Applications of recombinant DNA technology in gastrointestinal medicine and hepatology: Basic paradigms of molecular cell biology. Part B: Eukaryotic gene transcription and post-transcriptional RNA processing

Gary E Wild MDCM PhD FRCPC[1], Patrizia Papalia BSc[1], Mark J Ropeleski MDCM FRCPC[1], Julio Faria MDCM FRCSC[1], Alan BR Thomson MD PhD FRCPC[2]

GE Wild, P Papalia, MJ Ropeleski, J Faria, ABR Thomson. Applications of recombinant DNA technology in gastrointestinal medicine and hepatology: Basic paradigms of molecular cell biology. Part B: Eukaryotic gene transcription and post-transcriptional RNA processing. Can J Gastroenterol 2000;14(4):283-292. The transcription of DNA into RNA is the primary level at which gene expression is controlled in eukaryotic cells. Eukaryotic gene transcription involves several different RNA polymerases that interact with a host of transcription factors to initiate transcription. Genes that encode proteins are transcribed into messenger RNA (mRNA) by RNA polymerase II. Ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) are transcribed by RNA polymerase I and III, respectively. The production of each mRNA in human cells involves complex interactions of proteins (ie, trans-acting factors) with specific sequences on the DNA (ie, cis-acting elements). Cis-acting elements are short base sequences adjacent to or within a particular gene. While the regulation of transcription is a pivotal step in the control of gene expression, a variety of molecular events, collectively known as 'RNA processing' add an additional level of control of gene expression in eukaryotic cells. *(Pour le résumé, voir page suivante)*

Key Words: *DNA; Eukaryotic gene transcription; Gene expression; RNA*

Cell and Molecular Biology Collaborative Network in Gastrointestinal Physiology. [1]Department of Medicine, Division of Gastroenterology, McGill University Health Centre and McGill University Inflammatory Bowel Disease Research Program, [2]Department of Medicine, Division of Gastroenterology, University of Alberta, Edmonton, Alberta

Correspondence: Dr Gary E Wild, Montreal General Hospital, 1650 Cedar Avenue, Montreal, Quebec H3G 1A4. Telephone 514-934-8308, fax 514-934-8411, e-mail gwild@is.mgh.mcgill.ca

Received for publication March 23, 1999. Accepted July 15, 1999

Applications de la technologie de recombinaison de l'ADN en médecine gastro-intestinale et en hépatologie : concepts de base de la biologie moléculaire de la cellule. Partie B : transcription des gènes eucaryotes et maturation de l'ARN

RÉSUMÉ : La transcription de l'ADN en ARN est le premier niveau d'expression génétique commandée par les cellules eucaryotes. La transcription des gènes eucaryotes exige la participation de plusieurs ARN-polymérases qui interagissent avec de nombreux facteurs de transcription pour amorcer la transcription. Les gènes qui encodent les protéines sont transcrits en ARN messager (mRNA) par l'ARN-polymérase II. L'ARN ribosomique (rRNA) et l'ARN de transfert (tRNA) sont respectivement transcrits par les ARN-polymérases I et III. La production de chaque mRNA dans les cellules humaines met en jeu des interactions complexes de protéines (facteurs trans-activateurs) ayant des séquences très précises sur la chaîne d'ADN (éléments cis-activateurs). Les éléments cis-activateurs sont de courtes séquences de base qui se trouvent à proximité d'un gène ou à l'intérieur de celui-ci. Tandis que la régulation de la transcription constitue une étape essentielle du réglage de l'expression génétique, une foule d'événements moléculaires, appelés « maturation de l'ARN », ajoutent un autre niveau de commande de l'expression génétique dans les cellules eucaryotes.

In contrast to prokaryotes (where all genes are transcribed by a single RNA polymerase that binds directly to gene promoter sequences), transcription in eukaryotic cells involves several different RNA polymerases that interact with a variety of transcription factors (TFs) to initiate transcription. This increased complexity characteristic of eukaryotic transcription facilitates the sophisticated and orderly regulation of gene expression that ultimately determines the activities of the diverse array of cell types seen in multicellular organisms.

Three distinct nuclear RNA polymerases are found in eukaryotic cells. Genes that encode proteins are transcribed into messenger RNA (mRNA) by RNA polymerase II. Ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) are transcribed by RNA polymerase I and III, respectively. Some small nuclear and cytoplasmic RNAs are transcribed by RNA polymerase II and III. Finally, mitochondrial genes are transcribed by a separate group of RNA polymerases. RNA polymerases are comprised of eight to 14 different subunits. Although they recognize distinct promoters and transcribe different classes of genes, these RNA polymerases share many common features, including a clear dependence on other proteins to initiate transcription.

The transcription of DNA into RNA is the primary level at which gene expression is controlled in eukaryotic cells. Only a fraction of the transcribed RNA is translated into polypeptides. This is explained as follows.

Some transcription units code for RNA molecules only, as in the case of ribosomal RNAs, transfer RNAs, and a host of small nuclear and cytoplasmic RNA molecules.

The initial transcription product of those transcription units, which do not encode polypeptides, is subject to events known as RNA processing. With RNA processing much of the initial RNA sequence is trimmed to yield smaller mRNA molecules.

Only the central region of mRNA is translated; variable portions of the 5′ and 3′ ends of mRNA remain untranslated.

Transcription is mediated by the enzyme RNA polymerase, using DNA as a template and ATP, CTP, GTP and UTP as RNA nucleoside precursors. RNA is synthesized in the 5′ to 3′ direction as a single strand molecule. Only one of the two DNA strands serves as a template for transcription. Because the growing RNA molecule is complementary to this template strand, the transcript has the same 5′ to 3′ orientation and base sequence (except that uracil replaces thymidine) as the opposite, nontemplate strand of the DNA double helix. Thus, the nontemplate strand is called the 'sense strand' and the template strand is called the 'antisense strand'. Gene sequences listed in various databases show only the sequence of the sense strand. The orientation of sequences relative to a gene sequence is dictated by the sense strand and by the direction of transcription (eg, the 5′ end of the gene is the sequences of the 5′ end of the sense strand, and upstream or downstream of the gene is sequences that clamp the gene at the 5′ or 3′ ends of the sense strands, respectively). The general features have been reviewed elsewhere (1-8)

CHROMATIN STRUCTURE AND TRANSCRIPTION

The DNA present in all eukaryotic cells is tightly associated to histones, forming chromatin. Moreover, the packaging of eukaryotic DNA into chromatin has important ramifications in terms of its availability to serve as a template for transcription. Thus, chromatin structure is a critical aspect of eukaryotic gene expression. Actively transcribed genes are situated in regions of decondensed chromatin. The tight coiling of DNA around the nucleosome poses a major obstacle to transcription; the tight coiling impedes the ability of TFs to bind to DNA and impedes the ability of RNA polymerase to gain access to the DNA template. This inhibitory effect of nucleosomes is overcome by the action of nucleosome remodelling factors. These remodelling factors disrupt chromatin structure, thus allowing TFs to gain access to nucleosome DNA and coordinate the assembly of the transcription complex with the promoter. A multiprotein complex, initially identified in yeast as the switch/sucrose nonfermenting (SWI/SNF) complex has been localized in mammalian cells. SWI/SNF disrupts the nucleosome array and facilitates the transcription of DNA that was previously unavailable to the transcription complex.

Eukaryotic transcriptional activators play dual roles in modulating gene expression. In addition to promoting transcription by interacting with basal TFs, they stimulate changes in chromatin structure that alleviate repression by histones. The ability of RNA polymerase to transcribe chromatin templates is facilitated through the acetylation of his-

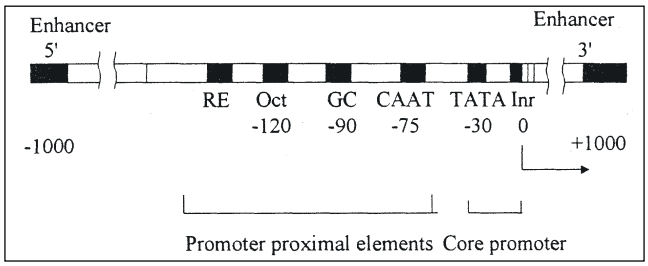

Figure 1) *The localization of cis-acting sequences in a typical human gene. The core promoter is situated with a TATA and initiator (Inr) sequences. The TATA sequence, located 30 base pairs upstream of the Inr sequence, is the binding site for the TATA-binding protein. The Inr sequence is where RNA polymerase II binds and initiates transcription. The promoter proximal elements are located 50 to several hundred base pairs upstream of the Inr site and include the common sequences CAAT, GC and octamer (Oct). These sequences are the binding sites for upstream transcription factors. Sequences in the promoter proximal regions are the response elements (RE), which are the binding sites for inducible transcription factors. Situated thousands of base pairs away, either 5′ or 3′ to the gene of interest, are enhancer elements that bind activators*

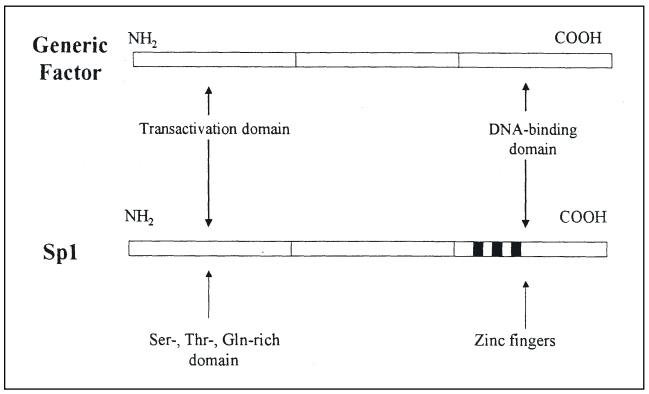

Figure 2) *Common functional domains in transcription factors (TFs). Many TFs contain two common functional domains. The transactivation domain is the amino acid sequence of the protein that interacts with other protein factors and is responsible for activating the transcription of genes. The DNA-binding domain is comprised of amino acid sequences that are responsible for interacting with and binding to specific DNA sequences. The upstream TFs (Sp1) binds to GC sequences through its DNA binding domain, which includes three zinc finger motifs. The transactivation domain of Sp1 is rich in the amino acids serine (Ser), threonine (Thr) and glutamine (Gln), and interacts with the TATA-binding protein-associated factor IID 110 subunit of TF IID*

tones and by the association of the nonhistone chromosomal high mobility group (HMG) proteins HMG-14 and HMG-17 with the nucleosomes of actively transcribed genes. The signals that target HMG-14 and HMG-17 to transcribe genes actively remain an enigma. The role of chromatic structure in transcription has been reviewed previously (9-13).

CIS-ACTING ELEMENTS

This discussion of the transcriptional control of gene expression focuses on the role of RNA polymerase II, the enzyme responsible for transcribing protein-encoding genes into mRNAs (9,12-14). The production of each mRNA in human cells involves complex interactions of proteins (ie, trans-acting factors) with specific sequences on the DNA (ie, cis-acting elements). Cis-acting elements are short base sequences adjacent to or within a particular gene. Alternatively, they can be sequences that occur several thousand base pairs from a particular gene. Cis-acting elements are sequences required for the recognition of a gene by RNA polymerase II. These sequences also serve as binding sites for the proteins that regulate the rate and specificity of transcription. The initiation of transcription is dictated by sequences that are present in each gene. The major cis-acting sequences of a gene are illustrated in Figure 1 and include the following.

- The core promoter element is situated 5′ to the gene and consists of the sequences where the transcription complex containing the RNA polymerase II assembles on the DNA molecule. There are two fixed sequence elements: the initiator element (Inr), which determines the transcription start site, as well as the TATA element, which is located 25 to 30 base pairs upstream from the Inr. The promoter initiation site defines the location and the direction of transcription.
- The promoter proximal elements are composed of two types of cis-acting sequences located 50 to a few hundred base pairs upstream from the start site. The first type of promoter proximal element comprises a class of base sequences (eg, CAAT or GC) found in many genes, and these sequences function as binding sites for proteins called upstream TFs. The second type of promoter proximal element is the response element (RE). The RE contains sequences that are found in promoters controlled by a particular stimulus (eg, genes that respond to particular glucocorticoid stimulation or iron response elements implicated in intestinal iron absorption).
- Promoter distal elements are cis-acting sequences found thousands of base pairs away from the start site of transcription. These distal sites are known as enhancers or silencers and are situated either upstream or downstream from the gene that they regulate. Enhancers, like promoters, act by binding TFs that subsequently regulate RNA polymerase. The looping of the DNA helix allows a TF bound to a distant enhancer to lie in relative proximity to the upstream promoter and interact with RNA polymerase or basal TFs at the promoter. The binding of specific transcriptional regulatory proteins to enhancers is responsible for controlling gene expression during development and cell differentiation. This mechanism also serves to mediate the response of cells to hormones and growth factors.

Transcription is initiated by the binding of a variety of TFs and the enzyme RNA polymerase to the promoter site. A large number of TFs serve to recruit the RNA polymerase to the promoter site. TFs bind to sequences in the promoter site on the DNA molecule or they can bind to one another in

TABLE 1
Different types of cis-acting sequences and the transcription factors that they bind

Cis-acting sequence	Transcription factor
Core promoter elements	
Initiator sequence	RNA polymerase II*
TATA	TATA-binding protein*
Promoter proximal elements	
Common elements	
CAAT	Common transcription factors
GC	Sp1
Octamer	Octamer transcription factors
Response elements	
Glucocorticosteroid response element	Glucocorticoid receptor
Thyroid response element	AP1
cAMP response element	cAMP response element binding protein
Promoter distal elements	
Enhancers	Activators
Silencers	Repressors

**General transcription factors bind to RNA polymerase and TATA-binding proteins at the core promoter*

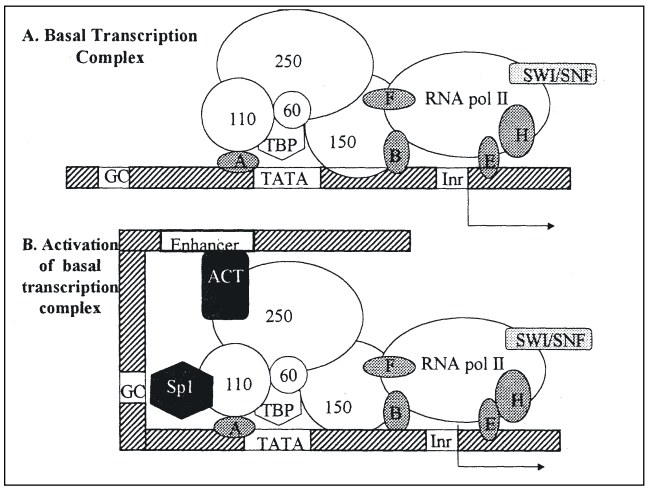

Figure 3) *Model of the initiation of transcription by RNA polymerase II (RNA pol II). The binding of the general transcription factors (GTFs) is illustrated in panel A, which depicts the formation of the basal transcription complex. RNA pol II binds to the core promoter. The TATA binding protein (TBP), a subunit of transcription factor (TF) IID, binds to the TATA sequence and facilitates the binding of the TBP-associated factors (TAFs). TBP and some of the TAFs are indicated as 250, 110, 150 and 60. Once TFIID is bound to the TATA sequence, the other GTFs (A, B, F, E and H) and RNA pol II bind to the core promoter thus forming the basal transcription complex. Also indicated is the switch/sucrose nonfermenting (SWI/SNF) multiprotein complex, associated with RNA pol II. This multiprotein complex is necessary for disruption of chromatin structure. The activation of the basal transcription complex is illustrated in panel B. The activation of transcriptional initiation by the TF Sp1 bound to the GC sequence and interacting with TAFIID 110. Further activation results from the binding of an activator protein (ACT) to an enhancer sequence located 1000 base pairs from the core promoter. The ACT is brought into close proximity with the basal transcription complex by looping away from the DNA between the enhancer sequence and the core promoter to allow the activator to interact with TAFIID 250*

several different areas to determine whether RNA polymerase will or will not transcribe a particular gene. The structural features of typical TFs are illustrated in Figure 2. TFs are characterized by the following shared features: binding to specific DNA sequences, interaction with other TFs to regulate transcription, a DNA binding domain made up of the amino acid sequences that recognize and bind specific DNA sequences, and a transactivation domain comprised of the amino acid sequences required for the activation of transcription.

TFs may have similar DNA binding domains but different transactivating domains. Thus, they bind the same sequence of DNA but activate transcription in a different manner. Alternatively, TFs have similar transactivating domains but different DNA binding domains. In this case, the TFs bind to different sequences of DNA, although the process of activation is similar. RNA polymerase catalyzes the formation of a phosphodiester bond by attaching the 5′ phosphate of the incoming ribonucleotide to the 3′ hydroxyl of the growing RNA chain. Multiple RNA transcripts may be synthesized from a single DNA molecule through the sequential binding of additional RNA polymerase to the promoter sequence.

TRANS-ACTING TFs

Trans-acting TFs bind to cis-acting elements on the DNA and interact with other TFs (9,12-18). These proteins control the initiation of transcription and comprise the following.

- General TFs are polypeptides that assemble at the core of the promoter site and recruit RNA polymerase II to that site to form the preinitiation complex.
- Upstream TFs are proteins that bind the common cis-acting sequences proximal to many promoters, such as the sequences CAAT and GC.
- Inducible TFs are proteins that respond to external stimuli that activate them, and in turn promote their binding to the RE sequences. This results in increased transcription of genes containing the particular RE sequence.
- Activator proteins are TFs that bind enhancers and increase transcriptional initiation of a particular gene.
- Repressor proteins are TFs that silence and inhibit transcriptional initiation of a particular gene.

The various types of cis-acting sequences and the TFs that they bind are listed in Table 1.

The ability of proteins to bind DNA is a reflection of their amino acid sequences and the formation of specific motifs. A well characterized DNA binding domain is the zinc finger domain, which contains repeats of cysteine and histidine residues that bind zinc ions within the DNA binding domain. Zinc finger domains are common among TFs that regulate RNA polymerase II promoters, including the com-

mon TF (Sp1), the general TF TFIIA and the glucocorticoid receptors. The helix-turn-helix motif is found in eukaryotic cell proteins including the homeodomain proteins. These play a central role in the regulation of gene expression during embryonic development. The molecular cloning and analysis of these genes have shown that they contain conserved sequences of 180 base pairs (homeoboxes) that encode the DNA binding domains (homeodomains of TFs). Homeobox genes are highly conserved across a variety of species. Finally, leucine zipper and helix-loop-helix proteins are two other families of DNA binding proteins that contain DNA binding domains formed by dimerization of two polypeptide chains. They appear to play important roles in regulating tissue specific and inducible gene expression.

INITIATION OF TRANSCRIPTION BY RNA POLYMERASE II

A set of basal TFs interact with the cis-acting core promoter sequences to form a basal transcription complex (Figure 3) during initiation of transcription by RNA polymerase II (9-22). These TFs are named TFII for TFs associated with RNA polymerase II followed by a letter (A, B, D, E, F or H). Other TFs bind to DNA sequences that control the expression of distinct genes and are thus responsible for regulating gene expression.

TFIID is the first TF to bind to the core promoter sequence and is made up of a variety of proteins, including a TATA-binding protein (TBP) that recognizes the TATA sequence at all promoter sites. The remaining proteins in TFIID are called TBP-associated factors. Once TFIID is bound at the TATA sequence, a preinitiation complex is formed with the recruitment of TFIIA, TFIIB, TFIIF/RNA polymerase II, TFIIE and TFIIH (Figure 3). The synthesis of mRNA then proceeds with the movement of RNA polymerase II away from the promoter region, and elongation of the mRNA transcript.

ACTIVATION OF TRANSCRIPTION

A variety of short cis-acting sequences (Figure 1, Table 1) that are located upstream of the TATA sequence facilitate the efficient and specific recognition of the core promoter by the basal transcription complex (12,13). The sequences include the common sequences found in RNA polymerase II promoters, including CAAT, octamer and GC. Specific upstream TFs recognize these sequences and bind to the DNA through a set of interactions among the DNA binding domain of the TF, the DNA sequence and the amino acid sequence of the TF. For example, the upstream TF, Sp1, binds to GC sequences and subsequently interacts with the TFIID bound at the TATA box to activate transcription.

The activation mechanism for transcription of some classes and families of genes is shared under specific conditions. For example, exposure of cells to glucocorticoids or phorbol esters elicits a specific induction of the transcription of all of the genes induced by these molecules. These inducible responses are attributed to upstream RE sequences in special promoters and function as binding sites for specific inducible TFs. An example of inducible control is the binding of the factor AP1 (made up of subunits encoded by *fos* and *jun*) to the 12-*O*-tetradecanoyl-13-acetate-responsive element sequence (TGACTCA) in genes that are activated by phorbol esters, growth factors or cytokines. In the absence of phorbol ester, AP1 is phosphorylated and then cannot bind to DNA (ie, inactive). The activation of AP1 involves its dephosphorylation such that it may bind to promoters containing TRE sequences. The binding of AP1 increases the rate of initiation of transcription.

Another example is steroid hormones that bind to specific receptors to form an activated complex that is capable of binding to RE sequences found in specific genes. Steroid receptor proteins comprise a DNA-binding domain that contains zinc finger motifs and a hormone-binding domain. Activated steroid-receptor proteins are essentially TFs that, when bound to RE sites in the DNA, activate transcription of a specific class of genes through activation of the initiation of transcription by RNA polymerase II. All genes that contain the common RE sequence are simultaneously activated. This allows the cell to coordinate the inducible expression of multiple genes collectively in response to specific hormone signals.

One important class of membrane protein receptors has intrinsic tyrosine kinase activity. The ligands of these receptors include growth factors and cytokines, both of which regulate cell growth. Important to this class of receptors are the signal transducers and activators of transcription (STATs). STATs are TFs that reside in the cytoplasm in an inactive form. The binding of cytokines to membrane-bound receptors leads to phosphorylation of the receptor by activation of the receptor tyrosine kinase activity. This provides a binding site for the STAT proteins. The bound STAT proteins are phosphorylated on tyrosine residues and undergo dimerization before migration to the nucleus. There they act as TFs by binding to specific DNA sequences upstream of the TATA sequence.

EUKARYOTIC REPRESSORS

Gene expression in eukaryotic cells is regulated by repressors as well as by activators (23). Repressors bind to specific DNA sequences and inhibit transcription through a variety of mechanisms. In some instances the repressors simply interfere with the binding of other TFs to DNA. Other repressors have been shown to compete with activators for binding to specific regulatory sequences. As a result, their binding to a promoter or enhancer blocks the binding of the activator, thereby inhibiting transcription. Other repressors contain specific functional domains called repression domains that inhibit transcription through protein-protein interactions.

The regulation of transcription by repressors as well as by activators extends the repertoire of mechanisms that control the expression of eukaryotic genes. One important role of repressors is the inhibition of expression of tissue-specific genes in appropriate cell types. Other repressors play key roles in the control of cell proliferation and differentiation in

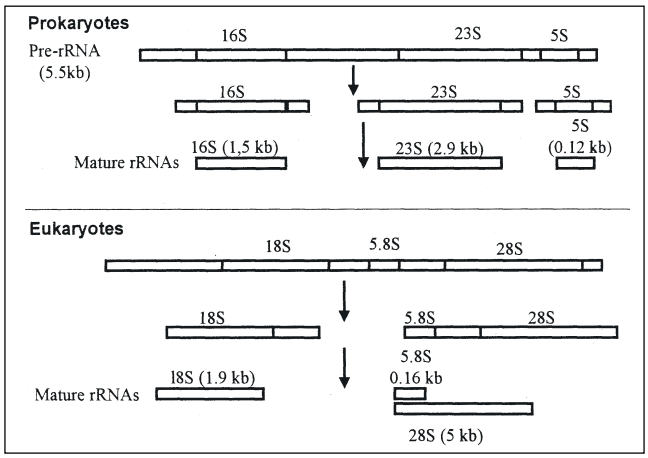

Figure 4) *The processing of ribosomal RNA (rRNA). Prokaryotic cells contain 3 rRNAs (16S, 23S and 5S) that are formed through cleavage of the pre-rRNA transcript. Eukaryotic cells contain four rRNAs. One of these (5S rRNA) is transcribed from a separate gene; the other three (18S, 28S and 5.8S) are derived from a common pre-rRNA. Following cleavage, the 5.8S rRNA (which is unique to eukaryotes) becomes hydrogen bonded to 28S rRNA*

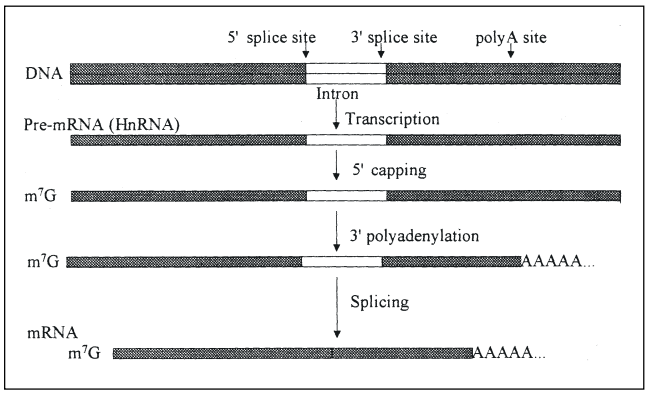

Figure 5) *The processing of eukaryotic messenger RNA (mRNA). The processing of mRNA involves modification at the 5′ end by capping with 7-methylguanosine (m7G), modification at the 3′ end by polyadenylation and removal of introns by splicing. The 5′ cap is formed by the addition of a GTP in reverse orientation to the 5′ end of the mRNA, forming a 5′ to 5′ linkage. The added guanine is then methylated at the N-7 position, and the methyl groups are added to the riboses of the first one or two nucleotides in the mRNA. HnRNA Heterogenous RNA*

response to growth factors as well as hormones. Such intricate control is especially important when considering the coordination required for maintaining the vertical crypt-villus and horizontal jejunoileo-colonic axis of the gut.

POST-TRANSCRIPTIONAL PROCESSING AND THE REGULATION OF EUKARYOTIC GENE EXPRESSION

The human genome contains coding information for approximately 100,000 different RNA molecules. However, within a single cell, different genes are expressed at different times through a process known as differential gene expression. Differential gene expression occurs in response to signals that occur during development, proliferation and differentiation. The orderly, programmed expression of every gene thereby plays a central role in cellular and whole organ homeostasis. Thus, it is not surprising that cells have evolved elaborate mechanisms that specifically control gene expression for particular genes. The pivotal step in all cells for the regulation of gene expression is at the level of transcription. The complex task of regulating gene expression in the many differentiated cell types in higher eukaryotes is a reflection of the combined actions of a diverse array of transcriptional regulatory proteins.

While the cellular events associated with the regulation of transcription are the predominant step in the regulation of eukaryotic gene expression, additional levels of control include the following (24-29).

- Controlling the processing of mRNA by determining which exons present in the initial mRNA transcript are retained in the mature and fully functional mRNA. Control mechanisms include either the alternative splicing of exons or the differential polyadenylation of the initial mRNA transcript.
- Controlling the stability or the rate of degradation of the mature mRNA transcript. As well, the packaging of DNA into chromatin and its modification by methylation add further dimensions to the control of eukaryotic gene expression.

RNA PROCESSING

The majority of newly synthesized RNAs are subsequently modified in a variety of ways to be converted to their functional forms. The regulation of the processing of RNA adds an additional level of control in eukaryotic gene expression (29).

RNA polymerase I is devoted to the transcription of rRNAs in the nucleolus. The processing of the 45S initial transcript, pre-rRNA, involves methylation of the RNA as well as ribonuclease-mediated cleavage of segments of the initial transcript to yield the 28S, 18S and 5.8S rRNAs (Figure 4).

The 5S tRNA is transcribed from a separate gene by RNA polymerase III, and the large precursor (pre-tRNAs) undergoes cleavage and methylation. The processing of the 3′ end of tRNA involves the addition of a CCA terminus, such that all tRNAs have the sequence CCA at the 3′ end. This sequence is the site of an amino acid attachment to the tRNA during protein synthesis.

In eukaryotic cells, the mRNA synthesized in the nucleus by RNA polymerase II is exported to the cytoplasm before it can be used as a template for protein synthesis. The initial products of transcription in eukaryotic cells (pre-mRNAs) are extensively modified before export from the nucleus. The processing of eukaryotic mRNAs is illustrated in Figure 5. This processing involves the modification of both ends of the mRNA, as well as the removal of introns from its mid-portion. The 5′ end of pre-mRNA is modified by the addition of a 7-methylguanosine cap. The 5′ cap has several

putative functions, including protecting from 5′ to 3′ exonuclease degradation, facilitating transport to the cytoplasm, facilitating RNA splicing and assisting in the alignment of mRNAs on the ribosomes during translation.

The 3′ end of most eukaryotic mRNAs is modified by a processing reaction called polyadenylation. The signal for polyadenylation is the hexanucleotide sequence AAUAAA. AAUAAA is recognized by a protein complex that cleaves the RNA chain 15 to 30 nucleotides further downstream. Subsequently, a poly A polymerase adds a poly A tail of approximately 200 nucleotides to the transcript. The initiation of polyadenylation heralds termination of transcription by RNA polymerase. The poly A tails have been envisaged to have several potential functions, including facilitating transport of mRNA molecules to the cytoplasm, stabilizing mRNAs to prevent degradation and facilitating translation by enhancing the recognition of the mRNA by the translational machinery.

Untranslated regions (UTRs) are found at both the 5′ and 3′ ends of the mRNA. UTRs are sequences in the exons that remain in the mRNA but are not translated into protein. The 5′ and 3′ UTRs contain signals that are necessary for the processing of the RNA and subsequent translation into protein.

SPLICING MECHANISMS

Most genes contain multiple introns, which account for about 10 times more pre-mRNA sequences than do the exons. Thus, the most striking modification of the pre-mRNAs involves the removal of introns by a process known as splicing (28,31,32). Splicing involves endonucleolytic cleavage and removal of intronic RNA, and end-to-end ligation (ie, splicing) of exonic RNA segments (Figure 6). The mechanism of RNA splicing is critically dependent on the GT-AG rule – introns start with GT and end with AG. The sequences adjacent to the GT and AG dinucleotides are highly conserved, and an additional conserved sequence situated just before the terminal AG at the end of the intron is the so-called branch site. The splicing mechanism is depicted in Figure 6, and involves the following steps:

1. Cleavage at the 5′ splice junction.
2. Joining of the 5′ end of the intron to an A within the intron (ie, branch site) to form a lariat-shaped structure.
3. Cleavage at the 3′ splice site leading to the release of the lariat-like intronic RNA and splicing of the exonic RNA segments.

Splicing occurs in large complexes called spliceosomes. The RNA components of the spliceosomes are small nuclear RNAs (snRNAs). These snRNAs range in size from approximately 50 to 200 nucleotides and are complexed with protein molecules to form small nuclear ribonucleoprotein particles (snRNPs). SnRNPs play an important role in the splicing process. The snRNA part of the snRNP carries out the 'intellectual task' of recognizing the splice and branch sites of the larger RNA molecule. In contrast, the protein part of the snRNP does the 'manual labour' of cutting and sticking the RNA molecule.

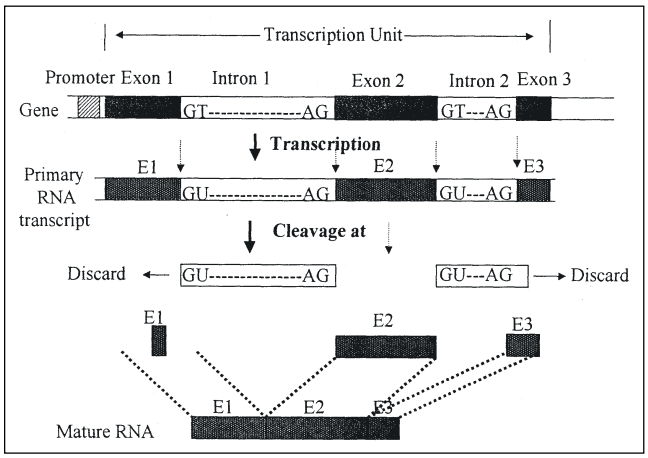

Figure 6) *Splicing of primary RNA transcripts. RNA splicing involves endonucleolytic cleavage and removal of intronic RNA segments, and splicing of exonic RNA segments*

The central role that splicing plays in the processing of pre-mRNA affords another mechanism for regulation of gene expression by the control of the activity of the cellular splicing machinery. Because most pre-mRNAs contain multiple introns, different mRNAs can be produced from the same gene by different combinations of the 5′ and 3′ splice sites. The possibility of joining exons in various combinations provides a novel mechanism for the control of gene expression through the generation of multiple mRNAs (and thus multiple proteins) from the same pre-mRNA. This process is termed alternative splicing and occurs frequently in genes of higher eukaryotes. Alternative splicing affords an important mechanism for the tissue-specific and developmental regulation of eukaryotic gene expression. In the case of transcriptional regulatory proteins, alternative splicing of pre-mRNAs yields products with dramatically different functions (ie, the ability to act as either activators or repressors of transcription). An important variation of the theme of splicing is a phenomenon known as trans-splicing, where exons originating from two separate transcripts are ligated together. The biological significance of trans-splicing remains to be elucidated.

EXON SELECTION DURING SPLICING

An additional level of control of gene expression occurs through the process of exon splicing during the processing of the pre-mRNA (33-35). The cell determines which exons present in the pre-mRNA are conserved in the final mRNA. This allows the production of more than one protein from the same gene. For example, the same gene encodes calcitonin and the calcitonin gene-related peptide (CGRP). These proteins differ with respect to their amino acid sequence, function and tissue localization. The synthesis of these different proteins using the same genetic information occurs by a combination of alternative polyadenylation and differential exon selection (Figure 7).

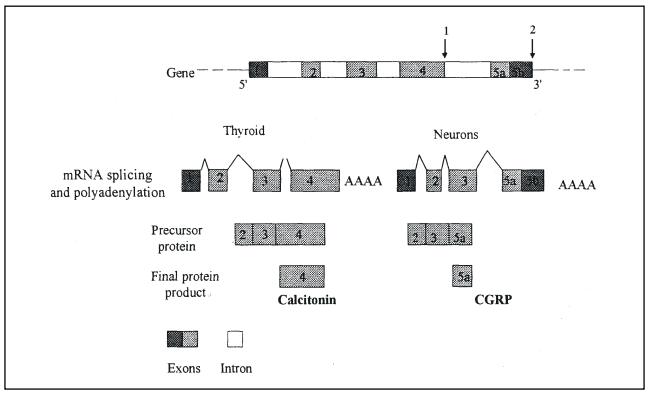

Figure 7) *The role of exon selection in the production of two proteins from the same gene. The calcitonin gene contains two polyadenylation signals and six exons (1, 2, 3, 4, 5a and 5b). In the thyroid, the upstream polyadenylation signal (arrow 1) is recognized, resulting in cleavage and polyadenylation of the messenger RNA (mRNA) at the 3′ end of exon 4 to produce a precursor mRNA containing exons 1, 2, 3 and 4. These four exons are spliced together, forming the mature mRNA, which codes for the calcitonin precursor peptide. This peptide is processed to yield calcitonin that contains amino acid sequence information only from exon 4. In neurons, the downstream polyadenylation signal (arrow 2) is recognized, resulting in cleavage and polyadenylation of the mRNA at the 3′ end of exon 5b to form a precursor mRNA containing exons 1, 2, 3, 4, 5a and 5b. During the splicing process, exon 4 is deleted, and the mature mRNA contains exons 1, 2, 3, 5a and 5b, which code for the calcitonin gene-related peptide (CGRP). The final processing gives CGRP, which contains the amino acid information that is found in exon 5a*

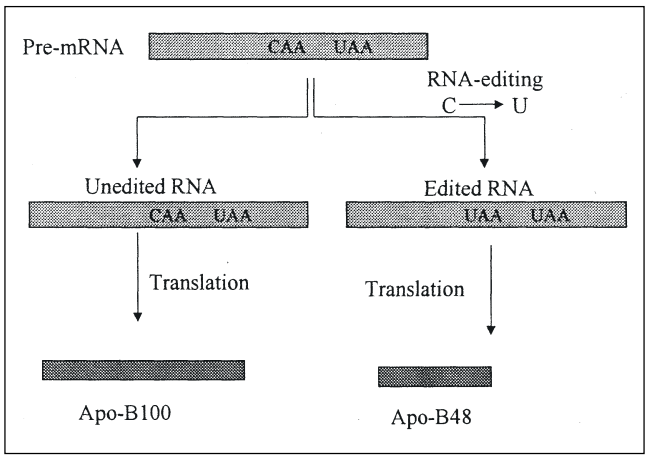

Figure 8) *The editing of apolipoprotein (apo) B mRNA. In the human liver, unedited messenger RNA (mRNA) is translated to yield a 4536-amino acid protein called apo-B100. In the human intestine, however, the mRNA is edited by a base modification that changes a specific cytosine to uracil. This modification changes the codon for a glutamine (CAA) to a termination codon (UAA), resulting in the synthesis of a shorter protein (apo-B48, consisting of only 2152 amino acids)*

RNA EDITING

The protein-coding sequences of some RNAs are altered by RNA processing events other than splicing. The best characterized example is the editing of the mRNA for apolipoprotein (apo) -B, where tissue-specific RNA editing gives rise to two different forms of apo-B (Figure 8) (36,37). Apo-B100 is synthesized in the liver by translation of the unedited

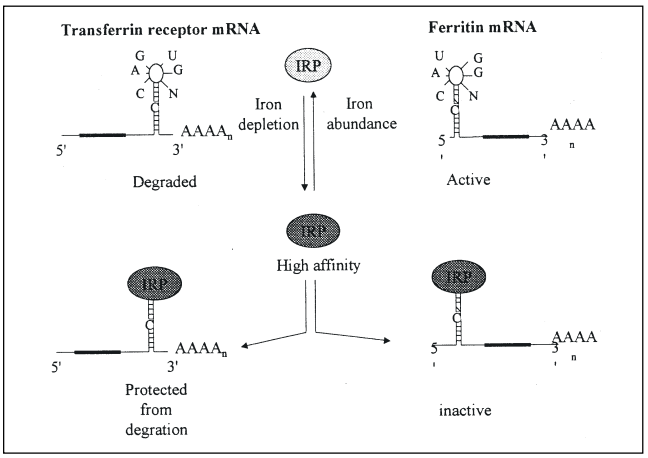

Figure 9) *The role of iron in the regulation of protein synthesis in the liver. A stem-loop structure is located at the 3′ end of the transferrin receptor messenger RNA (mRNA). Additional structures located at the 3′ end include the iron response element, which binds, and iron regulatory protein (IRP) when the cell is depleted of iron. The binding of the IRP to the 3′ end of the transferring receptor mRNA protects the mRNA from degradation and results in an increase in the level of the transferrin receptor mRNA and a corresponding increase in the level of the transferrin protein. At the 5′ end of the ferritin mRNA molecule is a stem-loop structure that binds IRP when iron is depleted in the cell. Binding of the IRP at the 5′ end of the ferritin mRNA inhibits the translation of this mRNA and results in a decreased level of ferritin protein. When iron levels are abundant, the ferritin mRNA no longer binds IRP and actively translated ferritin protein. At the same time, iron abundance inhibits IRP from binding to the 3′ end of the transferrin receptor mRNA, and the mRNA is degraded. This results in a reduction of the level of transferrin receptor protein*

mRNA, whereas a smaller intestinal protein, apo-B48, is synthesized as a result of translation of an edited mRNA where cytosine in a single codon has been changed to a uracil. This nucleotide substitution alters the codon for glutamine (CAA) in the unedited mRNA, to a translation termination codon (UAA) in the edited mRNA. This results in the synthesis of the shorter apo-B protein. This tissue-specific editing of apo-B results in the expression of structurally and functionally different proteins in the liver and intestine. The full length apo-B100 produced by the liver transports lipids of the circulation whereas apo-B48 mediates the absorption of dietary lipids by the small intestine.

RNA DEGRADATION

The final aspect of the processing of an RNA molecule is its eventual degradation (26,27,30). The intracellular level of any particular RNA species reflects a balance between synthesis and degradation. In this way, the rate at which particular RNAs are degraded constitutes another potential level at which gene expression can be controlled. In eukaryotic cells, different mRNAs are degraded at different rates, and this allows for the differential regulation of eukaryotic gene expression.

The degradation of most mRNAs is initiated by the trimming of the poly A tail. This is followed by the removal of the 5′ cap and degradation of the RNA by nucleases. The mRNA half-life varies from 30 mins to about 24 h. The

mRNAs with short half-lives usually encode for regulatory proteins.These mRNAs often contain specific AU-rich sequences situated near the 3′ end that appear to signal rapid degradation by promoting deadenylation at the 3′ poly A tail.

This stability of some mRNAs can be regulated in response to extracellular signals. For example, the level of abundance of the mRNA encoding the transferrin receptor, a cell surface protein involved in iron uptake, is regulated by the availability of iron (Figure 9). This occurs by modulation of the transferrin receptor mRNA stability. When iron is replete, the transferrin receptor mRNA is rapidly degraded by specific nuclease cleavage that occurs at a sequence near the 3′ end. When the supply of iron is rate-limiting, the transferrin receptor mRNA is stabilized, and this leads to an increased synthesis of transferrin receptor. Thus, more iron is transported into the cell. The regulation of the transferrin receptor is mediated by a protein that binds to specific sequences, called the iron response element, which is located near the 3′ end of the transferrin receptor mRNA. Binding protects the transferrin mRNA from cleavage and is controlled by the levels of intracellular iron.

PROMOTER SELECTION

The presence of more than one promoter within a particular gene can result in different amounts of the same gene product being produced in different tissues (38,39). Furthermore, tissue-specific availability of certain TFs also contributes to this process. For example, the alpha-amylase gene contains two promoter sites that control the expression of this gene in a tissue-specific manner. Salivary gland cells have very high levels of alpha-amylase, whereas hepatocytes have very low levels. The relative difference in amounts of alpha-amylase is controlled at the transcriptional level. In salivary gland cells, the first promoter site, located just 5′ to the first exon of the alpha-amylase gene, determines the start of transcription as well as the rate of gene transcription. This is a strong promoter because it has the ability to transcribe the gene at a high transcriptional rate. By contrast, in hepatocytes the available TFs do not recognize the first strong promoter of the gene and divert the RNA polymerase II to the second and weaker promoter located just 5′ to the second exon of the alpha-amylase gene. When the pre-mRNA is later spliced to form the mature mRNA, this results in the same alpha-amylase protein being transcribed, albeit at lower levels. This 5′ untranslated exon in each cell type is spliced to the first exon containing the amino acid sequence information. The final result is that the mature mRNA in hepatocytes differs from that found in salivary gland cells with respect to the 5′ untranslated sequence only (the amino acid coding regions are identical).

ALTERNATIVE POLYADENYLATION SITES

The differential production of the membrane or secreted form of immunoglobulin M (IgM) depends on the structure of the heavy chain component of the antibody molecule. The membrane form of IgM contains a heavy chain with a carboxy terminal amino acid sequence rich in hydrophobic amino acids that facilitates its interaction and binding to the cell membrane. In contrast, the secreted form of the antibody contains a heavy chain devoid of this carboxy terminal amino acid sequence and is unable to bind to the plasma membrane.

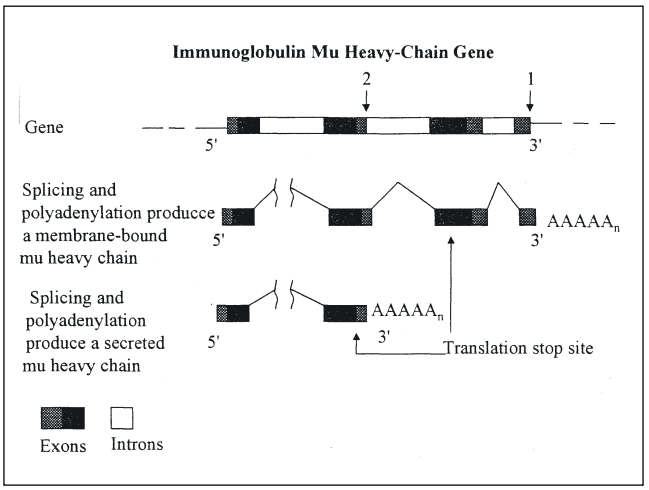

Figure 10) *Alternative polyadenylation in the immunoglobulin mu heavy chain gene. Only some of the exons of the mu heavy chain gene are represented. In cells in which the membrane-bound form of the antibody is produced, a polyadenylation signal present at the distal 3′ end of the messenger RNA (mRNA) indicated by arrow number 1, determines the site of cleavage and polyadenylation of the mRNA. During the splicing process of this mRNA, all of the exons, including the 3′ exon coding for the hydrophobic amino acids found at the carboxy terminal end of the membrane-bound form of the mu heavy chain, are spliced together to produce the mature mRNA. In cells in which the secreted form of the antibody is produced, the upstream polyadenylation of the mRNA found in mature B cells. The mRNA produced after splicing is missing the exons located 3′ to this polyadenylation signal. This results in the production of a mu heavy chain devoid of a hydrophobic tail, and is thus secreted*

By using alternative polyadenylation signals within the gene, the precise type of heavy chain mRNA is determined during B cell development (Figure 10) (25). When the mRNA encoding the membrane form of the heavy chain is produced, a polyadenylation signal present at the distal 3′ end of the message determines the site of cleavage and polyadenylation of the mRNA. After polyadenylation of the mRNA, all of the exons are spliced, including the 3′ exon that codes for the hydrophobic amino acid sequence located at the carboxy terminal end of the membrane-bound form of the heavy chain. This yields the mature mRNA. Translation of this particular mRNA produces a form of the heavy chain that has a hydrophobic tail and is found in the membrane-bound form of IgM. In contrast, the cells in which the secreted form of the IgM molecule is produced, a second polyadenylation signal, which is recognized by the cell-specific polyadenylation system of mature B cells, is located further upstream of the distal 3′ polyadenylation signal. In these cells, the cleavage and polyadenylation of the mRNA occur at this second site, and the exons located 3′ to this site are no longer present in the mRNA produced. Following poly-

adenylation, the remaining exons are spliced together to yield an mRNA that encodes a heavy chain that is lacking the hydrophobic tail. Translation of this mRNA produces the heavy chain found in the secreted form of IgM.

DNA METHYLATION AND THE CONTROL OF TRANSCRIPTION

Not only is methylation important in DNA synthesis and repair, but it is also another general mechanism associated with the control of eukaryotic gene transcription (24). Cytosine residues in eukaryotic DNA are modified by the addition of methyl groups. DNA is methylated specifically at the cytosines that precede guanines (CpG dinucleotides). Methylation is correlated with reduced transcriptional activity of several genes. Distinct patterns of methylation are seen in different tissues. The DNA of inactive genes is more heavily methylated than the DNA of genes that are actively transcribed. Moreover, some genes contain high frequencies of CpG dinucleotides in the region of their promoters. Transcription of these genes is repressed by methylation through the action of a protein that binds specifically to methylated DNA and inhibits transcription.

CONCLUSIONS

In the present review, we have examined the salient features pertaining to the molecular structure and function of the transcriptional complex and the mechanisms that modulate this pivotal step in the control of gene expression in eukaryotic cells. The subsequent processing of newly transcribed mRNA through events that include splicing, editing and polyadenylation, underscore the importance of post-transcriptional RNA processing as a major theme in eukaryotic gene expression.

ACKNOWLEDGEMENTS: This work was supported by operating grants from the Medical Research Council of Canada and the Crohn's and Colitis Foundation of Canada. Dr Gary E Wild is a chercheur boursier clinicien of Les Fonds de la Recherche en Sante du Québec. Dr Wild extends his appreciation to Drs David Fromson, John Southin, Howard Bussey and Bruce Brandhorst of the McGill Biology Department. Their tireless efforts in the area of undergraduate science education fostered a sense of inquiry and collegiality that guided a cohort of students through the early Recombinant DNA era.

REFERENCES

1. Lodish H, Baltimore D, Berk A, Zipursky SL, Matsudaira P, Darnell J. Molecular Cell Biology, 3rd edn. New York: WH Freeman, 1995.
2. Alberts B, Bray D, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Essential Cell Biology – An introduction to the Molecular Biology of the Cell. New York: Garland Publishing 1998.
3. Cooper GM. The Cell – A Molecular Approach. Washington: ASM Press, 1997.
4. Lewin B. Genes VI. New York: Oxford University Press, 1997.
5. Glick BR, Pasternak JL. Molecular Biotechnology – Principles and Applications of Recombinant DNA, 2nd edn. Washington: ASM Press, 1998.
6. Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD. Molecular Biology of the Cell, 3rd edn. New York: Garland Publishing, 1994.
7. Jameson JL. Principles of Molecular Medicine. New Jersey: Humana Press, 1998.
8. Strachan T, Read AP, Human Molecular Genetics. New York: Wiley-Liss, 1996.
9. Buratowski S. Mechanisms of gene activation. Science 1995;270:1773-4.
10. Grunstein M. Histones as regulators of genes. Sci Am 1992;267:78-84.
11. Paranjape SM, Kamakaka RT, Kadonaga JT. Role of chromatin structure in the regulation of transcription by RNA polymerase II. Annu Rev Biochem 1994;63:265-97.
12. Tijan R. Molecular machines that control genes. Sci Am 1995;272:54-61.
13. Tjian R, Maniatis T. Transcriptional activation: a complex puzzle with few easy pieces. Cell 1994;77:5-8.
14. Goodrich JA, Cutler G, Tjian R. Contacts in context: promoter specificity and macromolecular interactions in transcription. Cell 1996;84:825-3.
15. Beato M, Herrlich P, Schultz G. Steroid hormone receptors: many actors in search of a plot. Cell 1995;83:851-7.
16. Gehring WJ, Qian YQ, Billeter M, et al. Homeodomain-DNA recognition. Cell 1994 Jul 29;78:211-23.
17. Maniatis T, Goodbourn S, Fischer JA. Regulation of inducible and tissue-specific gene expression. Science 1987;236:1237-45.
18. Pabo CO, Sauer RT. Transcription factors: structural families and principles of DNA recognition. Annu Rev Biochem 1992;61:1053-95.
19. Buratowski S. The basics of basal transcription by RNA polymerase II. Cell 1994;77:1-3.
20. Conaway RC, Conaway JW. General initiation factors for RNA polymerase II. Annu Rev Biochem 1993;62:161-90.
21. Young RA. RNA polymerase II. Annu Rev Biochem 1991;60:689-715.
22. Zawel L, Reinberg D. Common themes in assembly and function of eukaryotic transcription complexes. Annu Rev Biochem 1995;64:533-61.
23. Hanna-Rose W, Hansen U. Active repression mechanisms of eukaryotic transcription repressors. Trends Genet 1996;12:229-34.
24. Bird A. The essentials of DNA methylation. Cell 1992;10;70:5-8.
25. Staudt LM, Lenardo MJ. Immunoglobulin gene transcription. Annu Rev Immunol 1991;9:373-98.
26. Beelman CA, Parker R. Degradation of mRNA in eukaryotes. Cell 1995;81:179-83.
27. Foulkes NS, Sassone-Corsi P. More is better: activators and repressors from the same gene. Cell 1992;68:411-4.
28. Green MR. Biochemical mechanisms of constitutive and regulated pre-mRNA splicing. Annu Rev Cell Biol 1991;7:559-99.
29. Keller W. No end yet to messenger RNA 3′ processing! Cell 1995;81:829-32.
30. Klausner RD, Rouault TA, Harford JB. Regulating the fate of mRNA: the control of cellular iron metabolism. Cell 1993;72:19-28.
31. Maniatis T. Mechanisms of alternative pre-mRNA splicing. Science 1991;251:33-4.
32. McKeown M. Alternative mRNA splicing. Annu Rev Cell Biol 1992;8:133-55.
33. Bennett MM, Amara SG. Molecular mechanisms of cell-specific and regulated expression of the calcitonin/alpha-CGRP and beta-CGRP genes. Ann N Y Acad Sci 1992;657:36-49.
34. Zandberg H, Moen TC, Baas PD. Cooperation of 5′ and 3′ processing sites as well as intron and exon sequences in calcitonin exon recognition. Nucleic Acids Res 1995;23:248-55.
35. Lou H, Cote GJ, Gagel RF. The calcitonin exon and its flanking intronic sequences are sufficient for the regulation of human calcitonin/calcitonin gene-related peptide alternative RNA splicing. Mol Endocrinol 1994;8:1618-26.
36. Chan L. Apolipoprotein B messenger RNA editing: an update. Biochimie 1995;77:75-8.
37. Chan L, Chang BH, Nakamuta M, Li WH, Smith LC. Apobec-1 and apolipoprotein B mRNA editing. Biochim Biophys Acta 1997;1345:11-26.
38. Schibler U, Hagenbuchle O, Wellauer PK, Pittet AC. Two promoters of different strengths control the transcription of the mouse alpha-amylase gene Amy-1a in the parotid gland and the liver. Cell 1983;33:501-8.
39. Sierra F, Pittet AC, Schibler U. Different tissue-specific expression of the amylase gene Amy-1 in mice and rats. Mol Cell Biol 1986;6:4067-76.