Review Article
Chemical Approaches for Structure and Function of RNA in Postgenomic Era

Tae Suk Ro-Choi and Yong Chun Choi

Departments of Pharmacology and Biochemistry, College of Medicine, Dong-A University, 3-1 Dong Dae Shin Dong, Seo Gu, Busan 602-714, Republic of Korea

Correspondence should be addressed to Tae Suk Ro-Choi, tsrochoi37@yahoo.com

Received 18 October 2010; Revised 21 June 2011; Accepted 5 July 2011

Academic Editor: Dmitry A. Stetsenko

Copyright © 2012 T. S. Ro-Choi and Y. C. Choi. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In the study of cellular RNA chemistry, a major thrust of research focused upon sequence determinations for decades. Structures of snRNAs (4.5S RNA I (Alu), U1, U2, U3, U4, U5, and U6) were determined at Baylor College of Medicine, Houston, Tex, in an earlier time of pregenomic era. They show novel modifications including base methylation, sugar methylation, 5′-cap structures (types 0–III) and sequence heterogeneity. This work offered an exciting problem of posttranscriptional modification and underwent numerous significant advances through technological revolutions during pregenomic, genomic, and postgenomic eras. Presently, snRNA research is making progresses involved in enzymology of snRNA modifications, molecular evolution, mechanism of spliceosome assembly, chemical mechanism of intron removal, high-order structure of snRNA in spliceosome, and pathology of splicing. These works are destined to reach final pathway of work “Function and Structure of Spliceosome” in addition to exciting new exploitation of other noncoding RNAs in all aspects of regulatory functions.

1. Introduction

A key element in the study of cellular RNA metabolism is the molecular characterization of RNA. This characterization requires accurate determination of the RNA sequence. It is imperative to understand how RNA structure complements the functional definition of RNA. Cellular RNAs are posttranscriptionally modified at various points in the primary RNA transcript as well as processed. In cellular RNA metabolisms, RNA maturation is performed through various structural alterations that include chemical modifications of constituent components. A most representative modification is observed in chain shortening, rearrangements by transfer of phosphodiester linkages involved in splicing mechanisms (pre-mRNA), deletions (pre-rRNA), and transsplicing (trypanosomal mRNA). Another is chain expansion demonstrated by modifications observed on polyadenylation, U-addition at 3′ ends, 5′-cap formation at 5′ ends, and insertions within trypanosome RNA. Other examples of modifications are base modifications, such as deaminations, methylations, hypermodifications, and ribose methylations.

The most modified RNAs are tRNAs containing approximately 2–22 modified nucleotides per molecule of ~75 nucleotide length, and there have been more than 130 different signature modified nucleotides reported [1]. The discovery of snRNA and m32,2,7G caps occurred within the last 50 years. They also contain their own specific modified nucleotides such as Ψ, m6A, m2G, and 2′-O-methylated nucleotides (Table 1).

The next class is the ribosomal RNAs which contain 204–209 modified nucleotides within 18S (1,869 nt) + 28S (5,035 nt) RNA in eukaryotes. The mRNAs contain the least modified nucleotides, with the exception of the 5′ end cap structure and occasional m6A in the molecule.

In ensuing years, massive scale DNA sequencing was advanced to accommodate the “Human Genome Project.” Two groups published the genomic map where the coding genes were cataloged. It was conservatively estimated that there are 25,000 genes and 50,000 proteomes involved in cell metabolism. It was also envisioned that processing mechanisms could be discerned by comparing the genomic structure with the RNA sequence determined using cDNA methods. Based
Table 1: Signature sequences and modifications of major snRNAs. The 5′ cap and 3′ nucleosides, base modified nucleosides, and alkali resistant oligonucleotides were determined by many methods described in the text. The table provides a summary of individual RNA characteristics of rat Novikoff hepatoma cells.

<table>
<thead>
<tr>
<th>RNA Subspecies</th>
<th>Subspecies</th>
<th>Localization</th>
<th>5′ End</th>
<th>3′ End</th>
<th>Modified Nucleotides</th>
</tr>
</thead>
<tbody>
<tr>
<td>5S RNA</td>
<td>I</td>
<td>Nucleoli, nuclei and cytoplasm</td>
<td>pppG</td>
<td>U-OH</td>
<td>3Ψ, m5A, m7G, AmA, AmG, AmGmC, CmC, CmA, CmCmU AmA, GmA, 2GmG, m7G, m8A, 3Ψ</td>
</tr>
<tr>
<td>IV</td>
<td>Nucleoli, nuclei and cytoplasm</td>
<td>pG</td>
<td>U-OH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>Nucleoli, nuclei and cytoplasm</td>
<td>pC, pG</td>
<td>U-OH, G-OH</td>
<td>AmC, 2Ψ</td>
<td></td>
</tr>
<tr>
<td>VI</td>
<td>Extraneucleolar nuclei</td>
<td>m3,2,7GpppAmUmAC</td>
<td>U-OH, A-OH</td>
<td>GmGmC, GmG, GmA, m7AmG, CmΨ, UmA, CmU, 13Ψ, (m8A, m7G)</td>
<td></td>
</tr>
<tr>
<td>VII</td>
<td>Extraneucleolar nuclei</td>
<td>m3,2,7GpppAmUmC</td>
<td>C-OH, A-OH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VIII</td>
<td>Extraneucleolar nuclei</td>
<td>m3,2,7GpppAm(m)AG</td>
<td>A-OH, U-OH, C-OH</td>
<td>2Ψ</td>
<td></td>
</tr>
</tbody>
</table>

on the ever-increasing number of RNA sequences, it was determined that most coding RNAs mature as a result of alternative splicing. Aberrant splicing is attributed to point mutations in the genetic code and splicing code [2]. It is noted that RNA sequencing can aid the determination of the molecular pathogenesis of diseases.

2. Historical Venture of RNA Research

Detailed nucleic acid chemistry began with discoveries of the DNA helix by Watson and Crick [3] and DNA polymerase by Lehman et al. [4, 5]. With DNA being the genetic material providing a blue print for living creatures, it moved genomic era thinking away from the earlier notion that protein, carbohydrate, and lipid were the only essences of living things.

DNA is there to provide information needed to build the cells, tissues, organs, and whole individuals. It took a long time to move from the histochemical presentation of DNA in the nucleus and RNA in the nucleolus and cytoplasm [14] to the isolation of nucleoli, nuclei, mitochondria and ribosomes, facilitating the elucidation of their components, their structures, and their functions. Even within the same species, no two individuals are identical. Disarray in DNA structure can determine whether one is healthy or diseased. In the quest to conquer cancer, differences in cellular morphology and uncontrolled growth became and remain a major research consideration when one compares normal cells with cancerous cells and tissues. Cancer cells with pleomorphic, hypertrophic nuclear, and nucleolar morphology remain a useful pathological criterion for a cancer diagnosis. The information within genes is transferred to RNA and then to proteins made on ribosomes that define a cell phenotype. The fractionation of cells into various components includes nucleoli, nuclei (Figure 1), ribosomes, mitochondria, cytosol and others.

The main interest among these compartmental components was the RNA. The RNA has its own exclusive properties which are not found in DNA.

The discovery of RNA polymerase I in the nucleoli [31] is the landmark of RNA research in these cellular compartments. It was not until 1968, with the introduction of gel electrophoresis into RNA research [32], that subspecies of 4–8S RNAs could be separated from high-molecular-weight RNAs (>18S RNA). Until then, the 4–8S RNAs were considered as tRNAs and their precursors. Different from the prokaryotic cells, eukaryotic cells were shown to have a variety of small RNAs in their nuclei (Figure 2). These RNAs used to be called LMWN RNA (low-molecular weight nuclear RNA) and now the name is unified as snRNA (small nuclear RNA).

These include U1 RNA, U2 RNA, U3 RNA, (named as such because these RNAs contain a high proportion of uridylic acid), 5S RNA III (U5 RNA), 4.5S RNA I (Alu RNA), 4.5S RNA II (U6), and 4.5S RNA III. All of these snRNA species and many more have been sequenced and their functions elucidated in pre-rRNA processing [33] and pre-mRNA splicing [34, 35].
Figure 1: Nuclear and nucleolar morphology. Normal rat liver nuclei have 1–6 round nucleoli which are less than 2 μm in diameter. In regenerating liver, cells contain enlarged nucleoli. In tumor cells (hepatocellular carcinoma), the nucleoli are not only enlarged but also they become pleomorphic in morphology. Nuclei were isolated by homogenization in 2.3–2.4 M sucrose containing 3.3 mM CaCl₂. Nuclei were sonicated in 0.34 M sucrose and layered on 0.88 M sucrose for purification by centrifugation. Isolated nuclei and nucleoli had high purity, and morphologies were well preserved [6].
Figure 2: Polyacrylamide gel electrophoretic separations of nuclear 4–7S RNAs of rat liver and Novikoff hepatoma cells [7]. (a) The 8% gel electrophoretic patterns of 4–7S RNA from various cell organelles of rat liver. The gel was stained with methylene blue for RNA visualization. (1) Nuclear 4–7S RNA, (2) ribosomal 4–7S RNA, (3) mitochondrial 4–7S RNA, and (4) soluble cytoplasmic sap 4–7S RNA. (b) The 10% slab polyacrylamide gel electrophoretic separation of [32P]-labeled 4–7S RNA from Novikoff hepatoma cell nuclei. The gel was autographed with X-ray film.

Table 2: Characteristics of RRM/RNP/RBD domain.

1. ~90–100 amino acids domain and most abundant in vertebrates
2. Many RNA binding proteins contain more than one RRM
3. Contains 2 conserved RNP1 (RGQAFVIF in β3) and RNP2 (TIYINNL in β1) in 4 antiparallel β-sheets of βαβαβαβ-fold
4. Binds 2–8 nucleotides of RNA (2 in CBP20, nucleolin and 8 in U2B"
5. A typical RRM containing 4 nucleotide binding sites (UCAC)
6. 3 conserved aromatic amino acids (Y, F, W, H or P) in central β-strands (2 in RNP1 of β3 and 1 in RNP 2 in β1)
7. 2 RRMs in a protein are separated by small linker and provide a large RNA binding surface or RNA binding surface point away from each other
8. RNA bases are usually spread on the surface of protein domains while the RNA phosphates point away toward the solvent
9. Binding surface of the protein is primarily hydrophobic in order to maximize intermolecular contact with the bases of the RNA
10. Few intramolecular RNA stacking and many intermolecular stacking mediated by aromatic amino acids
11. RNA recognition is a two-step process, in which any RNA is attracted approximately equally well. However, if stacking and hydrogen-bond interactions that “lock” the interaction cannot be properly established, the complex redissociates quickly (large koff), which results in overall weak affinity for RNA oligonucleotides of the wrong sequence.
12. Many ssRNA binding proteins recognize RNA in the loop (stem-loop) better than in ssRNA (kon ~ 3 fold & koff ~ 590 fold, therefore, overall affinity ~2000 fold differences) due to higher entropy loss with ssRNA binding than stem-loop binding and stabilizing interactions of stem

The most interesting discoveries in the midst of sequencing were the very unusual trimethylguanosine cap structure in U1 RNA (m2,2,7GpppAmUmAC), U2 RNA (m2,2,7GpppAmUmC), U3 RNA (m2,2,7GpppAmAm(m)-AGC), and 5S RNA III (U5 RNA) (m2,2,7GpppAmAmAC) [36]. Afterwards, myriads of cap structures in viral RNA and mRNA were discovered [37].

The history of RNA sequence work has occurred in three eras. The pregenomic era was devoted to the small RNAs and commenced with the sequence of large RNAs as technology developed for cDNA synthesis, amplification, cloning, and sequencing. The DNA technology was explosive and paved the way toward establishment of sequence technology not only for RNA and cDNA but also for genomic DNA.

In addition to sequence study, the secondary and tertiary structures have also been determined. A representative study was the crystallographic study of RNA-protein interactions. For example, the most well-worked-out motif is RRM (RNA recognition motif) which is most abundant in hnRNP [40] and splicing factors [41]. The summary of characteristics of RRM is in Table 2.

It has been known for a long time that pre-mRNA (hnRNA) is cotranscriptionally assembled into beads on a string consisting of 30–50S (20–30 nm) particles [42]. The
RNP (hnRNP) has usually 48 hnRNP proteins and ~700–800 nucleotide long RNA string [43]. More recently, most hnRNP proteins have been found to have 1-2 RRM motifs for RNA binding. From these characteristics, the primary RNA transcripts have been folded from the 5′ end with the following rules: a minimum of 3 nucleotides in the loop and a minimum of 3 base pairs at the stem. According to stacking and loop energy rules, two nucleotide loops cannot exist. The number of base pairs needed for stabilization with the most stable stacking energies by CCC/GGG or GGG/CCC is 3 base pairs with −9.8 kcal and the highest loop destabilizing energy is +8.4 kcal [44]. In addition, protein binding to RNA has been shown to have −ΔG ≈ 10–13 Kcal/mol [45] which can overcome the loop destabilizing energies of any size. With this rule, folding the hnRNA in GC, AU, and GU pairings was carried out as the RNA was transcribed, extending contiguous base pairing until it comes to a base pair mismatches. Accordingly, small simple RNA hairpins have been constructed with the aid of a computer [46] from the 5′ end (transcription start sites). Consensus patterns for folding characteristics have been observed (Table 3).

The transcripts form one stem loop for every 15–18 nucleotides which is consistent with ~15–17 nucleotides per hnRNP protein (700–800 nucleotides per 48 hnRNPs in one hnRNP particle) reported earlier [43]. The thermodynamics of RNA folding was consistent with the order of splicing in ovomucoid pre-mRNA [47]. From the point of view that suprasplicosomes contain hnRNP proteins (personal communication), it may be that this cotranscriptional formation of hnRNP string particles [47–49] may contribute to a role in the formation of suprasplicosomal RNP (Figure 3) [8].

The postgenomic era is the present day era or the second generation genome era. With the recent discovery that there is a paradox [50, 51] in the cellular transcript number, which is 2-3-fold in excess and that 50% of the cellular transcripts are ncRNAs, the second generation genomic era is in the process of resequencing the genome for ncRNAs. It is anticipated that there will be a revision in the first generation genomic picture. In this era, work is proceeding that will probe and dissect the RNA metabolism in which aberrant processing should be elucidated by RNA sequencing. To dissect the molecular pathology of RNA metabolism, it is also necessary to study higher-order structures based on the sequence studies involved in the assembly of macromolecular machinery. It is natural to hope that therapeutic interventions will be discovered that can correct errors in the genetic code and its product splicing.

**Table 3:** Frequency of stem loops in primary pre-mRNA transcripts. The simple stem loops with minimal 3 nucleotides in the loop and minimal 3 base pairs in the stem consisting of AU, GC, and GU pairs have been constructed with the aid of a computer [46].

<table>
<thead>
<tr>
<th>Transcript</th>
<th>nt/loop</th>
<th>nt/stem</th>
<th>nt in spacer</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Insulin</td>
<td>4.6</td>
<td>7.4</td>
<td>3.5</td>
<td>15.5</td>
</tr>
<tr>
<td>(2) HDHGT</td>
<td>5.8</td>
<td>7.0</td>
<td>4.9</td>
<td>17.6</td>
</tr>
<tr>
<td>(3) FMR1</td>
<td>5.0</td>
<td>6.8</td>
<td>3.4</td>
<td>15.3</td>
</tr>
<tr>
<td>(4) Ovomucoid</td>
<td>5.6</td>
<td>7.0</td>
<td>3.7</td>
<td>16.0</td>
</tr>
</tbody>
</table>

**Figure 3:** The suprasplicosome model from the article by Sperling et al. [8]. (a) It was stated that pre-mRNA which is not being processed is folded and protected within the native spliceosome. (b) With different staining protocol, it was possible to visualize the RNA strands and loops emanating from the suprasplicosome. These complexes were found to contain hnRNP proteins (personal communication).
Table 4: Paradoxical characteristics of ncRNAs in humans and mice [50, 51]. The excessive number of transcripts than anticipated for 25,000 genes indicates that the ncRNAs which were not detected due to scarce abundance have been detected by more sensitive methods. Some of these characteristics are summarized.

<table>
<thead>
<tr>
<th>Human</th>
<th>Mouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Number</td>
<td>75,000, 84,000 or 140,000 (cDNA identified)</td>
</tr>
<tr>
<td>Transcripts</td>
<td></td>
</tr>
<tr>
<td>Population</td>
<td>50% Poly-A RNAs (of 16% genome)</td>
</tr>
<tr>
<td>Intron</td>
<td>30% genome</td>
</tr>
<tr>
<td>Processing</td>
<td>Polyadenylation, 5′ cap, splicing, nucleotide modification</td>
</tr>
<tr>
<td>Transcripts from</td>
<td>Intergenic, Intron regions and antisense strand</td>
</tr>
<tr>
<td>Short ncRNAs</td>
<td>miRNA, siRNA (tasiRNA, natsiRNA), piRNA, rasiRNA (pitRNA), PARs (PROMTs, PARS, TSSs-RNAs, tIRNAS), MSY-RNA, snoRNA, sdRNA, moRNA, tel-sRNA, crasiRNA, hSRNA, scARNAs, AluRNA, YRNA, tRNA-derived RNAs</td>
</tr>
<tr>
<td>Long ncRNA (lncRNA)</td>
<td>Cancers, disorders in skin, heart, brain, cerebellum, and so forth. TR/TERC, NEAT RNA (NEAT1v-1, NEAT1v-2, NEAT2/MALAT1), PINC RNA, DD3/PCA3, PCGEM1, SPRY4-1T1, xiRNAs (Xist RNA, Tsix RNA, RepA RNA), AIR, H19, KCNQ1ot1, HOTAIR, BORG, CTN RNA, ANRIL RNA, LINE, CSR RNA, satellite DNA transcripts and so forth</td>
</tr>
<tr>
<td>(0.5–100 kb)</td>
<td></td>
</tr>
<tr>
<td>Function</td>
<td>Regulatory function in all aspects of metabolism [52]</td>
</tr>
</tbody>
</table>

The RNAs have been classified according to the following diverse basis of criteria:

(i) cell biology: cell types, subcellular origins,
(ii) molecular weight: high molecular weight (HMW) and low molecular weight (LMW/small),
(iii) S value: 5S rRNA, 7S RNA, 18S RNA, and others,
(iv) linearity: linear, cyclized, and branched (Y shaped),
(v) metabolism: precursor, processed intermediates, and mature,
(vi) standard: hnRNA, rRNA, mRNA, tRNA, and ncRNA (snRNA, snoRNA, miRNA, and others as in Table 4).

3. Preparation of RNA from Isolated Subcellular Compartments

RNA can be extracted from purified nucleoli, nuclei, ribosomes, mitochondria, and cytosol by the SDS-phenol procedure. The procedure involves the suspension of organelles in 0.3–0.5% SDS (sodium dodecyl sulfate), 0.14 M NaCl, and 0.05 M sodium acetate buffer at pH 5.0 and deproteinization by phenol containing 0.1% 8-hydroxyquinoline at 65°C [53]. The extracted RNA is precipitated with 2–2.5 volumes of ethanol containing 2% potassium acetate. The RNA is washed by ethanol and dissolved in appropriate buffer for the analysis. The DNA and protein contaminations are less than 3% by weight. The purified RNA is separated into individual RNA species using sucrose density gradient centrifugation, gel electrophoresis, and column chromatography [38].

4. Structure Determination

4.1. Structural Characteristics of Various RNAs Bearing Signature Sequences and Modifications. The RNA is composed of basic 4 nucleosides of guanosine, adenosine, uridine, and cytidine linked by 5′-3′ phosphodiester bonds between two ribose moieties. In addition, some of these nucleotides are modified in base as well as in ribose moieties and contain unusual pyrophosphate bonds at their 5′ ends and 2′ O-methylated 3′ end.

Mature RNAs are synthesized in the nuclei and directed by the posttranscriptional processing machineries. Because of these specific modifications, there is a general consensus on the presence of specific signature sequences and modifications for the identity of RNA classes. Based on extensive sequence work, it is possible to classify RNAs according to structural modifications. Figure 4 provides an outline for characteristics of RNA, and its modifications and brief examples are given in Table 5.

4.2. General Scheme of RNA Sequencing. The very first RNA sequence was obtained from the work of yeast alanine tRNA in 1965 [54]. In this work, the prerequisites for RNA sequence work were developed and described. Since then, it is a fundamental approach to establish oligonucleotide catalogs using specific RNases. One set is the catalog of T1 oligonucleotides produced by RNase T1. The other is the catalog of oligonucleotides produced by RNase A. The analytical method was based on UV spectral absorption in the earlier years. Subsequently, since 1970, isotopic labeling methods were widely used which are 1,000-fold more sensitive. Furthermore, many other improvements in RNA
Deletion
Splicing
intron5
\(-\text{end}\)
Cleave off
Insertion
\[m_{32,2,7}G\text{-cap}\]
\[m_{2,7}G\text{-cap}\]
m\(^7\)G-cap
my\(^\gamma\)G-cap
pppN, pN,

Figure 4: Summary of RNA modifications. Cotranscriptional and posttranscriptional RNA modifications are summarized.

Table 5: Examples of modification. All RNA species including high- and low-molecular-weight RNAs have their own signature sequences and modifications.

<table>
<thead>
<tr>
<th>RNA</th>
<th>Sequence and signature modification</th>
</tr>
</thead>
<tbody>
<tr>
<td>hnRNAs (Exons + Introns)</td>
<td>[m_{7}G\text{ cap}, m_{6}A, Poly-A, splicing codes}]</td>
</tr>
<tr>
<td>mRNAs (mainly exons)</td>
<td>[m_{7}G\text{ cap, m}_{6}A, Poly-A, Editing (C \rightarrow U &amp; A \rightarrow I)]</td>
</tr>
<tr>
<td>tRNAs</td>
<td>[TPC, CCA, Many hypermodified bases]</td>
</tr>
<tr>
<td>45S pre-rRNA</td>
<td>[Repeated U sequences at 5(^{-})-end spacers]</td>
</tr>
<tr>
<td>18S rRNA</td>
<td>[m_{3},^{2,7}G\text{ cap, my}\text{G cap}]</td>
</tr>
<tr>
<td>28S rRNA</td>
<td>[NmNmNmN, NmNmN]</td>
</tr>
<tr>
<td>snRNAs, snoRNAs</td>
<td>[Types I, II, and III caps by ribose methylations]</td>
</tr>
<tr>
<td>mRNA (Trypanosome)</td>
<td>[m_{3}\text{Gpppm}^{5,6}\text{AmpAmpCmpm}^{3}\text{UmmpAp}] Insertion of repeated U sequence</td>
</tr>
<tr>
<td></td>
<td>[Deletion of U sequence]</td>
</tr>
</tbody>
</table>

sequence technique have made it possible to advance the rate of RNA sequence work greatly (Table 6).

Improvement was observed in the following areas: (1) RNA labeling techniques, (2) fractionation procedures (chromatography, electrophoresis, and gel procedures), (3) use of various RNases, (4) contig seeking, and (5) ladder sequence gel analysis. For example, based on labeling at the 5\(^{-}\)-end with \[^{32}\text{P}\]-\gamma\text{-ATP}\ by polynucleotide kinase \[^{56}\text{]}\, it has become feasible to read a 150 nucleotide sequence using an endonuclease assisted ladder gel from the 3\(^{-}\)-end. Also, based on labeling at the 3\(^{-}\)-end with \[^{32}\text{P}\]-5\(^{\prime}\)-pCp by RNA ligase \[^{57}\text{]}\, it has become feasible to read approximately 150 nucleotides from the 5\(^{-}\)-end. Together, these enhancements make it readily feasible to sequence RNA with approximately 300 nucleotides. In contrast to success in the sequence work for small RNAs, two challenges remained. One challenge is related to RNA size and the other is concerned with scarce abundance of RNA in the cell. With the discovery of reverse transcriptase, heat stable DNA polymerase, and recombinant technology, it became possible to produce cDNA, amplify, and clone by RT-PCR methods.

With high-efficiency RT-PCR, high-molecular-weight RNA with 10,000 nucleotides in length can be readily sequenced \[^{59}\text{]}\,. A remaining shortcoming of this approach is the inability to fully characterize modified nucleotides. However, ability to deal with long chain lengths and scarce abundance outweighs this limitation. cDNA-based methods clearly dominate any RNA sequence work that involves long RNA length or low RNA abundance. Examples are observed in the direct gene isolation for cleavage controlled processing RNAs (Pre-rRNA and rRNA) and cDNA method for pre-mRNA and mRNAs. Therefore, as a result of accumulated methodologies, it becomes common that RNA sequence can be obtained through more than one scheme or type of technique, such as straight chemical approaches \[^{60}\text{]}\, or biotechnology-mediated approaches.

4.3. Outlined Steps of Sequence Work. Brief outlines are described for sequencing RNAs. It may be divided into two methods although combined methodology is in fact feasible.

4.3.1. Direct Method of RNA Sequencing

(a) Preliminary Examination of External Glycol Structures. In some cases, a rapid diagnostic examination is required. Most convenient procedures employ the use of specific antibodies against different forms of 5\(^{-}\)-cap structure (m\(^7\)G cap or m\(^{3,2,7}\)G cap) and a oligo-dT column for poly-A affinity chromatography. Alternatively, a \[^{3}\text{H}\]-derivative method can be useful. The radioactive labeling of terminals was performed using the periodate oxidation method, followed by reduction with \[^{3}\text{H}\]-borohydride. T\(_2\) RNase digestion and fractionation by paper chromatography reveal the presence of the 3\(^{-}\)-terminal and 5\(^{-}\)-cap.

(b) Selection of Labeling Methods. RNA can be labeled in vivo (prelabeling) or in vitro (postlabeling).

In vivo labeling is carried out by incubation of living cells in the presence of \[^{32}\text{P}\]-phosphate in a phosphate-free medium. RNA is uniformly labeled by this method.

In vitro labeling is called postlabeling because it labels the isolated RNA with isotopic agents such as \[^{32}\text{P}\]-phosphate or \[^{3}\text{H}\]-borohydride. \[^{32}\text{P}\]-labeling can be carried out using kinase enzymes. The 5\(^{-}\)-labeling is done with \[^{32}\text{P}\]-ATP by
Table 6: General schemes of RNA sequencing. The direct and indirect methods of RNA sequencings are briefly outlined. The cDNA and DNA pathways are considered indirect methods.

<table>
<thead>
<tr>
<th>Starting material RNA</th>
<th>Starting material RNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>[\text{Endonuclease}]</td>
<td>[\text{RT-PCR}]</td>
</tr>
<tr>
<td>Catalogs of oligonucleotides</td>
<td>cDNA library</td>
</tr>
<tr>
<td>Contig seeking</td>
<td>GenomicDNA library</td>
</tr>
<tr>
<td>Large fragments</td>
<td></td>
</tr>
<tr>
<td>Endonucleases assisted sequencing</td>
<td></td>
</tr>
<tr>
<td>Ladder sequencing gel</td>
<td></td>
</tr>
<tr>
<td>Sequence</td>
<td></td>
</tr>
<tr>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Modifications</td>
<td></td>
</tr>
<tr>
<td>RT-PCR</td>
<td></td>
</tr>
<tr>
<td>cDNA library</td>
<td></td>
</tr>
<tr>
<td>Genomic DNA library</td>
<td></td>
</tr>
<tr>
<td>Screening</td>
<td></td>
</tr>
<tr>
<td>Clone</td>
<td></td>
</tr>
<tr>
<td>Restriction mapping</td>
<td></td>
</tr>
<tr>
<td>Shot gun</td>
<td></td>
</tr>
<tr>
<td>Subclone</td>
<td></td>
</tr>
<tr>
<td>Chain termination</td>
<td></td>
</tr>
<tr>
<td>Chemical degradation</td>
<td></td>
</tr>
<tr>
<td>Sequence</td>
<td></td>
</tr>
</tbody>
</table>

polynucleotide kinase, that is, provided the 5′-end is free from phosphate. If the 5′-end is blocked by the presence of a 5′-cap structure, the pyrophosphate moiety must be removed by a pyrophosphatase and phosphatase. And then the kinase method can be employed to introduce the tracer. Labeling at the 3′-end is done with \[^{32}\text{P}^\text{-pCp}\] by RNA ligase. The \[^{3}\text{H}\]-derivative (nucleotide diol) with \[^{3}\text{H}\]-borohydride indicates that the 3′-end is free from phosphate or any other blocking structures. A shortcoming of \[^{32}\text{P}\]-labeling is the short half-life of the isotope which provides a working period of approximately 4 half-lives. The main limitation of the \[^{3}\text{H}\]-labeling method is weak energy of the tritium isotope. This can make the reading of the autoradiograph for a ladder sequencing gel very difficult.

(c) Initial Reading of Sequence by Ladder Sequencing Gel. To obtain the nucleotide sequence of RNA quickly without characterization of modified nucleotides, it is common to use the endonucleases-dependent sequencing technique [61]. Terminal labeled RNA (5′-end or 3′-end) is partially digested with specific endonucleases (T_1, U_2, A, phy I, and others), and each product is loaded in parallel on a 10–15% denaturing polyacrylamide gel. Note that if crude acrylamide is used, the running temperature of the gel can quickly rise to 60–70°C. Since the mode of cleavage is known, it is possible to discern G (T_1), A (U_2), U and C (A) and C-resistance (Phys I). It is not uncommon to read an RNA sequence using this method within one day.

(d) Base Composition. There are two technical approaches that can be used to determine RNA base composition (levels of nucleotides or of nucleosides).

RNase T_2 or alkali (0.3 N KOH) is used to complete hydrolysis. But alkali (0.3 N K/NaOH) is not preferred because
it destroys 7-methyl purines. Prelabeled \([^{32}P]-RNA\) is hydrolyzed, and its products are separated by 2-dimensional paper chromatography followed by autoradiography \([62]\). Since the standard separation pattern is known, various modified nucleotides are readily identified by comparison \([56]\).

Alternatively, after cold RNA is digested into constituent nucleotides, which are subsequently dephosphorylated by phosphatase, the resulting nucleotides are converted into \([^{3}H]-\)derivatives and separated by thin layer chromatography. The separated nucleotides (including all modified nucleotides except 2′-O-methylated nucleosides) are detected by autoradiography and identified based upon a standard migration pattern (Figure 5) \([9]\).

(c) Catalogs of Oligonucleotides. Two types of catalogs are made. One is an RNase T\(_1\) catalog, and the other is an RNase A catalog.

To map oligonucleotides, two necessary procedures are essential. The first is to prepare labeled oligonucleotides and the second is to fractionate two-dimensionally.

To obtain labeled oligonucleotides, three approaches are possible.

1. Use of prelabeled \([^{32}P]-RNA\) for specific endonuclease digestion.
2. 5′ labeling after enzyme digestion using \([^{32}P]-ATP\) and polynucleotide kinase.
3. 3′ labeling after endonuclease digestion and removal of resultant 3′-phosphate by phosphatase. Then the labeled derivatives can be formed by \([^{32}P]-5′-pCp\) and RNA ligase or periodate oxidation followed by \([^{3}H]-\)borohydride reduction.

To Map Oligonucleotides. There are a number of different techniques. However, most common are a combination of high voltage paper electrophoresis on cellulose acetate at pH 3.5 and high voltage DEAE paper electrophoresis (7% formic acid) or high voltage electrophoresis on cellulose acetate at pH 3.5 followed by DEAE homochromatography at 60–70°C. Another method that can be used is two-dimensional thin layer (PEI) chromatography using two-solvent systems \([63]\). Detection is performed by autoradiography. It is notable that T1 oligonucleotides from 45S prerRNA can be fractionated into approximately 200 spots by homochromatography \([64]\).

To Sequence Oligonucleotides. Several enzymatic digestions can be exploited.

The recovered \([^{32}P]-\)oligonucleotides (prelabeled) are subjected to secondary digestions with RNase U\(_2\) for placement of A residues, RNase T\(_1\) for G residues, RNase A for U, and C residues plus other endonucleases. Treatment with exonucleases (spleen phosphodiesterase, snake venom phosphodiesterase), and partial digestion with the enzymes above is required to sequence RNA. In each step, nucleotide composition is determined.

To Determine the Sequence of 5′-Labeled \([^{32}P]-Oligo-Nucleotides\). A mobility shift test can be applied \([56]\). After partial hydrolysis with snake venom phosphodiesterase the product is fractionated by homochromatography or PEI thin layer chromatography. The mobility shift pattern is produced according to the step-wise loss of each nucleotide from the 3′-end. The resulting pattern can be used to read the sequence of the oligonucleotides.

To Determine the Sequence of \([^{3}H]-Oligonucleotides\). The procedures used for prelabeled \([^{32}P]-\)oligonucleotides are applicable. Secondary digestion methods and accompanying \([^{3}H]-\)derivative methods for the determination of nucleotide composition can be carried out.

It may be necessary to strengthen the catalog of oligonucleotides. Generally this involves the expansion of the catalog to provide contiguous overlapping sequences. A feasible approach is to produce large fragments (purified on 10–15% denaturing polyacrylamide gel electrophoresis) and identify the overlapping oligonucleotides. Usually a limited fragmentation by a diluted endonuclease at low temperature or water hydrolysis may produce large overlapping fragments \([63]\). Examination of large fragments, as done above for ladder gel sequencing and catalogs, can often clarify any ambiguity encountered. An excellent example of one hit hydrolysis is observed in the work on tRNA structure \([63]\). Based on these very same methods, it can be summarized that many small RNAs have been sequenced. These include tRNAs, pre-tRNAs, 4.5S RNA I, 5S rRNA, 5.8S rRNA, snRNAs, snoRNAs, 7S RNA, and some fragments of pre-rRNA, 28S rRNA, and 18S rRNA.
4.3.2. Indirect Method of RNA Sequencing. The indirect method of RNA sequencing using cDNA or DNA gene analysis was developed as part of explosive advancements with DNA biotechnology. The direct RNA sequencing method proved useful for the characterization of small RNAs (~100–300 nt). However, sequencing high-molecular-weight (HMW) RNAs proved to be too difficult. Moreover, HMW RNAs that are scarce abundance often do not meet the sample amounts required by the former methods. The search for a solution to this dilemma was successful. One solution involved the isolation of the gene that codes for a specific RNA and the other is to synthesize cDNA which can also be used to isolate a specific RNA gene. Using DNA biotechnology, it proved possible to scale up and solve “The Human Genome Project.” Several genomes have been sequenced, specifically the human (2.9 Gb) and mouse (2.5 Gb) genomes [65–67]. In well equipped laboratories, it is possible to sequence DNA at the rate of 10^6–10^7 nt/day. This technology has been widely commercialized and is currently available as kits for cDNA cloning, sequencing, along with enzymes and equipment that supports automatic sequencing. The principal objective of the genomic approach was to determine the sequences of the coding genes. Vast collections of sequence data were compiled for RNAs, cDNAs, and genomic structures, revealing the base sequences for a number of RNAs. As a result of this work.

(a) Unidentified proteins have been predicted to number 25,243; whereas the known protein number is 15,337.

(b) A majority of mRNA species (95%) mature through alternative splicing mechanisms.

(c) Disease genes are estimated to be 2,577 in number.

(d) Point mutations are 31,250 in number; half of disease-causing mutations are attributed to aberrant splicing (disruption of splicing codes) whereas other forms of mutation include disruption of the genetic code.

(e) Disruption of splicing code occurs at the splice site and enhancer/silencer sites of exonic and intronic sequences.

(f) Pathogenic sequences that occur as a result of splice code mutations (transition and transversion) cause aberrant modifications of a variety of RNAs [68, 69].

Recently, evidence has been accumulating that suggests a need to revise earlier estimates of the number of transcriptional products arising from the genomic information. Paradoxical findings were obtained that contradicted earlier and more conservative estimates of the proteasomes size (50,000), in fact, the cellular transcripts are 2-3 times higher than estimated earlier [50, 51]. Also, 50% of the transcripts were comprised of noncoding RNA, some of which are polyadenylated. This paradoxical manifestation has led to the second generation of genomic work, strictly based on RNA characterization. It is worth emphasizing that this has become the second genomic frontier where a reevaluation of the first genomic work is necessary. The present task is more daunting than the “First Generation Genome Project.”

The task at hand is to resequence the genome and then categorize and catalogue the ncRNA species by utilizing all available sequence means, including direct sequencing and DNA microarray techniques.

The next step is to construct secondary structures according to enzyme susceptibility and computer-aided base pairing. Interacting proteins will need to be defined by biochemical, NMR, X-ray, and cryo-EM methods.

5. Reagent and Procedures Required for Sequencing

5.1. RNA-Specific Cleavage Reactions (2'-OH Required Reaction)

(1) Mild alkaline hydrolysis (0.3 N KOH) produces 3' monophosphorylated nucleotides.

(2) T₁ RNase cleaves phosphodiester bonds after G base producing 3' GMP at 3' ends.

(3) RNase A cleaves phosphodiester bonds after pyrimidines (U and C) producing 3' phosphates at 3' ends.

(4) T₂ RNase cleaves all phosphodiester bonds with a preference for A residues, producing 3' monophosphates.

(5) U₂ RNase cleaves phosphodiester bonds after A base, producing 3' monophosphates.

The mechanism catalyzed by alkaline hydrolysis, RNase A, T₁ RNase, T₂ RNase and U₂ RNase involves a S₈(2,p) mechanism attacking 2'-hydroxyl groups on the adjacent internucleotidic phosphodiester bond to displace the 5'-hydroxyl group of the neighboring nucleotides and generate a 2', 3'-cyclic nucleotide intermediate. A subsequent hydrolysis of the 2', 3'-cyclic nucleotide yields a final product, a 3' mononucleotide (Figure 6).

5.2. The Enzymes Cleaving All Phosphodiester Bonds Including 2'-O-Methylated Ribose

(1) P₁ RNase: the enzymatic digestion by P₁ RNase cleaves all phosphodiester bonds (except pyrophosphate linkages), producing 5' monophosphorylated nucleotides.

(2) The enzymes acting from the ends for sequencing fragments

(a) Snake venom phosphodiesterase (phosphodiesterase I) cleaves phosphodiester bonds, as well as pyrophosphate bonds producing 5' monophosphorylated nucleotides. It cleaves single-stranded RNA or DNA from the 3' end in a progressive manner.

(b) Spleen phosphodiesterase (phosphodiesterase II) produces 3' monophosphorylated nucleotides cleaving from nonphosphorylated 5' ends of single-stranded RNA or DNA.
Stereochemistry of the reaction catalyzed by RNase A

Figure 6: Stereochemistry of the reaction catalyzed by RNase A [10]. The intermediary 2',3’ cyclic nucleotide (cNp or cNMP) is hydrolyzed to a 3’ phosphorylated mononucleotide. Other 2’-OH requiring enzymatic and alkaline hydrolysis may go through the same path.

5.3. Other Enzymes Utilized for Sequencing

(1) Alkaline phosphatase removes phosphate from 3’ and 5’ ribose moieties.

(2) Pyrophosphatase will only cleave pyrophosphate linkages. There are pyrophosphatases from tobacco and potato as well as from Crotalus adamanteus venom type II.

Using varying combinations of fragmentation methods, it becomes possible to obtain fragments that range in size from nucleosides to very large fragments.

5.4. Chemical Modifications Used for Sequencing

5.4.1. CMCT Reaction. Originally reported by Gilham [73], the adduct formation of uridine and guanosine components of RNA with CMCT made uridine residues resistant to RNase A. In addition it has been shown that CMCT reacts with pseudouridine and to a lesser extent with inosine. This reaction takes place on U, C and modified residues in the fragments or in the molecule. These 4 bases have specific UV spectra and chemical reactivity to identify the nature of the bases in comparison with known standards. The unusual nucleoside, trimethylguanosine, has its specific UV absorption spectra (Figure 8) and mass spectrometric characteristics (Figure 9).

5.4.2. DMS (Dimethylsulfate). This has been used to identify secondary structures as well as for the synthesis of standard m32,2,7G. The properties of DMS modifying adenosine (N1) and cytosine (N3) make modified nucleotides unable to base-pair. For this reason RT-PCR stops one nucleotide before the modified nucleotide before the modified nucleotide enabling the location of a modified nucleotide as well as differentiating the single-stranded from double-stranded regions of RNA. DMS also has been used for synthesis of m32,2,7G from N2,N2-dimethylguanosine. For this synthesis, the reaction has been carried out by the methods of Saponara and Enger [76]. Twenty milligrams of N2,N2-dimethylguanosine were suspended in 400 μL of dimethylacetamide containing 10 μL dimethylsulfate. The mixture was shaken for 15 hours at room temperature and then centrifuged to remove insoluble products. The supernatant was adjusted to pH 8.0 with concentrated ammonia and then placed on a phosphocellulose column (1 × 50 cm) at pH 7.0 (0.001 M ammonium acetate). A linear gradient of 0.001–0.3 M ammonium acetate was used to elute the samples. One major peak of the product (m32,2,7 trimethylguanosine) was found between two minor peaks (corresponding to N2,N2-dimethylguanosine and 7-methylguanosine). The product was lyophilized and identified as m32,2,7G by mass spectrometry [12]. The summary of reagents and procedures required for sequencing is provided in Table 7.

The nucleotides or nucleosides obtained can be separated by column chromatography, paper electrophoresis or thin layer chromatography to determine the number of G, A, U, C and modified residues in the fragments or in the molecule. These 4 bases have specific UV spectra and chemical reactivity to identify the nature of the bases in comparison with known standards. The unusual nucleoside, trimethylguanosine, has its specific UV absorption spectra (Figure 8) and mass spectrometric characteristics (Figure 9).

6. The Major snRNA Sequenced

The first nuclear small RNA sequenced was 4.5S RNAI [77] shown in Figure 37. This RNA contains the RNA polymerase III promoter box A and box B like motifs and shows interesting enhancer motif elements resembling the Alu element transcript. The RNA polymerase III promoter areas are underlined and the first nucleotide of the enhancer motif is marked by colored letters. The red color is SF2/ASF (4 motifs), blue color is SC35 (3 motifs), green color is SRp40 (6 motifs), and yellow color is SRp55 (1 motif) (Figure 10(a)). It also exhibits 3’-splice sites marked by [AG] as well as branch sites with the highest score marked by (CACCUAU) (Figure 10(b)). The ESE (exonic splice enhancer), splice sites (Figure 10(c)), and branch sites were examined by ESEfinder 3.0 [13].

In comparison with known Alu elements in the FMR1 gene, the resemblance of 4.5S RNA I in ESE, 5’SS, BS, and 3’SS distribution (Table 8) suggests that 4.5S RNA I is more likely derived from an Alu gene expressed in Novikoff hepatoma cells.

The Alu element has been shown to have many different functions in transcription, splicing, exonization [78], gene insertions (transposons), and DNA replication. It is interesting to observe that the (+) oriented Alu has more
5’ splice sites and the (−) oriented Alu has more 3’ splice sites. It may suggest that exonization may occur from the 5’ side of (+) Alu elements and 3’ side of (−) Alu elements. The SRP RNA (7SL RNA) has Alu elements in its sequence [79]. Whether the Alu is derived from 7SL or Alu is exonized to 7SL is not clear. Subsequently, other snRNAs have been sequenced.

The sequences of the capped snRNAs are described in Figure 11. The pivotal sequences needed for functions are marked by colors.

In the course of any sequence work, there are always challenges in resolving unknown structures at the 5’ end portions which contain the 5’-cap structure and various modified nucleotides. The experimental steps required to discern this complicated region are described.

7. Nucleotide Composition and Modified Nucleotides in snRNAs

The compositional analyses were carried out by UV analysis as well as isotope labeling analysis. For example, UV analysis required ~10 mg of U2 RNA.

7.1. RNA Terminal Labeling with [3H]-KBH₄. The purified nuclear RNAs were separated by sucrose gradient centrifugation which separates 4–8S RNA, 18S RNA, 28S RNA, 35S RNA, and 45S RNA isolated from nuclei of rat liver, Walker tumor, or Novikoff hepatoma cells. As an initial step for the structural characterization, 3’ end nucleosides were labeled by the procedure of sodium periodate (NaIO₄) oxidation and potassium borohydride ([3H]-KBH₄) reduction.
Detection of external modification

- Oligo-dT column
- Borate column
- Antibodies against m7G, m32,2,7G, and others

Labeling

- Prelabeling (in vivo) with [32P]-phosphate
- Postlabeling (in vitro)
  - 5′ [32P] labeling
  - 3′ [32P] labeling
  - 3′ and 5′ [3H]-derivative labeling

Enzymes

- RNase T1 (Gp↓N), RNase A (Up↓N or Cp↓N), RNase U2 (Ap↓N), PhysI (C-resistance), RNase T2 (Np↓N), RNase P1 (pNm↓pN, pN↓pN), Debranching enzyme (N2p↓5p)

- Exonuclease
  - Spleen phosphodiesterase (Np↓pNp↓pNp)
  - Snake venom phosphodiesterase (N↓pN↓pNp)

- Phosphatase (p↓p)
  - Pyrophosphatase
  - Phosphatase (p↓N and N↓p)

Chemicals

- CMCT
  - React with U>Ψ>G>I on ssRNA
- DMS
  - Methylate G>A>C on ssRNA
- Diethyl pyrocarbonate
  - React with A>G
- Hydrazine
  - React with U>C>T

Limited hydrolysis

- Formamide
  - Hot water (80°C)
- 2D acrylamide gel electrophoresis
- 2D homochromatography
- 2D electrophoresis
- 2D paper chromatography
- 2D thin layer chromatography

The reaction was carried out in 0.1 M sodium acetate buffer at pH 5 with freshly prepared NaIO4 in the dark for 1 hour and precipitated the RNA with ethanol. The RNA was redissolved in the same buffer and treated with ethylene glycol to destroy excess NaIO4. The RNA was precipitated with ethanol and redissolved in 0.1 M sodium phosphate buffer, pH 7.7, and treated with radioactive [3H]-KBH4 [38]. These reaction products would have tritium labeling in cis-alcohols from cis-aldehyde oxidation products of the 2′ and 3′ hydroxyls of ribose, assuming all 3′ ends of RNA have accessible 2′ and 3′ OH groups (Figure 12).

The labeled 4–8S RNAs were separated by preparative polyacrylamide gel electrophoresis (Figure 13) and DEAE-Sephadex column chromatography (Figure 14) to purify individual snRNAs (U1, U2, U3, 4.5S RNA I, II, and III, 5S RNA I, II, and III).

Alkaline hydrolysis of these RNAs produced 3′ end nucleoside trialcohol derivatives (Table 9) which were subsequently identified by paper chromatography.

The RNA that appeared to be pure for sequencing was 4.5S RNA I which had 87.4% U at the 3′ terminus and only 6.5% unknown radioactivity at the origin. Unexpectedly, U1, U2, U3, 4.5S RNA II, and some of 5S RNA (5S RNA III/U5) had ~50% labeling in alkaline-resistant fragments that did not move as nucleoside derivatives. The 4.5S RNA III was not labeled by this procedure suggesting a blocked 3′ end (Figure 14). The U1, U2, and U3 RNAs were labeled with tritium, digested with RNase A, and separated on a DEAE-Sephadex column (Figure 15).

The oligonucleotides were digested with T1 RNase and rechromatographed, and only the U3 oligonucleotide was shortened by one nucleotide, indicating the presence of one G adjacent to RNase A susceptible pyrimidine [80]. In the course of sequencing U1, U2, U3 RNAs, it was found that the oligonucleotides with m32,2,7G was coming from the 5′ end segments. The only way 2′3′ hydroxyls could be at 5′ end was 5′-5′ pyrophosphate linkage to the rest of the RNA molecules [36]. The RNase A and T1RNase resistant oligonucleotides were digested with various enzyme combinations including snake venom phosphodiesterase, alkaline phosphatase, P1 RNase, T2 RNase, and U2 RNase into nucleosides. The component nucleosides were identified by mass spectrometry, U.V. spectroscopy, HPLC (high pressure liquid chromatography), paper chromatography, and thin layer chromatography. [12, 16, 37, 58].
7.2. Tritium Labeling of Nucleosides. The purified RNAs were digested with RNase A, snake venom phosphodiesterase, and alkaline phosphatase at pH 8.0, 37°C for 6 hours into nucleosides. The digest was treated with a 2X molar excess of NaIO₄ and labeled with [³H]-KBH₄ at pH 6 for 2 hours in the dark to produce trialcohol derivatives of nucleosides. All nucleosides with base modifications, except 2′-O-ribose modified, were labeled with tritium. The tritium-labeled trialcohol derivatives were separated by two-dimensional TLC (thin layer chromatography) on cellulose thin layers (Figure 5) [81]. The first dimension used a solvent of acetoni-trile, ethylacetate, n-butanol, isopropanol, 6 N aqueous ammonia (7:2:1:1:2:7); the second dimension used a solvent of tert-amyl alcohol, methyl ethyl ketone, acetoni-trile, ethylacetate, water, formic acid (sp.gr. 1.2) (4:2:1.5:2:1.5:0.18) [81, 82].
Presence of ESE, branch sites and splice sites in 4.5S RNA I (Novikoff)

Table 9: The snRNA 2' and 3'-OH labeling by NaIO₄ oxidation and [³H]-KBH₄ reduction [38]. The total 4–7S RNA from rat Novikoff hepatoma cell nuclei was labeled with [³H] by oxidation with NaIO₄ followed by [³H]-KBH₄ reduction (Figure 12). Individual RNA species were purified by gel electrophoresis (Figure 13). The RNA samples were hydrolyzed with 0.3 N KOH, and hydrolysates were chromatographed on whatman 3MM paper according to de Wachter and Fiers [55]. The radioactivities at the origin (22% for 5S RNA, 54.1% for U1 RNA, 49.7% for U2 RNA, and 50.6% for U3 RNA) represent % of total radioactivity applied and they represent the 5' end labeling which was later elucidated by many enzymatic methods described in the text. The radioactivities moved by chromatography with standard nucleoside derivatives are the % of total in nucleosides derivatives. The A' U' G' C' represent triakohol derivatives of nucleosides.

<table>
<thead>
<tr>
<th>RNA Species</th>
<th>Radioactivity at origin (5')</th>
<th>A'</th>
<th>U'</th>
<th>G'</th>
<th>C'</th>
</tr>
</thead>
<tbody>
<tr>
<td>4S RNA</td>
<td>10.9</td>
<td>89.0</td>
<td>3.2</td>
<td>3.9</td>
<td>3.8</td>
</tr>
<tr>
<td>4.5S RNA</td>
<td>15.8</td>
<td>11.2</td>
<td>79.7</td>
<td>4.0</td>
<td>5.1</td>
</tr>
<tr>
<td>4.5S RNA I</td>
<td>6.5</td>
<td>6.1</td>
<td>87.4</td>
<td>4.7</td>
<td>1.8</td>
</tr>
<tr>
<td>4.5S RNA II</td>
<td>30.9</td>
<td>13.1</td>
<td>80.2</td>
<td>4.7</td>
<td>2.4</td>
</tr>
<tr>
<td>5S RNA</td>
<td>22.0</td>
<td>11.4</td>
<td>75.5</td>
<td>6.0</td>
<td>7.0</td>
</tr>
<tr>
<td>U1 RNA</td>
<td>54.1</td>
<td>6.0</td>
<td>13.4</td>
<td>77.7</td>
<td>3.0</td>
</tr>
<tr>
<td>U2 RNA</td>
<td>49.7</td>
<td>61.5</td>
<td>6.0</td>
<td>4.3</td>
<td>28.2</td>
</tr>
<tr>
<td>U3 RNA</td>
<td>50.6</td>
<td>53.8</td>
<td>22.6</td>
<td>9.8</td>
<td>13.7</td>
</tr>
</tbody>
</table>

7.3. [³²P] Labeling of RNA. The Novikoff hepatoma cells were transplanted intraperitoneally into male albino rats of the Holtzman strain weighing 200–250 g, obtained from Cheek Jones Company (Houston, Tex). After 5–6 days, the cells were harvested and washed with NKM solution (0.13 M NaCl, 0.005 M KCl, and 0.008 M MgCl₂). Twenty milliliter (packed volume) of cells was incubated with 500 mCi of [³²P]-orthophosphate in 1 liter of medium (phosphate free modified Eagle's minimal essential medium) for 9–16 hours [83]. Nuclear RNA was purified by sucrose gradient centrifugation, gel electrophoresis, and column chromatography [38]. The purified RNA was hydrolyzed with 0.3 N KOH, and alkaline-resistant oligonucleotides were separated on DEAE-Sephadex. The alkaline resistant dinucleotides were collected, treated with alkaline phosphatase, and identified by two-dimensional chromatography (Figure 16).

The summary of modified nucleotides is in Table 1 [84].

8. Structural Determination of 5' Oligonucleotides

The structures of the 5' ends of U1 RNA, U2 RNA, U3 RNA, and 5S RNA III (U5) are determined by the characteristics of chemical reactions and enzymatic susceptibilities (Figure 17).

8.1. U1 RNA 5' End Oligonucleotide. The U1 RNA labeled with [³H] by NaIO₄ and [³H]-KBH₄, digested with RNase...
(1) Human HeLa U1A RNA (2) Rat Novikoff U1 RNA (Involved in pre-mRNA splicing)

(1) \[ \text{m}_3^{2,2,7} \text{GpppAmUmAC'T'FACCU} \]
\[ \text{GCGAGGGGAG AUACCAUGAU} \]

(2) \[ \text{m}_3^{2,2,7} \text{GpppAmUmAC'T'FACCU} \]
\[ \text{GCGAGGGAGA UACCAUGAU} \]

(1) CACGAAGGUG GUUUUCCCG AGCGAGGCUG AUUCAUUUGAm CUCGGGAAGU
(2) AGCAAGGUGUU UUUUCUGCG CCGGUGGCUC UAUCCAUGA GAGCGAmCUCC

(1) GCUGACCCUCU GCGAUAUCUC CAAAUGUGGG AAACUGGACU GCAUAAAUUUG
(2) GUGGAUGCGU ACCCCUGCGC UUUCUCCCAA AUGCGGAGAA CUGCAUGCA

(1) UUGGAUGGGG GGAGCGGCUU CGCGCUCUCC CCUG–OH (164) [15]
(2) UAAUUGUGG UAGUGCGGGG GACUGUGUCG GCUCGGUG–OH (170) [16]

Pre–mRNA 5' splice site binding region Sm protein binding region

(3) Human HeLa U2 RNA (4) Rat Novikoff U2 RNA (Involved in pre-mRNA splicing)

(3) \[ \text{m}_3^{2,2,7} \text{GpppAmUmCGCUUCUC GmGmCUUUUGmG CUAAGmAUCAAm} \]
(4) \[ \text{m}_3^{2,2,7} \text{GpppAmUmCGCGmGmCUC CYYFYUUAGmG CUAAGmAUCA} \]

(3) GUAGUAUCm WWGUUUAAUAC mUGAUAGCU
(4) \[ \text{m}_3^{5} \text{GuGUGYAG'F} \text{WmG'YPYUCUmAU CAGUYYUAA'W AUCmUUCAUA} \]

(3) CUCUAUCCGGA GAACAUAUA UUUAAUGGA UUUUGGAGCA GGGAGAGGA
(4) CUGGCUUUA CGAGAGCAAA UAAYUAYUAA UGGAAAAUUG GAAUCAUGG

(3) AUAGGACCUU GCUCGCUCCA CUCACGCAU CGACCUGUA UUGCAUGACC
(4) UUGGAUAGG AGCUUGCUCC GUCAACGUCA CGAUCGACC UGGUAUUGCG

(3) UCCAGGAAGC GUCCACCA–OH (188) [17]
(4) CAGAACCUCG AGGAACGGUG CACCA–OH (195) [18]

Branch point binding region Sm protein binding region

(5) Human Placenta U3 RNA (6) Rat Novikoff U3B RNA (Involved in rRNA production)

(5) \[ \text{m}_3^{2,2,7} \text{GpppAmAmGACUAUAC UUUCA} \]
\[ \text{GGAU CAUUUCUAUA UGUGUUACU} \]
(6) \[ \text{m}_3^{2,2,7} \text{GpppAmAm(m)AGCUUAAC UUUCA} \]
\[ \text{GGAU CAUUUCUAUA UGUGUUACU} \]

(5) AGAGAAGUUAU CUCUGAACGU GUAGAGCACC GAAACACAG AGGAAGAGAG
(6) AGAGAAGUUAU CUCUGAACGU GUAGAGCACC GAAACACAG AGGAAGAGAG

(5) GUAGCGUUUUU CUCCUG AAGCGGCG CCUGCUUGCCU UCGUUGGCUG
(6) AUAGCCUUCCCU CUGCCG UGAAGCCGGC UCUGAGGGCU GCUCGGUGCG

(5) CAACUGCGCU GACGCAUUUGA UACUGGCUUU UCUCCUCCUA UUGGGAGAG
(6) AGCGUGCCUU UGGCCUAAGA GAUGCUUU CUUCCUCCUG GGAGGGUAAG

(5) AGAGGGAGAG AACCGCGGCUU GAGGU–OH (217) [19,20]
(6) AGGAGGGAGAA CGAUCGUUGA GUAGGA–OH (215) [21]

Box A Box B Box C Box D

(7) Human HeLa U4A RNA (8) Rat Novikoff U4 RNA (Involved in pre-mRNA splicing)

(7) \[ \text{m}_3^{2,2,7} \text{GpppAmGmC'UUGmGmGC AGUGCGAUA UCGUGACCAAA UGAGGUUAAU} \]
(8) \[ \text{m}_3^{2,2,7} \text{GpppAmAC'UUGGCA GUGCGAUAUC GUAGCCAAUU AGGUAAAUCC} \]

(7) CCGAGGCGGG AUUAUUGCGAU AUUGAmAAACUU UYYUCCAAA'A CCCCAGCGUG
(8) GAGCGCGCGAU AUUGUCUAAU UGmAAACUUU UCCCAAYACC CCCCGCGGAC

(7) AGCAGUGCA AUAGUGCCGA CAUUGGCAAU UUUUGACAGU CUCUACCGAG
(8) GACUGUAmAAU AUAGUUGCGA UUGGCAAAUU UUGUUCUACG GAGACUG

(7) ACUGG–OH (144–145) [22,23]
(8) (138) [24]

Figure 11: Continued.
Figure 11: Sequences of major snRNAs (see [15–30]). The sequences of major snRNAs from human and rat involved in splicing and processing are aligned for comparison. The sequence elements in major spliceosomal snRNAs and processosomal snRNAs are highlighted in the sequences. Those are the pivotal motifs for the function. The numbers in parenthesis are the chain length of the RNAs.

A, showed enzyme-resistant oligonucleotide eluting close to the pentanucleotide region on a DEAE column (Figure 15). The 5′ oligonucleotide was analyzed by UV, [3H], and [32P] methods.

8.1.1. The UV Analysis. The 5′ oligonucleotides from U1 RNA, obtained by RNase A and RNase T1, were digested with snake venom phosphodiesterase and alkaline phosphatase. The nucleosides produced were separated on HPLC (high pressure liquid chromatography) [strongly basic cation exchange (quaternary amine)]. As shown in Figure 18, the amount of nucleoside ratio was 1.0, 1.2, 1.2, 0.7, and 0.9 for Am, A, Um, m32,2,7G, and C, respectively, for U1 5′ oligonucleotide.
8.1.2. The \([^3H]\) Method. The \([^3H]\)-labeled U1 RNA 5′ oligonucleotide, following digestion with snake venom phosphodiesterase and alkaline phosphatase, was separated by chromatographic methods with standards. Two-dimensional TLC (thin layer chromatography) and paper chromatography demonstrated that the \([^3H]\) labeled compound is a trimethylguanosine derivative (Figure 19).

8.1.3. \(^{32}P\)-Labeled 5′ Oligonucleotide from U1 RNA. The \(^{32}P\)-labeled RNA was digested with T\(_2\) and U\(_2\) RNase, and digestion products were separated by two-dimensional electrophoresis. The first dimension was on cellogel at pH 3.5, and the second dimension was on DEAE paper at pH 3.5 (Figure 20).

Spot “a” was eluted and treated with alkaline phosphatase and chromatographed with GMP, GDP, and GTP standards. The \(^{32}P\)-labeled 5′ oligonucleotide was chromatographed in the GTP region on a DEAE-Sephadex column (Figure 21).

The oligonucleotide peak from the GTP region was digested with snake venom phosphodiesterase and separated by electrophoresis in the first dimension followed by chromatography on second dimension (Figure 22).

The \(^{32}P\) activity ratio was 1.00, 1.11, 1.25, 0.53, and 1.14 for pm\(_{3,2,7}G\), pAm, pUm, pA, and Pi, respectively. The peak from the GTP region in Figure 21 digested with RNase P1 produced pUm, pA (peak a in Figure 23), and cap core m\(_{3,2,7}GpppAm\) (peak b in Figure 23). Table 10 shows the radioactivity distribution in peaks a and b in Figure 23.

For the analysis of a number of phosphates in cap core (peak b), the cap core was treated with NaIO\(_4\) and aniline to remove m\(_{3,2,7}G\) by \(\beta\)-elimination reaction (Figure 24).

The product was chromatographed on a DEAE column with standard AMP, ADP, and ATP. The product was eluted close to ATP, indicating that it is pppAm. This experiment proved that the 5′ oligonucleotide structure is m\(_{3,2,7}GpppAmpUmpApCp\).
8.2. U2 RNA 5′ End Oligonucleotide. The U2 RNA labeled with NaIO₄ and [³H]-KBH₃ was digested with RNase A. The labeled oligonucleotide eluted around the tetranucleotide region (Figure 15). The 5′ oligonucleotide was analyzed by UV, [³H], and [³²P] methods.

8.2.1 UV Analysis. The 5′ oligonucleotide obtained by complete RNase A digestion was analyzed for its base composition. The purified 5′ oligonucleotide was digested with snake venom phosphodiesterase followed by alkaline phosphatase. The digestion product (nucleosides) was separated by HPLC. The composition was Am, Um, C, and m₃²,⁷G in a ratio of 1.0, 1.3, 1.1, and 0.96, respectively (Figure 18) [12]. These nucleosides were also separated by two-dimensional TLC in a borate system. Um and Am migrated through the butanol-boric acid while the m₃²,⁷G and C, which form complexes with borate, were retarded in the butanol-boric acid phase (Figure 25).
8.2.2. $^3$H-Labeled U2 RNA 5′ Oligonucleotide. The $^{32}$P-labeled U2 RNA was digested with T1 RNase or RNase A. Half of each 5′ oligonucleotide was digested with alkaline phosphatase. Oligonucleotides were subsequently digested with snake venom phosphodiesterase, and the resulting 5′ nucleotides were separated first by electrophoresis and second by chromatography (Figure 27). The ratio of $^{32}$P counts is shown in Table 11.

From these data the 5′ end oligonucleotide from U2 RNA has been deduced to be m$_3^2,2,7$GpppAmpUmpCpGp.

8.3. U3 RNA 5′ End Oligonucleotide. The $^3$H-labeled U3 RNA was digested with RNase A and or T1 RNase. The $^3$H-labeled 5′ oligonucleotide obtained by RNase A digestion was eluted in the hexanucleotide region (Figure 15). The $^{32}$P-labeled U3 RNA digested with T2 and U2 RNA produced 2 spots that were separated by two-dimensional electrophoresis (Figure 29).

8.3.1. UV Analysis. The 5′ oligonucleotide obtained from U3 RNA by digestion with RNase A and T1 RNase was isolated by column chromatography. The purified 5′ oligonucleotide was digested with snake venom phosphodiesterase and alkaline phosphatase. The nucleosides obtained were subjected to HPLC. The molar ratios of m$_3^2,2,7$G, Am, A, and G were 1.0, 1.7, 1.1, and 1.0, respectively (Figure 18).

8.3.2. $^3$H Analysis. The intact U3 RNA, labeled with NaIO$_4$ and $^3$H-KBH$_4$ methods, was digested with RNase A and

Table 10: The $^{32}$Pcpm radioactivity distribution in nucleotides and cap core produced by RNase P1 digestion of U1 5′ oligonucleotide [16]. The $^{32}$P-labeled U1 5′ oligonucleotide obtained by digestion of U1 RNA with RNase T2, RNase U2, and alkaline phosphatase was treated with P1 nuclease which cleaves all phosphodiester bonds but not pyrophosphate bonds. The products were separated on a DEAE column (Figure 23). The radioactivity in peak a (mononucleotides pUm, pA) and peak b (cap core m$_3^2,2,7$GpppAm) were determined by Packard liquid scintillation spectrometer.

<table>
<thead>
<tr>
<th>Peak a (pUm, pA)</th>
<th>Peak b (m$_3^2,2,7$GpppAm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>U1 5′ cap</td>
<td>65,280</td>
</tr>
<tr>
<td></td>
<td>87,600</td>
</tr>
</tbody>
</table>

The UV spectra of pm$_3^2,2,7$G were typical of a trimethyl G nucleotide (Figure 8). The mass spectrometry of the unknown nucleoside from U2 RNA 5′ fragment was identified as m$_3^2,2,7$ trimethylguanosine (Figure 9).

8.2.3. $^{32}$P-Labeled U2 RNA 5′ Oligonucleotide. The $^{32}$P-labeled U2 RNA was digested with T1 RNase or RNase A. Half of each 5′ oligonucleotide was digested with alkaline phosphatase. Oligonucleotides were subsequently digested with snake venom phosphodiesterase, and the resulting 5′ nucleotides were separated first by electrophoresis and second by chromatography (Figure 27). The ratio of $^{32}$P counts is shown in Table 11.

The U2 RNA 5′ oligonucleotide obtained by RNase A was subjected to digestion with pyrophosphatase (Crotalus adamanteus venom type II, Sigma). The remaining oligonucleotide did not have m$_3^2,2,7$G, indicating that the m$_3^2,2,7$G is linked by pyrophosphate linkage (Figure 28).

From these data the 5′ end oligonucleotide from U2 RNA has been deduced to be m$_3^2,2,7$GpppAmpUmpCpGp.
Trimethyl guanosine caps of Novikoff hepatoma cell snRNAs cleavage sites by various enzymes

**CAP structure of U1 RNA**
\[ m_3^{2,2,7}GpppAmpUmmpApCp \]

**CAP structure of U2 RNA**
\[ m_3^{2,2,7}GpppAmpUmmpApCp \]

**CAP structure of U3 RNA**
\[ m_3^{2,2,7}GpppAmpA(m)pApGpCp \]

**CAP structure of 5S RNA III (U5 RNA)**
\[ m_3^{2,2,7}GpppAmpUmmpApCp \]

**Nucleases**
- T1 RNase
- Pl nuclease
- Snake venom phosphodiesterase
- T2 RNase
- U2 RNase
- RNase A
- Bacterial alkaline phosphatase

**Figure 17:** Characterization of \( m_3^{2,2,7}G \) caps of snRNAs U1, U2, U3, and U5. The enzyme susceptible bonds are indicated with arrows. The split arrows indicate that some bonds without 2'-O-methylation can be cleaved but the ones with 2'-O-methylated ribose are not.

**Table 11:** The analysis of \([^{32}P]\)-labeled U2 RNA 5' oligonucleotide [12]. The 5' ends obtained from uniformly \([^{32}P]\)-labeled U2 RNA were digested with T1 RNase or RNase A and isolated by two-dimensional electrophoresis (cellulose acetate at pH 3.5 followed by DEAE paper electrophoresis). The 5' oligonucleotides were digested with snake venom phosphodiesterase before and after removal of 3' phosphate with bacterial alkaline phosphatase. The products were separated as in Figure 27. The radioactivity ratios are listed.

<table>
<thead>
<tr>
<th>5' fragment from U2 RNA</th>
<th>Pi</th>
<th>pUm</th>
<th>pAm</th>
<th>pm3^{2,2,7}G</th>
<th>pm3^{2,2,7}G</th>
<th>pC</th>
<th>pG</th>
<th>pGp</th>
</tr>
</thead>
<tbody>
<tr>
<td>5' oligo from T1 digestion</td>
<td>1.58</td>
<td>1.33</td>
<td>0.91</td>
<td>0.90</td>
<td>1</td>
<td>1.48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alkaline phosphatase digested T1, 5' oligo</td>
<td>1.22</td>
<td>1.43</td>
<td>0.98</td>
<td>0.90</td>
<td>1</td>
<td>0.93</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5' oligo from RNase A digestion</td>
<td>1.74</td>
<td>1.25</td>
<td>1</td>
<td>1.08</td>
<td>1</td>
<td>1.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alkaline phosphatase digested A, 5' oligo</td>
<td>0.93</td>
<td>1.45</td>
<td>0.94</td>
<td>0.77</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From these data, obtained by UV, \([^{3}H]\), and \([^{32}P]\) experiments, the U3 RNA 5' oligonucleotide sequence has been deduced to be \( m_3^{2,2,7}GpppAmpA(m)pApGpCp \).

8.4. 5S RNA III (U5 RNA) 5' End Oligonucleotide. The oligonucleotide sequence was deduced as in the case of U1 RNA. The structure is identical to the U1 5' oligonucleotide \( m_3^{2,2,7}GpppUmpApCp \).
Table 12: Tgs1 interacting proteins [70–72]. The Tgs1 (trimethylguanosine synthase 1) interacting proteins are listed. It is interacting with proteins involved in many aspects of RNA metabolism such as transcription, spliceosome assembly, maturation, and modification.

<table>
<thead>
<tr>
<th>Tgs interacting proteins [70]</th>
<th>Spliceosome assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcription apparatus (regulators of RNA polymerase II transcription)</td>
<td>(1) Mud1 (yeast homolog of U1A)</td>
</tr>
<tr>
<td>(1) Rpn4 (TF; Proteasome subunits and U2 snRNA gene)</td>
<td>(2) Mud2 (yeast homolog of U2AF65)</td>
</tr>
<tr>
<td>(2) Spt3 (SAGA complex)</td>
<td>(3) Nam8 (Mud15) (homolog of TIA-1)</td>
</tr>
<tr>
<td>(3) Srb2 (mediator complex)</td>
<td>(4) Brr1 (snRNP)</td>
</tr>
<tr>
<td>(4) Soh1 (Med31; mediator complex)</td>
<td>(5) Lea1 (U2A; U2snRNP)</td>
</tr>
<tr>
<td>(5) Swr1 (ATPase; binds to 5' end of yeast transcription units)</td>
<td>(6) Ist3 (Snu17; U2 snRNP)</td>
</tr>
<tr>
<td>(6) Htz1 (H2AZ) (binds to 5' end of yeast transcription units)</td>
<td>(7) Isy1 (Ntc30; interacts with Prp19)</td>
</tr>
<tr>
<td>(7) CBP (binds to human PIMT/Tgs1)</td>
<td>(8) Cwc21 (component of Cef1 complex)</td>
</tr>
<tr>
<td>(8) P300 (binds to PIMT/Tgs1)</td>
<td>(9) Bud13 (Cwc26) (Cef1 complex)</td>
</tr>
<tr>
<td>(9) PRIP (binds to PIMT/Tgs1; [71])</td>
<td>(10) SMN (HeLa cell) [72]</td>
</tr>
<tr>
<td>RNA end processing and decay</td>
<td>RNA modifying factors</td>
</tr>
<tr>
<td>(1) Trf4 (Pap2; catalytic subunit of TRAMP)</td>
<td>(1) SmB (snRNP); interacts with YNL187/Swt21</td>
</tr>
<tr>
<td>(2) Lsm1 (decapping complex)</td>
<td>(2) SmDL (snRNP)</td>
</tr>
<tr>
<td>(3) Pat1 (decapping complex)</td>
<td>(3) Cbf5 (snoRNP)</td>
</tr>
<tr>
<td>(4) Prp (binds to PIMT/Tgs1)</td>
<td>(4) Nop58 (snoRNP)</td>
</tr>
<tr>
<td>(5) PRIP (binds to PIMT/Tgs1; [71])</td>
<td>(5) Mrm (RNA 2'-O-methyltransferase)</td>
</tr>
</tbody>
</table>

9. RNA Signature Modifications for Different RNA Classes

9.1. End Modifications

9.1.1. 5' End

(a) According to Chemical Nature of Caps

5' Trimethylguanosine cap for the snRNA,
5' 7-methylguanosine cap for the mRNA,
5' 2,7 dimethylguanosine cap of virus and nematode RNAs
5' mpppG of U6 RNA.

(b) According to Flanking Nucleotide Modification of Caps.
(See Table 13).

(c) 5' End Uncapped RNA. (pppNp) for primary transcripts such as 4.5S RNA I, 5S RNA, and Alu RNA. (pNp) 5' end for processed RNAs such as Alu RNA, 5S RNA, tRNA, YRNA.

9.1.2. 3' End

3' 2'-O-methylated; 4.5S RNA III
3' poly-A; mRNA, IncRNA
3' poly-U; polymerase III transcripts such as 4.5S RNA I, 5S RNA, and others
3' CCA; tRNA, U2 RNA.

9.2. Internal Modifications. The most colorful modifications are in tRNAs that contain methyl, formyl, acetyl, isopentyl, threonyl, carbamoyl, and other groups and modifications by pseudouridylation, deamination, reduction, or thiolation. Focusing on recent findings for snoRNA, m32,2,7G capping reactions are very interesting because trimethylguanosine is found only in noncoding RNA cap structures, although some nematode mRNA species also contain m32,2,7G caps. The snoRNAs are less abundant (10^5 copies) than ribosomal RNA or tRNA (10^6 copies). Isolating large amount of RNA can be a hurdle to overcome. Massive preparative procedures and syntheses were pivotal for the thorough analysis of these modifications. The 2' -O-modifications occur mostly internally, and 3' Um was also found in 4.5S RNA III. The RNA ribose with 2'-O-methylation confers resistance to enzymatic digestion by such enzymes as RNase A, RNase T1, RNase U2, and RNase T2. They are also resistant to alkaline hydrolysis, and the alkaline hydrolysates can be separated into di-, tri-, and tetr nucl eto ide s by column chromatography and then by two-dimensional paper chromatography (Figure 16). Other enzymes which can cleave 2'-O-methylated nucleotides are snake venom phosphodiesterase, P1 nuclease, and spleen phosphodiesterase. These are valuable tools for sequencing.

10. Presence of m32,2,7G Caps in RNA Species

10.1. Nucleolar RNA. Initially, the m32,2,7G cap containing snoRNA was found in U3 RNA [36]. Since then C/D snoRNA and H/ACA snoRNA have been discovered exponentially. The snoRNAs are transcribed from monocistronic as well
as polycistronic independent positions as well as intronic regions of mRNA, especially the genes coding ribosomal proteins. In vertebrates, there have been >76 snoRNAs that have been reported, but only U3, U8, and U13 snoRNAs have been reported to have m$_2^{2,7}$G caps [33, 88]. In yeast, there are at least 17 m$_2^{2,7}$G cap containing snoRNAs out of more than 76 snoRNAs. It was also reported that some snoRNA precursors, such as pre-snoRNAs 50, 64, and 69, have the m$_2^{2,7}$G cap, but mature snoRNA 50, 64, and 69 do not have m$_2^{2,7}$G caps. The maturation process cleaves the 5' fragment by Rnt1 (RNase III like enzyme), and trimming is performed by 5' → 3' exonuclease Xrn1 and Rat 1 [89].

10.2. Spliceosomal snRNAs. These include U1, U2, U4, U5, and U6 snRNAs. All of these except U6 contain the m$_3^{2,7}$G cap, and U6 has the mpppG cap instead. They are present in complexes as RNP with proteins specific for each RNA as well as some common snRNP proteins such as the Sm proteins. Functionally, U1 RNP acts at 5' splice sites and U2 RNA at branch sites including 3' splice sites. U4, U5, and U6 snRNAs enter the spliceosomal intermediate as a tri-snRNP complex.

10.3. Human Telomerase RNA (hTR). Human telomerase RNA has a structure containing the H/ACA motif with 8 conserved regions (CR 1–8) [92]. The CR7 contains the CAB...
Table 13: Cap variations of flanking nucleotide modifications.

<table>
<thead>
<tr>
<th></th>
<th>Trimethylguanosine cap</th>
<th>7-Methylguanosine cap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 0</td>
<td>( m_3^{2,2,7}GpppN^- )</td>
<td>( m_7^7GpppN^- )</td>
</tr>
<tr>
<td>Type I</td>
<td>( m_3^{2,2,7}GpppNnM^- )</td>
<td>( m_7^7GpppNnM^- )</td>
</tr>
<tr>
<td>Type II</td>
<td>( m_3^{2,2,7}GpppNmNnM^- )</td>
<td>( m_7^7GpppNmNnM^- )</td>
</tr>
<tr>
<td>Type III</td>
<td>( m_3^{2,2,7}GpppNmNmNm^- )</td>
<td>( m_7^7GpppNmNmNm^- )</td>
</tr>
<tr>
<td>Type IV</td>
<td>( m_3^{2,2,7}GpppNmNmNmNm^- )</td>
<td>( m_7^7GpppNmNmNmNm^- )</td>
</tr>
</tbody>
</table>

![Figure 19: Identification of \( m_3^{2,2,7}G^- \) [16].](image)

(a) Fluorogram of two-dimensional thin layer chromatography of \([^3H]\)-labeled nucleoside trialcohol derivative (N') released from U1 RNA 5' fragment. It was identified as \( m_3^{2,2,7}G^- \) with standard. (b) Fluorogram of chromatographic separation of \([^3H]\)-labeled nucleoside trialcohol derivative on Whatman 3MM paper. The \([^3H]\)-labeled compound was identified as \( m_3^{2,2,7}G^- \) with standard.

Figure 20: 2D map of U1 RNA digested with \( T_2 \) RNase and \( U_2 \) RNase [16]. The U1 RNA uniformly labeled with \([^{32}P]\) was digested with \( T_2 \) RNase and \( U_2 \) RNase. The resistant 5' fragment (spot "a") was separated from the rest of the hydrolysate by two-dimensional electrophoresis. The first dimension was on cellogel at pH 3.5, and the second dimension was on DEAE paper in 5% acetic acid-NH$_4$ acetate at pH 3.5.

10.4. *C. elegans* SL RNA. *C. elegans* has mRNA with the \( m^7G \) cap as well as \( m_3^{2,2,7}G \) cap, and the expression is regulated differentially. The genes for protein coding are monocistronic as well as polycistronic, and introns are much smaller than observed in mammalian cells. The polycistronic genes contain 2–8 operonic genes regulated by the same promoters. Some gene products are not processed, and others are spliced by cis-splicing as well as transsplicing. The transsplicings are carried out by SL RNA 1 or SL RNA 2. The approximately 110 SL RNA 1 genes are in tandem in chromosome V. The SL RNA 2 is derived from SL RNA 1 and there are \(~18\) dispersed genes with a variety of variant SL2 RNAs (some are called SL3, SL4, etc.). They are all 100–110 nucleotide long and contain \( m_3^{2,2,7}G \) caps and Sm protein binding sites. These pre-mRNAs, containing 5' outron (monocistronic and 5' first gene in polycistronic operonic genes), are transspliced by SL RNA 1 and internal operonic pre-mRNAs are mostly transspliced by SL RNA 2 and these genes have typically U-rich sequence containing \(~100\) bp spacers between two cleavage sites. The internal
mRNA gene of polycistronic operonic genes, lacking a spacer, is transspliced always by SL RNA I [95, 96]. The transspliced mRNA contains a m$^{3,2,7}$G cap containing 22 nucleotides of SL RNA at their 5' ends. The SL RNA (splice leader RNA) has a m$^{3,2,7}$G cap and Sm protein binding sites. The nematode C. elegans has 5 eIF4E isoforms of cap binding proteins. They are IFE-1 (m$^7$G cap and m$^{3,2,7}$G cap binding), IFE-2 (m$^7$G cap binding, but competed by the m$^{3,2,7}$G cap), IFE-3 (m$^7$G cap binding only), IFE-4 (m$^7$G cap binding only), and IFE-5 (m$^7$G cap and m$^{3,2,7}$G cap binding). The homolog amino acids W36 and W102 stacking the m$^7$G cap in mice elf4E are W51 and W97 in IFE-3 and W28 and W74 in IFE-5 (Figure 32).

The differences in 3-4 loop configuration between IFE-5 and IFE-3 are N64Y/V65L. The changes in IFE-5 amino acids isoleucine 64 to tyrosine and valine 65 to leucine change binding properties more to m$^7$G cap binding than to m$^{3,2,7}$G cap binding. IFE-5 has 4 cysteines, and its conformation is governed by disulfide bond formation. It is suggested that the cap binding cavity is altered to produce a smaller cavity that discriminates against the m$^{3,2,7}$G cap binding [85]. These may provide translational regulation of m$^7$G cap mRNA and transspliced m$^{3,2,7}$G cap mRNA in C. elegans.

11. Synthesis of m$^{3,2,7}$G Cap

Modifications. Trimethyl G caps are present in snRNAs involved in splicing and also in snoRNA involved in rRNA processing and modifications such as Ψ formation (H/ACA snoRNA) or 2'-O-methylation (C/D snoRNA). These include U1, U2, U4, and U5 spliceosomal RNAs, and U3, U8, and U13 nucleolar RNAs. Recently, telomerase RNA (S. cerevisiae TLC1) has also been reported to have
Figure 24: The β-elimination of cap core [16]. The cap core (peak b from Figure 23) was treated with NaIO₄ and aniline to remove m₃,2,7G by β-elimination reaction. The remaining nucleotide was chromatographed in the ATP region indicating it is pppAm. This proved that the cap core was m₃,2,7GpppAm.
Figure 25: Two-dimensional chromatography in borate system [12]. The nucleoside mixture from U2 RNA 5′ oligonucleotide (RNase A product) was obtained by digestion with snake venom phosphodiesterase and alkaline phosphatase. The resulting nucleosides were separated with the borate system. Um and Am migrated through the butanol-boric acid while the m$^{2,2,7}$G and C, which form borate complexes, were retained in the butanol-boric acid phase.

a trimethylguanosine cap structure. The trimethyl-G caps are formed on cap 0 or cap I of m$^7$G caps of pre-mRNAs by dimethylation of N2 position by trimethylguanosine synthase (Tgs1). The Tgs1 has been found to be in the Cajal body and cytoplasm. The U3 snRNA is hypermethylated in the Cajal body, and U1, U2, U4, and U5 snRNA have been reported to be hypermethylated in the cytoplasm.

11.1. The m$^7$G Cap Formation. The RNA polymerase initiates the RNA transcription with 5′ triphosphate nucleotides and in a majority with purine nucleotides of ATP or GTP. The capping reaction in a polymerase II system occurs cotranscriptionally within the nascent transcript of ~30–50 nucleotides. The guanylyltransferase is attached to heptad (YSPTPS) repeats of CTD of RNA polymerase II. It was reported with cloned mouse guanylyltransferase and synthetic heptad repeats that the serine 5 phosphorylated 6 heptad repeats stimulates guanylyltransferase activity 4-fold. Serine 2 phosphorylation also binds the guanylyltransferase but did not stimulate enzyme activity [97]. The capping enzymes contain RNA triphosphatase and RNA guanylyltransferase in the same molecule, but methylating enzymes are in different protein and occurs in separate steps.

The enzymes involved are RNA triphosphatase and RNA guanylyltransferase, which can be found in the same enzyme, catalyze removal of one phosphate from pppNp initiation nucleotide, and transfer GMP from GTP through intermediary GMP-lysine phosphamide enzyme complex. The RNA guanyl 7 methyltransferase methylates the guanine at N7 position. The RNA 2′-O-methyltransferase methylates penultimate nucleotide 2′ OH, producing the cap 1 structure. In rat liver, it has been reported that 2′-O-methylation may precede the guanosine N7 methylation [98].

The capping reactions by mammalian and shrimp capping complexes (HeLa cell, rat liver, calf thymus, and shrimp) [98] have been reported as below:

**RNA Triphosphatase and Guanylyltransferase.** The monomer of the 69–73 kDa protein has functions of RNA triphosphatase and RNA guanylyltransferase activity.

\[
\begin{align*}
\text{pppNpNpNpNp} - & \xrightarrow{\text{RNA triphosphatase}} \text{ppNpNpNpNp} - \\
\text{GTP + RNA guanylyltransferase} & \xrightarrow{\text{Mg}^{2+}} \text{GMP-(phosphamide)-E} + \text{ppi} \\
\text{GMP-E + ppNpNpNpNp} - & \xrightarrow{\text{Mg}^{2+}} \text{GpppNpNpNpNp- + RNA guanylyltransferase} \\
\text{GpppNpNpNpNp- + AdoMet} & \xrightarrow{\text{RNA 2′-O-methyltransferase}} \text{GpppNmpNpNpNp- + AdoHcy} \\
\text{GpppNmpNpNpNp- + AdoMet} & \xrightarrow{\text{RNA guanyl 7-methyltransferase}} \text{m}^7\text{GpppNmpNpNpNp + AdoHcy}
\end{align*}
\]
Some of the capping enzymes (vesicular stomatitis virus, spring viremia of carp virus) use the substrate monophosphorylated 5′ end (pNpNpNpNp-) [99, 100], and 7-methylation occurs after the 2′-O-methylation has taken place.

Trimethylguanosine cap synthesis is carried out by multiple steps involving From HeLa cells, two enzymes forming cap I from cap 0 and cap II from cap I have been purified and characterized [101].

11.2. Cap I Methyltransferase. This enzyme is present in both the nucleus (29.3 units/mg) and cytoplasm (3.74 units/mg) and cap II methyltransferase is exclusively in the cytoplasm (4.62 units/mg). Cap I methyltransferase uses GpppA(pA)n, m7GpppA(pA)n, m7GpppApGp, m7GpppApGpUp, and RNA with type 0 cap as substrates but not m7GpppA or GpppA. The substrate required for cap I formation should be at least a trinucleotide.

The order of 7-methylation of ultimate G nucleotide and 2′-O-methylation of penultimate nucleotide is uncertain, and both pathways may occur.

11.3. Cap II Methyltransferase. This enzyme is present only in the cytoplasm and converts cap I to cap II. The mature mRNA with 5′ m7G cap and 3′ polyadenylation is then transported into the cytoplasm as a complex with CBC20/80, PHAX, and Crm1-RanGTP. The m7G cap binds to CBC20 (156 amino acids) in complex with CBC80 (790 amino acids). The crystal structure of the CBC20/80 complex in association with m7G cap has been reported [86, 87]. The CBC20 is in an unfolded form in the absence of CBC80. The CBC80 has 3 domains, each containing consecutive 5-6 helical hairpins resembling the MIF4G (middle domain of eIF4G). The CBC20 has a typical RRM motif and binds between domains 2 and 3 of CBC80. The m7G cap is sandwiched between Tyr 43 and Tyr 20. And Phe 83, Phe 85, and Asp 116 have essential role for m7G cap binding. Asp 116 and Trp 115 interact with the N2 amino group and confer specificity of the m7G cap for other structures (Figure 33).

In the cytoplasm, the m7G cap plays a role in the initiation of translation by binding to eIF4E which complexes with eIF4A and eIF4G. The exact mechanism of exchange is not known but CBC80 has binding capacity for PHAX or eIF4G and dissociation of CBC80 from CBC20 makes CBC20 become disordered [86, 87].

11.4. Maturation of snRNAs. The snRNAs synthesized by RNA polymerase II with m7G cap structures are transported into the cytoplasm in complex with CBP20/80, PHAX (phosphorylated adaptor for RNA export), the CRM1 (export receptor, chromosome region maintenance 1) or exportin 1 and RanGTP (Ras-related nuclear antigen). The snRNPs in the cytoplasm are trimethylated and processed. The mature RNA is reimported into the nucleus in a complex with the trimethyl G cap-specific binding protein snurportin 1 and snRNA binding proteins of Sm RNP and SMN proteins.

Despite immunofluorescent staining of U1 and U2 RNA exclusively in the nucleus [102], biochemical analyses have
Figure 27: Autoradiograph of nucleotides from $^{[32]P}$-labeled U2 RNA 5' fragment [12]. The $^{[32]P}$-labeled U2 RNA 5' fragment (T1 RNase digestion) was treated first with alkaline phosphatase and then with snake venom phosphodiesterase. This mixture of mononucleotide products was separated by electrophoresis followed by chromatography. Approximately equal amounts of pm$_2,2,7$G, pAm, pUm, pC, and pG were observed (Table 11).

Figure 28: The susceptibility of 5' cap to pyrophosphatase [12]. The $^{[32]P}$-labeled 5' oligonucleotide obtained from U2 RNA by RNase A was digested with pyrophosphatase and base composition was analyzed by snake venom phosphodiesterase digestion. This digestion released m$_2,2,7$G from the 5' fragment indicating that m$_2,2,7$G is linked by a pyrophosphate linkage.

Figure 29: 2D map of U3 RNA digest [58]. The $^{[32]P}$-labeled U3 RNA was digested with T$_2$ RNase and U$_2$ RNase. It produced two 5' fragments "11A" and "11B".

demonstrated that trimethylation and maturation of some snRNA takes place in the cytoplasm.

The U1 snRNA [103] and U2 snRNA [104] have been shown to be hypermethylated in the cytoplasm in a Sm protein binding dependent manner. The Xenopus laevis U1 RNA, with the m$^7$G cap, has been shown to be hypermethylated in HeLa cell cytoplasmic extracts and Sm binding site in U1 RNA is required [103]. The Tgs1 has been shown to bind to Sm proteins of Sm B and Sm D. The Xenopus laevis U2 RNA with m$^7$G cap has been shown to be hypermethylated into the m$_2,2,7$G cap structure in enucleated xenopus oocytes [104]. In yeast and human HeLa cells, the Tgs1 for U3 RNA is localized in the nucleolar body of the nucleolus and Cajal bodies, respectively [105]. In the absence of Tgs1 or inactive Tgs1 in yeast, m$^7$G capped unprocessed U1 RNA is retained in the nucleolus and splicing becomes cold temperature sensitive. The same enzyme is responsible for the U3 nucleolar RNA hypermethylation [106].

The consensus between yeast and human cells is the presence of a nucleolar body in yeast and Cajal body in HeLa cell. The hypermethylation and processing during maturation take place in the nucleolar body in yeast and Cajal body in HeLa cells [105, 106]. The sequence element “UGAG” (also found in the U3 RNA B box) has been reported as a CAB box (Cajal-body-specific localization signal). U3 RNA trimethylation is somewhat different from other snRNAs. The U3 RNA, which does not have Sm protein binding sites, has been shown to require an intact 3' terminal stem structure for trimethylguanosine cap formation [107].

In HeLa cells, transfected U3 RNA gene products are trimethylated and mature U3 RNA is localized in the nucleolus. Immature U3 RNA, with both m$^7$G and 3' extension of 10–15 nucleotides, is detected in Cajal bodies [105, 106, 108].

12. The Tgs1 (Trimethylguanosine Synthase 1)

12.1. Human Tgs1. The Tgs1, trimethylguanosine synthase in human, protein is 110 kDa and 852 amino acids in chain.
The gene is located in chromosome 8q11. The mRNA is 3.2 kb in length and produces a 110 kDa protein and ~65–70 kDa protein that is proteasome processed. The long form is in the cytoplasm, and the short isoform has been reported to be localized in the Cajal body within the nucleus. The Tgs1 has S-AdoMet methyltransferase signature motifs of X, I, II (include post 1 motif), III, IV, V, and VI [70, 106, 109, 110].

The human Tgs1 motifs are the following.

motif X is (a.a.665)DREGWESVTPEKIAHIF/A(a.a.682),
motif I is (a.a.693)VVVDAFCGVGGN(a.a.704),
motif II is (a.a.714)RVIALIDPV/IKI(a.a.725) and post 1 motif is VIAID which is responsible for S-AdoMet binding to the enzyme,
motif III is (a.a.740)KIEFICGDFLLAS(a.a.753),
motif IV is (a.a.758/759)VVFLSPPWGGPDYA(a.a.771/772),
motif V is (a.a.785/786)DGFEIFRLSK(a.a.794/795),
motif VI is (a.a.798/799)NNIVYFIPRNASD(a.a.810/811),

It was reported that trimethylation catalytic activity is located in the C-terminal region (amino acids 631–852) and this region contains the S-AdoMet-dependent methyltransferase motifs. The tryptophan in motif 4 is involved in π stacking with m7G guanosine of the substrate. The motif I and post 1 motif are reported to interact with S-AdoMet. [110]. The C-terminal domain is localized in the Cajal body and binds to C/D-snoRNA- and H/ACA-snoRNA-associated proteins such as fibrillarin, Nop56, as well as dyskerin [110].

The N-terminal portion of the molecule (amino acids 1–477) has been reported to contain GXGXGXXI, a K-homology domain for RNA binding, and a motif for SmB and SmD1 binding. The Tgs1 has also been shown to interact with PRIP (proliferator-activated receptor-interacting protein), and the N-terminal portion (amino acids 1–384) of Tgs1 has been shown to have stimulatory effects on transcription of PPARγ and RXRα [109, 110].

The human Tgs1 (618–853) has been crystallized for structural analysis. The one monomer consists of 11 α-helices and 7 β-strands. It is composed of 2 domains, the core domain (Glu675-Asp844) and N-terminal extension (Leu34-Ser671) connected by 3 amino acids—Val672, Thr673, and Ser674. The core domain consists of 7-β-strands in topology of B61β71β81β41β11β21β31 with a classical class 1 methyltransferase fold resembling the Rossmann-fold AdoMet-dependent methyltransferase superfamily [90]. The N-terminal α-helices form a separate small globular subdomain involved in recognition and binding of both substrates. The residues Glu667 and Phe670 in motif X as well as Pro765, Trp 766, and Pro769 in motif IV are in proximity permitting the top of their binding clefts to be close together. Tryptophan 766 and m7G are stacked in a coplanar manner with a 3.2Å distance providing a tight π-π interaction between them (Figure 34).

Figure 30: m$_3^{2,2,7}$G’ identification from U3 RNA 5’ end fragment [58]. The [3H]-labeled U3 RNA 5’ fragment obtained by RNase T$_1$ and RNase A was digested with snake venom phosphodiesterase and alkaline phosphatase. The nucleoside mixture was separated by two-dimensional thin layer chromatography with standard nucleoside mixture in (a) and with the trialcohol derivative of m$_3^{2,2,7}$G’ in (b). The released nucleoside trialcohol derivative was identified as m$_3^{2,2,7}$G’ by fluorography.

The catalytic mechanism of methylation is by an S$_n$2 substitution reaction. The N2 of m7G does the nucleophilic attack on an activated methyl group of the AdoMet (Figure 35).

Dimethylation is not processive. After formation of m$_2^{2,7}$G both products (m$_3^{2,2,7}$G and AdoHcy) dissociate from the enzyme. Tgs1 can use m$_2^{2,7}$G as a substrate, and newly bound AdoMet can methylate at the same position by the same mechanism to form the m$_3^{2,2,7}$G cap structure.
and m^7GppA as substrates without preassembly of snRNP containing Sm proteins. The Tgsl of S pombe is 239-amino-acid long and m^7G is the pre-requisite for this reaction [112].

12.4. The G. lamblia Tgsl and Tgs2. The lamblia has 2 enzymes, Tgsl and Tgs2. Tgsl is not a processive enzyme but distributive and produces m^5,2,7G in excess of AdoMet and enzyme. However, Tgs2 produces only m^2,7G, and some G. lamblia RNAs contain dimethylG caps. The G lamblia Tgsl has 300 amino acids and Tgs2 is 258 amino acids long. They all have landmark motifs for Ado-Met-dependent methyltransferase motifs [113].

13. Parasite Capping Enzyme (Trypanosoma brucei)

The parasite Trypanosoma brucei SL RNA (splicing leader) has the biggest 5’ oligonucleotide, type IV, of m^7Gppp6,6AmAmpAmCmpm3UmpAp [114, 115]. Enzymes involved in the synthesis of this cap structure are TbCgm1, TbCet1, TbMTr1(cap1 2’OMTase), TbMTr2/TbCm1/TbMTr48(cap2 2’OMTase), TbMTr3/TbMTr57(cap3 2’OMTase). However, m^6,6A and m^3Um methylating enzymes have not been identified as yet [115].

13.1. TbCgm1 (T. brucei Cap Guanylyltransferase Methyltransferase 1). There exist two enzyme systems for 5’ cap formation. The first is the system composed of separate independent enzymes which are TbCet1 (Trypanosoma brucei triphosphatase, 253 amino acids), TbCe1 (Trypanosoma brucei guanylyltransferase, 586 amino acids), and TbCmt1 (Trypanosoma brucei m^7G Cap methyltransferase 1, 324 amino acids). The second is a set of fused enzymes possessing dual activities. It is TbCgm1 (Trypanosoma brucei cap guanylyltransferase and methyltransferase 1) that has 1050 amino acids [116] with dual activities of guanylyltransferase and guanine N-7 methyltransferase [117]. The TbCe1 guanylyltransferase has 250 amino acids at its N-terminal region which is not found in fungal or metazoan guanylyltransferase and has homology with the phosphate binding loop found in ATP- and GTP-binding proteins [118]. Silencing TbcCe1 and TbcCmt1 had no effect on parasite growth or SL RNA capping, but TbCgm1 was essential for parasite growth and silencing TbCgm1 increased the amount of uncapped SL RNA. The protein TbCgm1 has guanylyltransferase activity in N-terminal 1–567 amino acids and methyltransferase activity in C-terminal 717–1050 amino acids. The N-terminal guanylyltransferase portion contains 6 colinear guanylyltransferase motifs: I(KADGTR), II([FVDAELM], IIa(LGCFDVFYRI), IV(DGFI), V(QLXWKPSMLVD), and VI(WSIERLNRDK). The C-terminal methyltransferase portion contains regions homologous to m^7G methyltransferase from T. cruzi and L. major [117].

13.2. Cap Methylating Enzymes: TbMTr1, TbMTr2(TbCm1/TbMTr48), and TbMTr3(TbMTr57). They contain a K^95-D^207,K^248-E^285 tetrad critical for AdoMet-dependent

12.2. Drosophila Tgs1. In Drosophila melanogaster, DTL (Drosophila TAT-like) has been reported to exhibit trimethylguanosine cap formation activity for both U2 and U4 snRNAs. The mRNA for the protein Tgs1 is polycistronic and 2,600-nucleotide long with upper and downstream U4 snRNAs. The mRNA for the protein Tgs1 is polycistronic and 2,600-nucleotide long with upper and downstream U4 snRNAs. In yeast S. cerevisiae, Tgs1 is in the nucleolus and U3 RNA is also in the nucleolus. In the absence of Tgs1, the pre-U3 RNA was found within the nucleolar body and U1 RNA was retained in the nucleolus. S. cerevisiae. S. pombe, and Giardia lamblia Tgs1 can methylate m^7GTP, m^5GDP, and m^7GppA as substrates without preassembly of snRNP containing Sm proteins. The Tgs1 of S pombe is 239-amino-acid long and m^7G is the pre-requisite for this reaction [112].

12.3. Yeast Tgs1. In yeast S. cerevisiae, Tgs1 is in the nucleolus and U3 RNA is also in the nucleolus. In the absence of Tgs1, the pre-U3 RNA was found within the nucleolar body and U1 RNA was retained in the nucleolus. S. cerevisiae. S. pombe, and Giardia lamblia Tgs1 can methylate m^7GTP, m^5GDP,
methyltransferase and can convert cap type 0 of *Trypanosoma* SL RNA and U1 snRNA into type 1 cap [115]. The KDKE mediates S2 type transfer of methyl groups that involve 2′-OH deprotonation. The U1 snRNA 2′-O-methylation takes place before Sm protein binding to the RNA and it is prerequisite for the dimethylation at the N2 position to make \( m_7^3A^2pGpppAm \) cap structures. Other \( m_7^3A^2pGpppAm \) cap-containing snRNAs such as U2, U-snRNA B (U3 snRNA homolog), and U4 snRNAs were reported to be synthesized by RNA polymerase III in *Trypanosomes*.

The TbMTr2 and TbMTr3 are responsible for second and third nucleotides 2′-O-methylations. The enzymes that perform \( m_7^3A^6A \), \( m_7^3U \) base methylations, and fourth nucleotide 2′-O-methylating enzymes are not known yet.
Figure 34: Crystal structure of hTgs1 [90]. The crystal structure of hTgs1 with substrate m^7GTP and AdoHcy (at the site for AdoMet). The m^7G is stacked between tryptophan 766 and serine 671 [90]. (a) Relative orientation of substrates m^7GTP and AdoHcy at binding pockets. (b) Detailed view of the binding pocket for m^7GTP (shown W766 and S671). (c) Detailed view of the binding pocket for AdoHcy.

Figure 35: The active site of hTgs1 [90]. Proposed mechanism of methyltransferase activity of hTgs1 is shown. The AdoMet methyl group is in close proximity to N2 position of m^7GTP. The prerequisite as a substrate for hTgs1 is m^7G moiety.

14. Transport of Mature RNAs

The snurportin1 is a specific trimethyl G cap binding protein with an importin β binding site at its N-terminus (amino acids 1–65) and trimethyl G cap binding site at amino acids 95–300 forming a cap binding pocket. This protein has more resemblance to mRNA guanylyltransferase. The snurportin 1 binds the trimethyl G cap forming π-stacking with tryptophan 276 and the penultimate purine nucleotide G (Figure 36). The tryptophan 107 is in close proximity to dimethylamine of N2 G suggesting a cation-π interaction and has a role in discriminating between m^7G cap and m^3,2,7G cap [91].

15. Tgs 1 Interacting Proteins

Genetic and biochemical analysis of Tgs1 interacting proteins reveals a wide range of proteins involved in RNA metabolism.
34 Journal of Nucleic Acids

Snurportin1

CBP 20

Trp107 Ser105 Leu104 Trp276

m<sub>3</sub>GpppG

Asp116 Tyr20 Asp114 Arg112 Tyr43

m<sub>7</sub>GpppG

Figure 36: The comparison between m<sub>3</sub>,<sup>2</sup>,<sup>2</sup>,<sup>7</sup>G binding mode to snurportin1 and m<sub>7</sub>G cap binding mode to CBP 20, eIF4E and VP39 [91]. The m<sub>3</sub>,<sup>2</sup>,<sup>2</sup>,<sup>7</sup>G is stacked between tryptophan 107 and penultimate nucleotide G of cap core m<sub>3</sub>,<sup>2</sup>,<sup>2</sup>,<sup>7</sup>GpppG.

pppGGUCGAGAGG AUGGCU CAGC CG UUAA GC

UA GGCC

AAAA A UAA{CAC CUA

U}AAGAGUUC G GUUCCC

AG

C

A C CACGGCUGU CC UUCCAGCA CCUUUU-OH

Figure 37: The sequence of 4.5S RNA I (Novikoff hepatoma cell nuclei).

It interacts with proteins in the transcriptional apparatus, RNA end processing and decay, spliceosomal assembly and RNA modifying factors (Table 12).

Structurally, it is distinct from the m<sub>7</sub>G cap, and the specificity of binding proteins may determine the precision of its functional role in the RNP complex. The m<sub>3</sub>,<sup>2</sup>,<sup>2</sup>,<sup>7</sup>G cap structures are present only in nuclear snRNAs and snoRNAs which confer the function within the nucleus in transcription, splicing, modification, processing, and maturation of different RNA species.

16. Conclusion

16.1. General Consideration. In the present postgenomic era, study of the structure and function of noncoding RNAs is supremely important. It is estimated that ncRNAs are probably involved in all aspects of cell metabolism. Therefore, RNA-based information will contribute greatly to understanding various cell metabolisms. In the process of exploring ncRNAs, there may be many surprises awaiting us. They may include

(1) new species of RNA,
(2) new mechanism of RNA processing,
(3) new mechanism of transcription,
(4) new disease caused by RNAs with pathogenic sequences,
(5) new function for ncRNA.

16.2. The Problem of Unknown Modified Nucleotides. In the process of oligonucleotides cataloging, it is natural that an examination of base composition will reveal modified nucleotides or nucleosides in addition to unmodified standard nucleotides or nucleosides. In routine work, identification of modifications can be readily made by
two-dimensional paper chromatography for nucleotides or thin layer chromatography for nucleosides. However, there may be an occasion where chromatographic identification is not sufficient. Of course, it is best to have collaboration with outside specialists. For the sake of structural microanalysis, it is highly recommendable to determine molecular weight of the unknown nucleotide or nucleoside by mass spectrometry [119]. The required quantity is approximately 5 μg/nt where chromatographic identification of isotopically labeled sample requires 0.5 μg/nt. A difficulty may be confronted with purine bases that are fused to an imidazole ring (Queuosine) which is not suited for mass spectrometry. It is convenient to probe chemical complexity based on mass. The detailed analysis may require an unpredictably large amount of samples. There are 135 modified nucleosides listed, among which 6 nucleosides are not thoroughly identified [1].

16.3. Significance of Sequence Work. Past sequence work has permeated numerous significant areas of research providing a better understanding of cellular metabolism. The information obtained thus far is RNA-based information which is not seen in DNA, proteins, and others. As sequence work continues to make enormous progress, the postgenomic era will shape the direction of research in the area of molecular mechanisms of RNA metabolism. They are briefly as follows.

In RNA maturation, knowledge of structural modifications is necessary to discern between various mechanistic options. For example, there are two molecular mechanisms mediated by catalysis. One is mediated by RNA enzymes (snRNAs and snoRNAs) involved in splicing of pre-mRNA and processing of pre-rRNA. The other is protein enzymes involved in 5' cap formation. Currently, the higher order structural analysis is in progress. There is a need to elucidate the details of molecular mechanisms.

Along with the study of splicing physiology, splicing pathology is making significant progress. Aberrant modifications can generate disease causing alterations in structure. The aberrations cause problems in reading both genetic codes and splicing codes. Studying the regulation of alternative splicing will clarify the selective rules in intron removal and pathogenic rules in splicing code. From these studies, corrective strategy will evolve. The present sequence work is engaged in definition of ncRNAs diversity and their functional roles [120]. Since it is suggested that ncRNAs are involved in all aspects of regulations in cell metabolism, there may be opportunities to study various paths in cell metabolism, not limited to transcriptional and posttranscriptional events. It is this gigantic task, to reevaluate the genomic work, that holds excitement and promise.

**Abbreviations Used in Table 4**

**Short Noncoding RNA (Usually Shorter Than tRNA and Some Are Longer but Excluding snRNA Such As U1–U13)**

- miRNA: MicroRNA (imperfect base pairing)
- siRNA: Small interfering RNA (perfect base pairing)
- tasiRNA: Transacting small interfering RNA
- natsiRNA: Natural antisense transcribed small interfering RNA
- piRNA: PIWI interacting RNA (RNA precipitated by PIWI protein antibody)
- rasiRNA/piRNA: Repeat-associated small interfering RNA/pi-target RNA
- PARs: promoter associated RNAs
  - PROMTs: Promoter upstream transcripts (sense and antisense transcript)
  - PASRs: Promoter-associated small RNAs
  - TSSa-RNAs: Transcription-start-site-associated RNAs
  - tRNAs: Transcription initiation RNAs
- MSY-RNA: MSY2-associated RNAs (MSY; Y chromosome male-specific protein)
- snoRNA: Small nucleolar RNA (C/D box RNA, H/ACA RNA)
- sdRNA: sno-derived RNAs
- moRNA: MicroRNA-offset RNAs
- tel-sRNA: Telomere small RNAs
- crasiRNA: Centrosome-associated small interfering RNAs
- hsRNA: Heterochromatin small RNA or hairpin small RNA
- scaRNAs: Small Cajal-body-associated RNAs
- Y RNAs: Cytoplasmic small RNA Y1, Y3, Y4, and Y5
- tRNA-derived RNAs: Small RNA processed from tRNA by RNase (angiogenin)
- Alu/SINE RNA: Alu restriction enzyme cleaved repeat gene transcript/short interspersed nucleotide element RNA.

**Lnc RNA: Long Noncoding RNA (~0.5 to 100 kb)**

(1) **Specific Long Noncoding RNA**

- TR/TERC: Telomerase RNA/telomerase RNA component
- NEAT RNAs: Nuclear enriched abundant transcript 1 RNAs
  - NEAT1v-1: NEAT1 variant 1
  - NEAT1v-2: NEAT1 variant 2
  - NEAT2/MALAT1: Metastasis associated in lung adenocarcinoma transcript 1
- PINC RNA: Pregnancy-induced noncoding RNA
- DD3/PCA3: Prostate-cancer-associated RNA 3
- PCGEM1: Prostate cancer gene expression marker 1
- SPRY4-1T1: Sprouty homolog 4 gene transcript 1 (melanoma specific).
(2) Imprinting-Associated lncRNAs

- xiRNAs: X chromosome inactivating RNAs
- Xist: X chromosome inactivating sense RNA
- Tsix: Antisense transcript of Xist
- RepA: Repeat A RNA
- AIR RNA: Igf2r imprinting region RNA
- H19: Igf2 imprinting region RNA
- KCNQ1ot1: Antisense RNA from intron 10 of Kcng1 gene imprinting region.

(3) Regulatory lncRNAs

- HOTAIR: Homeogene inactivating RNA
- BORG: BMP/OP-responsive-gene-associated RNA
- CTN RNA: Cationic amino acid transporter protein coding region RNA
- ANRIL RNA: Antisense noncoding RNA in INK4 locus.

(4) Gene-Recombination-Associated lncRNA

- LINE: long interspersed nucleotide element
- CSR-RNA: Immunoglobulin class switch recombination region RNA.

(5) Satellite DNA Transcripts.

Abbreviations for Table 12

- Mud: Mutant U1 die
- RES complex: Heterotrimeric RNA retention and splicing complex composed of Bud13, Ist3/Snu17, and Pml1
- Swt21: Synthetic with Tgs1 number 21
- TRAMP complex (Trf4, Air2, Mtr4p polyadenylation complex): Interacts with exosome in the nucleus and involved in 3' end processing of rRNA, snoRNA, and U1, U4, and U5 snRNA;
  - Trf4 or Trf5: poly(A) polymerase(PAP);
  - Mtr4: RNA helicase;
  - Air1 or Air2: Zinc knuckle protein
- Cbf5(YLR175W): Centromere binding factor;
- Pseudouridine synthase catalytic subunit of box H/ACA snoRNP complex
- PRIMP: PRIP-interacting protein with methyltransferase domain, PRIMP is a Tgs1 (trimethylguanosine synthase 1) cloned from human liver cDNA library
- PRIP: PPAR interacting protein
- PPAR: Peroxisome proliferator-activated receptor
- PBP: PPAR binding protein.

Acknowledgment

The authors are indebted to Professor Lynn Yeoman of Baylor College of Medicine for helpful discussion in the preparation of this paper.

References


