RNA Processing Comes of Age

ROBERT P. PERRY

During the past two decades, an awareness of the importance of RNA processing has evolved as part of the quest to understand how living cells express the information encoded in their genes. As the knowledge of gene expression has expanded, we have come to realize that the old central dogma of DNA → RNA → protein is embellished with elegant and intricate design features, many of which are revealed in the processing of primary gene transcripts into the functional forms of RNA. The production of the two large structural RNA components of the ribosome (rRNAs), the synthesis of transfer RNA (tRNA), and the formation of messenger RNA (mRNA) in higher organisms all involve rather elaborate processing reactions, including nucleolytic cleavages, ligations, terminal additions, and nucleoside modifications.

A raison d'être for processing is readily apparent in the case of the coordinate production of rRNA components from a single transcriptional unit and the synthesis of mRNA from noncontiguous genetic elements. However, the purpose of the polyadenylate and methylated cap structuresthat are added to the termini of mRNA and the modification of internal nucleotides in most RNA species is less clear. These structural alterations may serve to improve the stability of the RNA and the efficiency of its function, but they might also be implicated in more subtle forms of discriminative regulation that are yet to be discovered. In any event, it is clear that RNA processing constitutes a major cellular activity and an integral part of the mechanism of gene expression.

In this essay I shall try to trace the evolution of our concepts of RNA processing in relation to the contemporary issues of cellular and molecular biology and to the introduction of key experimental tools which were critical to the development of these concepts. My idea for treating the subject in this way came from an engaging article on the nucleolus by a former colleague and source of inspiration, Jack Schultz (1). It is my intention to provide both a historical and a reasonably up-to-date overview of the subject without the burden of extensive detail. Fortunately, there are several recent reviews to which the reader can refer for a more comprehensive coverage of particular aspects of RNA processing (2-7).

Conceptual Origins

By the late 1950s, the idea that RNA plays the role of principal intermediary in information transfer between DNA and proteins was generally accepted (8). This concept became very popular and was considered part of a central dogma with which all phenomena dealing with genetic expression should be interpreted (cf. [9] for details). However, there was considerable confusion as to how RNA served this role until it was realized that there were several distinct classes of RNA, each with a different function in the overall process. The problem was being attacked in both prokaryotic and eukaryotic systems, the prokaryotes offering ease of genetic and nutritional manipulation, the eukaryotes the advantage of cellular compartmentalization and microscopic visibility. Several powerful new techniques were applied: autoradiography with tritiated nucleosides to localize the intracellular sites of RNA synthesis; multiphase extraction with phenol and detergent to obtain undegraded preparations of RNA; ultracentrifugation and sucrose gradient sedimentation to fractionate various RNA molecules and subcellular particles according to size; and electrophoresis and chromatography to measure RNA base composition.

It soon became evident that the bulk of the stable RNA in all types of cells consists of two homogeneous components, both of which are associated with the ribosome, a structure already known to be implicated in protein synthesis. The rRNAs were given names according to their sedimentation coefficients—16S and 23S for the bacterial species and 18S and 28S for the mammalian species. The other abundant RNA species identified at that time was transfer RNA, then called soluble RNA or 4S RNA because of its nonparticulate nature and small size. The biosynthesis of ribosomal RNA in Escherichia coli was studied by elegant isotope incorporation experiments (10), which served as models for future kinetic studies of RNA processing. However, there was not yet any evidence for precursors that were larger than the mature rRNAs or even for the existence of rRNA genes in bacterial cells. Nevertheless, by 1960 there was substantial evidence from both autoradiographic and cell fractionation studies to indicate that in higher organisms cytoplasmic RNA was derived from the nucleus, and in particular from nucleoli, which appeared to be especially active sites of synthesis (cf. [11] for references). Recalling earlier cytogenetic evidence that invoked the existence of the nucleolus organizer locus, one began to deduce that nucleoli were in fact the sites of the rRNA genes. This was conclusively demonstrated in the early to mid-1960s by a large variety of experiments which exploited both genetic and biochemical tools (12). Important in this regard was the introduction of the very powerful nucleic acid hybridization technique (13), which not...
only provided evidence for the existence of rRNA genes in both prokaryotes and eukaryotes, but allowed their enumeration, as well.

In 1962, when radioactively labeled nuclear (nucleolar) precursors of rRNA were examined by sucrose gradient sedimentation analysis, an unexpected fact was revealed (14, 15). Instead of the precursors being the same size as the mature cytoplasmic rRNA, they were substantially larger. The use of actinomycin D to block RNA synthesis (16), while allowing some processing to occur, helped overcome the difficulty of doing an effective chase experiment in mammalian cells, and provided compelling evidence that these large RNA components were indeed rRNA precursors. The kinetics of labeling of the pre-rRNA components suggested the following sequence for rRNA processing events in mammalian cells: 45S → 35S → 28S + 18S (15). Thus, although there was no precedent for it, the notion arose that primary products of genes might need some sort of “transformation” or “processing” in order to convert them into functional entities. The idea that RNA could be altered post transcriptionally was concurrently being developed in studies which showed that the “minor nucleotides” in tRNA arise by the modification of previously synthesized polynucleotides (17).

By this time, a substantial number of convincing experiments with bacterial and bacteriophage systems had indicated the existence of messenger RNA (9), and considerable effort was being made to determine whether mRNA was also present in higher organisms. Fractions of a “DNA-like” heterogeneous RNA were extracted from nuclei (15, 18–20) and suspected of being related to mRNA, but because we knew relatively little about the properties of mRNA, there was still some uncertainty as to whether it was also subject to processing. When it became apparent that polyribosomal mRNA was indeed smaller than the heterogeneous nuclear RNA (hnRNA), one began to give serious consideration to the notion that mRNA, like rRNA, was derived from a larger primary transcript. Although this idea was spawned in the mid-1960s, it took more than a decade for it to gain universal acceptance (see reference [6] for details).

The resolution of RNA on sucrose gradients is barely adequate to distinguish the 20% average-size difference between newly made and mature tRNA. Hence, the late 1960s, when polyacrylamide gel electrophoresis came into use for RNA size analysis (21), its superior resolution made it relatively easy to discriminate the larger pre-rRNAs found in pulse-labeled mammalian RNA from mature tRNA (22). Thus, it became generally accepted that tRNAs are also processed from larger precursors in eukaryotes. Similar analyses of 5S ribosomal RNA failed to reveal any oversized precursors, and 5S rRNA was thought to be the “exception to the rule.” However, recent experiments using gels of higher resolving power and nucleotide sequence analysis (23) have demonstrated that 5S rRNA transcripts are 5–10% larger than mature 5S rRNA, and, hence, that they require processing as well.

For many years it was believed that the processing of large RNA precursors into smaller, mature forms was a distinctive property of eukaryotic cells. However, studies in the early to mid-1970s with prokaryotic systems carrying mutations that block or attenuate processing (24) revealed that processing also occurs in prokaryotes, the main difference being that the processing reactions follow transcriptional events much more closely than in eukaryotes, so that full-length transcripts are less readily observed. Other types of processing reactions, such as methylations and terminal additions, also exist in prokaryotes, although they are generally less elaborate than in higher organisms.

**Further Developments**

**Ribosomal RNA:** Several important technical developments helped extend our knowledge of rRNA processing. Detailed chromatographic analyses of rRNA derivatives revealed that methylations occurred on both base and ribose moieties, with the ribose methylations largely predominating (25). It was found that most of the methyl groups are added to the initial (45S) precursor and that essentially all are conserved during processing (26). This finding, together with improved cellular fractionation techniques, with which one could obtain highly purified nucleoli or nucleoplasmic fractions, and polyacrylamide gel electrophoresis, with which one could separate the various precursors and processing intermediates, led to the formulation of a fairly detailed processing pathway for rRNA (27). Additional facts were supplied by fingerprint analysis of methylated oligonucleotides (28) and by analysis of the patterns of base-paired loop structures that persist under partial denaturing conditions (29).

Valuable contributions to our understanding of rRNA processing have also come from studies of the organization of rRNA genes. Buoyant density/hybridization studies of rDNA (30, 31) and electron-microscope visualization of transcribing rRNA genes (32) provided convincing evidence for the linkage of the 18S and 28S gene elements into a single transcriptional unit, and revealed the existence of nontranscribed spacer DNA separating each transcriptional unit. The mutual reinforcement of the parallel studies with rRNA precursors and with rDNA greatly accelerated the general acceptance of a rather complicated and apparently “uneconomical” scheme for the production of ribosomal RNA.

Our current view of rRNA synthesis and processing in mammalian cells is diagrammed in Fig. 1 (2). The 18/28S transcriptional unit is of the general form: "spacer-18S rRNA-spacer-28S rRNA". Its primary product is a 45S molecule of about 12.5 kb. Lower eukaryotes have the same general organization, except that the lengths of the various segments, especially the spacers, are shorter. The processing pathway can be described in terms of four principal events involving cleavages at or near the sites numbered 1 to 4. The first cleavage at site 1 removes the 5'-terminal leader sequence. The second

![Figure 1](image-url)
cleavage can be either at sites 2 or 3, depending on the species of cell and to some extent on environmental conditions; the predominant pathway apparently is determined by the conformational state of the first intermediate. The final trimming near site 4 is usually the rate-limiting step in the processing pathway, thus causing a substantial accumulation of the proximal intermediate, the 32S component. This trimming involves at least two cuts in the polynucleotide backbone, one at the 5' end of the 28S component and another at the 3' end of the 5.8S component. The 5.8S component, 140 nucleotides in length, is a stretch in the 5' region of the 32S component which remains bound to 28S component by base pairing after the final cleavage. In a portion of the tRNA genes of Drosophila and in the mitochondrial and chloroplast rRNA genes of some primitive eukaryotes, the segment encoding the large rRNA components is interrupted by an intron (33-35). When such genes are expressed, the intron sequence is probably removed by a splicing mechanism akin to that used for the excision of mRNA or tRNA introns (vide infra).

The 5S rRNA genes, which are situated remotely from the 18/28S genes (usually on different chromosomes), are transcribed into molecules which have the 5' terminus of mature 5S RNA and a stretch of 8 or more extra nucleotides at the 3' end (23). Processing involves removal of these extra nucleotides and possibly the addition of a terminal U residue.

The organization of rRNA genes in the well-studied prokaryote, E. coli, resembles that in eukaryotes, except that the 5S rRNA gene and one or more tRNA genes are also included in the transcription unit, the 5S gene being located on the 3' side of the 23S gene and the tRNA genes being either in the spacer between the 16S and 23S genes or at the 3' end beyond the 5S gene (3). This difference in gene organization is accompanied by a basic difference in transcribing enzymes. In prokaryotes all genes are transcribed by a single species of RNA polymerase, whereas in eukaryotes polymerase I is used for 18/28S genes and polymerase III for the tRNA and 5S rRNA genes. Another basic difference, alluded to earlier, is that in prokaryotes processing cleavages usually occur before the polymerase is finished transcribing the gene, whereas in eukaryotes cleavages generally occur on complete transcripts.

Several enzymes are known to be implicated in the processing of E. coli rRNA (3). The best known is RNase III, which requires double-helical RNA as part of its recognition element. This enzyme is responsible for separating the precursor segments containing the 16S, 23S, and 5S components. These precursor segments are in turn acted upon by at least two additional ribonucleases. The tRNA segments are processed by a separate set of enzymes (vide infra). It is generally believed that some of the cleavages involved in the processing of eukaryotic rRNA also require double-helical specificity, although the isolation and characterization of the relevant nucleases has not yet been achieved (see reference [2] for details).

The development of the powerful tools of gene manipulation, e.g., the use of transducing phages and recombinant DNA methods, together with the development of techniques for rapid nucleotide sequencing, has now made it possible to describe the rRNA genes and their products at the level of nucleotide sequence. The sequence information enables one to construct plausible models of RNA secondary structure, and hence to examine possible substrate specificities for the various processing enzymes (3, 36). Such an analysis has led to the remarkable conclusion that the substrates for RNase III are a pair of hairpins with loops consisting of the entire 16S (~1,600 nucleotides) or 23S (~2,800 nucleotides) rRNA components (Fig. 2). These models also indicate additional cleavage sites for the other enzymes that may be involved in rRNA maturation. If eukaryotic rRNA precursors have an analogous structural organization, one could imagine that there may be multiple cleavages at each of the processing sites described in Fig. 1.

**TRANSFER RNA:** Although tRNA precursors were initially found in mammalian cells, the major progress in elucidating the details of tRNA processing has been made with bacterial and yeast systems. The key factors in this development were the discovery of mutants blocked in the processing of tRNA and the ability to manipulate the tRNA genes (see [2] and [3] for references). Processing-defective mutants with lesions in either the tRNA genes themselves or in key tRNA genes and their products at the level of nucleotide sequence. The sequence information enables one to construct plausible models of RNA secondary structure, and hence to examine possible substrate specificities for the various processing enzymes (3, 36). Such an analysis has led to the remarkable conclusion that the substrates for RNase III are a pair of hairpins with loops consisting of the entire 16S (~1,600 nucleotides) or 23S (~2,800 nucleotides) rRNA components (Fig. 2). These models also indicate additional cleavage sites for the other enzymes that may be involved in rRNA maturation. If eukaryotic rRNA precursors have an analogous structural organization, one could imagine that there may be multiple cleavages at each of the processing sites described in Fig. 1.

**TRANSFER RNA:** Although tRNA precursors were initially found in mammalian cells, the major progress in elucidating the details of tRNA processing has been made with bacterial and yeast systems. The key factors in this development were the discovery of mutants blocked in the processing of tRNA and the ability to manipulate the tRNA genes (see [2] and [3] for references). Processing-defective mutants with lesions in either the tRNA genes themselves or in key tRNA genes and their products at the level of nucleotide sequence. The sequence information enables one to construct plausible models of RNA secondary structure, and hence to examine possible substrate specificities for the various processing enzymes (3, 36). Such an analysis has led to the remarkable conclusion that the substrates for RNase III are a pair of hairpins with loops consisting of the entire 16S (~1,600 nucleotides) or 23S (~2,800 nucleotides) rRNA components (Fig. 2). These models also indicate additional cleavage sites for the other enzymes that may be involved in rRNA maturation. If eukaryotic rRNA precursors have an analogous structural organization, one could imagine that there may be multiple cleavages at each of the processing sites described in Fig. 1.

**TRANSFER RNA:** Although tRNA precursors were initially found in mammalian cells, the major progress in elucidating the details of tRNA processing has been made with bacterial and yeast systems. The key factors in this development were the discovery of mutants blocked in the processing of tRNA and the ability to manipulate the tRNA genes (see [2] and [3] for references). Processing-defective mutants with lesions in either the tRNA genes themselves or in key tRNA genes and their products at the level of nucleotide sequence. The sequence information enables one to construct plausible models of RNA secondary structure, and hence to examine possible substrate specificities for the various processing enzymes (3, 36). Such an analysis has led to the remarkable conclusion that the substrates for RNase III are a pair of hairpins with loops consisting of the entire 16S (~1,600 nucleotides) or 23S (~2,800 nucleotides) rRNA components (Fig. 2). These models also indicate additional cleavage sites for the other enzymes that may be involved in rRNA maturation. If eukaryotic rRNA precursors have an analogous structural organization, one could imagine that there may be multiple cleavages at each of the processing sites described in Fig. 1.
processing enzymes were used to work out the details of the pathways. In *E. coli*, the tRNA precursors have the general form $\text{5'-leader-(tRNA-spacer)-tRNA-trailer-3'}$. They may be either monomeric ($n = 0$) or multimeric ($n = 1-6$), and in cases such as the rRNA transcription unit, they may be joined to other gene products. In higher organisms, most, if not all, tRNA precursors are monomeric. The 5'-leader sequence is removed by an enzyme termed RNAse P. This enzyme appears to contain an RNA molecule, as yet not well characterized, that is essential for its activity (37). The substrate recognition for RNAse P seems to reside largely in the structural features of the mature tRNA product, rather than in the sequences surrounding the cleavage point. Removal of the trailer sequence apparently requires an additional endonuclease as well as an exonuclease. Sometimes the tRNA precursors do not contain the universal CCA 3' terminus; when this is the case, the CCA is generated by a specific terminal transferase. Nucleoside modifications can occur on intact precursor molecules as well as on cleaved products.

With the advent of recombinant DNA methodology, detailed investigations of various yeast tRNA genes were made. These studies have revealed the presence of introns in several (but not all) tRNA genes (3). Thus, the processing of yeast tRNA requires, in addition to the cleavages described above, a splicing activity to remove the introns. In vitro processing experiments have revealed that the splicing activity can be resolved into two steps: a nuclease activity, which occurs in the absence of ATP, and a ligase step, which requires ATP (38). Surprisingly, the nuclease catalyzes the cleavage of phosphodiester bonds so as to yield 3'-phosphates and 5'-hydroxyl groups. This is in marked contrast to RNAse P and RNAse III cleavages, which yield 3'-hydroxyls and 5'-phosphates. In higher eukaryotes, tRNA genes containing introns may be less common (39, 40), although the number of organisms studied in detail is too small to permit broad generalizations. Intron-containing yeast pre-tRNA can be properly processed in *Xenopus* oocytes, indicating that the tRNA processing enzymes are very similar in different species (41).

**Messenger RNA:** As mentioned above, the notion of mRNA processing began with the finding that heterogeneous nuclear RNA (hnRNA) and polyribosomal mRNA have strikingly different size distributions and yet very similar base compositions. Before the development of recent techniques that have enabled us to study the synthesis of individual species of mRNA, an investigator wishing to probe the relationship between hnRNA and mRNA had to employ methods that were suitable for complex mixtures of RNA sequences, and to exploit, whenever possible, features such as poly A that are common to a substantial fraction of the mRNA species. Nucleic acid hybridization is a technique that may readily be applied to complex mixtures of RNA, and thus enjoyed widespread popularity in the study of hnRNA and mRNA, beginning in the late 1960s and extending over more than a decade (see [2, 4, 6, 42] for references). The information gained from these studies, together with concurrently acquired knowledge of the general properties of eukaryotic DNA sequences, e.g., the existence of single-copy and repetitive sequence elements (43), provided a new framework for comparing the properties of hnRNA and mRNA. It was evident that moderately repetitive, as well as unique, sequences are transcribed into hnRNA and that at least a portion of these same sequences are processed into mRNA (44). However, the biological significance of these repetitive sequence transcripts was not obvious and, in fact, still remains one of the challenging mysteries in our understanding of eukaryotic gene expression.

Around 1970, two important discoveries helped accelerate progress on the problem of mRNA processing. First was the finding that the majority of mRNA molecules and a significant fraction of hnRNA molecules possess a 3'-terminal poly-A segment, 150-200 nucleotides long, which is constructed posttranscriptionally (see [5] for references). The poly-A tail represented an interesting new aspect of processing, but, even more importantly, it was rapidly exploited for purifying mRNA away from the bulk of the cellular RNA (45-47). Second was the discovery of reverse transcriptase (48, 49), which was later used to synthesize DNA complementary to mRNA (cDNA), thus providing a valuable probe for the study of mRNA frequency distributions and the homology relationships between mRNA and hnRNA (see [42] for references). Most cell types were observed to have a very broad distribution of mRNA abundances, ranging from a few species present at several-thousand copies per cell to thousands of species present in a few copies per cell. About 10-20% of the hnRNA sequences are homologous to mRNA.

In the mid-1970s, it was discovered that the mRNAs of eukaryotic cells and many types of viruses contain an unusual methylated "cap" structure (Fig. 3) at their 5' terminus and one or more internal 6-methyl adenine residues (see [7] for references). These modifications, like poly A, are added posttranscriptionally to the mRNA precursors, and then carried along through the rest of the processing stages. For a long time it was thought that the capacity to be methylated was a property confined to the structural RNAs, i.e., the RNAs that do not encode proteins. This idea persisted because the level of methylation in mRNA is almost an order of magnitude lower than in tRNA, and without a means for effectively separating these two RNA species, the mRNA methylation is entirely masked—especially if one doesn't know that it is there in the first place. However, when methods for isolating mRNA based on its unique poly-A structure came into use, one could obtain sufficiently pure preparations of mRNA so that an unambiguous identification of its methylated derivatives could be made (51, 52). The parallel development of efficient cell-free systems for the synthesis of certain viral mRNAs contributed similarly to the characterization of their modified components, and, moreover, provided an excellent means for studying the biochemistry of cap formation (53, 54).

The formation of a complete cap structure (Fig. 3) involves the participation of four to six different enzymes (see [7] for

![Figure 3](image-url)
polynucleotide acceptor, it is reasonable to expect that cap
explanation for the heterogeneity of cap structures on certain
happens with rRNA genes in E. coli (57). This might be an
at more than one site with different relative efficiencies, as
references [55, 56]).

regions of prokaryotic genes (65). Given an equivalence of
similar, although not identical, to the so-called Pribnow box
ing regions of genes coding for several cellular and viral
SV40 and polyoma mRNAs (66, 67), and for the ability of
certain SV40 mutants to survive deletions at a capsite (68).
However, if such imprecise initiation ever occurs with cellular
mRNAs, it presumably is confined to relatively sparse mRNA
species, because the various abundant mRNA species studied
to date all seem to have a homogeneous 5' cap (7).

Certain one of the most surprising developments in the
history of mRNA processing was the discovery of splicing. The
initial observations were made in 1977 during investigations of
adenovirus mRNA synthesis (69, 70), in which the relationship
between viral mRNA and the DNA that encodes it were
examined in the electron microscope using the powerful R-
loop technique (71, 71a). The striking multiloop structures were
correctly interpreted to mean that the mRNA was specified by
several noncontiguous genetic elements. In spite of its novelty,
this interpretation was readily accepted because it explained the
(them) puzzling observation that mRNAs made from distinct
portions of the adenovirus genome have the identical 5'
terminal capped sequence (72, 73). Moreover, it also seemed to
be a possible solution to the riddle of how a large hnRNA
molecule with a cap structure on one end and a poly-A tail on
the other could be processed into a smaller mRNA molecule
without losing either its cap or its poly A. Within a matter of
months, experiments employing restriction-enzyme analysis
with Southern's blotting technique (74) and R-loop or hetero-
duplex analyses of cloned gene fragments established the wide-
spread occurrence of split genes and gave some idea of their
organizational features. Studies of the organization of SV40
genoms (75, 76) and of cellular genes like globin (77, 78),
immunoglobulin (79), and ovalbumin (80, 81) indicated that
the interruptions, termed intervening sequences or introns, can
occur in the coding portions of the gene as well as in 5'
untranslated leader sequences.

The expression of split genes always seems to involve pro-
duction of a composite RNA transcript and subsequent exci-
sion of the intron sequences. This has been established first by
showing that there are large nuclear transcripts which are
colinear with the complete gene (82), and second by using
kinetics and pulse-chase experiments to demonstrate that the
large transcripts are actually processed into mRNA (83-86).
The tendency of hnRNA to aggregate because of intermolec-
ular base-pairing (87) makes it imperative to use rigorous
denaturation conditions in such studies, for example, fraction-
ation of the hnRNA by electrophoresis on methyl-mercury-
agarose gels (88). This technique coupled with a blotting
procedure by which the fractionated hnRNA is covalently
attached to diazotized paper (89), enables one to visualize
precursors of any mRNA for which a pure sequence probe is
available. In fact, by using an assortment of probes for struc-
tural and intronic sequences, one can in principle delineate the
processing pathway. An example of such an analysis for im-
imunoglobulin light chain mRNAs is shown in Figs. 4 and 5.

In a transcript containing multiple intronic sequences, there
may be a preferential order of excision, but in some cases the
order is not necessarily absolute (91). This is reminiscent of the
alternative temporal order observed in the processing of mam-
malian rRNA (2). In certain viral systems like adenovirus and
SV40, a given transcript can give rise to multiple mRNA
species, depending on the choice of different splicing modes.
In this case, processing can have a role in determining quali-
tatively which gene elements are utilized. Such qualitative
discrimination at the processing level could provide a basis for
certain types of cellular differentiation. An example of this
principle has been recently invoked for early B lymphocytes,
FIGURE 4 (a) Scheme for the formation and expression of a κ-chain immunoglobulin gene. In germ line DNA the several hundred genes coding for the variable part of the κ-chain (Vk genes), and the gene coding for the constant portion (Cκ gene) are separated by an unknown distance. During B cell differentiation, a site-specific deletion occurs between one of the Vk genes and any one of four Jκ segments located 2.4-3.9 kb upstream from the Cκ gene. This event creates a functional κ-gene the size of which depends on the particular Jκ segment being utilized. Such functional rearrangements are normally found on only one of the allelic pair of chromosomes. The intervening sequence between Jκ and Cκ (J-C intron) is transcribed and the corresponding RNA sequence excised during RNA processing. The 3'-untranslated sequence is contiguous to the Cκ sequence. The 5'-untranslated sequence and the sequence encoding the amino-terminal signal peptide are separated from the Vκ gene by a small (~0.1 kb) intron. (b) Four probes used in the analysis of κ-mRNA transcription and processing shown in Fig. 5. The Vκ and Cκ probes are obtained by restriction endonuclease digestion of a cloned cDNA sequence corresponding to the κ-mRNA produced by MOPC 321 myeloma cells. The Jκ and intervening sequence (IVS) probes are similarly obtained from a cloned fragment of germ line DNA containing the Jκ-Cκ region. (see reference 90 for details). (c) Schematic representation of the transcripts produced by various myeloma cells. The unrearranged (germ line) allele produces an 8.4 kb transcript which is not processed into any functional mRNA. The allele encoding the expressed κ-chain is transcribed into a component the size of which varies according to the Jκ segment being used (5.3, 5.0, 4.4, and 4.1 kb, respectively, for J1, J2, J3, and J4 expressors). These precursors are processed into a common 1.2 kb κ-mRNA. In J2, J3, and J4 expressors components are found which seem to arise by an asynchronous cleavage at the 5' boundary of the J-C intron (narrow bars).
in which two distinct mRNA species encoding the membrane-associated and secreted forms of the μ heavy chain are apparently produced from a single set of μ-gene elements by variations in the modes of splicing (92). The mRNAs are identical except for a region near the 3' end, which in one case encodes the carboxy-terminal tailpiece of the secreted μ chain, and in the other, a hydrophobic segment that apparently anchors the membrane μ chain to the lipid bilayer. During its ontogeny the B lymphocyte shifts from producing predominantly membrane μ chain to predominantly secreted μ chain, presumably by shifting its major mode of μ-mRNA processing.

Although the enzymes involved in mRNA splicing have not yet been characterized, some clues concerning the splicing mechanism have come from comparisons of the nucleotide sequences surrounding the splice junctions and from studies of the consequences of perturbations in gene organization. A compilation of a large number of junction sequences of both cellular and viral pre-mRNAs has resulted in the consensus sequence shown in Fig. 6a (93). The doubly underlined nucleotides at the extreme ends of the introns are almost ubiquitous, being present in more than 95% of the sequences. Thus, it seems reasonable to suppose that they are essential for the splicing reactions. An interesting complementarity has been noted between the consensus sequence and a 5'-terminal sequence of one of the small nuclear RNAs (snRNAs), termed U1 (94), suggesting a model in which U1-RNA helps juxtapose the two-splice junctions by appropriate base pairing interactions (Fig. 6b). A similar role has been invoked for the similarly sized VA-RNA in the splicing of adenovirus mRNA (95). If this speculation turns out to be correct, the function of at least

FIGURE 5 A "northern blot" of the poly A+ nuclear RNA from eleven different myelomas expressing distinctive κ-chains. The poly A+ nuclear RNA was electrophoresed on methylmercury hydroxide gels and blotted onto diazotized paper. The immobilized RNA was annealed with the four probes described in Fig. 4b, and the nuclear components containing the corresponding sequences were revealed by autoradiography. The tumor designation is shown at the top; the expressed J segment (from amino acid analysis of the κ-chain) is shown at the bottom. The size of the various bands (in kilobases) is at the left. The interpretation of these data is given in Fig. 4c. See reference (90) for other details.

FIGURE 6 (a) A consensus sequence obtained by comparing 36-splice junction sequences (93). To appear in this sequence a base must be the most common in that position and occur with a frequency of at least 45%; bases occurring in 75% of the sequences are underlined; those present with 95% or greater frequency are underlined twice. Y indicates pyrimidines. X marks the position of a single non conserved base in the consensus sequence. Vertical lines mark the intron-exon boundaries. (b) A possible alignment of intron-exon boundaries by base pairing between the 5'-terminal portion of U1 RNA and sequences at both ends of an intron. Processing would consist of cleavage of two G-G bonds and formation of a new one (asterisk).
one small nuclear RNA will have finally been elucidated. Although the snRNAs were discovered more than 10 years ago, it has not been possible to assign them any specific cellular function. Fortunately, the uncertainty about their physiological significance did not deter studies of their structural characteristics, and by the mid-1970s the complete nucleotide sequence of two snRNA species was known (96). Indeed, their novel, highly modified 5' termini served as a model for subsequent elucidation of the cap structures of mRNA (97, 98). In the snRNAs the 'mG moiety is replaced by \( \text{m}^{2,2',3'} \text{G} \); otherwise the structures are essentially the same. The snRNAs are evolutionarily conserved, a fact which may be related to the apparent conservation of mRNA processing systems (41, 99–101).

Studies with total pre-mRNA populations (55, 56) and individual pre-mRNAs (86, 91, 102–105) indicate that cap formation internal methylation and polyadenylation usually precede the splicing out of intronic sequences. Thus, in regard to the general order of processing reactions, mRNA seems to resemble RNA and tRNA, in that the cleavages are directed at molecules which have already been subjected to other types of post-transcriptional modification. Poly-A formation, catalyzed by a terminal transferase enzyme, consists of the sequential addition of 150–200 adenylate residues to the 3' end. The recognition signal for the terminal transferase seems to involve the hexanucleotide AAUAAA, because this sequence is approximately 11–30 nucleotides upstream from the 3' end of all poly A-containing mRNAs but absent from poly A-lacking mRNAs. In some cases, e.g., adenovirus late mRNAs and SV40 mRNAs, the poly-A addition site may be formed by endonucleolytic cleavage of the growing transcript rather than by termination of the RNA polymerase (106, 106a). To what extent this applies to cellular mRNAs is presently unclear.

**Future Directions**

It should be evident from this narrative that progress in our understanding of RNA processing, like that of other natural phenomena, is largely dependent on our ability to formulate fresh and meaningful questions and to develop the appropriate methodologies to help answer these questions. Two decades ago our concern was nuclear/cytoplasmic or genotype/phenotype relationships, and our approach was limited by the cytological and biochemical tools then available. Kinetics of incorporation of radioactively labeled RNA precursors, base-composition analyses, autoradiography, microspectrophotometry, and sedimentation analyses in the ultracentrifuge, determined the scope of our experimental protocols. Today, we seek to define gene organization and expression in terms of arrangements of and changes in nucleotide sequences, and our horizon is confined to the information obtainable from restriction analysis and various blotting procedures, from heteroduplex and R-loop analysis, from nucleotide sequencing, etc. To be sure, serendipity plays a large and unpredictable role in our progress. Who would have imagined that the discovery of 3' terminal poly A on mRNA would provide a basis for methods of purifying mRNA and pre-mRNA, which would lead, in turn, to the uncovering of many new structural and functional features of these molecules? Or that the reverse transcriptase enzyme found to be associated with RNA tumor viruses would become a key tool in the recombinant DNA cloning of mRNA sequences? Yet, such developments are not really rare, and we can be sure that many of today's discoveries will be the sources of techniques that will be critical for answering tomorrow's questions.

What are tomorrow's questions about RNA processing? One broad class of questions concerns the enzymatic mechanisms of processing. So far only a very few of the processing enzymes have been isolated or even purified away from nonspecific degradative enzymes, and, indeed, at present we would be happy to know just how many different cutting and splicing enzymes exist in a cell. If some of these enzymes operate with small RNA cofactors, the various interrelationships between enzyme, cofactor, and RNA substrate will have to be worked out. From the information accumulated to date (cf. reference [107]), it would seem that the substrate recognition sites for processing enzymes reside only partly in features of primary and secondary structure and that they also depend heavily on features of tertiary structures, i.e., on three-dimensional conformation. This aspect poses some formidable obstacles to progress in this area, because most of the current methods for conformational determination are not well suited for studies of the minute quantities of precursor RNAs that are normally available from cells. In lieu of any significant methodological advances, one can resort to specific genetic manipulations of the substrate and to approaches such as those used to study an enzyme that processes 5S RNA in *B. subtilis* (108, 109). In this case, an efficient in vitro processing system was developed and then used with pre-55 RNA that was specifically modified by partial nucleolytic digestion and ligation of synthetic polynucleotide appendages. With detailed secondary structure models of the precursor and its derivatives, the requirement for particular base-pairing interactions within the substrate molecules can be determined.

Another broad class of questions concerns the role of processing in regulating gene expression. Earlier, I cited examples in which alternative processing modes produce multiple and functionally distinct mRNA molecules from a single set of noncontiguous genetic elements. It will be interesting to know whether such qualitative regulation at the processing level is widespread amongst eukaryotic genes, and, if so, to know how a cell can exhibit preference for one processing pathway over another under different physiological and developmental situations. The role of processing in the quantitative regulation of gene expression is also a phenomenon that deserves further scrutiny. The wastage of pre-rRNA that occurs in certain resting-cell populations and disappears in growing populations has been known for more than a decade ([110], see also [111] for other references), but we still do not understand what actually determines whether a particular precursor molecule will be processed or degraded to its nucleotide constituents. Similarly, there is reasonably good evidence to indicate that mRNA abundance is regulated in part by variation in processing efficiency (112, 113), but the molecular basis of this regulation remains obscure. Answers to these questions will require improved knowledge of the processing mechanisms and also the development of new experimental systems (biochemical or genetic) that will enable us to probe the determinants of processing specificity and to generate protocols capable of revealing cause and effect relationships.

Given the large number of interesting unanswered questions and the vast number of biological systems that have yet to be studied, it seems clear that the field of RNA processing will remain an exciting one for many years to come. One can look forward to the repeated satisfaction that will come when various complex biological phenomena are explained and clarified in straightforward molecular terms. Moreover, on the basis of past developments in this field, we can confidently predict that...
future progress will come from unexpected ramifications of conceptual discovery and methodology and that serendipity will play a major role.

Addendum

A delay between the completion of this chapter (in March 1980) and its actual publication has given me an opportunity to consider some of the recent developments in the RNA processing field within the context of my general theme. As anticipated, several important new facts have been revealed by the combined use of recombinant DNA-cloning methodology and various powerful techniques of nucleic acid structural analysis. Some of these facts have clarified uncertainties and helped resolve unsettled issues, while others have given new insights into the ways in which processing can be implicated in the regulation of gene expression and cellular differentiation. In contrast, there has been less progress on the characterization of processing enzymes and the determinants of processing specificity. In these cases the aforementioned technical obstacles still constitute a rate determining step. The following selective survey should give the reader some idea of the pace in this field over the past 18 months.

Formation of 5'-Terminal Caps: It has now been conclusively shown for the genes specifying early and late SV40 mRNAs (114), adenovirus late RNA (115), and mouse β-globin (115) that the 5'-terminal cap site and the site at which transcription is initiated by RNA polymerase II are one and the same. Among the multiple initiation/cap sites ascribed to early SV40 mRNA, a major species is "mGpppC"U, thus demonstrating that eukaryotic RNA polymerase II, like the prokaryotic RNA polymerase, can sometimes initiate with a pyrimidine nucleotide. Certain cellular mRNAs may also have heterogeneous cap sites, as indicated by the finding of two 5' termini in ovalbumin (116) and liver α-amylase (117) mRNAs. Although not yet conclusively demonstrated, this heterogeneity could be the result of imprecise initiation, as was shown for the SV40 mRNA species (114).

A significant addition to our knowledge of processing mechanisms concerns the guanylyltransferase-catalyzed reaction of cap formation. Recent studies of the vaccinia-capping enzyme, one of the few processing enzymes that can be obtained in high purity, have shown that the transfer of GMP from GTP to a triphosphate-terminated polynucleotide involves an intermediate in which GMP is covalently linked to a subunit of the capping enzyme (118). The specificity for GTP as a nucleotide donor seems to reside in the formation of this covalent complex; the basis for specificity of the polynucleotide acceptor is still obscure.

Termination and Polyadenylation: The importance of the AAUAAA sequence in determining the site of 3'-terminal cleavage and polyadenylation of growing transcripts has been firmly established by experiments with deletion mutants of SV40 (119). The cleavage/poly A-addition site is completely abolished in mutants lacking this sequence and is moved proportionately downstream in mutants with deletions to the 3' side of it. Although the AAUAAA sequence (or a very close relative) is apparently indispensible for proper 3'-terminal cleavage and polyadenylation, it is also clear that other structural features in this region can modulate the efficiency of this process (119).

It is now apparent that the use of endonucleolytic cleavage for transcript termination can apply to cellular genes as well as viral genes. In the transcription of the β-globin gene in both mouse (120) and chicken (121) nuclei the RNA polymerase continues more than 1,000 nucleotides beyond the poly A site.

Splicing: Despite the fact that well over 100 individual splice junctions have now been identified in a variety of mRNAs, our understanding of the splicing mechanism is still rather primitive. Except for the GU on the intron side of the 5' splice site and AG on the intron side of the 3' splice site, there is considerable variability at the other positions of the consensus sequence (Fig. 6). Indeed, in a completely random sequence, the 5'- and 3'-consensus sequences should occur, on average, every 2,000 and 500 nucleotides, respectively, distances which are short compared to the lengths of many known introns (122). Nevertheless, the hypothetical implication of U1-RNA in the splicing reaction has received some support from experiments in which a lupus antiserum that precipitates ribonucleoproteins containing U1-RNA was shown to inhibit the proper splicing of adenovirus mRNA by isolated HeLa cell nuclei (123).

RNA transcripts of recombinant chimeras that contain a 5' splice site of an SV40 gene and a 3' splice site of a mouse β-gene can be accurately spliced in vivo (124). Such a result indicates that the two splice sites bracketing an intron need not constitute a unique pair. This has led to the consideration of a processive mechanism in which splicing always occurs between the most proximal pair of splice sites (122). Yet, we now know of several examples of both cellular and viral genes in which a choice between alternative 5' or 3' splice junctions must be made in order to produce the correct mRNA product. One way out of this dilemma is to have a stepwise mechanism in which new splice sites are regenerated by having similar consensus sequences at intron and exon boundaries. This mechanism appears to operate in the processing of collagen mRNA (125) and could conceivably work for SV40 early mRNA (122). However, it does not seem applicable to other situations, e.g., the excision of J1,γ-C introns in immunoglobulin mRNAs (see Fig. 4).

Additional examples of obligate splicing in the processing of ribosomal RNA in simple eukaryotes have been discovered (126-128). The list now includes the products of nuclear genes as well as those associated with the genomes of cytoplasmic organelles. A comparison of four sets of sequences surrounding the ribosomal RNA splice junctions have revealed two interesting features: common nucleotides (T and G) at the 5' sides of both 5' and 3' splice sites, and a short repeated sequence further upstream of these sites (129). The lack of resemblance of these sequence features and those associated with mRNA and tRNA splice junctions supports the idea that a distinctive splicing mechanism has evolved for each major class of RNA.

Role in Gene Expression: The importance of RNA processing in regulating the expression of cellular genes is coming to be increasingly appreciated. For example, the prediction (ref. 92) that the mRNAs encoding the membrane and secreted forms of μ-heavy chain are produced by alternative processing of transcripts from the same gene was confirmed by detailed analyses of the 3'-terminal sequences of the appropriate μ-mRNAs and the identification of corresponding sequences in the μ structural gene (130, 131). Moreover, the finding of two AAUAAA sequences, one on the 3' side of the genetic element encoding the secreted chain, and the other, on the 3' side of the element encoding the membrane-associated chain, suggested that the two mRNAs might be derived from primary transcripts that are terminated at two different sites (131). An analysis of μ-mRNA precursors in cells producing different proportions of membrane and secreted mRNAs tends to support this idea.
(132). An additional role for RNA processing in the developmental regulation of immunoglobulin gene expression is suggested by the simultaneous production in a single cell of mRNAs encoding the same variable region sequence and constant regions of either the $\mu$ or the $\delta$ class (133–135). The $C_\alpha$ gene is located between the $V_H$ gene and the $C_\delta$ gene. Therefore, in the processing of $\delta$-mRNA, the $C_\alpha$ sequences are treated as an intron rather than an exon.

An elegant example of the implication of RNA processing in cellular differentiation was provided by an analysis of the $\alpha$-amalyse genes in different tissues of the mouse (117, 136). Although the $\alpha$-amalyse mRNAs in salivary gland and liver cells are derived from the same gene, their relative abundance in these two tissues is markedly different. An examination of gene and mRNA structure revealed that the salivary and liver mRNAs possess different 5′ untranslated sequences which are encoded by two widely separated elements located several kb upstream from a common coding segment. One attractive idea is that the wide variation in expression of this gene in the two tissues is related to the differential use of two transcriptional promoters. Obviously without the flexibility afforded by RNA processing, this could otherwise be achieved only by using a duplicate gene, as is the case in pancreatic cells. A hint that such flexibility may be exploited by other cellular genes was provided by the recent observation that the L and L′ subunits of pyruvate kinase are encoded by two distinct mRNAs, presumably derived from a single gene (137).

A remarkable regulatory principle involving RNA processing has been found in the expression of the cytochrome b gene in yeast mitochondria (138). In this system the first two exons and part of the second intron of the cytochrome b gene code for an “mRNA matu rase” which is, itself, responsible for the splicing out of the second intron. This constitutes a negative feedback system in which the activity of the maturase eliminates the mRNA that encodes it, and concomitantly produces cytochrome b-mRNA. The 5′ cleavage involved in the removal of the second intron is unorthodox in that it occurs downstream, rather than upstream, of a GU doublet. An inability of the normal mRNA splicing system to cope with this situation might account for the maturase requirement.

One of the most striking recent examples of the necessity for RNA processing in the expression of genetic information has come from a detailed examination of the human mitochondrial genome (139, 140). In this case, the genes for tRNA and various mRNAs and tRNAs are immediately contiguous to each other, leaving essentially no space for conventional transcriptional start and stop signals. The evidence accumulated to date indicates that over 90% of the genome is transcribed as a single unit, and that the various discrete RNA components are generated by a series of precise endonucleolytic cleavages of the growing transcript (140). The tRNA sequences, which punctuate most of the other genes, could conceivably serve as recognition signals for these cleavages.

These exciting new discoveries further illustrate the crucial role that RNA processing plays in mediating gene expression in eukaryotes. It is hoped that they will soon be paralleled by some deeper insights into the mechanistic aspects of the various processing reactions. One cause for some optimism in this regard is the recent progress in the determination of RNA secondary structure, in which accessibility to digestion with selected nuclease is used to discriminate among computer-generated models based on primary structure and minimum energy considerations (141, 142). These methods are suitable for relatively small amounts of material, and are thus capable of providing crucial information about the secondary structure of processing substrates. Such structural information should help considerably to sharpen our concepts of processing mechanisms.

ACKNOWLEDGMENTS

The works of the author, cited in this paper, have been supported by grants from the National Science Foundation, National Institutes of Health, and an appropriation from the Commonwealth of Pennsylvania is hereby acknowledged.

REFERENCES

1. Schultz, J. 1966. Natl. Cancer Inst. Monogr. 23:1–9.
2. Perry, R. P. 1976. Ann. Rev. Biochem. 45:605–629.
3. Abelson, J. 1979. Ann. Rev. Biochem. 48:1035–1069.
4. Darnell, J. 1979. Prog. Nucleic Acid Res. Mol. Biol. 22:327–353.
5. Brawerman, G. 1981. CRC Crit. Rev. Biochem. 10:1–38.
6. Scherrer, K., M.-T. Imazumi-Scherrer, C.-A. Reynaud, and A. Therwath. 1979. Mol. Biol. Rep. 5:5–28.
7. Banerjee, A. K. 1980. Microbiol. Rev. 44:175–205.
8. Brachet, J. 1957. Biochemical Cytology. Academic Press, Inc., New York.
9. Judson, H. F. 1979. The Eighth Day of Creation: Makers of the Revolution in Biology. Simon & Schuster, Inc., New York.
10. Roberts, R. B., R. J. Britten, and B. J. McCarthy. 1963. In Molecular Genetics, Part I J. H. Taylor, editor. Academic Press, Inc., New York, pp. 291–352.
11. Perry, R. P., M. Errera, A. Helli, and H. Durward. 1961. J. Cell Biol. 11:1–13.
12. Nat. Cancer Inst. Monogr. 23. 1966. The Nucleolus-Its Structure and Function. 610 pp.
13. Hall, B. D., and S. Spiegelman. 1961. Proc. Natl Acad. Sci. U. S. A. 47:137–146.
14. Perry, R. P. 1962. Proc. Natl Acad. Sci. U. S. A. 48:2179–2186.
15. Scherrer, K., H. Latham, and J. E. Darnell. 1963. Proc. Natl Acad. Sci. U. S. A. 49:240–248.
16. Reich, E., R. M. Franklin, A. J. Shatkin, and E. L. Tatum. 1961. Science (Wash. D. C.). 134:356–357.
17. Flessner, E., and E. Borek. 1962. Proc. Natl Acad. Sci. U. S. A. 48:1293–1303.
18. Sibatai, A., S. R. de Kloet, V. G. Allfrey, and A. E. Mirsky. 1962. Proc. Natl Acad. Sci. U. S. A. 48:471–477.
19. Georgiev, G. P., and V. L. Manev. 1962. Biochim. Biophys. Acta 61:153–154.
20. Perry, R. P., P. R. Srinivasan, and D. E. Kelley. 1964. Science (Wash. D. C.). 145:504–507.
21. Loening, U. E. 1967. Biochem. J. 102:251–257.
22. Bernhardt, D., and J. E. Darnell. 1969. J. Mol. Biol. 42:43–56.
23. Hamada, H., M. Muramatsu, Y. Urano, T. Oniski, and R. Kominami. 1979. Cell 17:163–173.
24. Brown, G. M., and G. Attardi. 1965. Biochem. Biophys. Res. Commun. 20:298–302.
25. Greenberg, H., and S. Penman. 1966. J. Mol. Biol. 21:527–535.
26. Weinberg, R. A., U. Loening, M. Willems, and S. Penman. 1967. Proc. Natl Acad. Sci. U. S. A. 58:1088–1095.
27. Madden, B. E. H., M. Salim, and D. F. Summers. 1972. Nature (Lond.). 237:5–9.
28. Wellauer, P. K., and I. B. Dawid. 1973. Proc. Natl Acad. Sci. U. S. A. 70:2827–2831.
29. Brown, D. D., and C. S. Weber. 1968. J. Mol. Biol. 34:681–697.
30. Birnstiel, M., J. Spelts, P. Purdom, K. Jones, and U. E. Loening. 1968. Nature (Lond.). 219:454–463.
31. Miller, O. L., and B. R. Beatty. 1969. Science (Wash. D. C.). 164:955–957.
32. Glover, D. M., and D. S. Hoggness. 1977. Cell. 10:167–176.
33. Box, J. L., C. Heyting, and P. Borst. 1978. Nature (Lond.). 275:336–338.
34. Roca, J. D., and P. Malnoe. 1978. Cell 15:661–670.
35. Bram, R. J., R. A. Young, and J. A. Steitz. 1980. Cell 19:393–401.
36. Stark, G. C., R. Kole, E. J. Bowman, and S. Altman. 1978. Proc. Natl Acad. Sci. U. S. A. 75:3717–3721.
37. Knaup, G. R., C. Ogden, C. B. Pettis, and J. Abelson. 1979. Cell 18:37–45.
38. Hagenbuchle, O., D. Larson, G. L. Hall, and K. U. Sprague. 1979. Cell 18:1217–1229.
39. Muller, R., and S. G. Clarkson. 1980. Cell 19:345–353.
40. deRobertis, E. M., and M. V. Olson. 1979. Nature (Lond.). 278:137–143.
41. Perry, R. P., E. Bard, B. D. Hames, D. E. Kelley, and U. Schiller. 1977. Prog. Nucleic Acid Res. Mol. Biol. 19:725–729.
42. Britten, R. J., and D. E. Kohne. 1980. Science (Wash. D. C.). 161:529–540.
43. Greenberg, J. R., and R. P. Perry. 1971. J. Cell Biol. 50:774–786.
44. Edmonds, M., and M. C. Caramella. 1969. J. Biol. Chem. 244:1314–1324.
