Synthesis and processing of RNA in prokaryotic and eukaryotic cells

I. Opening remark

Productive transcription initiation results in the RNA polymerase clearing out the promoter and transcribing a gene, thus generating an RNA transcript which is an exact copy of the coding strand. Two main classes of RNA can be distinguished: non-coding and coding RNA. The former comprises two main types in both prokaryotes and eukaryotes: transfer RNA (tRNA) and ribosomal RNA (rRNA). Ribosomal RNAs that are the most abundant RNA in the cell are components of the ribosomes, the protein-synthesis factories. Transfer RNAs, which are small molecules involved in protein synthesis, carry amino acids to the ribosome during translation of the messenger RNA (mRNA). Coding RNA corresponds to only one class of molecules: mRNA. All the protein-coding genes are transcribed into mRNA, the latter’s makes up about 3% of the total RNA. The entire mRNA content in a given cell is called transcriptome, the RNA part that should be changed by different patterns of transcription initiation events, in response to signals that require an alteration in the biochemical properties of the cell. Although changes in transcriptome result mainly from control over transcription initiation, some are due to regulatory processes that act post-transcriptionally. Attention should be drawn to the fact that transcriptome is not synthesized de novo, rather it is maintained by the cell during all its lifetime. After cell division, each daughter cell inherits a part of its parent's transcriptome, the latter being maintained by transcription, which replaces mRNAs that have been degraded, and accounts for the required changes by switching on and off of different groups of genes.

II. Precursor RNA are modified in prokaryotes and eukaryotes

The RNA content of a given cell is divided into different RNA types, some are common to all organisms and others are specific of either prokaryotic or eukaryotic cell. Description of different classes of RNA has been provided elsewhere. In addition to mature RNA molecules, cells also have precursor molecules that should be processed to acquire their activities. These RNAs are initially synthesized as pre-RNAs that undergo various processing events before they can carry out their functions. All the processing events we will be dealing with are described in table 1. These include the following:

1. End-modifications that occur especially in eukaryotes (they were also reported in Archaea). These comprise capping of the 5'-end and addition of a poly(A) tail to the 3'-end of mRNAs.
2. Splicing of the RNA transcript of a split or interrupted gene\(^1\) in order to remove the introns. The process involves the precise breakage of a phosphodiester bond at each exon-intron junction followed by the formation of a phosphodiester bond between the exon at the

\(^1\) Split gene is a structural gene (encoding a protein, rRNA or tRNA) that contains one to many intervening sequences or introns, which do not contribute to the structure of the gene product.
5'-end of the intron and the one at the 3'-end. As described before, unspliced pre-mRNAs form the nuclear fraction called heterogeneous nuclear RNA (hnRNA).
3. Cutting events are essential to produce mature rRNA and tRNA, which are initially synthesized from transcription units that specify more than one molecule. In both prokaryotes and eukaryotes, the pre-rRNAs and pre-tRNAs should be processed by cutting to generate mature RNAs.
4. Chemical modification concerns all three classes of RNA: mRNA, rRNA and tRNA. The tRNAs and rRNAs of all organisms are modified chemically by the addition of new chemical groups to specific nucleotides within the RNA molecule. Chemical modification of mRNA is called RNA editing and is specific of eukaryotes. This aspect of mRNA processing, by changing the coding properties of the mRNA, can result in a single pre-mRNA being converted into two different mRNA that code for distinct proteins.

It is now accepted that processing events have a variety of functions and that gene expression and modulation can be regulated over processing reactions. One of the processing reactions called alternative splicing (see below) is particularly important in increasing the coding capabilities of the genome without the requirement for additional genes. Indeed, by alternative splicing, one pre-mRNA can give birth to two or more mRNAs by different combinations of exons.

III. Synthesis and processing of bacterial mRNA

In contrast with eukaryotic mRNA, bacterial mRNAs undergo no significant processing because the primary RNA transcript produced by RNA polymerase is itself the mature mRNA. The translation of the bacterial mRNA starts before transcription is complete. As we described in the former chapter coupling of transcription and translation allows special types of controls to be applied to the regulation of bacterial mRNA synthesis. The first type of these controls is called attenuation². The second type of control is called antitermination. In this process the RNA polymerase ignores a termination signal and continues the elongation until a second termination signal is reached. Whether or not the genes at the end of an operon are switched on depends on the recognition by the RNA polymerase of a termination signal located upstream of these genes. This provides a mechanism by which genes can be switched on or off in response to prevailing conditions. Antitermination is controlled by a protein called antiterminator that

² Attenuation is a regulatory mechanism in which the expression of downstream genes is prevented by the termination of transcription at the attenuator, which is a rho-independent terminator located in the leader region of the mRNA. Whether or not transcription of the structural genes of the operon is terminated at the attenuator depends on the secondary structure adopted by the leader region of the mRNA. In one conformation the hairpin structure of the terminator is formed and transcription is terminated, while in the alternative conformation the terminator structure is not formed and transcription proceeds into the structural genes. Attenuation is involved in the regulation of many operons concerned with amino acid biosynthesis. In these operons the leader encodes a short polypeptide containing high proportions of the amino acid whose synthesis is controlled by the operon. If the amino acid is available, the leader peptide is synthesized and the mRNA adopts the conformation that terminates transcription at the attenuator. On the other hand, if the level of the amino acid falls, the translation of the leader peptide by the ribosome stalls and the mRNA adopts the alternative conformation (also called antiterminator but do not confuse with antitermination which occurs through a different mechanism) and transcription continues through the attenuator and into the structural genes downstream.
attaches to DNA and then transfers to RNA polymerase as it moves past toward the first termination signal. The presence of the antiterminator will cause the RNA polymerase to ignore the terminator (either intrinsic or rho-dependent terminator), most probably by preventing its destabilizing properties. Bacteriophage λ provides one of the best-studied example of antitermination (its life cycle will be described elsewhere).

Table 1. Description of the different RNA processing events

<table>
<thead>
<tr>
<th>Processing reaction</th>
<th>Functions (or possible functions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mRNA end modification</td>
<td>Transport of mRNA from nucleus to the cytoplasm and initiation of translation</td>
</tr>
<tr>
<td>Capping</td>
<td></td>
</tr>
<tr>
<td>Polyadenylation</td>
<td>Transcription termination by RNA polymerase II, initiation of translation and mRNA turnover</td>
</tr>
<tr>
<td>Cutting events</td>
<td>Release mature rRNAs and tRNAs from the precursor molecules</td>
</tr>
<tr>
<td>Splicing</td>
<td>Removal of introns from pre-RNAs</td>
</tr>
<tr>
<td>Alternative splicing</td>
<td>Increases in the coding capacity of the genome</td>
</tr>
<tr>
<td>Chemical modification</td>
<td></td>
</tr>
<tr>
<td>rRNA modification</td>
<td>Increases the range of rRNA catalyzed reactions</td>
</tr>
<tr>
<td>tRNA modification</td>
<td>Recognition of tRNAs by aminoacyl-tRNA synthetases and wobbling3</td>
</tr>
<tr>
<td>RNA editing</td>
<td>Enables one pre-mRNA to code for two proteins and converts non-functional transcripts into functional ones</td>
</tr>
</tbody>
</table>

Transcription elongation involves polymerization of ribonucleotides in the order specified by the template strand, their identity being determined by the base-pairing rules. During this process the β- and γ-phosphates are removed from the incoming NTPs and the hydroxyl group is removed from the nucleotide at the 3'-end of the growing chain leading to a phosphodiester bond. During this elongation phase, the RNA polymerase moves along the DNA covering some 30bp and maintaining an open region called transcription bubble of approximately 15-20bp. Inside the transcription bubble the RNA nascent chain is held to the template strand by a stretch of 8-12 RNA-DNA base pairs (Fig. 1).

Termination of bacterial transcription occurs either at intrinsic or rho-dependent terminators. At present, transcription is viewed as a step-by-step competition between elongation and termination. According to this model, the RNA polymerase pauses at each position and makes a decision between dissociating from the template or continuing elongation. Termination takes place when the RNA polymerase reaches a position, such as an intrinsic terminator, where termination and thus dissociation from DNA template is more favorable than elongation.

3 Wobbling is the recognition of two or more codons by one tRNA. The wobble hypothesis was originally proposed by Francis Crick in 1966 to account for the observed pattern of degeneracy in the 3rd base of a codon; it explains how certain tRNAs can recognize more than one codon when the codons differ only in the 3rd position. This hypothesis will be described in details in Chapter VII entitled "Synthesis and processing of the proteome".
Figure 1. Scheme of the transcription elongation complex in *E. coli*. During the elongation process, the RNA polymerase comprising four subunits (2α, β and β’ with a molecular mass of 35, 150 and 160kD, respectively) covers about 30bp of the template DNA including the transcription bubble in which RNA synthesis occurs. In this model the RNA polymerase makes contact with the DNA/RNA through three distinct sites: i) The double-stranded DNA binding site (DBS), which forms a contact with 9bp of DNA ahead of the transcription bubble. ii) The heteroduplex binding site (HBS), which binds to the hybrid region formed between DNA and RNA. iii) the RNA binding site (RBS), which binds with a short region of about 8 nucleotides of RNA, as it emerges from the elongation complex. This model, also called sliding clamp, combines the required tightness of binding with the flexibility needed to allow movement of RNA polymerase along the template DNA until the end of the gene is reached.

IV. Synthesis and processing of eukaryotic mRNA

The overall chemical process of RNA transcription is identical in all organisms. Moreover, the contacts between bacterial or eukaryotic RNA polymerase(s) with DNA template are probably similar. This conclusion results from the structural relatedness of the three eukaryotic RNA polymerases to their bacterial counterpart, their largest subunits being functionally equivalent to the α, β and β’ subunits of the *E. coli* RNA polymerase.

The elongation phase is crucial in eukaryotic transcription because transcripts can be very long. While bacterial genes are rarely more than a few kb in length, which means that their transcription is a matter of minutes, in eukaryotes a single gene might take several hours to be copied. For instance, the human dystrophin gene is 2400 kb in length and requires about 20 hours to be transcribed. Inferred from the structural relatedness of bacterial and eukaryotic RNA polymerases, the stepwise competition between elongation and termination holds also in eukaryotes. However, in the latter’s, transcription pausing and stopping should be reduced and stability of the transcription complex should be increased. This is achieved by the action of several elongation factors, proteins that associate with the RNA polymerase and promote its stability over DNA template. These factors suppress pausing (P-TEFb and TFIIS) or prevent arrest (ELL, Elongin and TFIIF) of RNA polymerase II.
As described above, bacterial protein-coding RNAs are not processed at all, the primary transcripts are mature mRNAs and their translation start before transcription is completed. In contrast, eukaryotic mRNAs are extensively processed during transcription by capping, polyadenylation and removal of introns. In the following sections we will address each of the modifications events that concern eukaryotic mRNAs. Capping and polyadenylation will be discussed in the following sections, while intron splicing will be treated as a distinct part together with other splicing processes concerning introns in eukaryotic pre-tRNAs and pre-rRNAs.

IV.1. Capping

Almost all eukaryotic mRNAs are capped at their 5' end. Capping, which is a multistep process, occurs shortly after transcription initiation i.e. as soon as the 5' end of the pre-mRNA has emerged from the transcription initiation complex. The first step is the addition to the newly-formed 5' end of the transcript of an extra and inverted guanosine residue, the latter being linked 5'-to-5' to the 5'-triphosphate of the terminal nucleotide. This reaction is catalyzed by the guanylyl transferase enzyme. The second step, which is catalyzed by guanine 7-methyltransferase, involves the transfer of a methyl group to the N7-position of the guanine cap to form the cap structure: m^7G^5ppp^5N^3 p^5N^3 p^5N^3 p^5N... The 7-methylguanosine structure is said to be of type 0 cap. Unicellular eukaryotes, such as the yeast *Saccharomyces cerevisiae*, have only type 0 caps. In higher eukaryotes additional modifications take place: a second methyl group replaces the hydrogen at O^2'-position of the next nucleotide resulting in type 1 cap. Moreover, if the second nucleotide is an adenine, a methyl group might also be added to the N^6'-position of the purine ring. Type 2 cap contains, in addition to type 1, a methyl group at the O^2'-position of the third nucleotide (Fig. 2). All RNAs generated by RNA polymerase II (mRNAs and snRNAs) are capped. In mRNAs, the 5' cap enhances translation by promoting the formation of translation initiation complex (see chapter VII: Synthesis and processing of the proteome). The recognition of the 5' cap by cap binding protein(s) (CBP) is an important factor in the regulation of gene expression. Furthermore, the 5' cap of snRNA is one of the feature that is recognized by some proteins, called importins, that mediate their transport back to the nucleus after they have traveled to the cytoplasm to collect the proteins required for their function.

IV.2. Polyadenylation

Most eukaryotic mRNAs have a Poly(A) sequence at the 3'-end, about 200-250 adenosines being added to the 3'-end of the primary transcript by a template-independent Poly(A) polymerase. This enzyme does not act at the extreme 3'-end of the transcript, but rather at an internal site generated by endonucleolytic cleavage. In fact, little is known about signals that regulate the termination of the transcription in eukaryotes, and primary transcripts show heterogeneity at their 3'-end indicating that the termination site is non specific. Despite the lack of information, transcription does not terminate until the RNA polymerase II has transcribed past a Poly(A) addition site, 5'-AAUAAA-3'. Addition of Poly(A) does not occur at
this conserved site, but immediately after the 5'-CA-3' dinucleotide located 10-30 nucleotides downstream. The dinucleotide is followed 10-20 nucleotides later by a GU-rich region. The Poly(A) addition site and GU-rich region are the binding sites for the cleavage and polyadenylation specificity factor (CPSF) and the cleavage stimulation factor (CstF), respectively (Fig. 3). The Poly(A) polymerase and at least one additional factor, called polyadenylate-binding factor, associate with bound CPSF and CstF in order for polyadenylation to occur. The polyadenylate-binding factor is thought to help the polymerase add the adenosines and influence the length of the Poly(A) tail.

Figure 2. Outline of capping of eukaryotic mRNAs. As shown in the figure, the reaction between the terminal triphosphate nucleotide and GTP involves the removal of the γ-phosphate of the terminal nucleotide and the β- and γ-phosphates of the GTP, resulting in the 5'-5' bond. Subsequent addition of a methyl group to nitrogen 7 of the purine ring yield type 0 cap (A). Asterisks indicate the positions where additional methylation produces type 1 (B) and type 2 (C) caps.

It was suggested that polyadenylation is the termination process for RNA polymerase II, however this does not hold true because some RNAs transcribed by the RNA polymerase II are not polyadenylated. The function of the poly(A) tail remains an open question; it does not appear to affect the stability of transcripts since some mRNAs (e.g. mRNAs encoding histones) are not polyadenylated. A possible role for Poly(A) tail in translation has been suggested, this proposal still has to be confirmed.
**Figure 3.** Polyadenylation of eukaryotic mRNAs. Poly(A) addition occurs at an internal site of the primary transcript defined by the consensus AAUAAA, located 10-30 nucleotides upstream from the dinucleotide CA. A 3'-OH end is generated by endonucleolytic cleavage of the nascent transcript after the CA dinucleotide, and Poly(A) polymerase adds 200-250 adenines using ATP as substrate. GU indicates a GU-rich region rather than the dinucleotide GU. For further details see text.

V. Intron splicing

Most nuclear structural genes in higher eukaryotes are split into coding sequences, termed exons, and non-coding regions, termed intervening sequences or introns. In eukaryotic microorganisms, such as *S. cerevisiae*, fewer nuclear genes contain introns, the latter's being smaller than those in higher eukaryotes. For example, the yeast *S. cerevisiae* genome contains 240 introns in total, whereas many individual human genes contain more than fifty introns. Table 2 summarizes the exon-intron content of some eukaryotic genes. Both introns and exons are found in the primary transcript, which subsequently should be processed by removing the introns and joining the exons to generate the mature RNA. Obviously, the mechanism by which the introns are removed and different exons are spliced together to produce translatable mRNA must be precise. The lengths of introns in eukaryotic genes vary considerably and range, in vertebrates, from a minimum of 60 bases to more than 10 kb.
Table 2. Introns and exons in some eukaryotic genes.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Size (kb)</th>
<th>Introns</th>
<th>exons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yeast</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Actin</td>
<td>1.0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Chicken</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pro α-2 collagen</td>
<td>40.0</td>
<td>50</td>
<td>51</td>
</tr>
<tr>
<td>Human</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insulin</td>
<td>1.4</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Serum albumin</td>
<td>18.0</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>Dihydrofolate reductase</td>
<td>31.0</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Type VII collagen</td>
<td>31.0</td>
<td>117</td>
<td>118</td>
</tr>
<tr>
<td>Factor VIII</td>
<td>186</td>
<td>25</td>
<td>26</td>
</tr>
</tbody>
</table>

V.1. Features of split or interrupted genes

Despite extensive variation in the split genes, some rules can be established for the distribution of introns and exons in the protein-coding genes.

1. Introns are less common in lower eukaryotes and much more frequent in higher eukaryotes, such as mammals.
2. The order of exons in the mRNA is the same as in the pre-mRNA and the DNA. Whenever the coding regions are separated by non-coding sequences, the order of exons is fixed and no random arrangement is observed.
3. The pattern and size of exons and introns are the same in all tissues and cells in a given organism. The no cell-specific rearrangement rule is applicable to all genes except those of the immune response and the major histocompatibility complex.
4. Most nuclear introns have non-sense codons in all three reading frames, so they are untranslatable.
5. When the same gene is compared in related species, some introns are found at the same positions but that each species has one or more unique introns. This indicates that some introns remain in place for millions of years while species diversify and other introns appear and disappear during that period. This led to the intron late hypothesis, which suggests that introns have evolved relatively recently and are gradually accumulating in the eukaryotic genomes.
6. Alternative splicing, which is the mechanism for production of multiple forms of mature mRNAs from the same gene, clearly increases the coding capacity of the genome (see below).

Despite the fact that the existence of introns has not been suspected until 1977 when the sequencing of DNA was devised and applied, we now recognize different types of introns in eukaryotes. All the seven types of introns are summarized in table 3. GT-AG introns, AT-AC introns, group I, group II and pre-tRNAs introns will be described in details in the following sections.
<table>
<thead>
<tr>
<th>Intron</th>
<th>Found in</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT-AG introns</td>
<td>Eukaryotic nuclear pre-mRNA</td>
</tr>
<tr>
<td>AT-AC introns</td>
<td>Eukaryotic nuclear pre-mRNA</td>
</tr>
<tr>
<td>Group I</td>
<td>Eukaryotic nuclear pre-rRNA, organelle RNAs and few bacterial RNAs</td>
</tr>
<tr>
<td>Group II</td>
<td>Organelle RNAs and some prokaryotic RNAs</td>
</tr>
<tr>
<td>Group III(^4)</td>
<td>Organelle RNAs</td>
</tr>
<tr>
<td>Twintrons(^5)</td>
<td>Organelle RNAs</td>
</tr>
<tr>
<td>Pre-tRNAs</td>
<td>Eukaryotic nuclear pre-tRNA</td>
</tr>
</tbody>
</table>

V.2. Nuclear mRNA introns

As described before protein-coding genes are transcribed by RNA polymerase II in the form of hnRNA (or pre-mRNA) which appear within the nucleus as hnRNP through association with a set of nuclear proteins. These proteins interact with the nascent chain as soon as it emerges from the transcription complex and, in doing so, maintain the hnRNA in untangled and accessible conformation. As a result, the substrate that enters the splicing pathway is the capped, polyadenylated hnRNA in the form of an hnRNP complex. Splicing occurs in the nucleus while the mature mRNA product is transported to the cytoplasm to be translated. Intron-exon borders must be recognized precisely in order for the splicing machinery to join accurately the exons. With most of the pre-mRNA introns the first two nucleotides of the intron sequence are 5'GT3' and the last two are 5'AG3'. These introns are therefore called GT-AG introns (are also written GU-AG) and all members of this type are spliced in the same way. The GU-AG motifs are part of longer consensus sequences that span the 5' and 3' splice sites as shown in figure 4. In addition to the splice junctions, a third conserved sequence within the intron, called the branch point, is essential to pre-mRNA splicing (Fig. 4). The branch point that lies 18 to 40 bp upstream of the 3'-splice site is represented, in higher eukaryotes, by the consensus YNYRAY where Y is any pyrimidine, R is any purine and N is any nucleotide.

The conserved sequences at the 5' and 3' exon-intron junctions indicate that these regions (in addition to the branch point) act as recognition regions for RNA-binding proteins playing a central role in splicing. A universal mechanism for the splicing of all nuclear pre-mRNAs has been proposed (outlined in figure 5). In terms of chemical reactions, the splicing pathway can be divided into two steps, which entail two transesterification reactions\(^6\).

1. Cleavage of the 5' splice site by the first transesterification reaction promoted by the hydroxyl group attached to the 2'-carbon of an invariant adenosine, located within the...

---

\(^4\) Group III introns that are found in organelle genomes are spliced autocatalytically (see self-splicing in section “Group II introns”) through a mechanism similar to that of group II introns. However, group III introns are smaller and have their own secondary structure.

\(^5\) Twintrons are composite structures made up of two or more group II and/or group III introns. Individual introns within a twintron are spliced in a defined sequence. For more details see Copertino D.W. and Hallick R.B. (1993) Group II and group III introns of twintrons: potential relationships with nuclear pre-mRNA introns. Trends Biochem. Sci., 18: 467-471.

\(^6\) In transesterification reaction, an OH group reacts with a phosphodiester bond, displacing an –OH to form a new phosphodiester link.
intron. In yeast, this adenosine is the last nucleotide of a conserved sequence 5′UACUAAC3′. The hydroxyl attack cleaves the phosphodiester bond at the 5′ splice site and is followed by the formation of a new 5′-2′ phosphodiester bond linking the intron's invariant first nucleotide G with the invariant branch site A. This results in a covalently closed loop of RNA, the lariat.

2. **Cleavage of the 3′ splice site and joining of the exons.** This results from a second transesterification reaction promoted by the 3′-OH attached to the 3′-end of the 5′ exon (upstream). The hydroxyl attack leads to cleavage of the phosphodiester bond at the 3′-splice site and the release of the lariat structure. At the same time, the 3′ end of the 5′ exon joins to the newly-generated 5′ end of the 3′ exon (downstream), thus completing the splicing process.

---

**Figure 4.** Features and conserved sequences of exon-intron boundaries in mammalian genes. In this scheme the pre-mRNA contains two exons separated by one intron. The GU-AG motifs are part of longer consensus: the 5′ splice site (also called donor site) and the 3′ splice site (also termed acceptor site) are shown. A pyrimidine-rich region, termed polypyrimidine tract, is found in higher eukaryotes just upstream of the 3′ end of the intron. Note that the branch point contains an invariant A, which is involved in the first transesterification reaction (see below).

Note that the reactions lead to no net change in the number of phosphodiester bonds, therefore, no energy input, such as ATP, is required. Moreover, the lariat product is unstable and is converted to linear excised intron by cleavage of the 2′-5′ phosphodiester branch. The debranched intron is subsequently degraded. In a chemical sense, intron splicing reactions are not more complicated than other biochemical reactions, however, two major topological problems can account for the complexity of the splicing apparatus that has evolved to deal with intron splicing. The first topological problem arises from the need to bring the splice sites into close proximity; these splice sites can be separated by a substantial distance, possibly several tens of kb. The second topological problem is related to selection of the correct splice sites. Selection of the wrong splice sites or cryptic splice sites (that may lie within intron or exon) results in either exon skipping (the loss of an exon from the mature RNA) or in the production of untranslatable RNAs (or immature RNAs). Both problems are avoided because splicing depends on snRNPs (also pronounced snurps), which are central components of the splicing
machinery. In higher eukaryotes, each snRNP consists of a small RNA molecule\(^7\) and a set of about 10 proteins. Some of the different proteins form a core set common to all snRNPs, while others are unique to a specific snRNP. The snRNPs are very abundant and some are present at greater than 100,000 copies per nucleus. The RNAs of snRNAs are rich in uridine, hence the classification of particular snRNPs as U1, U2, U4, U5 and U6.

\[ \text{Spliced mRNA} \quad \text{Debranched and degraded} \]

**Figure 5.** Splicing of pre-mRNA comprising two exons separated by one intron with the 5', 3' splice sites and branch point site. The 2'-OH attached to the branch site A attacks and cleaves the phosphodiester bond at the 5' splice site generating the lariat structure and a free 3'-OH at the end of the upstream exon. The latter is involved in the second transesterification reaction. This reacts with the 5' phosphate at the 3' splice site (the 5' end of the downstream exon), ligating the two exons and releasing the lariat structure, which is subsequently converted back to a linear RNA and degraded.

\(^7\) The snRNAs in yeast are much bigger, more than 1000 nucleotides in length.
Figure 6. Assembly of spliceosome and the roles of snRNPs in splicing of GU-AG and AU-AC introns (see text for details).

SnRNPs form a series of complexes that carry out the two steps of the splicing pathway. These snRNPs come together with the pre-mRNA to form a multicomponent complex called the spliceosome. Figure 6 outlines the assembly of spliceosome, which requires ATP:

1. Formation of the commitment complex which is a critical step of the splicing process. This complex comprises U1-snRNP and U2AF protein factors. U1 binds to the 5' splice site and U2AFs (U2AF^{35} and U2AF^{65}) bind to the polypyrimidine tract (a pyrimidine-rich region located just upstream of the 3’ end of the intron in higher eukaryotes).
2. The attachment of U2-snRNP to the branch site generates complex A.
3. Complex B is produced when U4/U6-snRNP (a unique snRNP containing two snRNAs) and U5-snRNP attach to complex A. U4/U6-snRNP binds to U2-snRNP and U5-snRNP attaches initially to the 3’ exon and migrates to the 3’ splice site of the intron. At this
stage, association between U1-snRNP and U2-snRNP brings the 5' splice site and branch site into close proximity.

4. **Complex C** is formed by the departure of U1-snRNP and U4-snRNP from complex B. When U1-snRNP dissociates from the complex, its role in binding the 5' splice site is taken over by U6-snRNP. The repositioning of U5-snRNP onto the 3' splice site brings all the key sites (5', 3' and branch sites) into proximity and the two transesterification reactions, possibly catalyzed by U6-snRNP, occur in parallel to complete the splicing process.

The second class of eukaryotic nuclear pre-mRNA introns comprises the AU-AC introns, which, to date, have been found in about 20 different genes in higher eukaryotes. The AU-AC motifs together with the invariant branch site, which is the last adenosine of the consensus 5'UCCUUAAAC3', are key features for the splicing of this class of introns. Their splicing pathway is identical to that for GU-AG introns, but involves a different set of splicing factors. The role of U1-snRNP, U2-snRNP, U4/U6-snRNP is taken by U11-snRNP, U12-snRNP and U4atac/U6atac-snRNP. Only U5-snRNP is common to the splicing apparatus of both types of introns.

Three processes may produce different mRNAs from the same gene: transcription initiation from different promoters, use of different terminators, and alternative splicing. Constitutive splicing refers to the removal of all introns from the pre-mRNA and splicing of all exons together, thus leading to one species of mRNA. On the other hand, selective use of different splice sites may lead to exon skipping so that different mRNAs, each containing different combinations of exons, are produced; this process is called alternative splicing. This enables different protein isoforms to be generated from the same gene and increases, consequently, the coding capacity of the genome. Figure 7 illustrates this fact, a 4-exon gene may be spliced in 4 different ways producing 4 different species of mRNAs; these encode four protein isoforms. Hence, instead of designing 4 different genes for these protein isoforms, nature has devised a variant of the splicing pathway that allows a sort of regulation, which tailors the expression of a gene with respect to the cell-type need or developmental stage.

**V.3. Eukaryotic nuclear pre-rRNA introns belong to group I family**

Pre-rRNA introns are quite uncommon. These introns are members of group I family that are also found in fungal mitochondrial and chroloplast genomes, where they appear in pre-mRNA as well as pre-rRNA. They also occur in microbial eukaryotes, such as *Tetrahymena thermophila*, and a few were isolated in bacteria. They are characterized by a common splicing mechanism and by a number of conserved sequences. Interaction between these sequences folds the intron into a core structure in which the 5' and 3' splice sites are brought into close proximity. The splicing pathway for group I introns is similar to that of pre-mRNA introns in that two distinct and independent cleavage-ligation steps are involved. In the first transesterification reaction hydroxyl attack is induced by a free guanosine or guanine nucleotide (GMP, GDP or GTP), but not by a nucleotide within the intron. This attack results in the cleavage of the 5' splice site, followed by the formation of a phosphodiester bond between the 3'-OH of a guanine nucleoside or nucleotide and the 5'-end of the intron. The
second transesterification reaction in promoted by the 3'-OH at the 3’ end of the upstream exon which attacks the 3’ splice site. This results in the cleavage of the 3’ splice site and the ligation of the 5’ and 3’ exons. The released intron is linear, rather than the lariat observed with pre-mRNA introns, however, the excised intron may circularize by additional transesterification reaction in which the 3’-terminal residue forms a covalent bond at a site near the 5’ end. This generates circular products that are subsequently degraded.

The remarkable feature is that excision of this type of intron is an intrinsic property of the RNA itself, i.e., group I introns are self-splicing; the splicing pathway proceeds in the absence of proteins and hence is autocatalytic, the RNA itself possessing the enzymatic activity. This RNA is termed Ribozyme or RNA enzyme and the first example was discovered in early 1980s. Although, the RNA structure is sufficient for splicing, it is possible that some introns require a protein factor to promote its stability. Many organelle introns contain an open reading frame coding for a protein called maturase, which appears to play a role in splicing.

![Figure 7](image)

**Figure 7.** Alternative splicing. The pre-mRNA containing 4 exons and 3 introns can be spliced in different ways to produce different combinations of exons. As a result, 4 different species of mRNA are generated. Translation of these mRNAs produce slightly different proteins (isoforms) that are believed to fit better the need of different cell types or, alternatively, enter into play at different developmental stages.

V.4. Group II introns

These introns are distinct from group I introns in their conserved sequences and secondary structure. Group II introns occur in fungal mitochondrial protein-encoding genes and in *Euglena* chloroplast genes. They are also found in the organelle genomes of plants and some were isolated in prokaryotes. Splicing reaction does not require guanosine or any other nucleotide, and the excised intron occurs in the form of a lariat similar to that formed by nuclear pre-mRNA splicing. Moreover, group II introns also differ from group I family in having at the splice sites conserved sequences which are similar to those in nuclear pre-mRNA introns. Nevertheless, splicing of group II introns is similar to that of group I introns in being
self-spliced or autocatalytic. In addition, certain group II introns have an open reading frame encoding for a maturase. Thus, group II introns seem to be intermediate, in terms of splicing mechanism features, between group I and nuclear pre-mRNA introns.

Figure 8. Outline of the tRNA splicing pathways in eukaryotes. Endonuclease cleavage leads to the removal of the intron and generates two tRNA half molecules. These two tRNA's exons are ligated together via an unusual 3',5' phosphodiester, 2'-phosphomonoester linkage. The transfer of the 2' phosphate to an NAD produces Appr>\text{p}, which is hydrolyzed to Appr-1"\text{p} by the action of Cyclic Phosphodiesterase (CPDase). The biological function of Appr>\text{p} and Appr-1"\text{p} remains unknown. For further details, see text (taken from Nasr, F. and Filipowicz, W., 2000).

---

V.5. Eukaryotic nuclear tRNA introns

Some nuclear tRNA genes in yeast are split genes, the single intron they contain is short and located in the region corresponding to the anticodon arm of the pre-tRNA. Unlike other intron types, splicing of pre-tRNA introns does not involve any transesterification. Pre-tRNAs splicing initiated by endonuclease cleavage leads to the removal of the intron and the generation of mature tRNAs with an unusual 3',5'-phosphodiester, 2'-phosphomonoester linkage at the exon-exon junction.

The 2' phosphate is transferred by specific phosphotransferase to an NAD acceptor molecule to produce ADP-ribose 1''',2'''-cyclic phosphate (Appr>\(p\)) which is converted to ADP-ribose 1'''-phosphate (Appr-1''p) by the action of cyclic phosphodiesterase (CPDase). This pathway is conserved in Yeast, plants and vertebrates, although in the latter organisms most of the tRNA splicing proceeds through another route that joins the two tRNA exons by the regular 3',5'-phosphodiester linkage (Fig. 8). This conservation argues in favor of an important role for Appr>\(p\) and Appr-1''p molecules in the cell. To date, the biological role of Appr>\(p\) and Appr-1''p molecules remains unknown. However, the cloning of CPD1 gene in the yeast Saccharomyces cerevisiae, which hydrolyzes Appr>\(p\) into Appr-1''p, provided the possibility to investigate the role of these compounds in the cell.

Box 1 - Splicing and cancer

Alternative splicing is a crucial mechanism for generating protein diversity. Different splice variants of a given protein can display different and even antagonistic biological functions. Therefore, appropriate control of their synthesis is required to assure the complex orchestration of cellular processes within multicellular organisms. Mutations in cis-acting splicing elements or changes in the activity of regulatory proteins that compromise the accuracy of either constitutive or alternative splicing could have a profound impact on human pathogenesis, in particular in tumor development and progression.

The response of cells to environmental signals, as well as their differentiation, death or malignant transformation, involves changes in gene expression. For many years, the modulation of gene expression was thought to be restricted to, or at least dominated by, the control of transcription. The discovery of introns and the process of splicing introduced an exciting but also perturbing factor in our simplified conceptions of the flow of gene expression in eukaryotes. Introns must be excised precisely to generate bona fide mRNA molecules; otherwise, translational frame shifts would be introduced. Introns are much longer than exons and most of their nucleotide sequence seems to be irrelevant except for the ‘donor’ and ‘acceptor’ consensus sequences located at their extremities, and the branching sequence that precedes the acceptor sequence. The spliceosome, a sophisticated nuclear machine comprising five types of small nuclear ribonucleoprotein (snRNP) and hundreds of auxiliary proteins, takes care of intron excision and exon joining repeatedly at each and every one of the seven splicing events that occur on average per human pre-mRNA molecule. ‘What is all this for?’ We know now that the high adaptive value of introns resides in the fact that they permitted exon shuffling in the past and alternative splicing in the present.

---

10 Mutations in splicing elements, for example, have been found in genes such as LKB1, KIT, CDH17, KLF6 and BRCA1, and changes in trans-acting regulators can affect the expression of genes such as Ron, RAC1 and CD44.
11 According to the Jacob-Monod-Lwoff paradigm, established in bacteria and logically extended to eukaryotes, what made a cell a cell was the combinatorial turning on and off of genes.
However, the apparent ‘irrationality of introns’ was questioned early on following the discovery that mutations that affect splicing, quantitatively or qualitatively, are a widespread source of hereditary diseases. We will focus not only on mutations in cis-acting splicing sequences that are associated with cancer but also on the variations in normal splicing processes and the signals that may affect them in cancer cells.

Figure 1. Cis-acting sequences that control splicing. (A) Consensus sequences for the 5′ splice site (donor), branch site and 3′ splice site (acceptor). (B) Arrangement of donor (D) and acceptor (A) sites in the architecture of a typical eukaryotic gene. Whereas internal exons are limited by acceptor and donor sites, the first exon is limited by the Cap site and a donor, and the last exon is limited by an acceptor and the poly(A) site. (C) Examples of sequences for strong and weak acceptor sites. Red indicates the 3′ acceptor splice site; bold indicates deviations from the consensus sequence. (D) Schematic roles of exonic splicing enhancers (ESE), exonic splicing silencers (ESS), intronic splicing enhancers (ISE) and intronic splicing silencers (ISS) on the recognition of a weak acceptor by the splicing machinery; red indicates enhancing, blue indicates silencing. In A-D, exon sequences are boxed.

The basic language of splicing
Splicing occurs when a group of proteins and RNPs recognize specific RNA sequences conserved at the boundaries of introns (Fig. 1). The classical spliceosome, which acts on >99% of introns from genes transcribed by RNA polymerase II (Pol II), recognizes a 5′ donor splice site beginning

12 http://www.eurasnet.info/ian_eperon.shtml
13 No matter how absurd or energy consuming a biological process appears to be, if its disruption or perturbation causes disease, it must be important, and its conservation in evolution is paramount. The links between splicing and human disease have been extensively reviewed lately. The specific association between splicing and cancer has received less attention, perhaps because the field is still emerging.
with a GU dinucleotide and a 3' acceptor splice site ending with an AG dinucleotide (the GU-AG rule). A second type of spliceosome acts on a minor class of Pol-II-transcribed introns obeying the AU-AC rule. In the major class of introns, the 5' splice site, the branch site in between and the 3' splice site are recognized by U1 snRNP, U2 snRNP and the auxiliary factor U2AF, respectively. The splicing reaction is completed with the participation of the U4, U6 and U5 snRNPs. Every internal exon is flanked by upstream and downstream splice sites. The situation is different for the terminal exons: the upstream limit of the first exon is the Cap site or transcription initiation site, whereas the downstream limit of the last exon is the polyadenylation site (Fig. 1). These particular features are important considerations for alternative splicing (see below). When the sequence of a splice site deviates from the consensus shown in Fig. 1, the site can still be used, although less efficiently, depending on the number of base changes. These weak sites (Fig. 1) have less affinity for their spliceosomal protein or RNP partners and are the main cause of alternative splicing. A second class of cis-acting sequence can influence the recognition and use of weak sites by the splicing apparatus: the enhancers and silencers of splicing. These are short (~10 nucleotides) conserved sequences located in exons or introns, either isolated or in clusters, that stimulate or inhibit the use of weak splice sites (Fig. 1). An exonic sequence is defined as an exonic splicing enhancer (ESE) if its mutation reduces inclusion of the corresponding exon into the mature mRNA. Conversely, mutation of an exonic splicing silencer (ESS) increases inclusion of the exon. In most cases, the mechanisms of action of splicing enhancers and silencers involve the specific binding of regulatory proteins such as SR proteins (serine/arginine-rich proteins) or heterogeneous nuclear (hn)RNP s. Certain silencers, instead of binding regulatory proteins, form a particular pre-mRNA secondary structure that hinders the recognition of a neighboring splicing enhancer by SR proteins. Approximately 15% of mutations that cause genetic disease affect pre-mRNA splicing. Some disrupt or create splice sites. If a canonical splice site is completely disrupted, a cryptic splice site nearby is used instead, leading to aberrantly spliced mRNA molecules and failure to produce a functional protein. If the disruption is partial, the cryptic and mutated sites compete, leading to a mixed population of aberrant and normal mRNA molecules, with a reduction in normal protein levels. Creation of a new splice site has similar consequences: the new (aberrant) and the old site compete and there is a concomitant reduction in normal protein levels.14

Alternative splicing
When two or more splice sites compete, alternative splicing generates mRNA variants that yield different polypeptides from a single gene. Alternative splicing is more a rule than an exception: it affects an estimated 60% of human genes. Its regulation not only depends on the interaction of SR and hnRNP proteins with splicing enhancers and silencers, but is also coupled to Pol II transcription, as in the case of other pre-mRNA processing reactions. Fig. 2 illustrates the different modes of alternative splicing, which we distinguish from alternative transcriptional initiation.15

Cancer and mutations that affect splicing
Many examples of cancer-associated alterations in splicing are attributable to mutations that create or disrupt splice sites or splicing enhancers and silencers. However, in only a few cases has a cause-effect relationship been proved. LKB1 is a tumor suppressor gene that encodes a serine/threonine protein kinase involved in the control of several cellular processes, including cell-cycle arrest, p53-mediated apoptosis, Ras-induced transformation, cell polarity. Its second intron belongs to the minor spliceosome class that obeys the AU-AC rule. Mutations of LKB1 that lead to reduced levels of the protein are found in patients with Peutz-Jeghers Syndrome (PJS), an

14 Recently, the role of mutations that create or abolish splicing enhancers and silencers in disease has been examined. This is particularly important when one studies single base changes in exonic sequences. Traditionally, these base changes were assumed to produce nonsense, missense or silent substitutions that could only affect the quality of the encoded protein. However, we now know that in many cases they disrupt or create functional ESEs or ESSs, and provoke changes in the levels of inclusion of the exons to which they map.
15 Alternative transcriptional initiation generates mRNA diversity but is not alternative splicing because there are no competing splice sites.
autosomal dominant disorder associated with gastrointestinal polyposis and an increased cancer risk. Another interesting example involves the oncogene \textbf{KIT}. The encoded protein is a member of the type III receptor tyrosine kinase family whose constitutive activation is associated with gastrointestinal stromal tumors. Patients with deletions of an intron-exon segment encompassing the 3' splice site of intron 10 were reported; these deletions concomitantly create an intronic 3' splice site within exon 11. The resulting polypeptides remain in-frame but lack an internal stretch that is crucial for auto-inhibition of the kinase. Structural studies of the mutated kinase revealed a conformation consistent with constitutive activation. \textbf{p53} The delicate balance between cell proliferation, differentiation and death maintains tissue homeostasis within multicellular organisms. Deregulation of any of these processes can lead to tumorigenesis. Inactivation of the \textbf{p53} tumor suppressor gene is a very frequent event in human cancer. \textbf{p53} is a crucial protein involved in cell-cycle control, apoptosis and maintenance of genetic stability. It was thought to exist as a single isoform. However, multiple isoforms generated through the use of two different promoters and alternative splicing have recently been discovered, making \textbf{p53} similar to its relatives \textbf{p63} and \textbf{p73}. It has been shown that these isoforms are expressed in a tissue-dependent manner and that their expression pattern is altered in human breast tumors. The regulators of promoter selection and alternative splicing of \textbf{p53} still needs to be elucidated\textsuperscript{16}.

\textbf{Signal transduction, splicing and cancer}

Activation of a pathway involving \textbf{Ras}, \textbf{PI 3-kinase} (\textbf{PI 3-kinase/AKT} pathway) has been associated with multiple human cancers. This pathway is counterbalanced by the tumor suppressor phosphatase \textbf{PTEN}, which in turn is often inactivated in cancer. The \textbf{Ras/PI 3-kinase/AKT} pathway leads to changes in activity of \textbf{SR} proteins, in particular \textbf{SF2/ASF} and \textbf{9G8}. Consequently, it regulates alternative splicing of fibronectin transcripts. Furthermore, \textbf{AKT} can phosphorylate \textbf{SF2/ASF} and \textbf{9G8} \textit{in vitro}. Activation of this pathway by insulin regulates the activity of another \textbf{SR} protein, \textbf{SRp40}, stimulating the inclusion of an alternative exon in protein kinase C (\textbf{PKC} II) pre-\textbf{mRNA}. It is tempting to speculate that deregulation of the \textbf{Ras/ PI 3-kinase/AKT} pathway by activating mutations in its components or by \textbf{PTEN} inactivation would have dramatic consequences for the splicing pattern of any of the pre-\textbf{mRNAs} regulated by these splicing factors\textsuperscript{17}.

\textsuperscript{16} The \textbf{p53} regulators \textbf{MDM2} (and its human analog \textbf{HDM2}) and \textbf{MDMX} (and its human analog \textbf{HDMX}) also undergo alternative splicing. Binding of \textbf{MDM2} to \textbf{p53} inhibits its transcriptional function and also facilitates its degradation by the proteasome. \textbf{MDMX} (also named \textbf{MDM4}) heterodimerizes with \textbf{MDM2}, affecting \textbf{MDM2} activity. More than 40 \textit{MDM2/MDM2} transcripts, including alternatively as well as aberrantly spliced forms, have been identified both in tumors and normal tissue. Some of these variants encode proteins that possess transforming properties \textit{in vitro} and \textit{in vivo}. Genomic mutations that can account for the observed usage of cryptic splice sites and the generation of aberrantly spliced isoforms have not been found. Moreover, an aberrantly spliced and tumor-specific \textbf{HDMX} isoform has been recently described. This transcript, isolated from a thyroid tumor cell line, encodes a protein named \textbf{HDMX211}, which enhances \textbf{MDM2} protein levels and counteracts its \textbf{p53}-degrading function.

\textsuperscript{17} Several reports implicate other \textbf{Ras}-dependent pathways in the regulation of alternative splicing of the genes encoding \textbf{agrin}, \textbf{CD44} and \textbf{CD45}. \textbf{CD44} is one of the most studied alternatively spliced genes in cancer because inclusion of variable exons correlates with tumor development and metastasis. In T-lymphoma cells, phorbol-ester-dependent activation of extracellular-signal-regulated kinase (\textbf{ERK}) leads to phosphorylation of the nuclear RNA-binding protein \textbf{Sam68}, which regulates alternative splicing of variable exon 5 in \textbf{CD44} . Scientists recently identified \textbf{SRm160} as another \textbf{Ras}-regulated splicing co-activator responsible for inclusion of this exon in \textbf{CD44}. Furthermore, they demonstrated that silencing of this factor decreases cell invasiveness, providing another link between regulation of alternative splicing and tumorigenesis. Several questions remain to be answered. First, whether changes in the splicing patterns of different genes involved in neoplastic transformation result from the deregulation of a certain set of splicing factors. Second, whether a given extracellular signal can affect different alternative splicing events that cooperatively participate in tumor progression.
Figure 2. (A) Different modes of alternative splicing. (B) Alternative promoters provoke mRNA diversity but do not necessarily imply alternative splicing. There are two alternative promoters, P1 and P2. If promoter P2 is used, exons 2, 3 and 4 are constitutively included. If promoter P1 is used, exon 2 simply does not exist for the splicing machinery because, being a 5' terminal exon of the P1 transcription unit, it lacks an upstream 3' splice site. Therefore, the splicing machinery has no option but to join exon 1 directly to exon 3, generating a 1-3-4 mRNA.

Signaling to the splicing machinery
It is becoming increasingly clear that alternative splicing is regulated by extracellular signals through the activation of complex networks of transduction pathways. However, data addressing how these extracellular signals impinge upon splicing factor activity are scarce. Serine phosphorylation of the arginine- and serine-rich (RS) domain is an important modulator of SR protein activity and localization. Protein kinases that phosphorylate SR proteins and their antagonistic hnRNP proteins could play a crucial role in linking extracellular cues to regulation of alternative splicing. A recent study supports a model in which SF2/ASF is first phosphorylated in the cytoplasm by SRprotein-specific kinase 1 (SRPK1) on only a few serine residues within the RS domain. This hypophosphorylated SF2/ASF is imported into the nucleus and stored in nuclear speckles. Release of SF2/ASF from these storage sites and its recruitment to active sites of transcription requires a second round of phosphorylation, which is carried out by CDC-like kinase 1 (CLK). Dephosphorylation of SF2/ASF is required for its activity during mRNA splicing, and so it is conceivable that it must undergo dephosphorylation by as-yet-unknown phosphatases. Although SRPK and CLK have been identified as SR protein kinases, signaling pathways that involve these kinases have not been identified. As already mentioned, AKT can also phosphorylate SR proteins, in particular SF2/ASF, 9G8 and SRp40. ERK phosphorylates the RNA-binding protein Sam68, which is involved in the regulation of CD44 alternative splicing. ERK-dependent activation of Sam68 triggers the formation of a macromolecular complex that...
Conclusion and perspectives
Cancer is a multistep process that involves severe changes in gene expression. The more we learn about regulatory pathways that are disturbed during tumorigenesis, the more we realize about the complexity of tissue homeostasis. Mutations that alter cis-acting splicing elements can modify mRNA quality and therefore protein function. Activation of signaling pathways that can affect the activity of splicing regulatory factors or modify the balance between them can also change the proportions of mRNA splicing isoforms. Both can lead to the deregulation of crucial cellular processes such as adhesion, proliferation, differentiation, death, motility and invasion (Fig. 3), all of which contribute to cancer. Although still in development, splice-isoform-sensitive microarrays will undoubtedly contribute towards the analysis of changes in splicing patterns associated with cancer. Establishing which genes actively participate in different steps of tumor initiation and progression, or whether deregulation of their splicing is a consequence of this, will require both high-throughput techniques and reductionist approaches in culture and animal models.

Figure 3. Causes and consequences of splicing pattern alterations. For simplicity, only SR and hnRNP proteins are shown. However, other factors not belonging to any of these families could also be targets of these signaling pathways.

contains Sam68, Pol II and Brm (a component of the SWI/SNF chromatin-remodeling complex). This interacts with the nascent transcript, stalling Pol II. Such pausing could favor inclusion of the variable exons within the mature mRNA, which would be consistent with the proposed kinetic coupling between transcription and splicing. The balance of different kinase activities could modulate the function of splicing regulatory proteins and consequently modify splicing patterns. But there are still more questions than answers. What is the SR protein specificity, if any, of all these kinases? Is there a precise order in which each phosphorylation takes place? Which residues are involved in each case? Does each kinase affect different SR protein properties (e.g. RNA binding, protein-protein interaction, or protein localization) or functions (e.g. splicing, surveillance, or translation)?
VI. Chemical modification of RNAs

VI.1. Processing of rRNAs and tRNAs by chemical modifications

The final step in the processing of RNAs involves chemical modification of some nucleotides within the RNA transcript. This processing event occurs with pre-rRNA and pre-tRNA of both bacteria and eukaryotes, and with pre-mRNA of eukaryotes (where it is called RNA editing). Over fifty chemical modifications were identified with different pre-RNAs, some of them are described in Table 4.

Table 4. Chemical modification affecting nucleotides in both rRNAs and tRNAs.

<table>
<thead>
<tr>
<th>Modification</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methylation</td>
<td>Addition of CH3 groups to the base or the sugar</td>
</tr>
<tr>
<td>Deamination</td>
<td>Removal of an NH2 group from the base</td>
</tr>
<tr>
<td>Nucleotide replacement</td>
<td>Replacement of an existing nucleotide with a new one</td>
</tr>
<tr>
<td>Double bond saturation</td>
<td>Converting a double bond to a single bond</td>
</tr>
<tr>
<td>Sulfur substitution</td>
<td>Replacement of oxygen with sulfur</td>
</tr>
<tr>
<td>Base isomerization</td>
<td>Changing the positions of atoms in the ring component</td>
</tr>
</tbody>
</table>

The vast majority of these modifications is carried out directly on the nucleotides within the transcript, but two modified nucleotides, queosine and wyosine, replace the existing nucleotides; this occurs by cutting out an entire nucleotide and replacing it with the modified version. In tRNAs, these modifications are thought to mediate recognition of individual tRNAs by aminoacyl-tRNA synthetases on one hand, and recognition of two or more codons by the same tRNA on the other hand. In rRNAs, chemical modifications concern mainly those parts of rRNAs known to play a critical role in the activity of these molecules. Thus, modified nucleotides might increase the range of reactions catalyzed by the rRNAs molecules. We know little about how these changes in tRNAs are carried out and how the enzymes, which are involved in the modification process, are directed to the correct nucleotides. Nevertheless, chemical modification of pre-rRNAs has progressed with the discovery of short RNA molecules called snoRNAs, located in the nucleolus where the processing takes place; these snoRNAs are involved in the modification process by directing the processing enzyme to the nucleotide that should be modified.

VI.2. RNA editing

In contrast with rRNAs and tRNAs, with which chemical changes only affect their structural features, chemical modification with mRNA has the possibility to change the coding properties of the transcript. RNA editing is a post-transcriptionally process which essentially affects the pre-mRNA transcripts (i.e. of protein-coding genes). RNA editing of primary transcript can result in two mRNAs coding for two distinct proteins. This means that RNA editing is essential for correct gene expression: the proteins translated from edited transcripts are different from the ones deduced from the genes in DNA. RNA editing was initially
described in trypanosomes mitochondria where the mRNA for coxII gene was found to contain four uridines that are not coded by the DNA. Later, post-transcriptional modification of transcripts by RNA editing was detected in different genetic systems affecting nuclear, mitochondrial and chloplast RNAs. RNA editing discovered in trypanosome involved deletion and insertion of uridines, however, this post-transcriptional process can also produce transcripts that contain adenosines, guanosines, cytidines and inosines that are not present in the genomic coding sequence. All these modifications can be obtained either by insertion-deletion events or by conversion mechanisms (table 5).

VII. Turnover of mRNAs

At any time, the transcriptome results from the synthesis of new mRNAs by transcription and their removal by degradation, the latter being a potentially important means that regulates gene expression during organism development and as a response to environmental conditions. Moreover, mRNA turnover contributes to accurate translation by eliminating aberrant transcripts. In bacteria, mRNAs are degraded by a multienzyme complex termed degradosome, which contains endonucleases and exonucleases. This multiprotein complex cuts the molecules at internal sites and remove nucleotides from both the 5’ and 3’ ends. In yeast, three degradation pathways have been described:

1. **Deadenylation dependent pathway**, which first involves the removal of the poly(A) tail either by exonuclease cleavage or by the loss of the polyadenylate binding protein that stabilizes the tail. This 3’-to-5’ decay pathway involves a complex of multiple 3’-to-5’ exonucleases, which is termed exosome. In a second time, the 5’ cap is removed by decapping enzyme followed by rapid exonuclease digestion from the 5’ end. The 5’-to-3’ pathway is the major mRNA degradation pathway in the yeast *S. cerevisiae* (in contrast, mRNAs in mammals may be primarily degraded through the 3’ to 5’ pathway).

2. **Deadenylation-independent pathway** in which cap cleavage and exonuclease degradation occur while the poly(A) tail is in place. This pathway was primarily designed to account for the degradation of those transcripts containing termination codons that prevent the mRNA from being translated correctly. At present, we have evidence that this pathway is also involved in the turnover of normal mRNAs.

3. **Nonsense-mediated mRNA decay or nonstop mRNA decay or NMD pathway** is concerned with degradation of mRNAs that lack a termination codon, thus preventing the synthesis of truncated proteins (potentially inactive or deleterious) when translation terminates prematurely. When the 80S ribosome stalls at the 3’ end of a nonstop mRNA the ribosomal A site will be associated with zero to two nucleotides of the poly(A) tail. The

---

19 Benne, R. *et al.* (1986) Major transcript of the frameshift coxII from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell* 46: 819-826.

20 In *E. coli* the degradosome consists of polynucleotide phosphorylase (PNPase), a DEAD-box RNA helicase, RNase E, enolase, and other proteins.

ribosomal A site will then be recognized by a protein (called Ski7p in *S. cerevisiae*) that will recruit the exosome. Eventually, the nonstop mRNA is degraded from the 3’ end\(^{22}\).

### Table 5. RNA editing events in different genetic systems.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Type of editing</th>
<th>Cell compartment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trypanosomes</td>
<td>U insertion/deletion</td>
<td>Mitochondrion</td>
</tr>
<tr>
<td>Physarum(^{23})</td>
<td>Nucleotide insertion</td>
<td>Mitochondrion</td>
</tr>
<tr>
<td>Paramyxoviruses(^{24})</td>
<td>G insertion</td>
<td>Cytoplasm</td>
</tr>
<tr>
<td>Higher plants</td>
<td>C- to U- conversion</td>
<td>Mitochondrion</td>
</tr>
<tr>
<td>Mammals</td>
<td>C- to U- conversion</td>
<td>Nucleus</td>
</tr>
</tbody>
</table>

### Box 2- Nonsense-mediated mRNA decay in mammals\(^{25}\)

Nonsense-mediated mRNA decay (NMD) in mammalian cells generally degrades mRNAs that terminate translation more than 50-55 nucleotides upstream of a splicing-generated exon-exon junction. Notably, dependence on exon-exon junctions distinguishes NMD in mammalian cells from NMD in all other organisms that have been examined, including *Saccharomyces cerevisiae* and *Drosophila melanogaster*. NMD downregulates spliced mRNAs that prematurely terminate translation so production of the potentially toxic truncated proteins that they encode. NMD also downregulates naturally occurring mRNAs, such as an estimated one-third of alternatively spliced mRNAs, certain selenoprotein mRNAs, some mRNAs that have upstream open reading frames, and some mRNAs that contain an intron within the 3’ untranslated region. In fact, it is thought that NMD has been maintained throughout evolution not only because it degrades transcripts that are the consequence of routine abnormalities in gene expression but also because it is widely used to achieve proper levels of gene expression. Although disease-associated mutations that result in the premature termination of translation led to the discovery of NMD, it is not likely that this type of mutation ever drove significant evolutionary selection. Nevertheless, some of these mutations nicely illustrate the importance of NMD. For example, nonsense mutations within the last exon of the human β-globin gene do not elicit NMD because there is no downstream exon-exon junction. As a consequence, the resulting truncated β-globin has near-normal abundance, fails to properly associate with α-globin and causes a dominantly inherited form of what is otherwise (e.g. for nonsense codons located within exons other than the last exon) a recessively inherited thalassemia.

The importance of NMD is exemplified by the findings that mouse embryos that cannot perform NMD because they lack a key NMD protein, Upf1, resorb shortly after implantation. Furthermore, blastocysts that have the same defect, isolated 3.5 days post-coitum, undergo apoptosis in culture after a brief growth period. The inviability of NMD-deficient embryos and cells probably reflects the combined failure to regulate natural substrates properly and eliminate transcripts that were generated in error. Note that, Upf1 has been shown to function in other pathways, as well as NMD (see below), which may also contribute to the observed inviability. NMD in mammalian cells is a consequence of a pioneer round of translation. Precursor (pre)-

---


\(^{23}\) *Physarum* is a large genus of slime moulds of Myxomycetes class (*P. polycephalum* is widely used in laboratory studies). Myxomycetes are true or acellular slime moulds in which the major vegetative stage is a multinucleate diploid and migratory plasmodium. The latter is defined as a motile mass of protoplasm, which is usually naked (bounded only by a plasma membrane), and which is generally variable in size and form.

\(^{24}\) Paramyxoviridae is a family of enveloped viruses which have a non-segmented, negative-sense ssRNA genome. The family includes some important pathogens of humans, animals and birds. The viruses are transmitted via the air (aerosols).

mRNA in the nucleus is bound to by the major nuclear cap-binding protein (CBP) CBP80-CBP20 heterodimer and, after 3'-end formation, the major nuclear poly(A)-binding protein (PABP) PABPN1. PremRNA splicing generates spliced mRNA that is bound by CBP80, CBP20, PABPN1 and the major cytoplasmic PABPC as well as an exon junction complex 26 (EJC) of proteins that is deposited, as a consequence of splicing, ~20–24 nucleotides upstream of each exon-exon junction. The resulting mRNP constitutes the pioneer translation initiation complex. This complex is thought to undergo a ‘pioneer’ round of translation either in association with nuclei, in the case of mRNAs that are subject to nucleus-associated NMD, or in the cytoplasm, in the case of mRNAs that are subject to cytoplasmic NMD. If NMD occurs, it is the consequence of nonsense codon (NC) recognition during this pioneer round of translation. Upf1 may function as a component of the translation termination complex before it functions in NMD, considering that NMD requires translation termination and Upf1 associates with eukaryotic translation release factors 1 and 3. Upf1 might associate with mRNA regardless of whether termination occurs at a position that elicits NMD. If translation terminates at an NC that resides more than 50-55 nucleotides upstream of an exon-exon junction, then Upf1 is thought to elicit NMD by interacting with EJC associated Upf2. Consistent with a role for EJCs in NMD is the observation that NC-containing mRNAs that derive from intronless genes fail to undergo NMD. Once the mRNA is remodeled so that eukaryotic translation initiation factor (eIF) 4E replaces CBP80-CBP20 at the mRNA cap, PABPC replaces PABPN1 at the poly(A) tail, and EJCs have been removed from mRNA, the mRNA becomes immune to NMD. Translation has been reported to remove Y14, and it may remove other mRNA binding proteins as well. Although these conclusions derive largely from studies of mRNP structure, they are consistent with kinetic analyses indicating that NMD is restricted to newly synthesized mRNA and does not detectably target steady-state mRNA. Cell fractionation studies indicate that most nonsense-containing mRNAs are subject to nucleus-associated NMD. This means that mRNA decay occurs prior to the release of newly synthesized mRNAs into the cytoplasm. Nucleus-associated NMD has been proposed to occur within the nucleoplasm, but it is generally thought to take place during or after mRNA transport across the nuclear pore complex. A fraction of mRNAs is subject to cytoplasmic NMD. What destines some mRNAs for nucleus associated NMD and others for cytoplasmic NMD is currently unknown.

NMD in mammalian cells occurs both 5'-to-3' and 3'-to-5'; it thus involves decapping and 5'-to-3' exonucleolytic activities as well as deadenylating and 3'-to-5' exosomal activities. It remains to be determined whether NMD occurs in association with translating ribosomes or so-called cytoplasmic foci, which appear to be ribosome-free sites of general mRNA decay. Notably, the efficiency of NMD in mammalian cells is generally not influenced by NC position, indicating that a higher number of downstream EJCs does not lead to more efficient NMD. However, NMD can be augmented by additional mechanisms that are not well understood.

An understanding of how various factors function in NMD is far from complete. Upf1 is an ATP-dependent group 1 RNA helicase and phosphoprotein that, as described above, presumably triggers NMD by interacting with Upf2 at an EJC that resides sufficiently far downstream of an NC. Also, as noted above, Upf2 and either Upf3 or Upf3X, which appear to have distinct but overlapping functions, are components of the EJC. In fact, Upf3 and Upf3X consist of multiple isoforms that result from alternative pre-mRNA splicing. Whether or not Upf2, Upf3 and Upf3X are involved in Upf1 dephosphorylation, as are their orthologues in C. elegans, remains to be

26 Constituents of EJCs include Y14, RNPS1, SRm160, REF/Aly, UAP56, Magoh, Pmn/DRS, eIF4AIII, PYM and Barentsz/MLN51. The EJC also contains additional proteins, including the NMD factors Upf3 (also called Upf3a) or Upf3X (also called Upf3b), Upf2 and, presumably transiently, Upf1. Either Upf3 or Upf3X, each of which is mostly nuclear but shuttles to the cytoplasm and interacts with Upf2, is thought to recruit Upf2, which concentrates along the cytoplasmic side of the nuclear envelope.

27 For example, replacing exons 2-4 and flanking intron sequences of the triosephosphate isomerase (TPI) gene with the 383-bp VDJ exon and flanking intron sequences of the T-cell receptor β (TCR-β) gene, which generates mRNA that is more efficiently targeted for NMD than TPI mRNA, increases the efficiency with which TPI mRNA undergoes NMD >15-fold. The efficiency of NMD is increased only when the TCR-β sequence is located upstream of an NC.
determined. However, as in *C. elegans*, Upf1 phosphorylation is mediated by the PIK related kinase SMG1. Also as in *C. elegans* and, possibly, *D. melanogaster*, Upf1 dephosphorylation is mediated by SMG5 and, presumably, SMG6 and SMG7\(^\text{28}\).

As is evident from this short overview, many mechanistic details of NMD still require resolution. In the future, it will also be important for us to understand the extent to which NMD regulates the level of proper mRNA production as opposed to degrading mRNAs that produce aberrant and, therefore, potentially harmful proteins. How NMD is mechanistically linked to other cellular processes, some of which can also be viewed as a type of quality control, requires further study.

**

\(28\) Interestingly, factors that function in NMD have also been shown to function in other pathways. For example, SMG1 is an ATM-related kinase that is also involved in the recognition and/or repair of damaged DNA. SMG1 phosphorylates the tumor suppressor checkpoint protein p53 in response to UV and γ irradiation, and cells in which SMG1 has been downregulated accumulate spontaneous DNA damage and are sensitized to ionizing radiation. Providing another example, Upf1 is the δ helicase that partially co-purifies with DNA polymerase δ. Upf1 (unlike Upf2, the only other NMD factor tested) also appears to function in nonsense-mediated altered splicing (NAS), a poorly understood pathway by which NCs influence the efficiency or accuracy of splicing. In fact, Upf1 can be mutated so that it functions in NAS but not NMD, indicating that the two pathways are genetically separable. Furthermore, Upf1 has recently been found to function in a new pathway called Staufen 1 (Stau1)-mediated mRNA decay (SMD). In this pathway, the RNA-binding protein Stau1 interacts directly with Upf1 to elicit mRNA decay when bound sufficiently far downstream of an NC, including the normal termination codon. The results of microarray analyses indicate that there are a number of natural targets for SMD. Finally, Upf1 interacts with PABPC and forms distinct complexes of approximately 1.3 MDa and 400-600 kDa that appear to differ in their content. The functional significance of all these findings remains unknown. Multiple roles for NMD factors are also evident in the case of SMG5, SMG6 and SMG7, which are identical to the Ever Shorter Telomere (EST) proteins EST1B, EST1A and EST1C, respectively. Each associates with active telomerase and is involved in telomere integrity.