsulin. Proinsulin is further cleaved yielding active insulin, which is composed of two peptide chains linked together through disulfide bonds.

IV) Vitamin C-Dependent Modifications

Vitamin C acts as a cofactor in hydroxylation of proline and lysine and carboxy terminal amidation. The hydroxylating enzymes involved are prolyl hydroxylase and lysyl hydroxylase. Collagens are an important group of hydroxylated proteins. Hormones such as oxytocin and vasopressin have C-terminal amidation.

V) Vitamin K-Dependent Modifications

Vitamin K is a cofactor in the carboxylation of glutamic acid residues resulting in the formation gla residues. These residues are found in the proteins of the blood clotting cascades. They allow the protein to chelate calcium ions and thereby render an altered conformation and biological activity to the protein.

Suggested Readings

1. Anderson, W. F., Ohlendorf, D. H.,Takeda, Y., Matthews, B. W. (1981) Structure of the cro repressor from bacteriophage and its interaction with DNA. *Nature* **290**,754-58.
2. Bailey, M.J., Hughes, C., and Koronakis, V. (2000) In vitro recruitment of the RfaH regulatory protein into a specialised transcription complex, directed by the nucleic acid ops element. *Mol Gen Genet* **262**, 1052–1059.
3. Burke TW, Kadonaga JT. (1997) The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila. *Genes Dev.* **11**, 3020-31.
4. Burley S.K., Roeder R.G. (1996) Biochemistry and structural biology of transcription factor IID (TFIID) *Annu Rev Biochem*. **65,** 769-99.
5. Galli G, Hofstetter H, Birnstiel ML. (1981) Two conserved sequence blocks within eukaryotic tRNA genes are major promoter elements. *Nature*. **294**, 626-31.
6. Geiger JH, Hahn S, Lee S, Sigler PB. (1996) Crystal structure of the yeast TFIIA/TBP/DNA complex. *Science*. **272,** 830-6.
7. Lobo SM, Hernandez N. (1989) A 7 bp mutation converts a human RNA polymerase II snRNA promoter into an RNA polymerase III promoter. *Cell.* **58,** 55-67.
8. Martinez E. (2002) Multi-protein complexes in eukaryotic gene transcription. *Plant Mol Biol.* **50**, 925-47.
9. McKay, D. B., Steitz, T. A. (1981) Structure of catabolite gene activator protein at 2.9 Å resolution suggests binding to left-handed B-DNA *Nature* **290,**744~49 J.
10. Mendez, R. & Richter, J. D. (2001) Translational control by CPEB: a means to the end. *Nature Rev. Mol. Cell Biol.* **2**, 521–529.
11. Miller, J., McLachlan, A. D., Klug, A (1985) Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *EMBO J*. **4**, 1609-14.
12. Nechaev S, Severinov K. (2003) Bacteriophage-induced modifications of host RNA polymerase. *Annu Rev Microbiol*. **57,**301-22.
13. Nikolov DB, Chen H, Halay ED, Usheva AA, Hisatake K, Lee DK, Roeder RG, Burley SK. (1995) Crystal structure of a TFIIB-TBP-TATA-element ternary complex. *Nature* **377**, 119-28.
14. Nudler E, Gottesman ME. (2002) Transcription termination and anti-termination in E. coli. *Genes Cells*. **7**,755-68.
15. Oelgeschlager T, Chiang CM, Roeder RG. (1996) Topology and reorganization of a human TFIID-promoter complex. *Nature*. **382**, 735-8.
16. Ohlendorf, D. H., Anderson, W. F., Fisher, R. G., Takeda, Y., Matthews, B. W. (1982) The molecular basis of DNA−protein recognition inferred from the structure of cro repressor *Nature* **298,**718-23.
17. Pabo, C. O., Lewis, M. (1982) The operator-binding domain of lambda repressor: structure and DNA recognition. *Nature* **298**, 443-4753.
18. Rauscher FJ 3rd, Cohen DR, Curran T, Bos TJ, Vogt PK, Bohmann D, Tjian R, Franza BR Jr. (1988) Fos-associated protein p39 is the product of the jun proto-oncogene *Science* **240**,1010-16.
19. Roberts J, Park JS. (2004) Mfd, the bacterial transcription repair-coupling factor: translocation, repair and termination. *Curr Opin Microbiol.***7**, 120-5.
20. Ross W, Gosink KK, Salomon J, Igarashi K, Zou C, Ishihama A, Severinov K, Gourse RL. (1993) A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase. *Science.* **262**,1407-13.
21. Sakonju S, Bogenhagen DF, Brown DD. (1980) A control region in the center of the 5S RNA gene directs specific initiation of transcription: I. The 5' border of the region. *Cell.* **19**, 13-25.
22. Sharp S, DeFranco D, Silberklang M, Hosbach HA, Schmidt T, Kubli E, Gergen JP, Wensink PC, Soll D. (1981) The initiator tRNA genes of Drosophila melanogaster: evidence for a tRNA pseudogene. *Nucleic Acids Res*. **9**, 5867-82
23. Squires CL, Zaporojets D. (2000) Proteins shared by the transcription and translation machines. Annu Rev Microbiol. **54**,775-98.
24. Sukhodolets MV, Cabrera JE, Zhi H, Jin DJ. (2001) RapA, a bacterial homolog of SWI2/SNF2, stimulates RNA polymerase recycling in transcription. *Genes Dev* **15**,3330-41.
25. Tan S, Hunziker Y, Sargent DF, Richmond TJ. (1996) Crystal structure of a yeast TFIIA /TBP / DNA complex. *Nature.* **381**, 127-51.
26. Tsai FT, Sigler PB. (2000) Structural basis of preinitiation complex assembly on human pol II promoters. *EMBO J.* **19**, 25-36.

27. Verrijzer CP, Tjian R. (1996) TAFs mediate transcriptional activation and promoter selectivity. *Trends Biochem Sci*. **21**,338-42
28. Yudkovsky N, Ranish JA, Hahn S. (2000) A transcription reinitiation intermediate that is stabilized by activator *Nature*. **408**, 225-9.