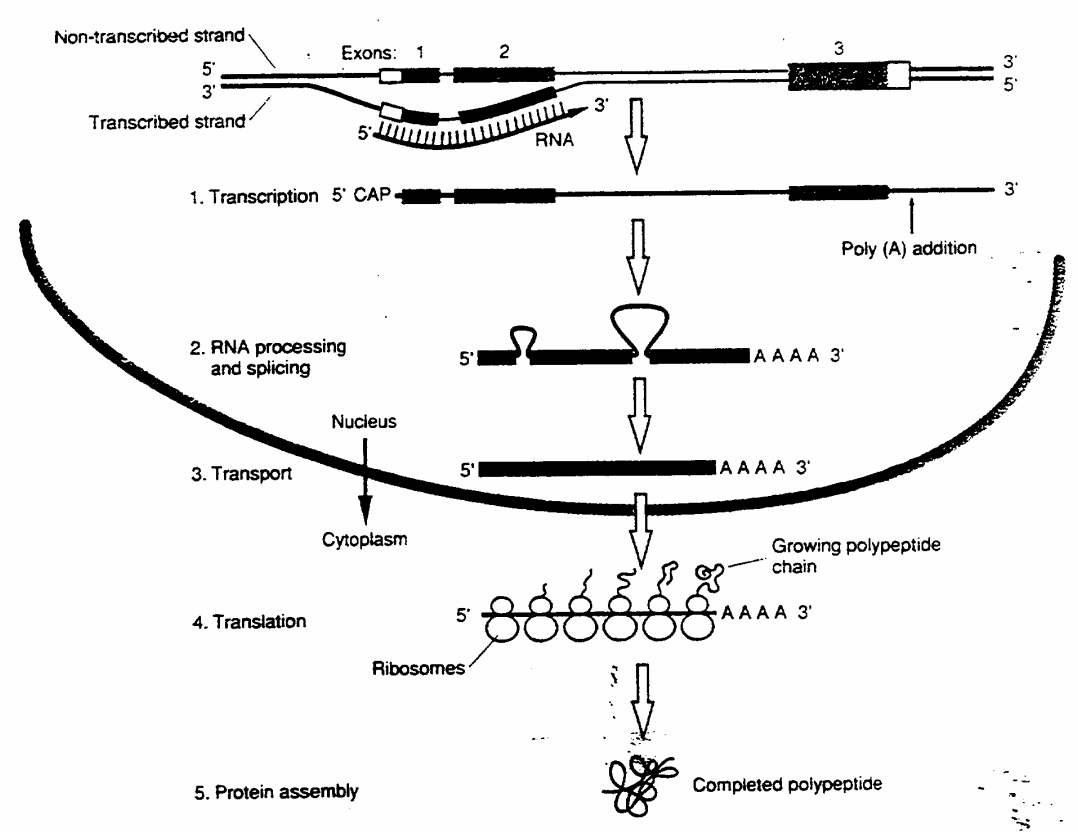
GENE EXPRESSION

The process of gene expression simply refers to the events that transfer the information content of the gene into the production of a functional product, usually a protein. Although there are genes whose functional product is an RNA, including the genes encoding the ribosomal RNAs as well as the transfer RNAs and certain other small RNAs, the vast majority of genes within the cell are protein-encoding genes.



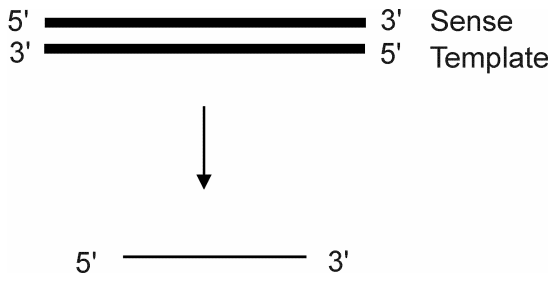
As shown in the figure above, the expression of a eukaryotic gene is a complex process involving a variety of steps prior to the actual synthesis of a protein. These include the transcription of the gene into the primary RNA product, processing of this initial gene transcript to remove intron sequences and create the mature 3' terminus, transport of the processed mRNA transcript to the cytoplasm, and then finally, translation of the messenger RNA into protein. With very few exceptions, all of the genes that encode proteins follow this pathway. This complexity is in sharp contrast to the relatively simple process of gene expression in prokaryotic cells where the product of the gene is the mRNA and there is no nucleus that separates the genetic material from the machinery necessary to synthesize proteins.

Thus, in a bacterium, protein synthesis actually begins on a nascent RNA molecule, well before the synthesis of the RNA is complete.

Transcription

The initial step in gene expression is the transcription of the DNA molecule into an exact RNA copy. As already discussed, the basic unit of heredity, the gene, is a double stranded DNA molecule and the information in the gene is encoded in the sequence of nucleotides. The transfer of information, to the ultimate synthesis of a protein, is accomplished via an RNA intermediate, the so-called messenger RNA. The mRNA molecule contains the exact same sequence of nucleotides as found in the DNA molecule (with U substituted for T). This occurs through the process known as transcription and is carried out by an enzyme termed DNA-dependent RNA polymerase.

The product of transcription is an RNA molecule that is identical in sequence content to one of the DNA strands (the sense strand) and complementary to the other DNA strand (the

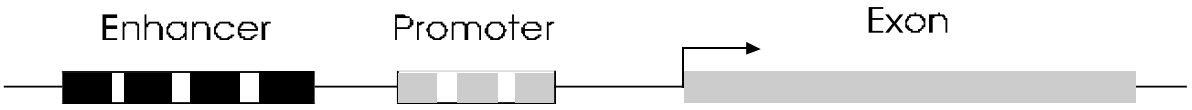


template strand). RNA differs from DNA in two respects. First, RNA contains a hydroxyl (OH) residue at the 2' position of the sugar moiety (ribose) whereas DNA contains an H at this position (deoxyribose). This changes some of the chemical properties of the two molecules. Second, RNA replaces the base thymine with uracil. Although they are chemically different, their base pairing properties (complementarity to adenosine) are the same. Thus, in DNA, A pairs with T whereas in RNA, A pairs with U. Transcription always proceeds in a 5' to 3' direction with respect to polarity of the nucleotides in the RNA. Thus, an unmodified primary transcription product would contain a 5' end with a triphosphate and a 3' end with a OH.

There are actually three distinct forms of RNA polymerase found in eukaryotic cells that are responsible for the transcription of three distinct types of genes: the genes encoding the ribosomal RNAs are transcribed by RNA polymerase I; the protein-encoding genes that produce

the messenger RNAs are transcribed by RNA polymerase II; finally, the genes encoding transfer RNAs as well as certain other small RNA molecules are transcribed by RNA polymerase III.

Transcription involves three distinct steps. First, there must be a recognition of the gene by the RNA polymerase. This is accomplished via the interaction of a variety of proteins called transcription factors that provide the recognition step, guiding the polymerase to the correct site. Moreover, the interaction of these transcription factors with their DNA recognition sequences represents a rate-limiting step in the process of transcription initiation. Sequences in the DNA which are recognized by these transcription factors, and which serve as binding sites for the transcription factors, are generally upstream (5') of the start site for transcription and are called ***promoters***. Additional transcription factors can bind to sequence elements called ***enhancers*** that may be located further upstream or even downstream of the gene. In either case, the DNA sequences that bind the transcription factors must be located on the same DNA molecule, thus the same chromosome, as the gene which is regulated. As such, these sequences are said to act in *cis*. The promoter is absolutely essential for transcription whereas the enhancer, as its name implies, increases the efficiency of transcription. A critical component of many promoters is the sequence element TATA which is located nearest to the site of transcription initiation (about 30 nucleotides away). The TATA element binds a general transcription factor known as TBP (TATA-binding protein) which is a component of a larger complex known as TFIID.



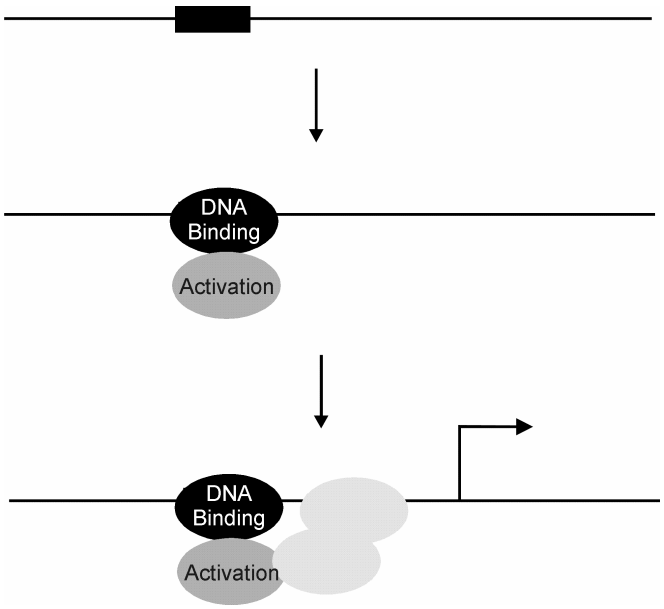
Second, the polymerase must initiate transcription. This does not involve a primer molecule as is the case for DNA synthesis but rather starts *de novo* at a specific site in the gene. This site is dictated as a result of the interaction of the RNA polymerase at a specific site, guided by the transcription factors that have bound to the promoter and enhancer sequences.

Third, the polymerase must complete the transcription of the gene and then terminate transcription. In some cases, the termination of transcription is precise whereas in other cases it can occur heterogeneously over a broad region of DNA. It is also true that transcription can terminate prematurely, prior to the complete synthesis of the RNA transcript, thus precluding the formation of a functional RNA molecule, and therefore serving as a mechanism to regulate transcription.

Cis-Acting Elements: As discussed previously, transcription is governed by DNA sequences usually located upstream (5') of the coding sequences. These are binding sites for proteins that stimulate transcription by allowing RNA polymerase to interact. They must be physically linked to the gene in order to have an effect and thus are said to function in *cis*.

Trans-Acting Transcription Factors: These are the proteins that recognize and bind to the cis-acting promoter and enhancer elements. There are usually multiple cis-acting elements, and thus multiple trans-acting factors, that are essential for transcription of a given gene. The binding of factors to these elements facilitates the interaction of RNA polymerase and thus the initiation of transcription. Thus, transcription initiation can be regulated by controlling the activity or the presence of these trans-acting factors.

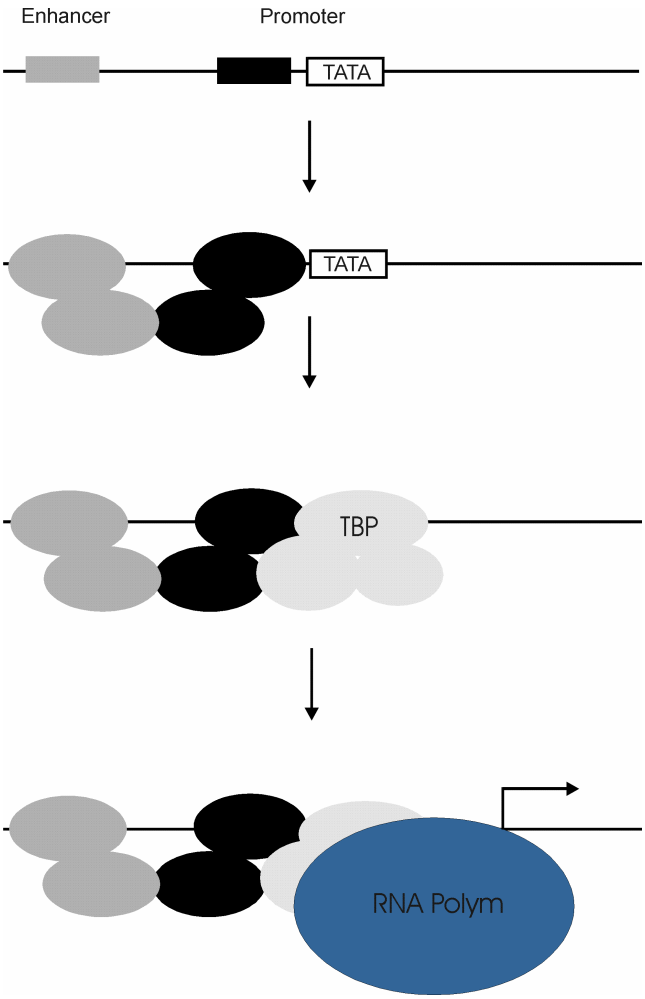
Transcription factors possess two essential properties - the ability to bind to DNA in a sequence specific manner and the ability to cause a stimulation of transcription. Generally, these two properties are a function of distinct and separable domains in the proteins. That is, one can often identify and define a domain of the protein, the transcription factor, that is essential for DNA binding and a separate, distinct domain of the protein that functions to activate transcription once bound to the DNA. The DNA binding domain is responsible for recognition and interaction with a specific linear DNA sequence in the enhancer or promoter. Typically, this involves a 4 to 8 nucleotide recognition sequence in the DNA. Although the interaction of the protein with the DNA is specific and can be detected by a variety of methods, it generally is a weak interaction that will readily dissociate. The capacity to stabilize this otherwise weak interaction is likely a critical aspect of transcription control whereby multiple factors must interact on the DNA to stabilize a functional complex. The domain of the transcription factors that is essential for transcription activation mediates protein-protein interactions with other components of the



transcription machinery. In particular, activation domains are often responsible for interaction with components of the basal transcription complex that includes the TFIID factor.

The Steps Involved in Transcription Initiation

Although variations in any step of gene expression can be regulatory, by far the most frequent form of gene control is the regulation of transcription initiation. Control of transcription involves the regulation of transcription factors that interact with the critical cis-acting sequences in the promoter and enhancer that then dictate the frequency of RNA polymerase binding to the gene and subsequent transcription. The process of promoter recognition and utilization involves a stepwise interaction of a complex series of transcription factors with the promoter to create a stable DNA-protein complex that allows RNA polymerase to initiate transcription.



Key in this process is the recognition of the specific sequences in the promoter DNA sequence (the cis-acting elements) by the trans-acting transcription factors. Given the fact that there are as many as 100,000 protein-encoding genes in the mammalian genome, it is obvious that the expression of each gene cannot be regulated by unique transcription factors. Rather, it would appear that a limited number of transcription factors are responsible and that the high degree of specificity is generated by specific protein-protein interactions that stabilize otherwise weak interactions on a promoter. Very likely, the ability of specific factors to interact provides the possibility of combinatorial interactions of promoter-specific factors that creates specificity to transcription control.

Although it was initially believe that the three types of eukaryotic genes (Pol I -ribosomal RNA, Pol II -mRNA, Pol III -tRNA were quite distinct with respect to factors involved in the

transcription initiation event, it is now clear that many of the activities involved in promoter recognition are shared. Moreover, although there are distinct RNA polymerases for the three types of genes, many of the subunits of the polymerases are also shared. What distinguishes the classes of genes most clearly is the complexity of regulatory elements and factors necessary for the transcription of the mRNA genes.

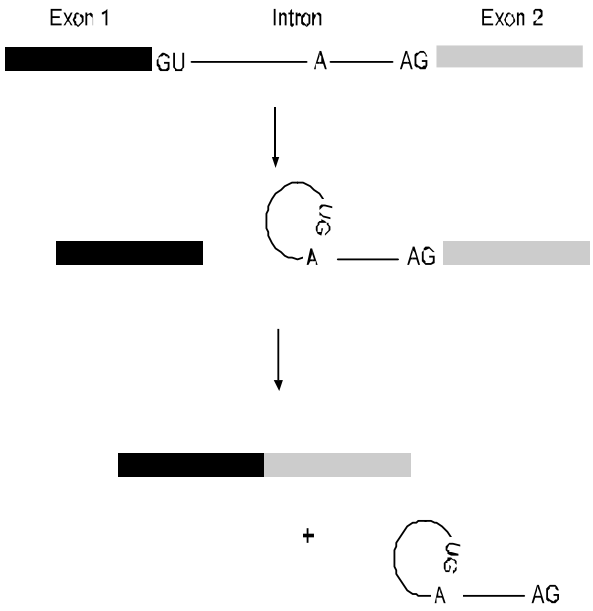
As stated before, transcription factors possess two essential properties - the ability to bind to DNA, recognizing the cis-acting elements, in a sequence specific manner and the ability to stimulate transcription. The fact that these two properties are a function of distinct and separable domains in the proteins has been exploited in a method to detect protein-protein interactions, the so-called yeast two-hybrid assay, in which the DNA sequences encoding the two functional domains of a yeast transcription factor have been separated in two vectors. Sequences are then cloned into these vectors to create chimeric fusion proteins as a method for selecting or assaying for sequences that will bring the two functional domains back together via protein-protein interaction.

Post-Transcriptional Events of Gene Expression

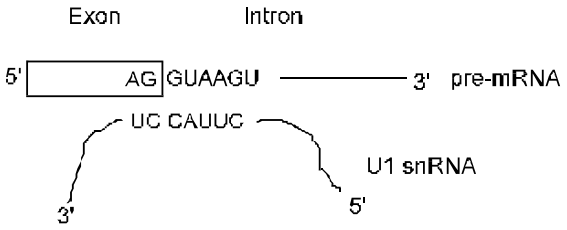
Whereas the initial transcript of a bacterial gene is the actual messenger RNA, the initial transcript of a eukaryotic gene must be altered in a variety of ways before it can function. Thus, post-transcriptional processing and modification events are critical to the formation of a eukaryotic mRNA.

RNA Processing

The products of all three types of eukaryotic genes are processed in a variety of ways. The most dramatic is the splicing of most of the protein coding transcripts, joining exon sequences together and removing intron sequences. Splicing of the pre-mRNA occurs via a step-wise series of cleavage and ligation events that remove the intron sequences and bring the exons together in a precise manner. The initial step involves the cleavage of the RNA at the exon 3' end/intron 5' end border. The free 5' end of the intron forms an unusual linkage with the 2' OH of an A residue near the 3' end of the intron, creating an intermediate that has a lariat structure. Subsequently, a cleavage is made at the 3' end of the intron, releasing the lariat intron, and then allow the two exons to be ligated.

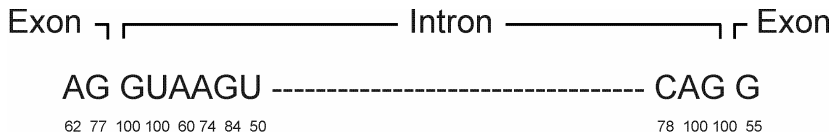


The splicing reaction is accomplished by a complex set of components that recognize specific sequences within the RNA at the splice site junctions. Many of the activities that mediate the splicing reaction are composed of RNA-protein particles, the so-called snRNPs. In fact, it appears that the RNA components are responsible for the recognition and perhaps the enzymatic cleavage of the pre-mRNA. This is particularly evident for the U1 snRNA that recognizes the 5' splice site and facilitates the initial cleavage. As shown below, the U1 RNA is



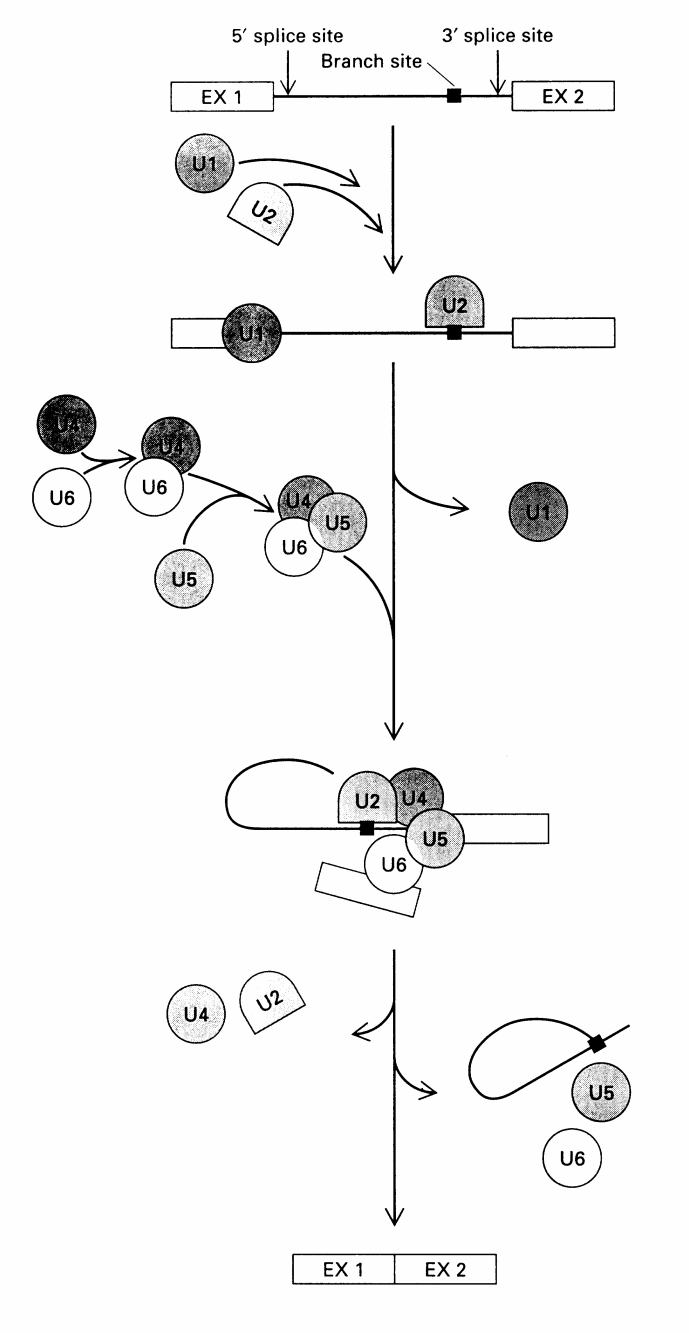
complementary to sequences at the exon-intron junction.

Indeed, the sequences found at the exon-intron boundaries in pre-mRNAs are very highly conserved as indicated in the figure below that summarizes the frequency of particular nucleotides at the exon/intron border of a very large number of genes.



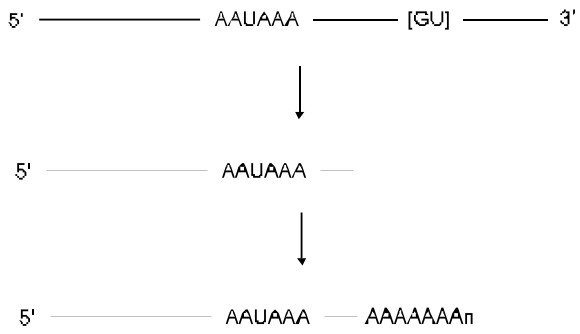
As can be seen, the two nucleotides at the ends of the intron, a GU at the 5' end and an AG at the 3' end, are absolutely conserved, found in 100% of the splice junctions. The nucleotides bordering these sequences, either in the exon or in the intron, are also conserved but less so.

Through the analysis of nuclear extracts that can carry out the splicing reaction in vitro, it has been possible to identify many of the components involved in splicing and to define a pathway for the splicing reaction. This involves a series of snRNPs that function at distinct points in the splicing reaction to recognize first the 5' splice site, then the 3' splice site together with the branch site, and carry out the various cutting and ligation events necessary to accurately splice the two exon sequences together.



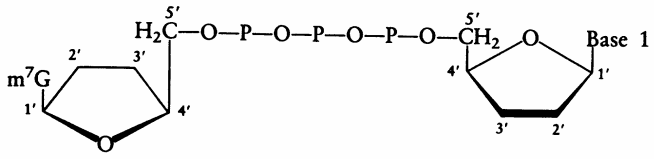
In addition to the splicing of exon sequences, the mature 3' terminus of the mRNA is created by a cleavage of the precursor RNA followed by the addition of a poly A tail. That is, transcription does not terminate at the sequence that represents the mature 3' terminus of the mRNA but rather continues some distance beyond. RNA cleavage then generates the correct 3' end. Poly A addition is post-transcriptional, being added after the cleavage has occurred and is

not directed by a poly T sequence in the DNA. As with splicing, this process involves a recognition of specific sequences in the pre-mRNA by a series of activities that then process the RNA and add the poly A tail.



RNA Modifications

In addition to the various steps that process the initial primary transcript, the mRNA is also modified in several ways. The 5' terminus of the transcript is capped by the addition of a modified GTP residue that forms a 5'-5' linkage. This probably serves to protect the RNA against degradation and it also clearly serves a role in the initiation protein synthesis by interacting with an essential factor. Internal adenosine residues in the RNA are modified by methylation but the function of these modifications is not known.



Finally, as already discussed, the 3' terminus of the RNA is modified by polyadenylation. This poly A tail of approximately 200 adenosine residues serves as a binding site for a protein that participates in protein synthesis, likely by facilitating the formation of an initiation complex with activities that bind to the 5? end of the mRNA. The poly A tail also serves to protect the RNA from degradation.

Nucleus-Cytoplasmic RNA Transport

Unlike bacteria, the eukaryotic cell is compartmentalized. Thus, the primary transcript is produced in one compartment (the nucleus) but it must function in another compartment (the cytoplasm). Therefore, the final processed product (mRNA) must be transported through the nuclear envelope to reach the cytoplasm and be engaged with the ribosomes for translation. It is now clear from a variety of studies that there are specific pathways for the transport of the various RNAs and that this involves recognition of the RNA by transport activities that allow the RNA to exit through the complex nuclear pore structure.

The understanding of macromolecular transport into and out of the nucleus has progressed at a rapid rate over the past several years. The mechanisms involved in RNA transport have progressed in large part through studies of Bryan Cullen and colleagues here at Duke that have focused on the events involved in transport of HIV RNA. They have shown that the product of the HIV rev gene is required for the efficient transport of unspliced viral RNA from the nucleus to the cytoplasm. This requires the specific binding of rev to the HIV RNA which then facilitates an interaction with the cellular transport machinery. In addition to the understanding of molecular mechanisms of HIV gene expression that results from this work, two very significant advances have been made. First, the study of Rev function has led to a dissection of the cellular machinery involved in nuclear/cytoplasmic RNA transport, including specific factors that mediate the transport of RNA out of the nucleus. Second, the analysis of Rev function resulted in the generation of a series of mutants of the Rev protein, one of which, termed Rev M10, had the property of acting in a dominant negative manner. That is, not only was the protein non-functional but it also interfered with the function of the wild type protein. As such, expression of this protein inhibits HIV replication and has been utilized as a gene therapy reagent to block AIDS infection which is currently in clinical trials.

Protein Synthesis

Upon successful transcription, processing, modification and then transport to the cytoplasm, the mRNA is finally competent to perform its function - namely, translation of the sequence encoded in the gene into the synthesis of a protein. As discussed previously, the genetic code is read as triplets of nucleotides. The reading of the sequence in the mRNA to direct the synthesis of a unique protein is the process termed translation.

The Components of Translation:

Ribosome: a multicomponent RNA-protein structure that serves as the framework upon which protein synthesis takes place and which also provides the enzymatic activity for formation of the peptide bonds.

tRNA: a set of small RNAs, each specific for a given amino acid. The tRNA carries the amino acid to the ribosome for insertion into the growing polypeptide chain. The anticodon in the tRNA is complementary to the codon in the mRNA.

Aminoacyl tRNA synthetases: enzymes that link an amino acid to a cognate tRNA in a two-step process. There is a unique synthetase for each amino acid but not one for every tRNA.

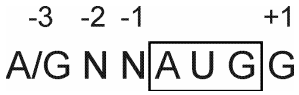
1) Amino acid + ATP + Enzyme ⟶ Enzyme [aminoacyl-AMP] + Ppi

2) Enzyme [aminoacyl-AMP] + tRNA ⟶ aminoacyl-tRNA + AMP + Enzyme

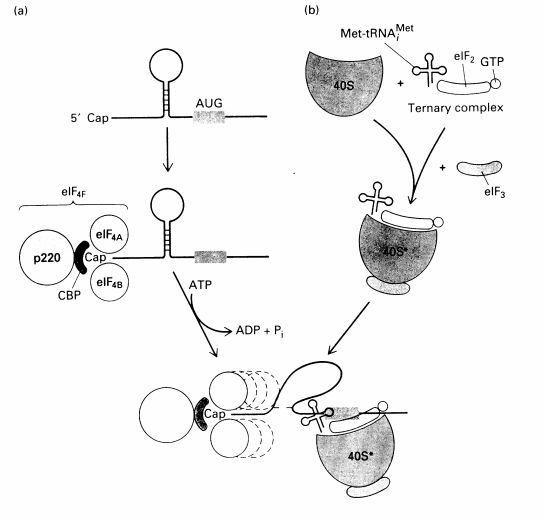
The Translation Process:

Polypeptide chains are synthesized in a sequential, step-wise fashion from N terminus to C terminus. Three distinct steps in protein synthesis can be defined: initiation, elongation, and termination. Each assembly step during the elongation process involves a peptidyl transferase reaction resulting in the formation of a peptide bond.

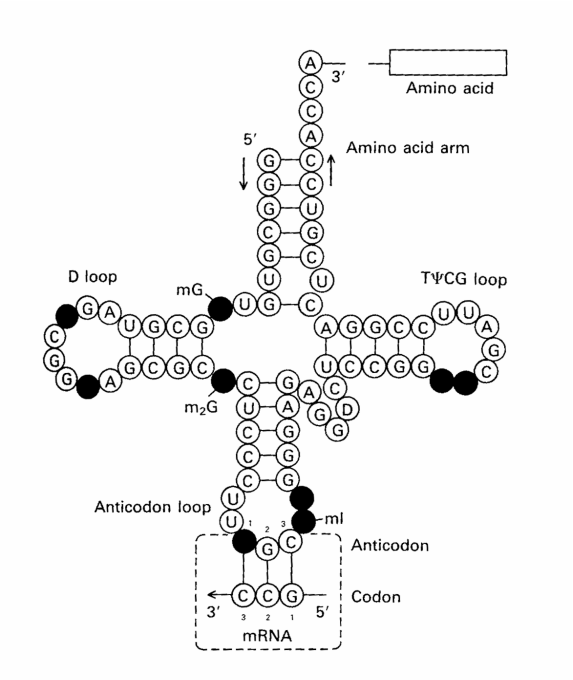
Initiation: Initiation of protein synthesis involves recognition of a methionine codon (AUG) in the mRNA by a special methionyl tRNA, different from the methionyl tRNA that is used for elongation. The initiation event is accomplished through the action of several initiation factors that allow interaction of the mRNA with the small ribosomal ubunit, GTP, and the initiator met-tRNA.



The recognition of the initiating AUG codon in the mRNA is facilitated by the RNA sequences that surround the codon. In particular, the three nucleotides preceding the AUG and the nucleotide immediately following the AUG are important for recognition of the initiation codon. Once the interaction takes place, the large subunit of the ribosome interacts and protein synthesis begins.

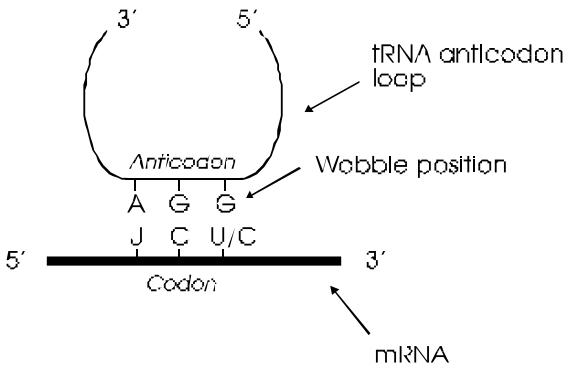


Codon Recognition: Codons (base triplets) in the mRNA are recognized by tRNAs which carry the appropriate amino acid to the translation machinery. Codon recognition involves base pairing between the codon in the mRNA and the anticodon in the tRNA.



Each tRNA is specific for one amino acid but many of the tRNAs can recognize more than one codon. There are 61 possible codons and approximately 50 tRNAs in animal cells.

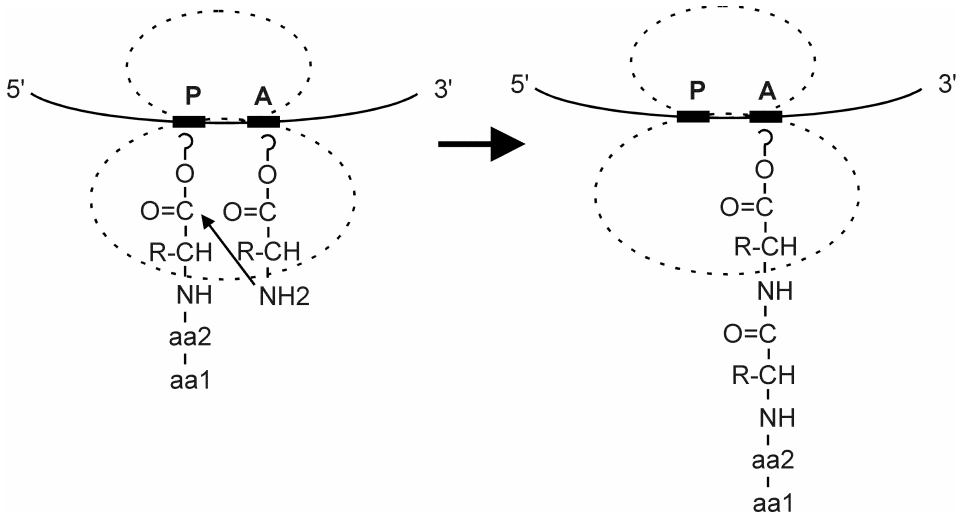
A single tRNA can recognize multiple codons because of *wobble* - a process of non-standard base pairing involving the first nucleotide of the anticodon in the tRNA and the third nucleotide of the codon in the mRNA.



Codon-Anticodon Pairings Allowed by Wobble:

5' Position of Anticodon	3' Position of Codon
G	U or C
C	G
A	U
U	A or G
I	U, C, or A

Elongation: There are two functional sites on the ribosome that are occupied by tRNA and that facilitate peptide bond formation. The P site (peptidyl) and the A site (aminoacyl).



Following formation of the peptide bond, the tRNA remaining in the P site leaves and the tRNA-peptidyl complex moves to the P site. A new aminoacyl-tRNA, specified by the mRNA codon, then moves into the A site and peptide chain elongation continues.

The catalysis of peptide bond formation is a function of the large ribosomal subunit. Very recent evidence indicates that it is the ribosomal RNA component of the large subunit that carries the enzymatic activity of peptidyl transferase.

Termination: Three codons (UAG, UAA, and UGA) do not specify an amino acid-tRNA and thus cause termination of translation. These codons signal the release of the peptidyl-tRNA complex when recognized by termination factors. This results in the release of an uncharged tRNA lacking an attached amino acid residue as well as the completed polypeptide chain. The ribosome then disengages from the mRNA and the subunits dissociate, ready to start the cycle over again.

Drugs and other agents that affect translation:

Interferon - a cellular protein produced in response to many viral infections. Functions to limit the spread of the infection by inducing the phosphorylation of the eIF2 initiation factor. This causes the inactivation of translation by preventing exchange between GDP bound form and GTP bound form.

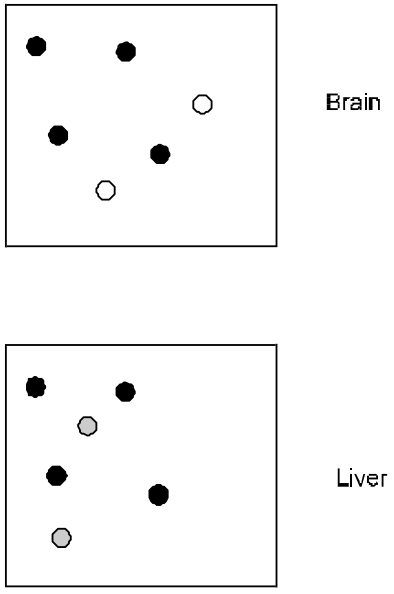
Diptheria toxin - catalyzes ADP ribosylation of EF2 (involved in the switch from the A to the P site) via NAD. Very small amounts of the toxin, even 1 molecule, can completely inactivate EF2 function and shut down protein synthesis.

GENE REGULATION

The phenotype of a cell as well as the organism as a whole, is the consequence of the regulated expression of a group of genes. Every cell in the organism contains the exact same complement of genes; nevertheless, there are unique proteins produced in the brain that are not produced in the liver; proteins are expressed at a particular time in the cell cycle; proteins are produced in response to hormones; etc. Clearly, an understanding of the molecular basis for the control gene expression is critical to an overall understanding of the basis for cell phenotype.

The Regulation of Gene Expression Is Responsible for Tissue Differences and Many Other Cellular Phenotypes

Since the expression of a gene is ultimately the production of the protein product of the gene, control must be defined as any process that alters the production of the protein. Control of



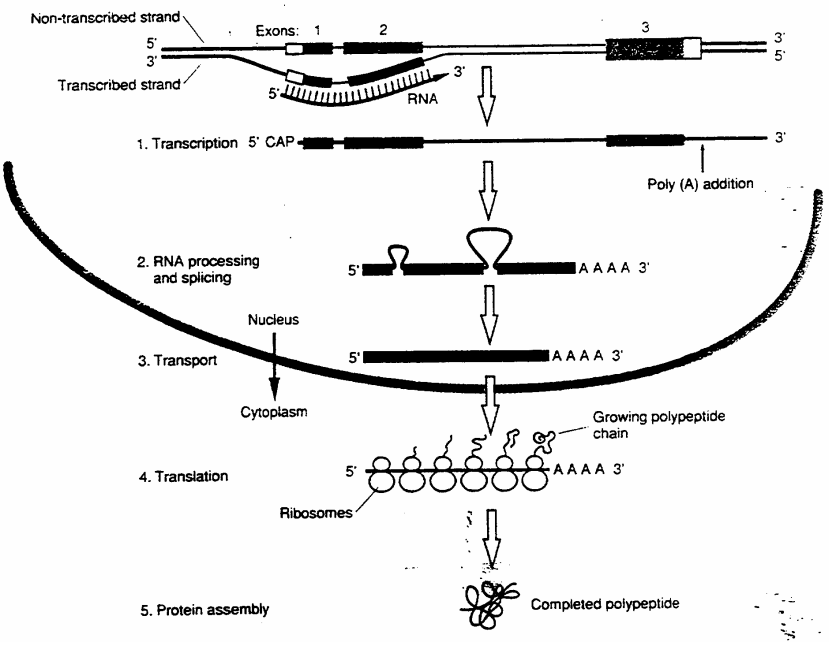
gene expression can be most easily visualized by the pattern of proteins produced in one circumstance versus another. For instance, as schematically depicted in the figure above, a two-dimensional gel analysis of proteins (a method that can separate thousands of individual proteins in a sample) in the brain versus in the liver reveals a number of proteins that are common (black spots) but a number of others that are unique to each tissue type. Thus, even though both tissues possess the exact same complement of genetic information, the *expression* of this information

differs - gene control clearly does exist. The question is - what is the basis for this control? What are the underlying mechanisms?

Gene control in prokaryotes and simple unicellular eukaryotes is largely a response to environmental signals - nutrients, etc. In higher eukaryotes (metazoans), the major form of gene control relates to cellular differentiation. Thus, in most cases it is long term and permanent. An example can be seen in the comparison of proteins synthesized in the brain versus in the liver, as analyzed by two dimensional gel electrophoresis. Although the majority of proteins that are synthesized are the same in each tissue, one can find examples of species that are unique to one or the other.

The Complexity of Eukaryotic Gene Expression Provides Multiple Opportunities for Gene Control

As discussed previously, the events associated with the expression of any given gene in a eukaryotic cell is a very complex process, involving multiple processing events as well as transport from the nuclear to the cytoplasm in order to achieve the final production of a



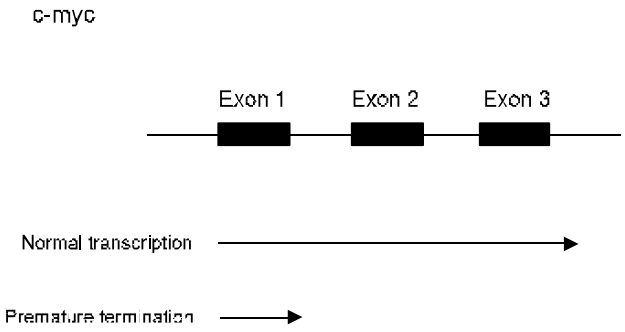
functional mRNA. Thus, alterations in any of the steps in mRNA biogenesis could alter the final concentration of functional mRNA. Moreover, control of gene expression could also result

from an alteration in the translation efficiency of the mRNA or alterations in the stability of the protein product.

Transcription Control

The initial step in gene expression is transcription of the gene and it is now clear from a variety of studies that the control of transcription is a critical regulatory step in the control of gene expression. In considering transcription control, and particularly when carrying out measurements of transcription, one usually defines the transcriptional unit which is that segment of the chromosome (DNA) that specificies the start and the end of transcription. This includes all of the signals necessary for proper transcription.

Transcription regulation could take the form of either initiation control or termination control. Clearly, the control of initiation will determine whether the primary transcript, and thus the funtional mRNA, will be produced and thus represents the most basic form of gene control. Termination can also be a factor if it occurs prior to the completion of the transcript (premature

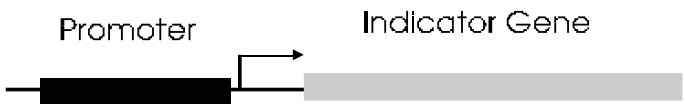


termination). This has in fact been demonstrated to be an important control in the expression of several oncogenes such as the c-myc gene.

A similar example of control of transcription elongation can be found in the control of HIV transcription. In the absence of viral regulatory proteins, HIV transcripts initiate properly but fail to efficiently elongate. One of the early viral proteins produced, known as *Tat*, functions to promote elongation and thus increase the efficiency of viral transcript production. Interestingly, the mechanism for the action of *Tat* is unique in that the protein recognizes sequences in the 5' end of the initiated RNA transcript rather than DNA promoter sequences.

How does one measure transcription, and thus determine transcription control? It is important in this regard to distinguish steady state RNA levels from synthesis which requires that the events of transcription must be separated from subsequent processing steps. Analysis of the RNA in a cellular extract provides a measure of the steady state level - thus, a combination of both transcription and subsequent events. Transcription must be measured by pulse-labeling, usually with a radioactive precursor to RNA, for short time such that no processing has taken place. In short, one is measuring the synthesis of the RNA not the accumulation of the RNA. In actual practice, this is accomplished by incubating cells with radioactive RNA precursors and then measuring the amount of radioactivity incorporated into the RNA, usually by hybridization to a DNA probe.

Once a gene has been isolated and the promoter/enhancer has been identified, it is also possible to study transcription control, particularly the identification of regulatory sequences, through the use of a reporter gene. The promoter to be studied is fused to a gene that can be easily assayed (a reporter) and then assayed by introduction into appropriate cells, or into animals, scoring for the expression of the reporter. For instance, if a suspected promoter element is placed upstream of a ?-galactosidase gene (the reporter), activity of the promoter, and thus transcriptional activity, can be assessed by measuring the production of ?-galactosidase activity.



In this way, transcription is being separated from RNA processing contributions by only analyzing the role of DNA sequences that contribute to transcription control.

Given the fact that transcription is the key first step in gene expression, and the fact that the binding of trans-acting factors to promoter elements is critical for transcription, gene expression can be regulated by controlling the activity of these trans-acting factors.

Mechanisms Regulating Transcription Factor Activity

1. Control of synthesis of the transcription factor. This is primarily the basis for tissue specific control; i.e., a key regulatory factor or factors is only found in the cell type that the target gene

is expressed. For example, the albumin gene is transcribed in the liver but not the brain because the necessary transcription factors are not found in the brain. Another example is the Myc transcription factor that functions to regulate the transcription of genes important for cell proliferation. The Myc protein is not found in quiescent cells because the Myc gene is inactive. Upon stimulation of cell growth, the Myc promoter is activated and transcription of the gene is induced. As already discussed, the control of termination of transcription of the Myc gene is also a factor in the regulation of Myc gene expression.

2. Control of the DNA binding activity of the factor. In this case, the protein (transcription factor) is present but it is not active in DNA binding. For example, the steroid hormone receptors are transcription factors. These are intracellular (cytoplasmic) proteins that bind specifically to the hormone when it enters the cell. Once the hormone binds, the receptor is then activated and can enter the nucleus, bind to the gene, and stimulate transcription.

3. Control of the transcriptional stimulatory activity of the factor. In this instance, the protein can bind to DNA but it is not able to stimulate transcription. For instance, the activity of the E2F transcription factor, which is responsible for the control of transcription of various genes important for DNA replication and cell growth, is regulated by interaction with the retinoblastoma (Rb) tumor suppressor protein. When Rb binds to E2F, the resulting complex can still bind to DNA but it is inactive in stimulating transcription. In fact, the complex can function is just the opposite fashion by serving as a repressor of transcription. The interaction of Rb with E2F is regulated by phosphorylation. That is, unphosphorylated Rb can bind to and regulate E2F but when Rb is phosphorylated by cell cycle regulated protein kinases, it loses the capacity to bind to E2F.

Practical Importance of Defining Transcription Control Elements

In considering strategies for gene therapy, one must be able to express the protein of interest (for instance, the cystic fibrosis gene product) in the right cell type, at the right time, and in the proper amounts. Thus, an understanding of the mechanisms controlling the expression of the gene to be used is essential in designing the gene therapy vector.

To understand the basis for alterations in transcription control that occur in disease conditions such as cancer, it is critical to know the normal mechanisms of function of the gene. Such alterations can take two general forms:

1. Mutation of trans-acting factors. This could involve an inactivation of a factor as the result of a specific mutation or deletion or it could involve the creation of a factor with altered properties. For instance, it might become constitutively active (no longer regulated) or it might acquire an altered specificity.

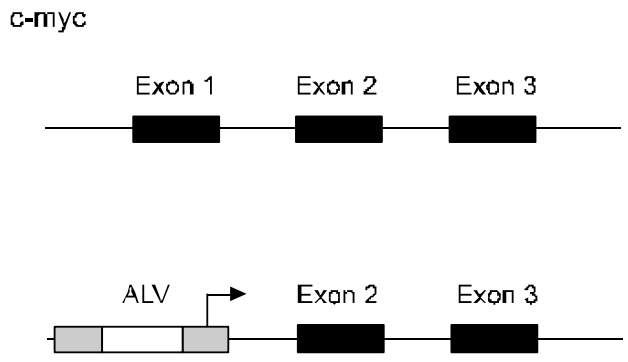
2. Alteration of cis-acting elements of a promoter/enhancer. This could involve mutation of the element resulting in a loss of binding of the transcription factor or it could involve chromosomal alterations that create new elements resulting in a change in the transcription of the gene.

Alterations of Transcription Regulation in Disease

By developing an understanding of normal gene structure, as well as the components of transcription regulation, it has been possible to define the molecular basis for alterations in gene control events that underlie certain disease states. Several such examples are given here.

Retrovirus mediated promoter insertion resulting in activation of the c-myc gene

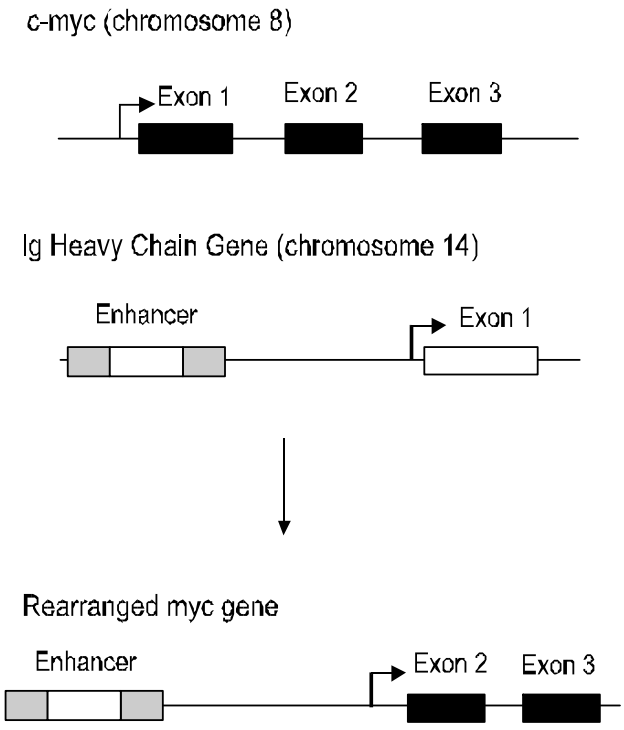
As discussed above, transcription of the Myc gene, which encodes a transcription factor that controls cell cycle progression, is normally tightly controlled by cell growth regulation. This normal control can be disrupted as the result of an insertion of a retrovirus (ALV) into the promoter region of the c-myc gene. As a result of this insertion, the myc gene is now controlled by the retrovirus promoter which does not respond to the cell growth regulatory signals.



Although this is a rare event, it does raise the potential danger of the use of retrovirus vectors in gene therapy protocols; that is, the inadvertant activation of an oncogene as a result of the retrovirus insertion. This was a critically important discovery that led directly to the discovery of Myc gene rearrangements in human tumors.

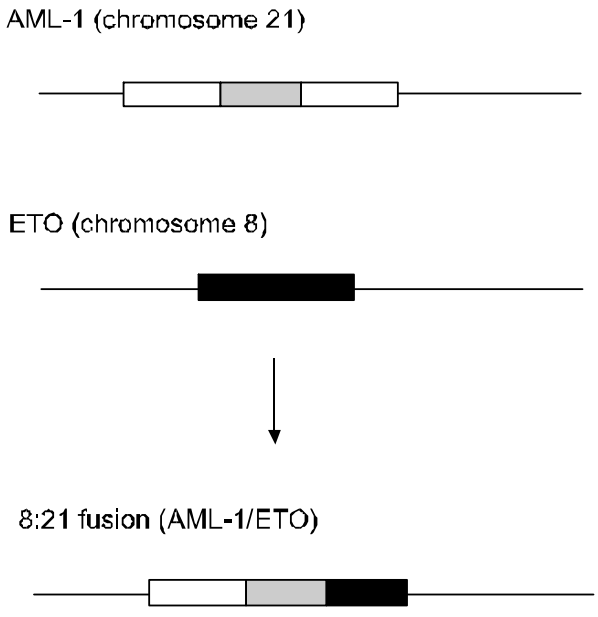
Rearrangement of the c-myc gene in B cell lymphomas

The expression of the c-myc gene is also deregulated in many tumors as the result of chromosome rearrangements, once again resulting from a change in the transcriptional regulatory sequences. For instance, many B cell lymphomas contain a translocation involving chromosome 8 and chromosome 14 that places the c-myc gene in the chromosomal environment of the immunoglobulin heavy chain gene enhancer. In this case, the normal regulation of the myc gene is disrupted with control now being directed by the immunoglobulin enhancer. This then confers a high level of transcription that is B cell-specific and non-cell cycle regulated.



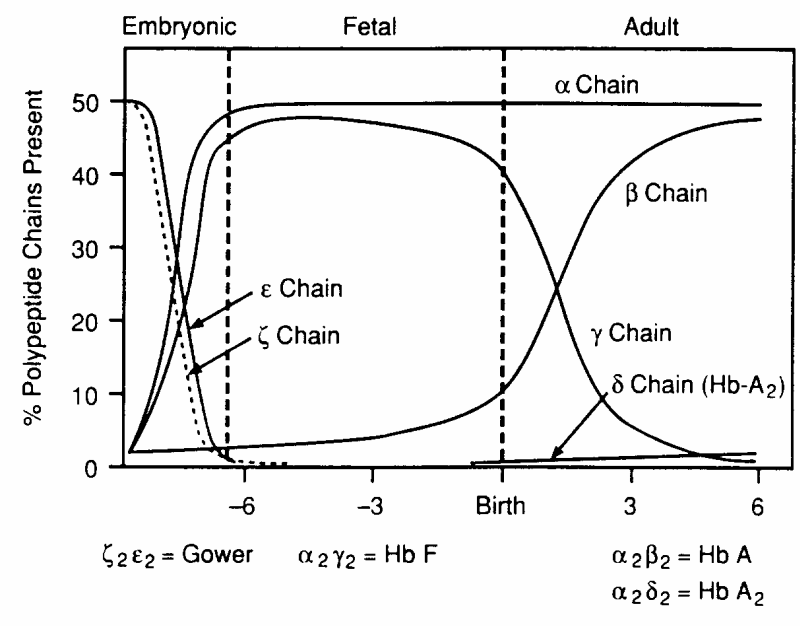
Creation of a chimeric transcription factor in AML by chromosome rearrangments

Whereas the changes detailed above regarding the myc gene result in alterations in regulation of expression of the gene, but still producing the normal protein, another form of transcriptional deregulation can be seen in a chromosomal rearrangement that alters the structure of the encoded protein. The most common chromosomal rearrangment seen in acute myelogenous leukemia is a translocation that fuses a portion of chromosome 8 with a portion of chromosome 21, a so-called 8:21 translocation. The breakpoints involve a gene on chromosome 21 known as AML-1 which encodes a transcription factor and a gene on chromosome 8 known as ETO of unknown function. As a result of the translocation, a new gene is created that encodes a chimeric protein containing sequence from AML-1, including the DNA binding domain, and sequence from ETO. Although the nature of the effect on AML-1 function is unknown, one presumes that some aspect of the specificity or the regulatory properties of the transcription factor has been altered.



Mutations in cis-acting motifs that regulate globin transcription

Hemoglobins are iron-heme-containing proteins that transport oxygen to tissues from the lungs. The binding and the release of oxygen are complex biochemical events that are enhanced by conformational changes in the hemoglobin molecule, a tetramer composed of alpha and non-alpha globin chains. Because oxygen transport requirements are different during the fetal and the adult developmental stages, hemoglobins have evolved into a family of molecules that meet these physiologic needs.



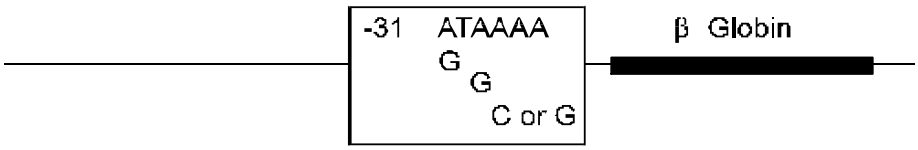
The primary difference between fetal and adult hemoglobin is the replacement of the gamma chain in the fetal hemoglobin with a beta chain in the adult hemoglobin, hemoglobin A.

This switch between fetal to adult globin production is accomplished by inactivation of the gamma globin gene and activation of the beta globin gene, both members of the beta globin gene domain, shown below.

Transcription of these genes is regulated by a typical TATA promoter which interacts with a distant enhancer called the LCR. In addition, the 5' flanking region of both genes contains cis-acting binding motifs that binds developmental stage specific transcription factors which regulate their developmental expression. Mutations in these cis-acting elements of the gamma and beta globin genes result in specific clinical phenotypes.

? **-thalassemia- point mutations**

Abnormalities in the synthesis of beta globin produce hereditary anemias called thalassemias. Although there are many different mutations that cause thalassemias, one of the more informative mutations involves the TATA box. In the beta globin gene, the normal TATA box sequence is ATAAAA. At least four distinct mutations have been found in which one of the A's has been changed to a G or C. These mutations result in a reduction of the beta globin mRNA production from that gene by approximately 70-80% in developing erythroid cells. Expression of the mutant cloned gene in tissue culture cells has confirmed this experimentally.



? **-thalassemia-enhancer deletion**

One type of beta thalassemias was quite puzzling when first discovered because the beta globin gene and the immediate flanking region were completely normal. Extensive mapping studies of the beta globin gene domain revealed a deletion that began almost 40,000 base pairs 5' of the beta globin gene and encompassed a strong DNase I hypersensitivity site. Cloning and characterization of this DNase I hypersensitivity site demonstrated that it was the enhancer for the entire beta globin gene domain. Linkage of this enhancer, called the locus control region or LCR, to a beta globin gene permits expression at normal levels in transgenic mice whereas the beta globin gene without a linked LCR is expressed at only 1% the endogenous level in transgenic

mice. This mutation established that a distant enhancer is absolutely essential for the expression of a linked gene family.

Hereditary persistence of fetal hemoglobin: point mutations in the gamma globin gene.

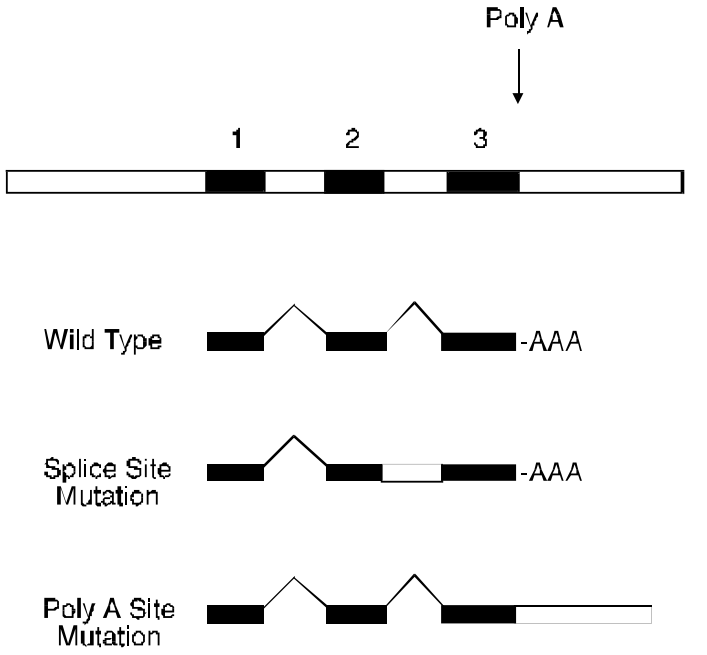
Typically, adult patients have less than 1% fetal hemoglobin. A small group of patients have been identified who have up to 25% fetal hemoglobin persisting into the adult period. At least 30 different mutations produce this phenotype, and they can be classified as point mutations within the fetal globin gene or large deletions involving the beta globin gene domain. The first class is exemplified by a point mutation in the gamma globin gene in a regulatory motif that is bound by the transcription factor GATA 1. Experimental studies have demonstrated that this gamma globin gene is expressed normally during the fetal period but is not repressed during the adult state. The single-base change in the GATA 1 binding site results in impaired binding of GATA 1 and failure to repress transcription during the adult state.

Post-Transcriptional Gene Control

Each one of the steps of mRNA biogenesis following transcription has been demonstrated to participate in gene regulation. Thus, splicing of the primary transcript, 3' end cleavage and polyadenylation, transport to the cytoplasm, and metabolism of the mRNA in the cytoplasm, including translation efficiency and stability of the mRNA, all can be altered to achieve a regulation of the production of the product of the gene.

A variety of transcription units have now been shown to possess the potential to produce more than one mRNA as a result of alternative processing. This can include both alternative splicing as well as polyadenylation. A selective use of exons in a primary transcript can thus define a distinct protein product. If this selection is regulated, changing under one circumstance or another, then then such changes are defined as events regulating the expression of the gene.

It is also clear that mutation of either critical splice site sequences or poly A site sequences can impair gene expression. Although such mutations would not alter the coding sequences directly, they can result in alterations that do affect the coding capacity. For instance, a splice site mutation that altered the splice donor following exon 2 in the example, would leave the intron sequence in the mRNA which would lead to a frameshift and likely a non-functional

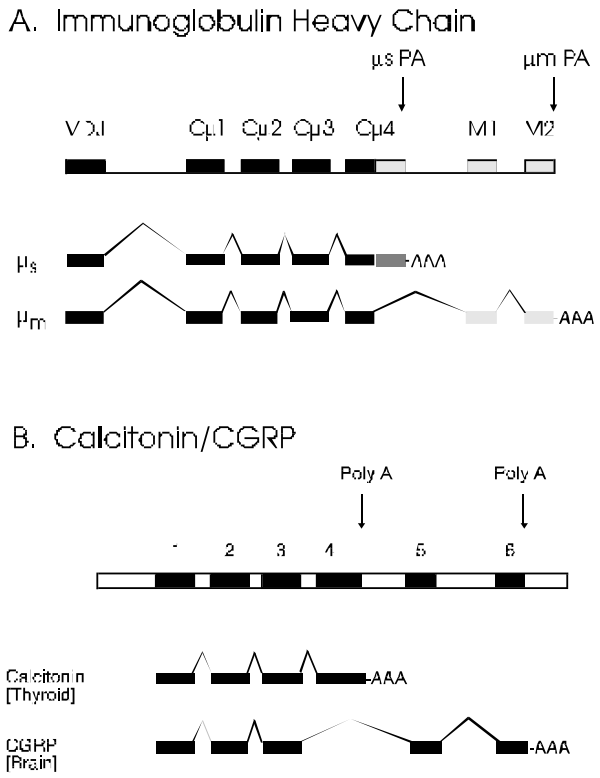


protein. Likewise, a poly A site mutation would prevent processing at the poly A site resulting in

an RNA with an extended 3' terminus and no poly A tail. Such an RNA would not be efficiently transported to the cytoplasm and would be very unstable.

Alternative Splicing and Polyadenylation as Gene Control Mechanisms

There are a variety of examples of gene control through alternative RNA processing events, both splicing as well as polyadenylation. Perhaps two of the best studies examples include the immunoglobulin heavy chain gene and the calcitonin/CGRP gene. The

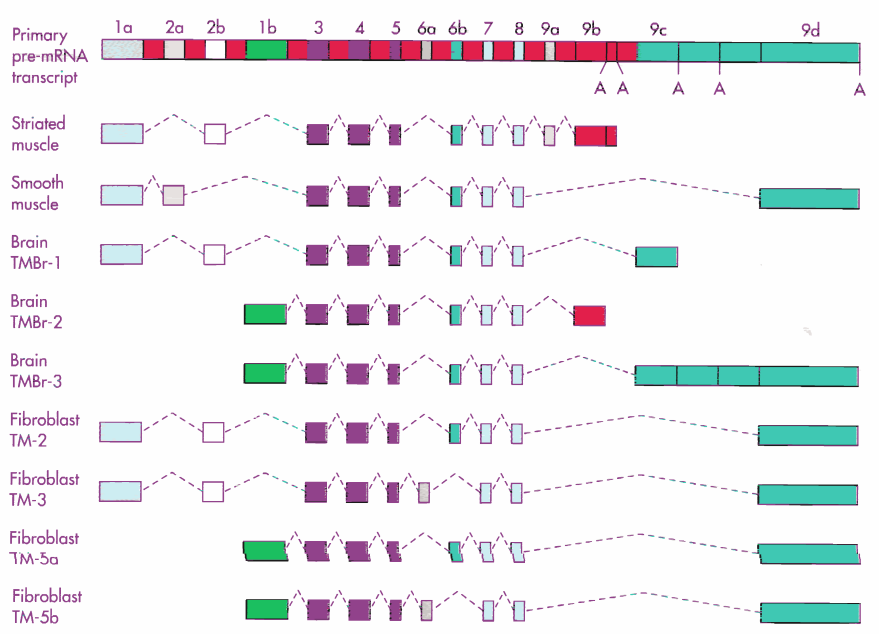


immunoglobulin heavy is composed of two protein molecules, a heavy chain and a light chain. Antibody diversity is determined by variation in the sequence in both the heavy chain and the light chain as a consequence of gene rearrangement as well as mutation. In addition, the production of the heavy chain is regulated during B cell differentiation. In a mature B cell, the heavy chain, together with the light chain, is inserted into the B cell membrane and serves as an antigen receptor. When antigen binds, the B cell is stimulated to mature to a plasma cell where the immunoglobulin molecule is now secreted as a antibody. This switch in immunoglobulin

expression is the result of alternative splicing as well as polyadenylation of the primary transcript of the gene as indicated below. This brings a different set of exon sequences into the 3' position of the transcript. The Cμ4 exon encodes the secreted form of the protein whereas the M1 and M2 exons encode the membrane bound form of the protein.

Another example can be found in the calcitonin gene which encodes a calcium regulating hormone that is produced in the thyroid. This locus also encodes a distinct separate product known as CGRP (calcitonin gene related peptide), a neuropeptide produced in the brain. Thus, alternative RNA splicing and polyadenylation, that occurs in a tissue specific manner (brain versus thyroid), results in the production of two distinct gene products with distinct function.

Finally, one very striking example of the role of alternative RNA processing in the control of gene expression can be seen in the large variety of products that can be generated from the alpha-tropomyosin locus in a tissue-specific fashion. As shown in the diagram below, the gene contains 14 distinct exons and differences in splice site selections as well as poly A site



selections can generate a large number of mRNAs that differ by the exons they contain, leading to the production of distinct forms of tropomyosin that presumably function differently in the various cell types, possibly via distinct protein-protein interactions. Moreover, it is also clear that alternative transcription initiation events can include the initial exon 1a or begin at exon 1b.

Alterations of Post-Transcriptional Regulation in Disease

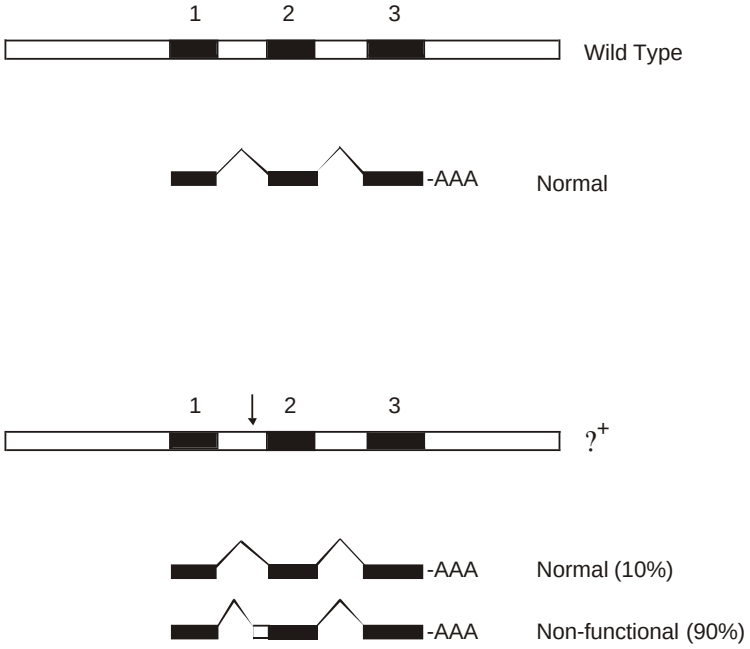
In addition to the mutations that affect transcriptional control of the globin genes, numerous mutations have been identified in thalassemia patients that alter post-transcriptional expression of the globin genes. For instance, shown below are mutations identified in a series of ?-thalassemias that alter the conserved sequence at the exon1/intron1 splice junction of the ? globin gene. The severity of the clinical phenotype correlates with the nature of the mutation – the more severe $?^{o}$ thalassemias can be seen as mutations within the highly conserved sequence at the splice junction whereas the less severe $?^{+}$ thalassemias are seen as mutations in sequence that make less of a contribution to splice site recognition.

Splice site mutations in ? thalassemia

Exon 1 ⌐ ⌐ Intron 1	
GCCAG GTTGGTAT	Normal
GCCAG **A**TTGGTAT	$?^{o}$
GCCAG **T**TTGGTAT	$?^{o}$
GCCAG GTTG**T**TAT	$?^{+}$
GCCAG GTTG**C**TAT	$?^{+}$

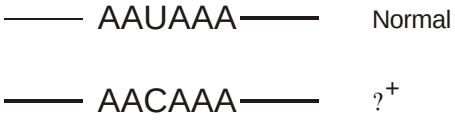
These mutations result in either a complete loss of splicing ($?^{o}$ mutations) or a reduction in the efficiency of making the proper splice ($?^{+}$ mutations); the latter leads to reduced levels of function ? globin mRNA.

It is also true that ?-thalassemia mutations can occur within an intron that *create* a splice site rather than abolish one. As shown in the figure below, a mutation (indicated by the arrow) within intron 1, that creates a functional splice acceptor site, would then allow splicing to occur to this site, thus including intron sequence as part of the second exon. This would then result in the formation of a non-functional mRNA. In this example, approximately 90% of the transcripts of this mutant globin allele splice to this incorrect site leaving 10% splicing to the correct site. Thus, there is not a complete elimination of ? globin production but rather a reduction ($?^{+}$ thalassemia).



In addition to these examples of mutations that affect RNA splicing, other cases of thalassemias exhibit mutations within the polyadenylation recognition sequence. An example is shown below.

Poly A site mutations in ? thalassemia



Given the critical nature of the polyadenylation signal in allowing the proper cleavage of the pre-mRNA and subsequent polyadenylation, these mutations drastically reduce the production of

globin mRNA. Generally, they do not completely eliminate globin production because transcription continues until an alternate polyadenylation signal is encountered in the genome.

Other Forms of Post-Transcription Gene Control

Control of RNA Transport: Although there are no clear examples whereby the nuclear/cytoplasmic transport of a cellular mRNA is regulated, there are at least two instances in viral infections in which RNA transport is affected. First, adenovirus infection results in the inhibition of transport of most cellular mRNAs - a specific viral gene product is required for this to occur and at the same time, this protein facilitates the transport of viral RNA.

Second, as indicated previously, studies of Bryan Cullen and colleagues here at Duke have shown that the product of the HIV rev gene is required for the efficient transport of unspliced viral RNA from the nucleus to the cytoplasm.

Control of mRNA Stability: The stability of mRNAs varies over a large range. Some RNAs are quite stable with half lives approaching the cell division time. Other RNAs turn over very rapidly (half lives of a few minutes). As a general rule, RNAs that are expressed in a transient fashion often are short lived. Many RNAs that encode cytokines as well as early responses to mitogens are unstable, dependent on specific sequences in the 3' untranslated region of the RNA. The unstable nature of the mRNA as a result of recognition of this sequence is associated with shortening of the poly A tail.

Translation Control: General control - alterations of translation factors can alter the translation efficiency of mRNAs. For instance, phosphorylation of eIF2 inhibits is action. Translation efficiency is also determined by cis acting sequences in the mRNA - particularly the sequences that surround the AUG initiation codon.