SEM1 CC2

Transcription

In 1958, Francis Crick enunciated the "central dogma of molecular biology. This scheme outlined the residue-by-residue transfer of biological information as encoded in the primary structure of the informational biopolymers, nucleic acids and proteins. The predominant path of information transfer, DNA → RNA → protein, postulated that RNA was an information carrier between DNA and proteins, the agents of biological function. In 1961, François Jacob and Jacques Monod extended this hypothesis to predict that the RNA intermediate, which they dubbed messenger RNA, or mRNA, would have the following properties:

1. Its base composition would reflect the base composition of DNA (a property consistent with genes as protein-encoding units).

2. It would be very heterogeneous with respect to molecular mass, yet the average molecular mass would be several hundred kD. (A 200-kD RNA contains roughly 750 nucleotides, which could encode a protein of about 250 amino acids—approximately 30 kD—a reasonable estimate for the average size of polypeptides.)

3. It would be able to associate with ribosomes because ribosomes are the site of protein synthesis.

4. It would have a high rate of turnover. (That is, mRNA would be rapidly degraded. Turnover of mRNA would allow the rate of mRNA synthesis to control the rate of protein synthesis.)

Since Jacob and Monod's 1961 hypothesis, it has been realized that cells contain three major classes of RNA—mRNA, ribosomal RNA (rRNA), and transfer RNA (tRNA)—all of which participate in protein synthesis (All of these RNAs are synthesized from DNA templates by DNA-dependent RNA polymerases in the process known as transcription

Transcription is the process of creating an equivalent RNA copy of a sequence of DNA. Both RNA and DNA are nucleic acids, which use base pairs of nucleotides as a complementary language that can be converted back and forth from DNA to RNA by the action of the correct enzymes. During transcription, a

DNA sequence is read by RNA polymerase, which produces a complementary, antiparallel RNA strand. As opposed to DNA replication, transcription results in an RNA complement that includes uracil (U) in all instances where thymine (T) would have occurred in a DNA complement.

Transcription can be explained easily in 4 or 5 simple steps, each moving like a wave along the DNA.

Step 1: DNA unwinds/"unzips" as the Hydrogen Bonds Break.

Step 2: The free nucleotides of the RNA, pair with complementary DNA bases.

Step 3: RNA sugar-phosphate backbone forms. (Aided by RNA Polymerase.)

Step 4: Hydrogen bonds of the untwisted RNA+DNA "ladder" break, freeing the new RNA.

Step 5: If the cell has a nucleus, the RNA is further processed and then moves through the small nuclear pores to the cytoplasm.

Transcription is the first step leading to gene expression. The stretch of DNA transcribed into an RNA molecule is called a *transcription unit* and encodes at least one gene. If the gene transcribed encodes a protein, the result of transcription is messenger RNA (mRNA), which will then be used to create that protein via the process of translation. Alternatively, the transcribed gene may encode for either ribosomal RNA (rRNA) or transfer RNA (tRNA), other components of the protein-assembly process, or other ribozymes.

A DNA transcription unit encoding for a protein contains not only the sequence that will eventually be directly translated into the protein (the *coding sequence*) but also *regulatory sequences* that direct and regulate the synthesis of that protein. The regulatory sequence before (upstream from) the coding sequence is called the five prime untranslated region (5'UTR), and the sequence following (downstream from) the coding sequence is called the three prime untranslated region (3'UTR).

Transcription has some proofreading mechanisms, but they are fewer and less effective than the controls for copying DNA; therefore, transcription has a lower copying fidelity than DNA replication.

As in DNA replication, DNA is read from 3' → 5' during transcription. Meanwhile, the complementary RNA is created from the 5' → 3' direction. Although DNA is arranged as two antiparallel strands in a double helix, only one of the two DNA strands, called the template strand, is used for transcription. The other DNA strand is called the coding strand, because its sequence is the same as the newly created RNA transcript (except for the substitution of uracil for thymine).

Transcription is divided into 5 stages: *pre-initiation*, *initiation*, *promoter clearance*, *elongation* and *termination*.

Transcription in Prokaryotes

In prokaryotes, virtually all RNA is synthesized by a single species of DNA-dependent RNA polymerase. (The only exception is the short RNA primers formed by primase during DNA replication.) Like DNA polymerases, RNA polymerase links ribonucleoside 5'-triphosphates (ATP, GTP, CTP, and UTP, represented generically as NTPs) in an order specified by base pairing with a DNA template:

$$n\ \text{NTP} \rightarrow (\text{NMP})_{n} + n\ \text{PP}_{\text{i}}$$

The enzyme moves along a DNA strand in the 3' → 5' direction, joining the 5'-phosphate of an incoming ribonucleotide to the 3'-OH of the previous residue. Thus, the RNA chain grows 5' → 3' during transcription, just as DNA chains do during replication. The reaction is driven by subsequent hydrolysis of PP_i to inorganic phosphate by ubiquitous pyrophosphatase activity.

The Structure and Function of *Escherichia coli* RNA Polymerase

The RNA polymerase of E. *coli* , so-called RNA polymerase holoenzyme**,** is a complex multimeric protein (450 kD) large enough to be visible in the electron microscope. Its subunit composition is $a_2$bb's. The largest subunit, b' (155 kD), functions in DNA binding; b (151 kD) binds the nucleoside triphosphate substrates and interacts with s (70 kD). Any of a number of related proteins, the sigma (σ) factors**,** can serve as the s subunit. Sigma subunits function in recognizing specific sequences on DNA called promoters that identify the location of transcription start sites, where transcription begins. Both b and b' contribute to formation of the catalytic site. The two a subunits (36.5 kD each) are essential for assembly of the enzyme and activation by some regulatory proteins. Dissociation of the s subunit from

the holoenzyme leaves the so-called core polymerase ($\alpha_2\beta\beta'$), which is catalytically competent but unable to recognize promoters.

Binding of RNA Polymerase to Template DNA

The process of transcription begins when the σ subunit of RNA polymerase recognizes a promoter sequence , and RNApolymerase holoenzyme and the promoter form a so-called closed promoter complex). Dissociation constants for RNA polymerase holoenzyme:closed promoter complexes range from 10^{26} to 10^{29} M. This stage in RNA polymerase:DNA interaction is referred to as the closed promoter complex because the DNA strands must be unwound so that the RNA polymerase can read and transcribe the DNA template strand into a complementary RNA sequence.

Once the closed promoter complex is established, the RNA polymerase holoenzyme unwinds about 14 base pairs of DNA (base pairs located at positions -10 to +2, relative to the transcription start site, forming the very stable open promoter complex. In this complex, RNA polymerase holoenzyme is bound very tightly to the DNA (K_D » 10^{-14} *M*).

PROPERTIES OF PROKARYOTIC PROMOTERS.

Prokaryotic promoters vary in size from 20 to 200 bp, but typically consist of a 40-bp region located on the 5'-side of the ranscription start site. Within the promoter are two consensus sequence elements. (A consensus sequence can be defined as *the bases that appear with highest frequency at each position when a series of sequences believed to have common function are compared.*) These two elements are the Pribnow box near -10, whose consensus sequence is the hexameric TATAAT, and a sequence in the 235 region containing the hexameric consensus TTGACA . The Pribnow box and the -35 region are separated by about 17 bp of nonconserved sequence. RNA polymerase holoenzyme uses its *s* subunit to bind to these sequences, and the more closely the -35 region sequence corresponds to its consensus sequence, the greater is the efficiency of transcription of the gene.

Note:-The highly expressed rrn genes in E. *coli* which encode ribosomal RNA (rRNA) have a third sequence element in their promoters, the upstream element (UP element), located about 20 bp immediately upstream of the -35 region. Whereas the *σ* subunit recognizes the -10 and -35 elements, the C-terminal domains (CTD) of α subunits of RNA polymerase recognize and bind the UP element.

In order for transcription to begin, the DNA duplex must be “opened” so that RNA polymerase has access to single-stranded template. The efficiency of initiation is inversely proportional to the melting temperature, T_m in the Pribnow box, suggesting that the A:T-rich nature of this region is aptly suited for facile “melting” of the DNA duplex and creation of the open promoter complex . Negative supercoiling facilitates transcription initiation by favoring DNA unwinding. The RNA polymerase s subunit is directly involved in melting the dsDNA. Interaction of the s subunit with the non template strand maintains the open complex formed between RNA polymerase and promoter DNA, with the s subunit acting as a sequence-specific single-stranded DNA-binding protein. Such association of the s subunit with the non template strand stabilizes the open promoter complex and leaves the bases along the template strand available to the catalytic site of the RNA polymerase.

Initiation of Polymerization

RNA polymerase has two binding sites for NTPs—the initiation site and the elongation site. The initiation site binds the purine nucleotides ATP and GTP preferentially; most RNAs begin with a purine at the 5'-end. The first nucleotide binds at the initiation site, H-bonding with the +1 base exposed within the open promoter complex. The second incoming nucleotide binds at the elongation site, H-bonding with the +2 base. The ribonucleotides are then united when the 3'-O of the first nucleotide makes a nucleophilic attack on the a-phosphorus atom of the second nucleotide. A phosphoester bond is formed, and PP_i is eliminated. Movement of RNA polymerase along the template strand (translocation) to the next base prepares the RNA polymerase to add the next nucleotide .Once an oligonucleotide 6 to 10 residues long has been formed, the *s* subunit dissociates from RNA polymerase, signaling the completion of initiation. The core RNA polymerase goes on to synthesize the remainder of the mRNA. As the core RNA polymerase progresses, advancing the 3'-end of the RNA chain, the DNA duplex is unwound just ahead of it. About 12 base pairs of the growing RNA remain base-paired to the DNA template at any time, with the RNA strand becoming displaced as the DNA duplex rewinds behind the advancing RNA polymerase.

Note: - Rifamycin B and its analog, rifampicin**,** are inhibitors of initiation. Despite their structural similarity (Figure 31.4), they act in different ways. Rifamycin binds to the b subunit of RNA polymerase and blocks binding of incoming NTP at the initiation site. Rifampicin allows the first phosphodiester bond to be formed, but it prevents the translocation of RNA polymerase along the DNA template. However, once the second phosphodiester bond is formed, creating an RNA trinucleotide, rifampicin is without effect.

Promoter clearance

After the first bond is synthesized, the RNA polymerase must clear the promoter. During this time there is a tendency to release the RNA transcript and produce truncated transcripts. This is called *abortive initiation* and is common for both eukaryotes and prokaryotes. Abortive initiation continues to occur until the σ factor rearranges, resulting in the transcription elongation complex (which gives a 35 bp moving footprint). The σ factor is released before 80 nucleotides of mRNA are synthesized. Once the transcript reaches approximately 23 nucleotides, it no longer slips and elongation can occur.

Elongation

Elongation of the RNA transcript is catalyzed by the core polymerase, because once a short oligonucleotide chain has been synthesized, the σ subunit dissociates. One strand of the DNA, the *template strand* is used as a template for RNA synthesis. As transcription proceeds, RNA polymerase traverses the template strand and uses base pairing complementarity with the DNA template to create an RNA copy. Although RNA polymerase traverses the template strand from 3' → 5', the coding strand and newly-formed RNA can also be used as reference points, so transcription can be described as occurring 5' → 3'. This produces an RNA molecule from 5' → 3', an exact copy of the coding strand (except that thymines are replaced with uracils, and the nucleotides are composed of a ribose (5-carbon) sugar).Unlike DNA replication, mRNA transcription can involve multiple RNA polymerases on a single DNA template and multiple rounds of transcription (amplification of particular mRNA), so many mRNA molecules can be rapidly produced from a single copy of a gene.

Chain elongation does not proceed at a constant rate, but varies between 20 to 50 nucleotides per second. The RNA polymerase slows down and even pauses in G:C-rich regions due to the greater difficulty in unwinding G:C base pairs. As the RNA polymerase moves along the template, the DNA double helix is unwound ahead of it and recloses after the polymerase has passed by. Only a short stretch of RNA:DNA hybrid duplex exists at any time. Two possibilities can be envisioned for the course of the new RNA chain. In one, the RNA chain is wrapped around the DNA as the RNA polymerase follows the template strand around the axis of the DNA duplex, but this possibility seems unlikely due to its

potential for tangling the nucleic acid strands. The more likely possibility involves supercoiling of the DNA, so that positive supercoils are created ahead of the transcription bubble and negative supercoils are created behind it .To prevent torsional stress from inhibiting transcription, topoisomerases act to remove these supercoils from the DNA segment undergoing transcription.

Termination

Two types of transcription termination mechanisms operate in bacteria: one that is dependent on a specific protein termination factor called ρ and another that is not dependent on this protein. In the latter, termination of transcription is determined by specific sequences in the DNA called termination sites. These sites are not characterized by a unique base where transcription halts. Instead, these sites consist of three structural features whose base-pairing possibilities lead to termination:

1. Inverted repeats, which are typically G:C-rich, so a stable stem-loop structure can form in the transcript via intrachain hydrogen bonding .

2. A nonrepeating segment that punctuates the inverted repeats.

3. A run of 6 to 8 As in the DNA template, coding for Us in the transcript..

Termination then occurs as follows: A G:C-rich, stem-loop structure, or "hair-pin," forms in the transcript. The hairpin apparently causes the RNA polymerase to pause, whereupon the A:U base pairs between the transcript and the DNA template strand are displaced through formation of somewhat more stable A:T base pairs between the template and nontemplate strands of the DNA. The result is spontaneous dissociation of the nascent transcript from DNA. The alternative mechanism of termination, ρ factor-dependent termination, is less common and mechanistically more complex. ρ Factor is an ATP-dependent helicase (hexamer of 50-kD subunits) that catalyzes the unwinding of RNA:DNA hybrid duplexes (or RNA:RNA duplexes). The ρ factor recognizes and binds to C-rich regions in the RNA transcript. These regions must lack secondary structure and be unoccupied by translating ribosomes for ρ factor to bind. Once bound, ρ factor advances in the 5' → 3' direction until it reaches the transcription bubble. There it catalyzes the unwinding of the transcript and template, releasing the nascent RNA chain. It is likely that the RNA polymerase stalls in a G:C-rich termination region, allowing r factor to overtake it.

Transcription in Eukaryotes

Eukaryotic cells have three classes of RNA polymerase, each of which synthesizes a different class of RNA. All three enzymes are found in the nucleus. RNA polymerase I is localized to the nucleolus and transcribes the major ribosomal RNA genes. RNA polymerase II transcribes protein-encoding genes, and thus it is responsible for the synthesis of mRNA. RNA polymerase III transcribes tRNA genes, the ribosomal RNA genes encoding 5S rRNA, and a variety of other small RNAs, including several involved in mRNA processing and protein transport. All three RNA polymerase types are large, complex multimeric proteins (500 to 700 kD), consisting of 10 or more types of subunits. Although the three differ in overall subunit composition, they have several smaller subunits in common. Further, all possess two large subunits (each 140 kD or greater) having sequence similarity to the large b and b' subunits of E. *coli* RNA polymerase, indicating that the fundamental catalytic site of RNA polymerase is conserved among its various forms.

In addition to their different functions, the three classes of RNA polymerase can be distinguished by their sensitivity to α-amanitin, a bicyclic octapeptide produced by the poisonous mushroom *Amanita phalloides*. α-Amanitin blocks RNA chain elongation. Although RNA polymerase I is resistant to this compound, RNA polymerase II is very sensitive and RNA pol III is less sensitive. The existence of three classes of RNA polymerases acting on three distinct sets of genes implies that at least three categories of promoters exist to maintain this specificity. All three polymerases interact with their promoters via so-called transcription factors, DNA-binding proteins that recognize and accurately initiate transcription at specific promoter sequences. For RNA polymerase I, its templates are the rRNA genes. Ribosomal RNA genes are present in multiple copies. Optimal expression of these genes requires the first 150 nucleotides in the immediate 5'-upstream region, but the precise locations and sequences of the promoter(s) are not known with certainty. RNA polymerase III interacts with transcription factors TFIIIA, TFIIIB, and TFIIIC. Interestingly, TFIIIA and/or TFIIIC bind to specific recognition sequences that in some instances are located *within* the coding regions of the genes, not in the 5'-untranscribed region upstream from the transcription start site. TFIIIB associates with TFIIIA or TFIIIC already bound to the DNA and in turn facilitates the association of RNA pol III to establish an initiation complex.

The Structure and Function of RNA Polymerase II

As the enzyme responsible for the regulated synthesis of mRNA, RNA polymerase II has aroused greater interest than RNA pol I and pol III. RNA pol II must be capable of transcribing a great diversity of genes, yet it must carry out its function at any moment only on those genes whose products are appropriate to

the needs of the cell in its ever changing metabolism and growth. The RNA pol II from yeast (*Saccharomyces cerevisiae*) has been extensively characterized. The yeast RNA pol II consists of 10 different polypeptides, designated RPB1 through RPB10, ranging in size from 220 to 10 kD. RPB1 and RPB2 functions are homologous to those of the prokaryotic RNA polymerase β and β' subunits: RPB1 has a DNA-binding site, RPB2 binds nucleotide substrates, and both contribute to the catalytic site. RPB3 is the functional homolog of the prokaryotic α; there are two RPB3 subunits per enzyme and RPB3 is essential for assembly of the polymerase. RPB4 resembles σ subunit in amino acid sequence. RPB 3, 4, and 7 are unique to RNA pol II, whereas RPB 5, 6, 8, and 10 are common to all three eukaryotic RNA polymerases. RPB4 and 7 readily dissociate from RNA pol II. The RPB1 subunit has an unusual structural feature not found in prokaryotes: Its *C-terminal domain* (CTD) contains 27 repeats of the amino acid sequence PTSPSYS. (The analogous subunit in RNA pol II enzymes of other eukaryotes has this heptapeptide tandemly repeated as many as 52 times.) The side chains of 5 of the 7 residues in this repeat have -OH groups, endowing the CTD with considerable hydrophilicity *and* multiple sites for phosphorylation. This domain may project more than 50 nm from the globular enzyme. The CTD is essential to RNA pol II function. Only RNA pol II whose CTD is not phosphorylated can initiate transcription. However, transcription elongation proceeds only after protein phosphorylation within the CTD, suggesting that phosphorylation triggers the conversion of an initiation complex into an elongation complex. Following termination of transcription, a phosphatase recycles RNA pol II to its unphosphorylated form.

Transcription Initiation by RNA Polymerase II

Promoters

RNA polymerase II promoters commonly consist of two separate sequence features, the core element, near the transcription start site, where general transcription factors bind, and more distantly located regulatory elements, known variously as enhancers or silencers. These latter elements are recognized by specific DNA-binding proteins that activate transcription above basal levels (enhancers) or repress transcription (silencers). The core region often consists of a TATA box (a TATAAA consensus element) and the transcription start site; the TATA motif is usually located at position -25. An important role of the TATA box is to indicate the site of the initiator element, or *Inr,* where transcription is initiated. The initiator element *Inr* encompasses the transcription start site. The sequence of *Inr* is not highly conserved between genes; a consensus *Inr* for one gene family is $_{-3}$YYCAYYYYY$+_{6}$ (where Y represents any

pyrimidine). Regulatory elements occurring near the core promoter (within 50 to 200 bp), the so-called promoter proximal elements, possess one or more binding sites for interaction with DNA-binding regulatory proteins and show great variation in sequence. Other regulatory elements, so-called distal enhancer (or silencer) elements, where another group of DNA-binding regulatory proteins bind, can be located far from the core promoter, either upstream or, rarely, downstream.

Initiation of Transcription in Eukaryotes

A universal set of proteins, called the basal apparatus, binds the core promoter and initiates transcription. The basal apparatus consists of RNA polymerase II and the *general transcription factors* (GTFs). There are six GTFs , five of which are required for transcription: TFIIB, TFIID, TFIIE, TFIIF, and TFIIH. One, TFIIA, stimulates transcription by stabilizing the interaction of TFIID with the TATA box. TFIID consists of TBP (TATA-binding protein), which directly recognizes the TATA box, and a set of TBP-associated factors (TAFs or TAFs), which have positive or negative effects on transcription; some are capable of recognizing core promoters lacking a TATA box. TBP binds to the core promoter through contacts made with the minor groove of the DNA, distorting and bending the DNA so that DNA sequences upstream and downstream of the TATA box come into closer proximity. In one model of transcription initiation, once TBP binds to core promoter, TFIIB joins it, followed by RNA polymerase IIA in association with TFIIF. Then other factors join, establishing a competent transcription *preinitiation complex.* Another model for transcription initiation suggests that RNA polymerase II holoenzyme (RNA polymerase IIA in association with various general transcription factors other than TBP or TFIID) assembles in the absence of any interaction with DNA and then binds to TBP/TFII. In either case, once RNA polymerase IIA and the GTFs have assembled into a preinitiation complex on DNA, an *open complex* then forms and transcription begins.

Elongation process in eukaryotes is more or less same like that of prokaryotes. Transcription termination in eukaryotes is less understood but involves cleavage of the new transcript followed by template-independent addition of As at its new 3' end, in a process called polyadenylation.

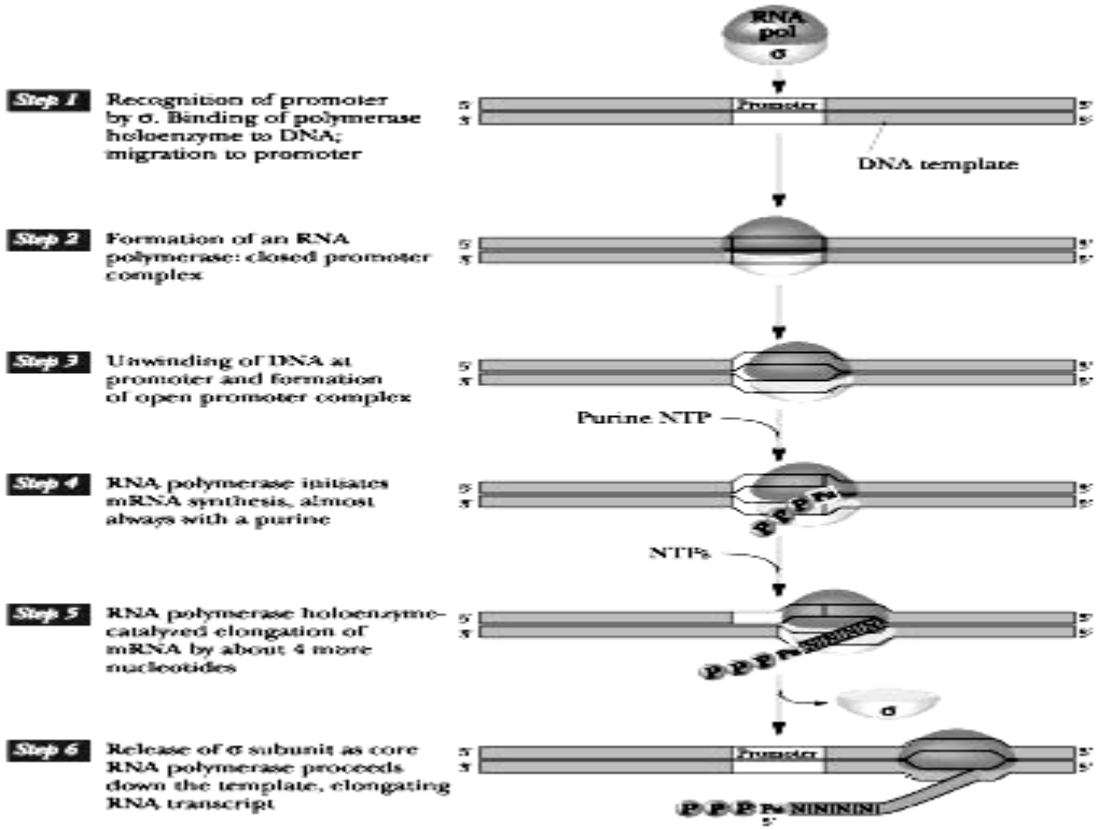

FIG-Sequence of events in the initiation and elongation phases of transcription as it occurs in prokaryotes

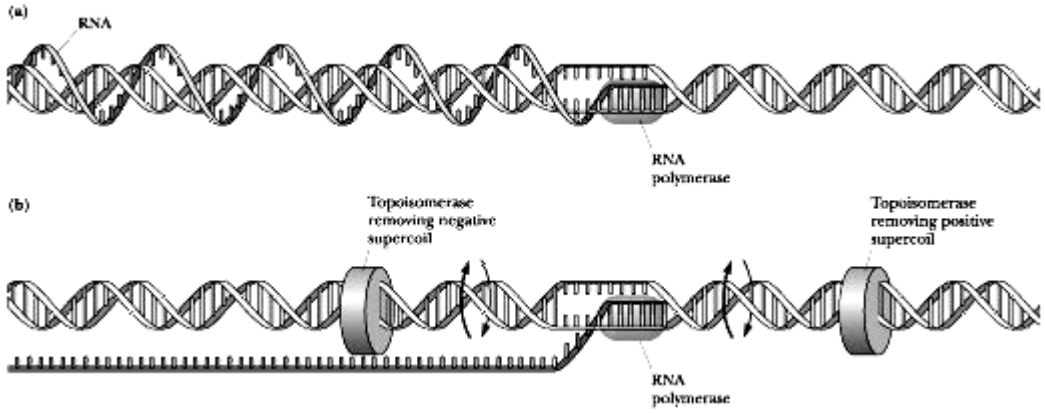

FIG- Supercoiling versus transcription. (a) If the RNA polymerase followed the template strand around the axis of the DNA duplex, no supercoiling of the DNA would occur, but the RNA chain would be wrapped around the double helix once every 10 bp. This possibility seems unlikely because it would be difficult to disentangle the transcript from the DNA duplex. (b) Alternatively, topoisomerases could remove the supercoils. A topo-isomerase capable of

relaxing positive supercoils situated ahead of the advancing transcription bubble would “relax” the DNA. A second topo-isomerase behind the bubble would remove the negative supercoils.

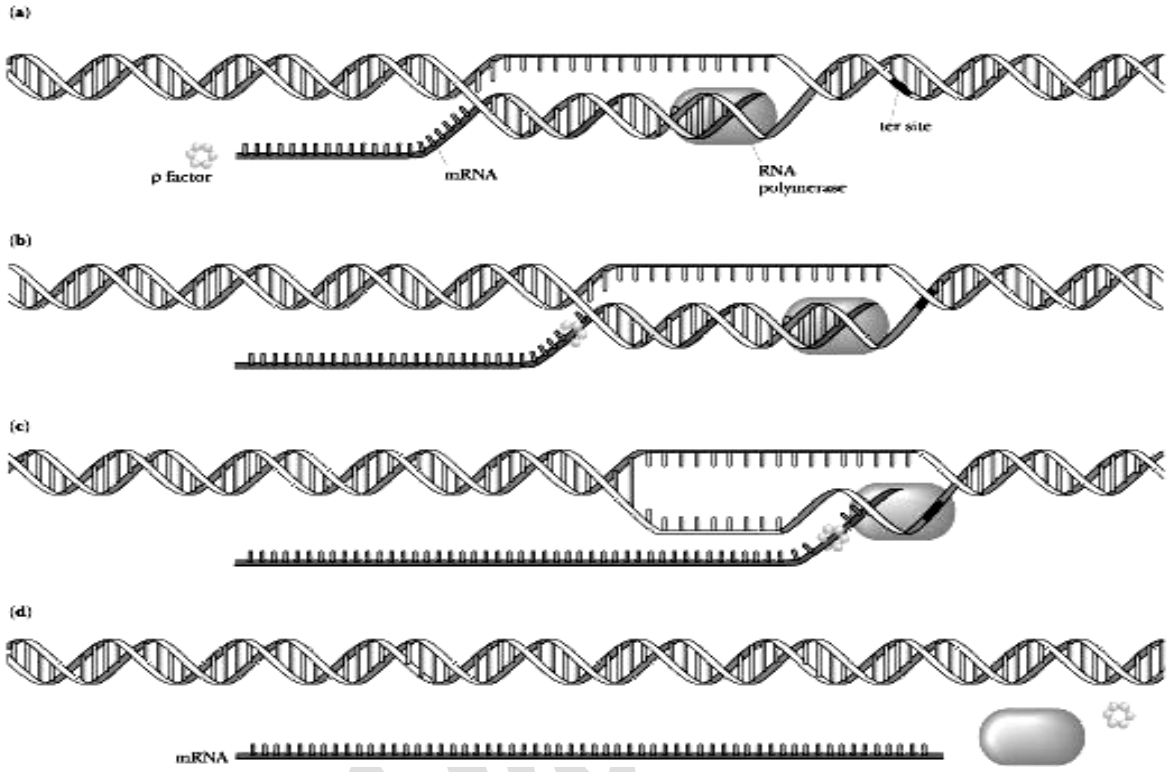

FIG- The ρ factor mechanism of transcription termination. ρ factor (a) attaches to a recognition site on mRNA and (b) moves along it behind RNA polymerase. (c) When RNA polymerase pauses at the termination site, ρ factor unwinds the DNA:RNA hybrid in the transcription bubble, (d) releasing the nascent mRNA.

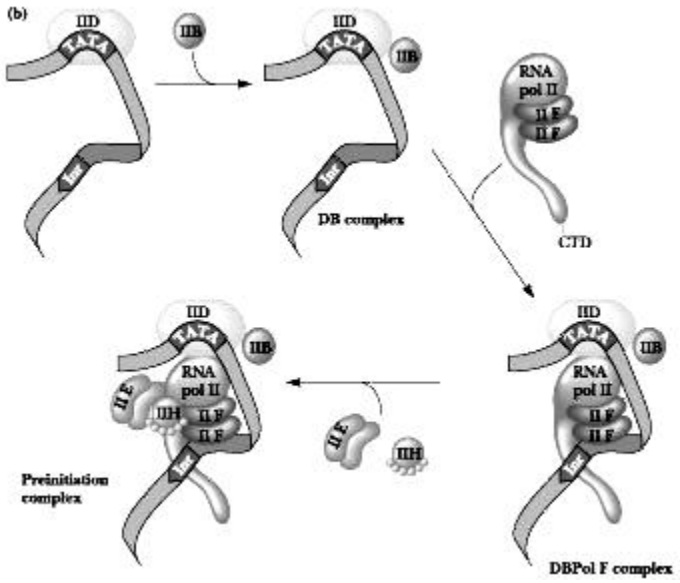

FIG- Formation of a preinitiation complex at a TATA-containing promoter. Binding of TFIID, the multisubunit protein (>100 kD) consisting of the TATA-binding protein (TBP) and other polypeptides, is stimulated by TFIIA. TFIID bound to the TATA motif recruits TFIIB, forming a DB complex. In association with TFIIF, RNA pol IIA (the nonphosphorylated form of RNA pol II) joins the DB complex to give the DBpol F complex. TFIIE and TFIIH then associate to yield the preinitiation complex. Melting of the DNA duplex around *Inr* generates the *open complex* and transcription ensues

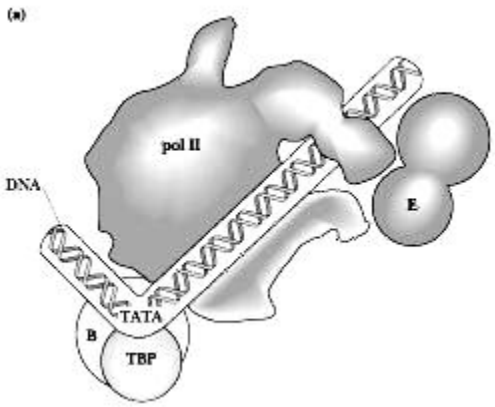

FIG- Structure of the preinitiation complex, showing the distortion of the DNA and the relative positions of RNA polymerase IIA (pol II), the TATA box, TBP, TFIIB (B), and the TFIIE dimer (E). Transcription initiation occurs at a site on DNA within the region encircled by RNA polymerase IIA next to TFIIE.